Department of Computer Architecture and Technology
Konputagailuen Arkitektura eta Teknologia saila (KAT)
Departamento de Arquitectura y Tecnología de Computadores (ATC)

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

University of the Basque Country

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Contributions to improve Human-Computer Interaction using machine learning

Ph.D. Dissertation presented by
Ainhoa Yera Gil

Supervised by
Olatz Arbelaitz Gallego
Javier Muguerza Rivero

Donostia, December 2019

Department of Computer Architecture and Technology
Konputagailuen Arkitektura eta Teknologia saila (KAT)
Departamento de Arquitectura y Tecnología de Computadores (ATC)

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

University of the Basque Country

INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA

Contributions to improve Human-Computer Interaction using machine learning

Ph.D. Dissertation presented by
Ainhoa Yera Gil

Supervised by
Olatz Arbelaitz Gallego
Javier Muguerza Rivero

Donostia, December 2019

This work was funded by the following units: firstly, by the University of the Basque Country UPV/EHU (PIF15/143 grant). Secondly, by the research group ADIAN that is supported by the Department of Education, Universities and Research of the Basque Government (grant IT980-16). Finally, by the Ministry of Economy and Competitiveness of the Spanish Government, co-founded by the ERDF (eGovernAbility project, TIN2014-52665-C2-1-R).

“Reserve your right to think,
for even to think wrongly
is better than not to think at all.”
Hypatia of Alexandria

Ikasten, ikasten, beti beti gure adimena zabaltzen,
ikasten, ikasten, momentu oro...
BTX

Abstract

This PhD thesis contributes on designing and applying data mining techniques targeting the improvement of Human Computer Interaction (HCI) in different contexts. The main objectives of the thesis are to design systems based on data mining methods for modelling behaviour on interaction and use data. Moreover, having to work often in unsupervised learning contexts has lead to contribute methodologically to clustering validation regardless of the context; an unsolved problem in machine learning. Cluster Validity Indexes (CVIs) partially solve this problem by providing a quality score of the partitions, but none of them has proven to robustly face the broad range of conditions. In this regard, in the first contribution several CVI decision fusion (voting) approaches are proposed, showing that they are promising strategies for clustering validation.

In the Human-Computer Interaction context, the contributions are structured in three different areas. The accessibility area is analysed in the first one where an efficient system to automatically detect navigation problems of users, with and without disabilities, is presented.

The next contribution is focused on the medical informatics and it analyses the interaction in a medical dashboard used to support the decision-making of clinicians (SMASH). On the one hand, connections between visual and interaction behaviours on SMASH are studied. On the other hand, based on the interaction behaviours observed in SMASH, two main cohorts of users are automatically detected and characterised: primary (pharmacists) vs secondary (non-pharmacists).

Finally, two contributions on the e-Services area are made, focusing on their interaction and use respectively. In the first one, potential students aiming to enrol the University of the Basque Country (UPV/EHU) are satisfactorily modelled based on the interactive behaviours they showed in the web of this university. The second one, empirically analyses and characterises the use of e-Government services in different European countries based on survey data provided by Eurostat.

Laburpena

Doktorego-tesi honek, hainbat testuingurutan, Pertsona-Konputagailu Elkarrekintzaren (PKE) hobekuntzarako datuen meatzaritzako teknikak diseinatzen eta aplikatzen laguntzen du. Tesiaren helburu nagusiak datu-meatzaritzako metodoetan oinarritutako sistemak diseinatzea da, elkarrekintza- eta erabilera-datuen portaera modelatzeko. Gainera, gainbegiratu gabeko ikasketa-testuinguruekin sarritan lan egin behar izanak, datuen testuinguru guztiei zuzendutako clusteringa baliozkotzeari buruzko ekarpen metodologikoa egitera bultzatu gaitu. Kluster baliozkotze indizeek (CVI) partizioen kalitate-neurri bat ematen duten heinean, arazo hau partzialki ebazten dute, baina horietako batek ere ez du erakutsi egoeren espektro handiari aurre egiteko gaitasuna. Ildo honetatik, lehen kontribuzioan CVIen arteko erabaki-fusioen (bozketa) hainbat sistema proposatzen ditugu, eta klusteringa baliozkotzeko estrategia eraginkorrak direla erakusten dugu.

Pertsona-Konputagailu Elkarrekintzaren testuinguruan, ekarpenak hiru arlotan egituratuta daude. Irisgarritasun arloa lehenengo kontribuzioan aztertzen da, sistema eraginkor bat aurkeztuz, desgaitasuna duten eta desgaitasuna ez duten erabiltzaileen nabigazio-arazoak automatikoki detektatzen dituena.

Hurrengo ekarpena informatika-medikoan zentratzen da eta medikuei erabakiak hartzeko jardueretan laguntzeko erabiltzen den osasun-arbela mediko baten (SMASH) elkarrekintza aztertzen du. Batetik, SMASH arbelean portaera bisualen eta interaktiboen arteko loturak aztertzen dira. Bestalde, SMASH arbelean antzemandako portaera interaktiboen arabera, bi erabiltzaile talde nagusi detektatu eta ezaugarritu dira: lehen mailakoak (farmazialariak) eta bigarren mailakoak (ez farmazialariak).

Azkenik, bi kontribuzio egiten dira zerbitzu elektronikoen (e-Zerbitzuen) arloan, elkarrekintza eta erabileran oinarrituz, hurrenez hurren. Lehenengoan, Euskal Herriko Unibertsitatean (UPV/EHU) izena eman nahi duten ikasle potentzialak modu eraginkorrean modelatu dira unibertsitate honen webgunean erakutsitako jokabide interaktiboen arabera. Bigarrenean, gobernuko e-Zerbitzuen erabilera aztertu da Europako hainbat herrialdetan, Eurostatek emandako inkesta-datuetan oinarrituz.

Resumen

Esta tesis doctoral contribuye al diseño y la aplicación de técnicas de minería de datos dirigidas a la mejora de la Interacción Persona-Computadora (IPC) en diferentes contextos. Los objetivos principales de la tesis son diseñar sistemas basados en métodos de minería de datos para modelar el comportamiento en datos de interacción y uso. Además, como los contextos de aprendizaje no supervisado han sido una constante en nuestro trabajo, hemos contribuido metodológicamente a la validación de clustering independientemente del contexto de los datos; problema no resuelto en el aprendizaje automático. Los índices de validación de cluster (CVI) resuelven parcialmente este problema al proporcionar un valor cuantitativo de calidad de las particiones, pero ninguno de ellos ha demostrado poder enfrentarse de manera robusta en una amplia gama de condiciones. En este sentido, en la primera contribución se proponen varios sistemas de fusión de decisiones (votaciones) entre CVIs, demostrando que son estrategias prometedoras para la validación de cluster.

En el contexto de Interacción-Persona Computador, las contribuciones están estructuradas en tres áreas diferentes. En la primera de ellas se analiza el área de accesibilidad, presentando un sistema eficiente para detectar automáticamente los problemas de navegación de los usuarios, con y sin discapacidad.

La siguiente contribución se centra en la informática médica y analiza la interacción en una pizarra médica web (SMASH) utilizada para asistir en la toma de decisiones de los médicos. Por un lado, se estudian las conexiones entre los comportamientos visuales y de interacción en SMASH. Por otro lado, en base a los comportamientos de interacción observados en SMASH, se detectan y caracterizan automáticamente dos grupos principales de usuarios: primario (farmacéuticos) y secundario (no farmacéuticos).

Finalmente, se realizan dos contribuciones en el área de servicios electrónicos, centrándose en su interacción y uso, respectivamente. En la primera, se modelan satisfactoriamente los estudiantes que potencialmente desean matricularse en la Universidad del País Vasco (UPV / EHU), en función de los comportamientos interactivos que muestran en la web de esta universidad. La segunda contribución, analiza empíricamente y caracteriza el uso de los servicios de gobierno electrónico en diferentes países europeos en base a datos de encuestas proporcionados por Eurostat.

Acknowledgements

Eskerrak eman nahi nizkioke pertsona askori baina bereziki hemen aipatuko ditudanei:

- Olatz eta Javi (orkestra honen zuzendariak): asko ikasi dut zuengandik eta tesi hau zuek gabe etzan posiblea izango, ongi dakizue, beraz mila esker!

- Bide hau hasi berria nintzela, ondotik joan zitzaizkidan bi pertsona inportanteei, Lola amonari eta Zezi osabari, beti bihotzean, eskerrik asko emandako maitasun guztiarengatik.

- Ama (Hermi): txikia bai baina makala ez! kulturaren maitalea eta despisteen erregina, eskerrik asko nire txapak entzun eta animatzeagatik.

- Aita (Iñaki): handia ez, erraldoia zara! baina batez ere bihotz onekoa eta primerako sukaldaria, zu gabe elikadura eskasa izango nuke, beraz, esker mila tupper horiengatik!

- Ahizpa (Itziar): eskerrak norbaitek konponketen genea atera zuen, zorionak! asko zor dizut eta eredu zara zentzu guztietan beraz muxu haundi bat. Bide batez Ion, zorionak hain gauza handia egin izanagatik, ea horrelako gehiago egiten dituzuen ;)

- Iloba (Hodeitxo): Bide hau bukatzen ari nintzela zu jaio zinen eta bizitza ederragoa egiten diguzu egunetik egunera, asko maite zaitut pitxi!.

- Inma izebari eta Gorka lehengusuari: berriz zutitzen irakasteagatik, taupadak itzuli daitezen guztion bihotzetara. Gora Azkoien!

- Pisukideak eta lagunak: Jere, Manex, Telmo, Amaia, Naiara, Irantzu, Ilargi, Inhar, Ibon, Txabo, Eneko, Paleta herria...eskerrik asko laborategitik kanpo mundu bat dagoela oroitarazteagatik, pintxopoa eta bermuak ez daitezela eten.

- Lankideak: Aizea, Igor, Iñigo eta Ugaitz, mila esker lan hau aurrera ateratzen laguntzeagatik eta bide luze hau arinago egiteagatik, tesiaren zati haundi bat zuena da.

- ALDAPA taldea: Ibai eta Txus zuek landatutako zuhaitza begira noraino hazi den, ”kontsolidatutako haritza” dudarik gabe ;), mila esker! osteguneko kafeak ez ditzagun galdu, eh Otzeta??. Natxo, zozketa noizko? Agus, ea libururen bat gomendatzen didazun!

- Informatika Fakultatea: Tabako eta pintxopo zaleei, Joselito eta Ander, deskonektatzen laguntzeagatik; laster bukatuko duten beste ikertzaile guztiei, Jere, Igor, Amaia, Garde, Edu, Haritz...animo eta eutsi gogor, bukaeran salda badago! Bukatu duten ikertzaile eta irakasleei, Leti, Izaro, Olatz, Igor, Itziar, Maider, Ekaitz, Mendi, Momo, Usue, Leti, Ekhiñe... mila esker aholkuengatik, Aristiri faktura eta arazo guztiekin lagundu izanagatik, burokraziari sua!; EGOKITUZ taldekoei Julio, Myriam, Xabi...lan hau borobiltzen lagundu izanagatik, SIA Masterreko irakasle eta ikaskideei, zeinen ondo pasa genuen, ahaztezina!, KAT departamentuari, Elenari sinadura eta tramite guztiekin lagundu izanagatik, ADIAN taldekoei, Pako’s kafetegiko neskei (Iosune, Amparo eta Elena), kafe, morenito eta zuen irribarrea ez dadila falta inoiz! garbitzaileak, nire kaos guztia txukuntzeagatik eta tratu atseginarengatik; atezaintzako langileak, laborategia mila aldiz ireki izanagatik.

- Manchester University: thank you so much to all the people from the lab, Markel, without your help this dissertation would not have been possible, thank you so much!, Aitor, Julio, Julia, Alan, Alaa and Deema (good luck girls, you can do it!), Simon...for making my international stay easier, for supporting me and of course for all the pints. Cheers!

- Atzerriko ikuskatzaileei: Hugo eta Markel eskerrik asko benetan lan honi egindako iradokizun eta ekarpenengatik.

- Epaimahaiko kideei: tesia irakurtzeko eta ebaluatzeko lana hartu izanagatik, mila esker benetan.

- Beste batzuk: Anari, hainbestetan entzun izanagatik eta burua ordenatzen irakatsi izanagatik, laborategitik pasa zaretenoi, Nikos, Ana... botatako barre guztiengatik.

- Zuri: mila esker gau ilunetan izarrak erakusteagatik, sinpleki ni naizen moduan maitatzeagatik, asko gutsut!.

Contents

I Introduction xix

1 Introduction to Human-Computer Interaction and machine learning 1

1.1 Organisation of this dissertation 6

II Background 7

2 Supervised learning 9

2.1 Naïve Bayes 10

2.2 Neighbour based classifiers 10

2.3 Decision trees 11

2.3.1 C4.5 12

2.3.2 CTC 14

2.4 Support Vector Machines 15

2.5 Artificial Neural Networks 16

2.6 Multiple classifier systems 18

2.6.1 Bagging 18

2.6.2 Boosting 19

2.7 Validation 20

2.7.1 K-fold cross-validation 21

2.7.2 Leave-one-out cross-validation 21

2.8 Performance metrics 21

2.9 Statistical tests 23

2.9.1 Student t-test 23

2.9.2 Kendall test 24

3 Unsupervised learning 25

3.1 Hierarchical clustering 26

3.1.1 Hierarchical agglomerative clustering 27

3.2 Partitional clustering 28

3.2.1 K-means 29

3.2.2 K-medoids 30

3.3 Clustering validation 32

3.3.1 Cluster Validity Indexes (CVIs) 32
3.4 Statistical Tests 39
3.4.1 Wilcoxon-signed rank 39
3.4.2 Friedman test 40

III Contributions 41

4 Unsupervised classification: Analysis of several decision fusion strategies for clustering validation. 43
4.1 Introduction 43
4.1.1 Related work 44
4.2 Experimental Setup 45
4.3 The proposed approach 47
4.4 Designed Strategies and Results 48
4.4.1 Global Voting 48
4.4.2 Selective Voting 49
4.4.3 Statistical Tests 57
4.4.4 Summary 58

5 Modelling the interaction of users with disabilities 59
5.1 Introduction 59
5.1.1 Related work 60
5.2 Description of the platform used for the experiments: RemoTest 61
5.3 Experiments with users 63
5.4 Tasks' characteristics 64
5.5 Feature extraction 65
5.5.1 Feature extraction for device detection 67
5.5.2 Feature extraction for problem detection 67
5.6 Automatic device detection: first approach 69
5.6.1 Description of the dataset 69
5.6.2 Results and analysis 70
5.7 Automatic device detection: adopted approach 74
5.7.1 Description of the new dataset 74
5.7.2 New results and analysis 75
5.8 Automatic problem detection 77
5.8.1 Clustering for pattern discovery 77
5.8.2 Use of the detected patterns to facilitate navigation . . . 85
5.8.3 Summary 87

6 Modelling the interaction with specific web platforms 91
6.1 Introduction 91
6.2 Context: The SMASH Intervention 93
6.3 User studies and metrics 95
6.3.1 Lab study 95
6.3.2 Observational study 97

6.3.3 Computed metrics 98
6.4 Inferring visual behaviour from interaction data 100
6.4.1 Description of the datasets 100
6.4.2 Results and analysis 101
6.5 Characterizing different types of users 106
6.5.1 Description of the datasets 106
6.5.2 Results and analysis 108
6.5.3 Discussion 112
6.6 Summary 114

7 Modelling the interaction and use of e-Services 117
7.1 Introduction 117
7.2 Modelling the interaction with e-Services 118
7.2.1 Context: The website of the University of the Basque Country 119
7.2.2 Preprocessing, session classification and description of the datasets 120
7.2.3 Automatic classifier system based on supervised learning techniques 123
7.2.4 Automatic classifier system based on unsupervised learning techniques 126
7.3 Modelling the use of e-Services 132
7.3.1 Introduction 132
7.3.2 Eurostat CSIS Surveys 134
7.3.3 Characterisation of extreme values of E-Government Use (EGU) 139
7.3.4 Comparison between EGUI$^+$ and other indexes 144
7.3.5 Discussion 148
7.4 Summary 150

IV Conclusions 153

8 Conclusions 155
8.1 Machine learning - Clustering validation. 156
8.2 HCI - Accessibility 156
8.3 HCI - Medical informatics 157
8.3.1 Synergies between visual and interactive behaviours . . . 157
8.3.2 User modelling based on interactive behaviour 158
8.4 HCI - E-Services 158
8.4.1 Modelling the interaction with e-Services 158
8.4.2 Modelling the use of e-Services 159
8.5 Further work 160
8.6 Related publications 161

Part I

Introduction

Chapter 1

Introduction to Human-Computer Interaction and machine learning

Due to the digital transformation of the society in the last decades, all kind of electronic devices, including smart phones, tablets or computers, have invaded our lives, completely transforming the way we interact with the world. This means that, since we get up until we go to sleep in order to accomplish every day tasks, we find ourselves forced to interact with a wide number of digital tools. As a result, interactive systems are constantly gathering information about the activities we perform and our personal preferences, which enable them to learn about our behaviour and thinking.

Human-Computer Interaction (HCI), the study of the interaction between humans and computers, has not been inherent to the popularisation of the use of digital devices. In the early 80s Card et al. (Card et al. 1983) referred for first time to Human-Computer Interaction using text edition tasks as a representative example and proposing cognitive models about human performance relevant to this area. In contrast, nowadays HCI comprises multiple disciplines contributing to the three elements involved in HCI: the user (e.g psychology, sociology and ethnography), the interaction (e.g computer science) and the device (e.g engineering, ergonomics and design).

The goal of HCI is to build usable systems, namely, those being effective, efficient (easy to use), safe, useful, easy to learn and easy to remember (Preece et al. 2001). Accordingly, usability is defined by the International Organisation for Standardisation (ISO), ISO 9241-11:2018 (ISO 2018), as the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

Some usability metrics can be used to measure these properties, such as completion rate (percentage of tasks successfully completed) for effectiveness and task duration for efficiency (ISO 2018). In addition, usability entails users to be satisfied, what can be evaluated asking the users about their experience when interacting with the system through formal questionnaires such as the Questionnaire for User Interaction Satisfaction (QUIS) (Shneiderman 1997). Figure 1.1 summarises usability and user experience (UX) goals (Preece et al. 2001).

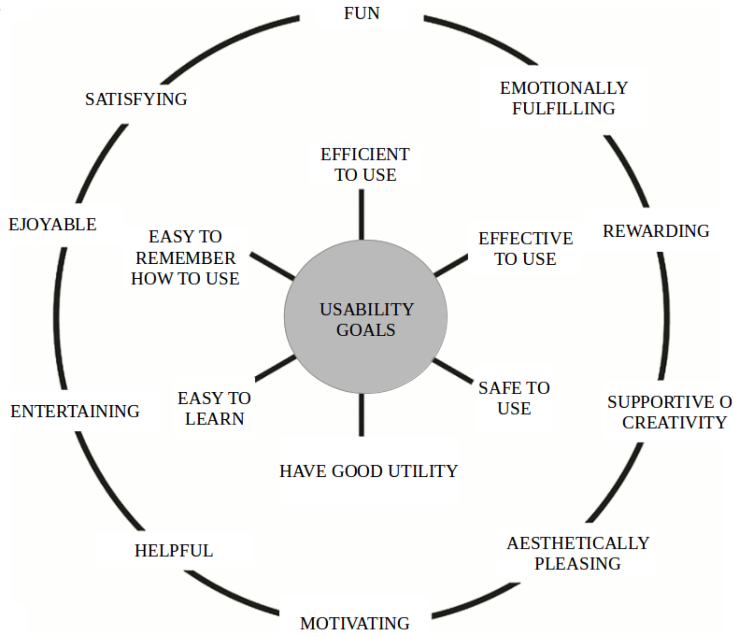


Figure 1.1: Usability and user experience goals to be considered in HCI.

Therefore, integrating usability in HCI systems requires users to be in the centre of their design and evaluation processes. Human-Centred Design (HCD) illustrates this idea, since it contributes to develop usable and useful interactive systems by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, and usability knowledge and techniques (ISO 2019). HCD is an iterative process where first, the context of use and the user requirements need to be understood and defined to then, produce design solutions (user, tasks and interface) and finally, the previous steps are repeated until the solution design does meet the user requirements. But the reality is that few HCI designers engage users in all the stages of the design process on the grounds that it is expensive, time-consuming, technically complex to undertake or difficult to manage. Even so, the analysis of real interaction of the users with the final system is crucial to extract helpful knowledge (user profiles, navigation problems...) about the usability of the solution design, thereby actions can be taken (adaptations, recommendations...).

In this context, machine learning becomes a powerful ally able to analyse massive interaction data and automatically identify meaningful patterns in a cheap and unobtrusive manner. Machine learning is a branch of Artificial Intelligence (AI) which enables machines to learn from past experiences (training data) to make independent decisions / predictions (e.g. classification) on new data without human intervention. In particular, this learning can be described as the acquisition of structural descriptions from the examples, that then can be used for prediction, explanation and understanding (Witten and E. Frank 2005). Data mining techniques, also known as pattern discovery from data (Han et al. 2011), provide these structural descriptions through mathematical models built based on sample data. As shown in Figure the 1.2 machine learning process involves five steps: data selection, data preprocessing (noise removal, feature extraction etc.), use of machine learning techniques and evaluation of the patterns obtained.

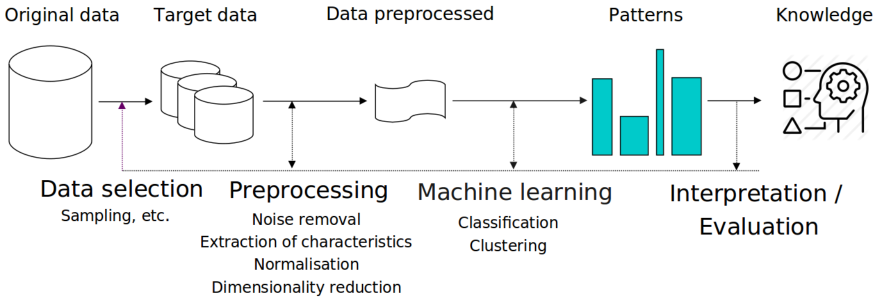


Figure 1.2: Machine learning process.

Using a generalised taxonomy, machine learning algorithms are divided in two main categories, supervised learning and unsupervised learning, which are characterised by the availability and lack of labelled data respectively. Specifically, supervised learning algorithms use labelled data to learn (training) and infer a function to determine the label (dependent variable) of new data (test). On the other hand, unsupervised learning algorithms (e.g clustering) infer a function to group the data into a number of clusters (groups), according to their proximity / similarity, somehow finding out a new unknown label for each group. In both types of learning the validation of the results is essential in order to legitimise the knowledge gathered. In supervised learning, also known as classification, the validation evaluates how good is the algorithm in classifying new data by using the trained model with the test data to determine to what extent the training model can be generalised. In contrast, in unsupervised learning, the validation measures how well does the output partition fit the underlying structures of the data.

In the HCI literature we find several approaches using machine learning with the goal of improving user experience. Yang et al. (Yang et al. 2018) provide a conceptual model where the contributions creating value for users made by these approaches are classified into four types:

- Self type: They are carried out by monitoring and logging the user's actions and provide personal knowledge, about the user or about a group of users showing similar behaviours.

- World type: Provide information about the user's current context, a distant context, or relevant information for a currently unfolding interaction. A representative example of this channel will be a robot that transforms data about the external world into machine intelligence.

- Optimal type: Provide information about an arbitrarily defined "optimal" or "better" status, such as optimal behaviours (e.g. active participation in a class). Intelligent tutoring systems that work to increase learning efficiency are a representative example of this channel.

- Utility and/or new capability type: Provide information to increase utility, including aspect such as interaction efficiency, availability, reduced cognitive and interaction efforts, and/or the acquisition of new capability. A representative example within this channel is an adaptive mobile user interface that minimise the users' navigation efforts.

In this dissertation four contributions valid to improve HCI have been drawn using machine learning techniques in the following contexts: clustering validation, modelling users with disabilities, modelling the interactions on a web platform from the medical area and modelling e-Services.

The first contribution is framed within the area of unsupervised learning and it focused on clustering validation. In any clustering procedure, finding the partition that best fits the underlying structure of the data, named clustering validation, is a difficult task because it implies to blindly group unlabelled instances, that is, labelling the instances without a certain criteria. Cluster Validity Indexes (CVIs) make this task easier by measuring the compactness and separation of the clusters using specific indexes that provide a quality metric for the output partition. Even so, one of the most extensive comparative works on CVIs to the date (Arbelaitz et al. 2013b) brings to light one of their major weaknesses, the instability of their performances depending on the clustering environments (e.g noise, dimensions, number of clusters etc.). To address this problem, in our contribution we propose several CVI decision fusion approaches that can improve the performances of individual CVIs in all the environments. This contribution facilitates the decision-making on the optimal number of clusters regardless of the nature of the data.

The second contribution deals with web accessibility and proposes a two-step system based on machine learning techniques to automatically detect possible user navigation problems. In the first step, supervised learning algorithms are used to detect the interaction device being used, whereas in the second step, possible user navigation problems are detected by means of unsupervised learning procedures. In addition some possible adaptations are discussed for the different navigation problems detected. Therefore, this contribution is included in the four categories described above: *self* type, as it analyses how does each participant perform particular tasks; *world* type, since it detects the device being used by each participant to enhance accessibility; *optimal* type as it determines which users may be experiencing different type of problems to accomplish a task; *utility/new capability provision* type insofar as it discusses suitable adaptations to alleviate the interaction problems detected.

The third contribution belongs to the area of medical informatics and it is concerned to improve the existing technology designed to support clinicians in decision-making activities, more precisely medical dashboards. In particular we analysed the use of the Salford Medication Safety Dashboard (SMASH) used in primary care across Salford, UK, (Williams et al. 2018), based on the data collected in two different studies: one of a lab nature with six clinicians and the other one of observational type with 35 clinicians. Two relevant questions wanted to be answered in this work using machine learning algorithms on the interaction and gaze data gathered in SMASH: whether visual behaviour can be inferred from the interaction behaviour shown by the users in the dashboard and whether is it possible to automatically classify and characterise the two main cohorts of users of the dashboard. The first question was studied using correlation metrics and clustering procedures initially on the gaze and interaction data gathered in the lab study and in a second instance, on the interaction data of both lab and observational studies. The second question was elucidated using supervised learning procedures on the interaction data captured in the observational study. In short, the contribution made in this context is included in the four categories above described: *self* type because it provides knowledge about different groups of clinicians sharing similar gaze and/or interaction behaviours; *world* type as it provides knowledge about clinician's usage of medical dashboards which can be used to improve the design of such decision support tools such; *optimal* type insofar it detects and characterises competences of secondary users who engaged less with SMASH than primary users; *utility/new capability provision* type since it monitors the interaction and gaze behaviour of users in a medical dashboard, which are related with competence and cognitive load respectively, to detect usability problems or lack of competence and inform adaptations.

In the last contribution two different approaches are presented within the area of e-Services, one for modelling the interaction of the users in the enrolment web information area of the University of the Basque Country (UPV/EHU) and the other one, for empirically analysing the real use of e-Government services in Europe. In the first analysis both, supervised and unsupervised learning algorithms were used on the interaction data of the enrolment area of the UPV/EHU to automatically classify and characterise two types of users: those obtaining enrolment information (potential users interested in enrolling) and those carrying out searching type tasks. In contrast, in the second work two indexes were defined to quantify the e-Government practical use in 26 EU countries. Based on survey data provided by Eurostat and using supervised learning procedures a characterisation of this factor was carried out for a selection of countries with different levels of e-Government use. Therefore, the contribution drawn in this context correspond to three of categories above explained: *self* type by providing knowledge about groups of users showing similar interaction behaviours in the enrolment web information area of the UPV/EHU, or regarding the use of e-Government services; *world* type by quantifying the use of e-Services for different countries (location context of use); *optimal* type by detecting and characterising users unsuccessfully seeking for information and users who use e-Government services at a very low level.

1.1 Organisation of this dissertation

After this first part with the introduction the rest of this thesis is divided in four different parts: Background, Contributions and Conclusions.

In Part II, Background, the key notions, principal techniques and particular notation used in this dissertation are provided. This part is divided into two chapters: Supervised learning techniques (Chapter 2) and unsupervised learning techniques (Chapter 3).

Part III gathers the four contributions of the dissertation which are divided in the following chapters: Contributions to clustering validation (Chapter 4), Contributions to modelling the interaction of users with disabilities (Chapter 5), Contributions to modelling the interaction with specific web platforms (Chapter 6) and Contributions to modelling the interaction and use of e-Services (Chapter 7).

Finally, in Part IV, Conclusions, the main conclusions of the dissertation are drawn (Chapter 8). This chapter discloses the contributions of the dissertation, discusses the main lessons learned from them and draws the future work to be addressed hereafter. The document concludes showing the referenced bibliography supporting the dissertation.

Part II

Background

Chapter 2

Supervised learning

As stated in the previous chapter, the goal of supervised learning is to predict the label (dependent variable or class) of new data based on a training process where labelled data are analysed. According to the general taxonomy, supervised learning is called regression when the data are of continuous type and classification when it is of discrete type. In this dissertation algorithms of the second type have been used, that is, those named classifiers. Inside this category two main types of algorithms can be distinguished, parametric when some parameters are assumed in the learning model (e.g data follow particular density of probability) and non parametric when no assumptions are made. The algorithms used in this dissertation to deal with the problems raised in each contribution are among the 10-top ranking presented by Wu et al. 2008. In particular, we used one parametric algorithm, Naïve Bayes (NB) (G.H. John and Langley 1995) and the non-parametric algorithms listed below:

- Neighbourhood based classifiers: IBK (Aha et al. 1991), which is a k Nearest Neighbour implementation (kNN).

- Decision trees: C4.5 (J.R Quinlan 1993) and CTC (Consolidated Tree Construction) (J.M. Pérez et al. 2007)

- Support Vector Machines (SVM): Sequential Minimal Optimisation (SMO) (J. Platt 1998).

- Artificial Neural Networks (ANN): Multilayer Perceptron (MLP) (Rumelhart et al. 1986).

- Multiple classifier systems: Bagging (Breiman 1996) and Boosting (Schapire 1999).

In the next sections the operating principles of these supervised learning algorithms will be described, which were run using the suite of machine learning free software Weka (M. Hall et al. 2009).

2.1 Naïve Bayes

Naïve Bayes algorithm (NB) (G.H. John and Langley 1995) is based on the Theorem of Bayes shown Equation 2.1 which assesses that, given the prior probabilities, $\mathrm{P}(w_i)$, and the class conditioned probability density functions $\mathrm{P}(x|w_i)$, it is possible to compute the posterior probability, $\mathrm{P}(w_i|x)$:

$$\mathrm{P}(w|x_i) = \frac{\mathrm{P}(x|w_i) * \mathrm{P}(w_i)}{\mathrm{P}(x)} \quad \mathit{where} \quad \mathrm{P}(x) = \sum_{i=1}^{C} \mathrm{P}(x|w_i) * \mathrm{P}(w_i) \tag{2.1}$$

Naïve Bayes algorithm assumes that the features or characteristics of the data are statistically independent although it also performs well when this condition is not met. If the data are independent the multivariate joint probability is the product of the marginal conditional probabilities ($\mathrm{P}(x_1, \ldots, x_F|w_i) = \mathrm{P}(x_1|w_i) \cdot \ldots \cdot \mathrm{P}(x_F|w_i)$). Equation 2.2 shows the operating principle of the NB classifier, which assigns to the pattern, x_f the class with highest probability, W_{NB}:

$$\mathrm{W}_{\mathrm{NB}} = \operatorname*{argmax}_{w_i \in \mathcal{C}} \mathrm{P}(x) \prod_{k=1}^{F} \mathrm{P}(x_f|w_i) \tag{2.2}$$

The NB algorithm is highly appreciated because of its computational simplicity and its high efficiency which in some applications can be similar to that of neural networks and decision trees.

2.2 Neighbour based classifiers

IBK (Aha et al. 1991) is a k Nearest Neighbour classifier (kNN) implemented in Weka that bases the classification in a distance function. It labels any test instance with the majority label among the k closest instances from the training set. Figure 2.1 illustrates a 3NN example.

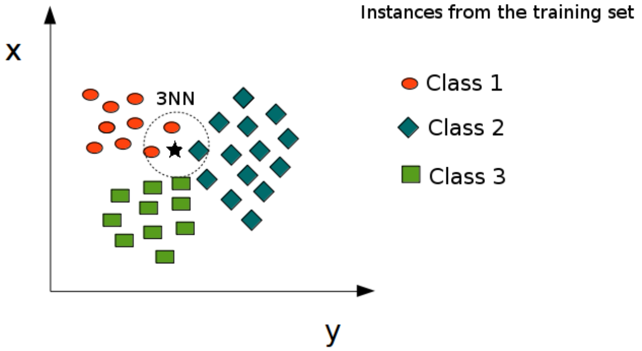


Figure 2.1: A three Nearest Neighbours (3NN) example.

In the example shown in Figure 2.1 the instances of the training set have three different labels or classes: circle, rhombus and square. According to the 3NN procedure the label assigned to the test instance, represented by a star, will be the majority label among the three closest instances (circle, circle, rhombus), circle in this case.

Algorithm 1 summarises the procedure followed by this algorithm which consists of three steps: first, the distances between the instance to be classified (test) and all the instances of the training set must be calculated; second, the k closest instances (neighbourhood) from the training set (minimum distance) must be selected; third, the majority class (label) among the k closest instances is assigned to the test instance.

Algorithm 1 kNN algorithm.

```
neighbourhood = {};
$x$ new test instance;
**for** each training instance $y$ **do**
    Compute the distance $d(x, y)$;
    **if** $d(x, y)$ is into the $k$ smallest distances; **then**
        Add $y$ to neighbourhood
    **end if**
**end for**
$x$.class = FindMajorityClass(neighbourhood);
```

Several distances can be used with kNN, perhaps, the most popular distance function used is the Euclidean distance. kNN algorithm is effective and simple and allows adding new examples to the training set at any time. However, its major drawback is its speed considering that the time required to classify a single test instance is proportional to the number of training instances. In addition, it does not deal very well with noise and redundant characteristics, and it has a null or very limited explanatory ability.

2.3 Decision trees

A decision tree can be defined as a graphical representation of a particular type of hierarchical analysis carried out on a set of data, separating the population in subgroups of individuals which differ from each other according to a discriminant criteria. A division function based on a discriminant criteria determines in each step the predictive variable or attribute selected to divide the node being treated, and the stratification of that variable to determine the different children nodes (building sub-populations of the parent node). There are multiple discriminant criteria when building a decision tree but the goal in all the cases is to generate children nodes as homogeneous as possible from the dependent variable point of view, that is, nodes with a minimal mixture of instances of different classes.

Figure 2.2 shows an example of a decision tree, where the leaf nodes in the bottom represent the classifications or decisions, the node in the top is called root node considered and the rest of nodes (set of predictors required for a final classification) are called the intermediate or split nodes.

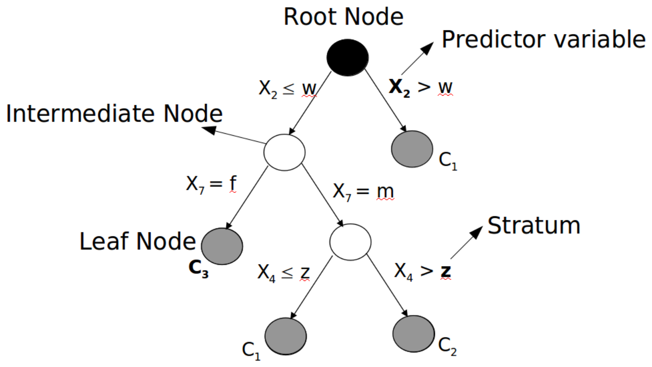


Figure 2.2: An example of a decision tree.

The decision trees classify or estimate the class belonging probabilities, providing an explanation of the decision made with each pattern. They have a very volatile behaviour (weak classifier) regarding the training set, due to the fact that the first divisions condition overmuch the final tree. Depending on the application they require methods to increase their stability.

2.3.1 C4.5

The C4.5 (J.R Quinlan 1993) algorithm, implemented as J48 in Weka, was designed by J Ross Quinlan who also is the author of its predecessor the Induction of Decision Tree ID3 (J.R. Quinlan 1986) algorithm, being both of them two of the most widely used decision trees.

Both algorithms use the Shannon Entropy (Shannon 1948) as split function. Equation 2.3 shows how the entropy or the amount of information of the dependent variable (class), C, the independent variable, V, and the contingency table are be computed (TC), $\mathrm{H_y}, y \in \{C, V, TC\}$).

$$\mathrm{H_C} = -\sum_{i=1}^{n_C} (\mathrm{p_i} \log_2 \mathrm{p_i}), \quad \mathrm{H_V} = -\sum_{j=1}^{n_V} (\mathrm{p_j} \log_2 \mathrm{p_j}), \quad \mathrm{H_{TC}} = -\sum_{i=1}^{n_C} \sum_{j=1}^{n_V} \mathrm{p_{ij}} \log_2 \mathrm{p_{ij}}$$
$$\mathrm{p_i} = \frac{\mathrm{M_i}}{\mathrm{T}}, \quad \mathrm{p_j} = \frac{\mathrm{M_j}}{\mathrm{T}}, \quad \mathrm{p_{ij}} = \frac{\mathrm{o_{ij}}}{\mathrm{T}} \tag{2.3}$$

- $\mathrm{p_i}$, $\mathrm{p_j}$ and $\mathrm{p_{ij}}$: distributions of the class, the independent variable and the contingency table respectively.
- $\mathrm{M_i}$ and $\mathrm{M_j}$: marginal distributions of the class and independent variable respectively (probabilities of the values of one of the variable without reference to the values of the other variable).

- o_{ij}: number of observed instances in the training set with values i and j for the class and independent variables respectively.

- T: grand total (total number of observations).

ID3 algorithm uses the Information Gain ($\mathrm{H_T}$) as a split function, shown in Equation 2.4.

$$\mathrm{H_T} = \mathrm{H_V} + \mathrm{H_C} - \mathrm{H_{TC}} \tag{2.4}$$

In the C4.5 algorithm the split function used is the Gain Ratio criteria ($\mathrm{G_R}$), shown in Equation 2.5.

$$\mathrm{G_R} = \frac{\mathrm{H_T}}{\mathrm{H_V}} \tag{2.5}$$

The procedure followed by the C4.5 algorithm is described in Algorithm 2.

Algorithm 2 C4.5 algorithm, based on (Zhu et al. 2019).

```
$Tree$= {};
$D$= feature-valued dataset;
if $D$ is TRUE or Stopping Criteria is TRUE; then
    Terminate
end if
for each attribute $a$ in $D$ do
    $subset$=spliton($a$);
    $a.G_R$ = FindGainRatio(subset);
end for
$a.best$ = Max($a.G_R$);
$Tree$=decision_node=spliton($a.best$);
$D_v$=Induce subsets from D based on $a.best$;
for all $D_v$ do
    $Tree_v$=C4.5($D_v$)
    Attach $Tree_v$ to the corresponding branch of $Tree$
end for
Return $Tree$
```

According to Algorithm 2 the procedure of the C4.5 consists of the following steps: selecting in the root node the attributes with the maximum information gain to split the training data into as many subsets as the values of a chosen attribute has; processing recursively for every subset until all of them are classified (stopping criteria). Other stopping criteria include reaching a maximum tree depth or a minimum number of instances in a leaf node (pruning threshold).

C4.5 present some improvements over ID3 in terms of methods to deal with numeric attributes (continuous data), missing values, noisy data, and generating rules (Witten and E. Frank 2005). In addition, C4.5 incorporates pruning (removal of sections of the tree with low predictive ability), which reduces the size of the tree and the over-fitting occurred when there is a is very high training set accuracy at the expense of a high test set error preventing the generalisation of the learning model.

2.3.2 CTC

The consolidated tree construction (CTC) algorithm (J.M. Pérez et al. 2007), implemented as J48 Consolidated in Weka, was designed to deal with a class imbalance problem. In contrast to C4.5 which uses a single sample to build the tree, CTC creates several sub-sets of samples which then uses to build the tree. The CTC algorithm carries out a voting procedure in order to select the variable splitting the node of the tree at each step of the tree's building process (Arbelaitz et al. 2013a). The same split criteria proposed by Quinlan in the C4.5 algorithm (J.R Quinlan 1993) is used in the CTC, that is, the Gain Ratio ($\mathrm{G_R}$) illustrated in Equation 2.5. Algorithm 3 summarises the iterative process to build a consolidated tree.

Algorithm 3 CTC algorithm, based on (J.M. Pérez et al. 2010).

```
$S$=training set
$N_S$=number of sub-samples to generate;
$R_M$=method used to generate sub-samples (Re-sampling_Mode);
$n$=number of examples to generate;
for $i$ in 1 to $N_S$ do
    $S^i$=$\{R_M(S)\}$;
    $LS^i$=$\{S^i\}$ // initialise $LS^i$ with $S^i$);
end for
$CurrentConsolidatedNode$=$RootConsolidatedNode$;
repeat
    for $i$ in 1 to $N_S$ do
        $CurrentS^i$=$First(LS^i)$ ;// first element of the list
        $LS^i$= $LS^i - CurrentS^i$;
        $(X, B)^i$=$BestSplit(CurrentS^i)$
    end for
    $(X_c, B_c)$= $Consolidatedpair(X, B)^i, 1 \leq i \leq N_S$
    if $(X_c, B_c) \neq Not_Splitt$ TRUE; then
        $Split(CurrentConsolidatedNode)\_basedon(X_c, B_c)$
        for $i$ in 1 to $N_S$ do
            $\{S^i_x, 1 \leq x \leq n\}$=$Divide(CurrentS^i)\_basedon(X_c, B_c)$;
            $LS^i$= $\{S^i_x, 1 \leq x \leq n\} \cup LS^i$
        end for
    else
        $LeafconsolidateNode$=$CurrentConsolidatednNode$
    end if
    $CurrentConsolidatedNode$=$NextNodeToConsolidate()$
until $\forall i \quad LS^i$=empty;
```

According to the algorithm, first, a set of sub-samples ($S^i, 1 \leq i \leq N_S$) are extracted from the training set based on a particular re-sampling technique (R_M), e.g bootstrap (random sampling with replacement). Then, all the sub-samples S^i are stored in a list in LS^i and the construction of the CT tree starts. The building process is commanded by $CurrentConsolidatedNode$ as it enables the function $NextNodeToConsolidate()$ to return the next node to be used. Similarly, $CurrentS^i$ is used as a pointer of the next data partition (related to one node) of S^i to be treated in the building process of the i^th tree.

In Algorithm 3 the split proposal for the first data partition in LS^i is represented by the pair $(X, B)^i$, where X is the feature selected to split and B represents the proposed branches (criteria) to divide the data in the current node. Then, in the consolidation step, a voting process based on all the proposals is carried out in order to determine the consolidated feature and branches (X_c, B_c). This process is repeated until LS^i is empty for all i, that is, the tree does not grow any more if in the last partition in all LS^i, the majority vote is not to split thus, to become a leaf node (stopping criteria).

The main strengths of the CTC algorithm are its good performance in imbalanced an noisy contexts (high accuracy), the comprehensibility of the classification it carries out, which is provided in a single tree and the stability of the explanation provided.

2.4 Support Vector Machines

The Sequential Minimal Optimisation (SMO) algorithm (J. Platt 1998) is categorised inside the group of Support Vector Machines which were first developed by Cortes and Vapnic for binary classification (Cortes and Vapnik 1995). The idea is to maximise the margin around the hyper-plane separating two classes, assuming a lineal separability between them. This hyper-plane is determined based on the subset of patterns defining the border between classes (quadratic optimisation problem), which are named support vectors. When the margin between the nearest points of the two classes is maximised, the points of the boundaries are defined as support vectors and the middle of the margin is the optimal separating hyper-plane.

Using a primary formulation the goal of SVM in the example will be to minimise the objective function, $1/2 \sum_{i=1}^{n} w_i^2$, given the restrictions $y_i(wx_i + b) \leq 1, \quad 1 \leq i \leq N$. Figure 2.3 illustrates an example of two dimensions dataset with two lineally separable classes ($y_i \in \{-1, 1\}$), where the points of the hyper-plane that divides the two classes (x_i) satisfy that, $wx_i + b = 0$, given $x : i \in R^n, 1 \leq i \leq l$; $y_i \in \{-1, 1\}$. In this case, all the training tuples allocated in any of the hyper-planes will be support vectors (the four points touching the two support vectors drawn in Figure 2.3.

Support Vector Machines can be applied to problems of high dimensions, and in addition to lineal separable problems (hard margin) they also are able to deal with non-linearity by using kernel procedures, where a projection of data points into an (usually) higher-dimensional space is carried out so they become linearly separable. SVM algorithms can also deal with overlapping classes by using soft margin, that is, applying low weights to the data points located in the incorrect side of the margin so that their influence is diminished.

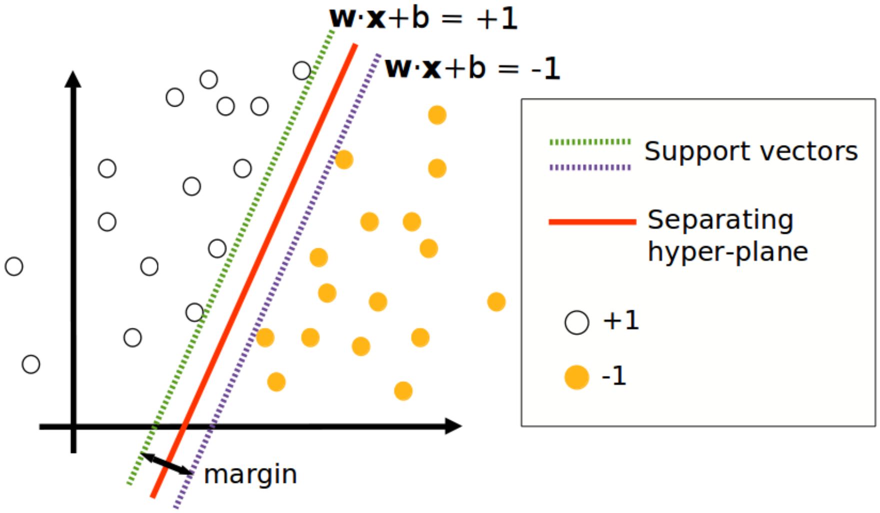


Figure 2.3: Two dimensions dataset with two lineally separable classes $y_i \in \{-1, 1\}$ divided by the hyper-plane.

2.5 Artificial Neural Networks

The Multilayer Perceptron (MLP) algorithm (Rumelhart et al. 1986) is categorised inside the group Artificial Neural Networks (ANN), which aroused from the idea of modelling mathematically the human intellectual abilities.

The basic structure of an ANN is a neuron, and Simple Lineal Perceptron (SLP) is the most simple ANN with a single one. As shown in Equation 2.6 the output of a SLP (o), obtained applying a nonlinear activation function (f), e.g sign, sigmoid (σ), to the network (net). The net is defined as the inner product between the input weights (w_j) of the neuron and the input pattern (x_j).

$$\mathrm{o} = f(\sum_{j=1}^{N} w_j * x_j + w_0) \tag{2.6}$$

In the training of the simple lineal perceptron first, the input patterns ($x_j, j \in \mathbb{N}, 1 \le j \le N$) are given to the network and their outputs are computed (o) and then, the weights ($w_j, j \in \mathbb{N}, 1 \le j \le N$) are updated depending on whether the output obtained is correct or not (t, target). This process is repeated until a good performance of the network is achieved or in case it does not converge, until a certain predetermined number of training runs. Equation 2.7 shows how the weights are updated:

$$\begin{aligned} \mathrm{w_{jnew}} &= \mathrm{w_{jprior}} + (\mathrm{t} - \mathrm{o}) * x \quad \text{t: desired output} \\ \mathrm{w_{0new}} &= \mathrm{w_{0prior}} + (\mathrm{t} - \mathrm{o}) \end{aligned} \tag{2.7}$$

This kind of network is not able to solve nonlinear problems. For such problems more complex models like the Multilayer Perceptron shown in Figure 2.4 are used.

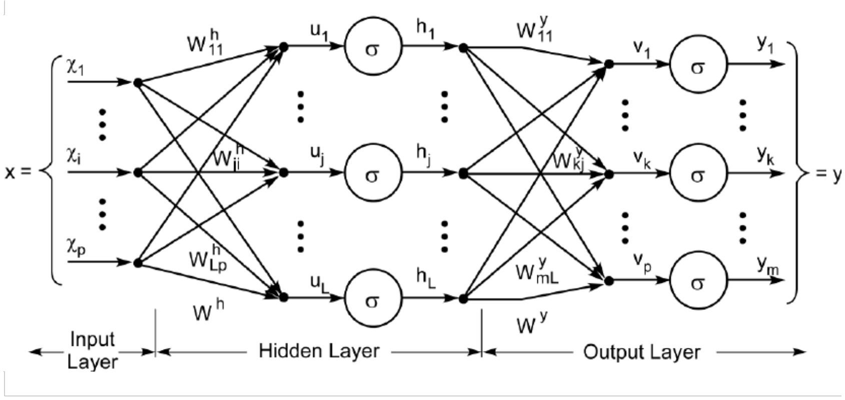


Figure 2.4: Structure of a MLP neuronal network (Faghfouri and Frish 2011)

MLP are *feedforward* type networks (all the connections between the neurons are forward), where as shown in Figure 2.4 all the neurons of a particular level are connected to all the neurons of the next level. As it can be observed in the figure this network has one input layer, one output layer and can have none or several hidden layers. Depending on the number of neurons (or internal levels) of the MLP it is possible to approximate more complex functions. The working principle of each neuron is the same described for the SLP and can is represented in Equation 2.8.

$$\mathrm{o_j^k} = f_j^k\left(\mathrm{net_j^k}\right), \quad \mathrm{net_j^k} = \sum_{i=1}^{N^{k-1}} w_{ij}^{k-1} \mathrm{o_i^{k-1}} \tag{2.8}$$

The learning in a MLP can be carried out using for example the Back Propagation (BP) algorithm. BP prevents the delta rule, the gradient descent learning rule for updating the weights of the inputs $\Delta_{w_{ij}}$, and minimises its derivation, the Least Means Squares (LMS) function error shown in Equation 2.9.

$$\mathrm{E} = \frac{1}{2}\sum_{i=1}^{N^L} (\mathrm{t_j} - \mathrm{o_j^L})^2 \tag{2.9}$$

The weights in an MLP are updated following the updated equation of the SLP described by Equation 2.10. As shown in the equation the compute of δ differs depending on the type of neuron, and ot is different for the output or from an intermediate layer (hidden).

$$\begin{aligned} \Delta w_{ij}^k &= \eta \delta_j^{k+1} o_i^k, \quad \eta = \text{learning coefficient} \\ \delta_j^L &= (t_j - o_j^L) o_j^L (1 - o_j^L), \quad \text{output layer neuron} \\ \delta_j^k &= o_j^k (1 - o_j^k) \sum_{l=1}^{N^{k+1}} \delta_l^{k+1} w_{jl}^k, \quad \text{hidden layer neuron} \end{aligned} \tag{2.10}$$

Among the advantages of using a MLP are that they are computationally efficient as they can easily be parallelised. In addition, some models with a finite number of patterns are able to approximate any discriminant function with high accuracy (universal approximation). However, they can not easily be scaled and the convergence can be slow.

2.6 Multiple classifier systems

In order to achieve models with less variance (caused by the training set) and bias (classification error caused by the algorithm), ensemble methods such as like bagging (Breiman 1996) and boosting (Schapire 1999) are widely used combining the output of different models (multiple classifiers). In the next lines we summarise two ensemble methods widely used, bagging and boosting, which accomplished simple (equally weighted) and weighted vote procedures between several classifiers respectively to make the final decision.

2.6.1 Bagging

The first bagging algorithm named from Bootstrap aggregating was proposed Breiman in 1996 (Breiman 1996) and consists of building classifiers based on boostrap samples (with replacement) where the final decision is taken according to the majority vote among all the individual classifiers. Algorithm 4 summarises the bagging procedure.

The main strengths of a bagging procedure are that it reduces the variance caused by the training and that it provides high accuracy, contributing to alleviate the over-fitting problem and improving the stability of the model. In addition, the independence of the models being combined allows to apply parallelisation techniques can if required. However, the model looses the explanation capabilities.

Algorithm 4 Bagging procedure.

```
S=training set;
T=number of boostrap samples ($\mathrm{B_k}, \quad 1 \leq \mathrm{k} \leq \mathrm{T}$))
L=inductor algorithm;
$\mathrm{C_k}$=classifier built with the sample $\mathrm{B_k}$;
$\mathrm{C}^*$= final classifier
for $k$ in 1 to $T$ do
    $\mathrm{B_k}$ = boostrap sample of S;
    $\mathrm{C_k}$=L($\mathrm{B_k}$);
end for
$\mathrm{C}^*(\mathrm{x}) = \operatorname*{argmax}_{w_i \in \mathcal{C}} \sum_{k:C_k(x)=w_i} 1$;
```

2.6.2 Boosting

The boosting algorithm was proposed by Schapire in 1990 (Schapire 1999) aiming to reinforce the performance of weak classifiers. Six years later Freund and Schapire presented the AdaBoost (Adaptive Boosting) algorithm (Freund and Schapire 1996), which has been used in this dissertation. In this algorithm T classifiers are built sequentially and each pattern of the sample is assigned a particular weight which vary in each step depending on whether the pattern is correctly classified or not. The final decision is the result of a weighted voting between all individual classifiers. Algorithm 5 shows the procedure used by AdaBoost.

As described in Algorithm 5 AdaBoost starts assigning equal weights to all the instances in the training data and then, uses a particular learning algorithm to build a classifier for this data. At this point based on the output of the classifier, the instances are assigned new weights (re-weighting) so that correctly classified instances (easy) are lowly weighted and the missclassified ones (hard) are highly weighted. This process is repeated several times and when the error on the weighted training data is higher than 0.5 or equal to 0, then, the boosting procedure deletes the current classifier and does not perform any more iterations. The logarithmic expression $\log(1-\epsilon_k)/\epsilon_k$ enables the correctly classified instances to be highly weighted and vice versa. The weights of all the classifiers that voted for a particular class are summed and the one with the highest total is chosen in the end ($\mathrm{argmax}_{w_i \in \mathcal{C}} \sum weightedvotes$).

Among the advantages of using a boosting procedure we can mention that it reduces the classification error caused by the algorithm (bias). However, unlike in bagging here we can not parallelise the computations to combine the model and thus, coordinating sequentially several complex models can be computationally expensive.

Algorithm 5 Boosting procedure.

```
S=training set;
T=number of individual classifiers built based on S weighted (S');
L=inductor algorithm;
$C_k$=classifier built with the sample $B_k$;
$C^*$= final classifier
for $k$ in 1 to $T$ do
    $C_k = L(S')$ ;
    $\epsilon_k = \frac{1}{n} \sum_{x_j \in S' : C_k(x_j) \neq w_i} weight(x)$; // weighted error in the training set
    if $\epsilon_k > \frac{1}{2}$; then
        Terminate;
    end if
    $N_{S'}$=size($S'$);
    for $x_j$ in 1 to $N_{S'}$ do
        if $C_k(x_j) \neq w_i$; then
            $weight(x_j) = \frac{weight(x_j)}{2\epsilon_k}$;
        else
            $weight(x_j) = \frac{weight(x_j)}{2(1-\epsilon_k)}$;
        end if
    end for
end for
$C^*(x) = \operatorname*{argmax}_{w_i \in \mathcal{C}} \sum_{k: C_k(x) = w_i} \log \frac{(1-\epsilon_k)}{\epsilon_k}$; // most voted class (weighted)
```

2.7 Validation

Validation in the supervised learning context evaluates the generalisation capacity of a predictive model on an independent data set. In particular, predictive models learn to perform predictions using the training dataset and then, their learning ability is tested on new data named test dataset. To accomplish a suitable validation it is important to separate training and test datasets, carrying out a hold-out procedure or performing a cross validation. The hold-out procedure splits the dataset into training and test disjoint sets but the performance of the classifier may be biased by the sets of data selected. In case limited data are available and the split is not possible, a cross-validation procedure can be used. Therefore, preferably random sub-sampling (repeated hold-out) is used, which rather than generating a single training/test partition, it splits the dataset several times by randomly selected instances in both types of sets so that the learning capacity of the model is given in terms of average values obtained in all the partitions.

2.7.1 K-fold cross-validation

A K-fold cross-validation divides the data into K number of partitions or folds of the same size. In order to predict the error rate of a learning algorithm usually an stratified 10-fold cross-validation is performed. This way, the dataset is randomly divided into 10 parts, trying to preserve the same original proportion of class instances in all of them. Then, 10 performance estimations are carried out, keeping nine of the parts for training and one for test. This way, the learning is carried out 10 times in each of them, and the total error is computed as the average of the 10 learning processes. The existing literature shows that using 10 folds is the best approach to estimate the error (Kuhn and Johnson 2013), although 5-fold or 20-fold cross-validations are also suitable (Witten and E. Frank 2005). In order to obtain reliable results, several runs of 10-fold cross-validation are usually required.

2.7.2 Leave-one-out cross-validation

Alternatively, leave-one-out cross-validation can be used, which is a fold cross validation with the same number of folds as instances has the dataset. Each of the n learning processes are carried out leaving one instance out (test), that is, with $n-1$ training instances. The final error is computed as the average of the n learning processes. Therefore, this process is computationally expensive as the n learning processes must be executed, which may not be possible in large datasets but can be very effective for small datasets. In addition, as just one instance is used as test in each learning, this procedure does ensure an stratification, which can be a critical problem for balanced binary class datasets.

2.8 Performance metrics

The majority of metrics used to evaluate the performance of classifiers need to be evaluated are based on the confusion matrix (Ron Kohavi and Provost 1998), where the number of instances which belong to each class are represented in rows and the number of instances classified as belonging to the each class are represented in columns.

In binary problems (two classes), the minority and majority classes are referred as positive and negative respectively and the confusion matrix is of 2x2 dimensions, providing therefore the four next listed distributed as shown in Table 2.1:

- **True Positive (TP)**: number of positive instances classified as positive.
- **True Negative (TN)**: number of negative instances classified as negative.
- **False Positive (FP)**: number of negative instances classified as positive.
- **False Negative (FN)**: number of positive instances classified negative.

		Prediction	
		P	N
Reality	T	TP	FN
	F	FP	TN

Table 2.1: Confusion matrix

Equations 2.11 and 2.12 show respectively the most widely known metrics to evaluate the performance of a classifier computed based on the values mentioned above: accuracy (Acc), percentage of correctly classified instances, and error rate (Err), percentage of wrongly classified instances.

$$Acc = \frac{TP + TN}{TP + FP + FN + TN} = \frac{T}{T + F} \tag{2.11}$$

$$Err = \frac{FP + FN}{TP + FP + FN + TN} = \frac{F}{T + F} = 1 - Acc \tag{2.12}$$

In many problems, false positive are critical (e.g false terrorism accusation) whereas in other problems false negative have more importance (e.g dismissing a correct tumour diagnosis). Therefore, alternative metrics considering different types of errors were proposed to evaluate the performance of the classifiers, such as the ones shown in Equations 2.13, 2.14 and 2.15: precision (Pr), percentage of instances that are actually positive among those who have been classified as such ; recall (Re), percentage of correctly classified positive instances; F-measure (Fm), harmonic mean of the precision and the recall.

$$Pr = \frac{TP}{TP + FP} \tag{2.13}$$

$$Re = \frac{TP}{TP + FN} \tag{2.14}$$

$$Fm = 2 \cdot \frac{Pr \cdot Re}{(Pr + Re)} \tag{2.15}$$

In addition there are some methods that graphically combine two of the above mention metrics over threshold values. The procedure of such methods is carried out by first, performing a test to obtain the probability of being a member of the positive class or the negative class for each instance and second, by fixing a threshold that enables to determine whether each instance is classified positively or negatively. Analysing these plots the best threshold value is selected keeping the crucial error to zero and the other as low as possible. In this dissertation the Area Under ROC (Receiver Operating Characteristic) Curve (AUC) graphic has been used, which graphically represents the recall (also named True Positive Rate) in the X axis and the False Positive Rate ($\frac{FP}{FP+NP}$) in the Y axis. Ideally, the area under the curve in a classifier would be 1, thus, the classifier with higher AUC is usually considered as the best classifier.

2.9 Statistical tests

The goal of statistical tests is to determine whether significant differences exists between the performances of different classifiers or other types of procedures (e.g indexes, CVIs...). Thus, the initial or null hypothesis of such tests is that the performances are not significantly different and accordingly, rejecting this hypothesis implies that significant differences exist.

Two types of statistical tests can be distinguished, parametric which assume that the data follow a particular probability distribution and infers its features and non-parametric, which do made such assumptions and use order statistics based on ranks of observations. In case the assumptions of parametric tests are correct, they can provide more accurate estimations and are statistically more powerful but in the opposite case they can be misleading.

In this dissertation on the one hand, the parametric statistical Student's t-test (Gosset 1908) was used in order to determine whether significant differences existed between the performance of two classifiers. In particular, in the Student's t-test the null hypothesis is that the statistic follows a Student's t-distribution (continuous probability distribution aroused when estimating the mean of a normally distributed population).

On the other hand, the non-parametric Kendall's rank correlation test was used in order analyse whether significant differences existed in the rankings provided by different indexes. In particular this test is used to compare the correlation on ranking type data being the tau-test a non-parametric test for statistical dependence based on the tau coefficient.

In the next lines we briefly describe both statistical tests.

2.9.1 Student t-test

In order to compare two population samples X_1 and X_2 of n instances with the Student t-test (Gosset 1908) the t statistic is computed as shown in Equation 2.16, where s_p is the pooled standard deviation for $n = n_1 = n_2$ (populations of equal sizes) and s^2x_1 and s^2x_1 are unbiased estimations of the variances of the two samples.

$$\begin{aligned} t &= \frac{\overline{\mathrm{X}}_1\overline{\mathrm{X}}_2}{sp\sqrt{\frac{2}{n}}}, \\ sp &= \sqrt{\frac{s^2x_1 + s^2x_2}{2}} \end{aligned} \tag{2.16}$$

The null hypothesis in this case is that the population means from the two groups are equal. Using the tables of the t-distribution to the resulting t value for the t_{n-1} distribution the p-value for the paired t-test can be obtained.

2.9.2 Kendall test

Kendall's tau (Kendall 1938) also named Kendall's correlation coefficient, τ, measures the rank correlation, that is, the similarity between different orderings of the same dataset. In particular, two ordinal variables are pairwise observed computing their correlation, which would be high if observations have equal ranks ($tau = 1$) and low in the opposite case ($tau = -1$). In Equation 2.17 the computed τ is described, which is the ratio between the difference of concordant (n_c) and non concordant (n_d) pairs and the binomial coefficient $n_0 = \binom{n}{2} = \frac{n(n-1)}{2}$ for the number of ways to choose two items from n items.

$$\tau = \frac{n_c - n_d}{n_0} \tag{2.17}$$

For data with excessive number of ties, Kendall's τ_b shown in Equation 2.18 is computed. In this case, the null hypothesis will be that the pairs are not correlated $\tau b = 0$ and the alternative hypothesis that they are correlated $\tau_b \neq 0$. If in the pairwise correlation test we obtain a p-value higher than a significant level $\alpha = 0.05$, the null hypothesis will not be rejected meaning that both variables are not correlated at 0.05 significance level, and the alternative hypothesis (variables are correlated) will be accepted in case $pvalue < 0.05$.

$$\begin{gathered}\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \quad n_1 = \sum_i \frac{t_i(t_i - 1)}{2}, \quad n_2 = \sum_j \frac{u_j(u_j - 1)}{2} \\ t_i : \text{number of tied values in the} i^{th} \text{group for the first quantity} \\ u_j : \text{number of tied values in the} j^{th} \text{group for the second quantity}\end{gathered} \tag{2.18}$$

Chapter 3

Unsupervised learning

In contrast to supervised learning, unsupervised learning deals with finding underlying structures of unlabelled data (lack of a dependent variable). One of the main methods used in unsupervised learning is clustering where similar instances of a dataset are grouped in the same cluster and dissimilar ones in different clusters, based on a particular similarity metric (e.g Euclidean distance). The clustering procedure consist of four steps (Xu and Wunsch 2008): feature extraction (easy to interpret, representative , not redundant etc.), selection of a clustering algorithm that best fits the data, clustering validation (evaluation of clustering structure) and result interpretation.

In this dissertation two main types of clustering algorithms were used: hierarchical, which provide a hierarchy of the partitions in a graph (dendrogram) and partitional, which provide a single partition of the data. The specific algorithms used are listed next:

- Hierarchical clustering: SAHN (Sneath and Sokal 1973) with average-linkage (Jain and Dubes 1988) and with Ward (Ward 1963) criteria

- Partitional clustering: k-means (Lloyd 1982) and PAM (k-medoids) (Kaufman and P. Rousseeuw 1990)

Not having labelled data makes clustering validation one of the main challenges of this area. In order to evaluate the suitability of the partition obtained, three main types of validation techniques can be distinguished: external, when the correct partition exists and the resulting one can be evaluated by comparison; internal, when the correct partition is not available and the compactness and separation of the clusters is measured to evaluate the partitions; relative, which combines external and internal validations. In this dissertation internal validation has been studied, more concretely, defining several voting approaches between Cluster Validity Indexes (CVIs) previously analysed in one of the most extensive comparative works existing in the literature (Arbelaitz et al. 2013b).

In the following sections the three types of algorithms mentioned (hierarchical and partitional) will be described first and then a summary of the internal validation indexes (CVIs) will be provided. For the majority of the procedures R (R Core Team 2017), the free software environment for statistical computing and graphics, was used.

3.1 Hierarchical clustering

Hierarchical algorithms produce a hierarchical structure of clusters usually in a dendrogram type diagram (see Figure 3.1), where instances at low levels are more tightly clustered than those joined at higher levels (Witten and E. Frank 2005). As shown in the figure (right) in the y-axis and x-axis of the dendrogram the similarity measure and the clustered instances are represented respectively. In this case, if the dendrogram is horizontally cut where the dashed line we obtain three clusters (k=3), marked as $C1$, $C2$ and $C3$, being the instances inside the second one (D,E) more similar between them (compact) than the ones inside the other two.

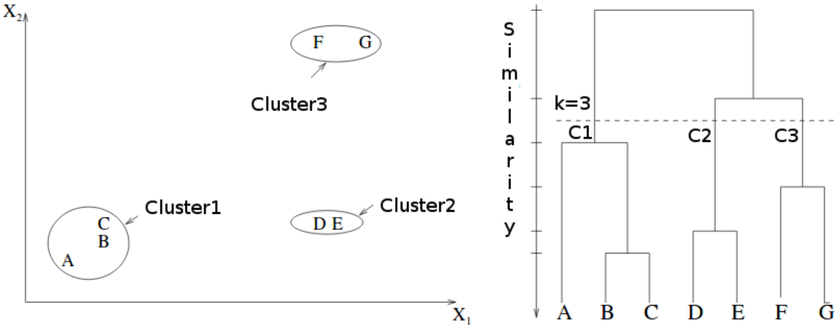


Figure 3.1: Dendrogram (left) obtained from a hierarchical clustering algorithm applied to seven instances of a two dimensional dataset (right) (Jain et al. 1999).

The general taxonomy divides the hierarchical algorithms into two main groups (Hastie et al. 2009) :

- Agglomerative (bottom-up): the starting point is in the bottom and at each level a selected pair of clusters are recursively merged into a single one so that the grouping at the next higher level has one less cluster. The selection of the pair of clusters that will be merged is done according to the smallest inter-group distance.

- Divisive (bottom-up): the starting point is in the top and at each level one of the existing clusters is split into two new clusters. The split decision is made so that the two new groups have the largest inter-group distance.

In this dissertation SAHN (Sneath and Sokal 1973) agglomerative type clustering algorithm has been used, which is described in the next lines.

3.1.1 Hierarchical agglomerative clustering

SAHN is an acronym to designate clustering methods that are Sequential, Agglomerative, Hierarchical and Non-overlapping (Sneath and Sokal 1973). In these clustering methods the distance between each pair of instances in the set of instances to be clustered must be quantitatively specified, using for example a distance matrix. The number of rows (i) and columns (j) of this matrix is given by the number of instances (N) of the dataset and in each cell the distance between each instance pair is provided.

Algorithm 6 summarises the procedure of SAHN, where initially each of the N instances of the training set (S) in one cluster (a partition S_1 of N clusters) and then the distance matrix (M_1) of the N clusters is computed. In the third step, the two nearest clusters are joined (i, j) in the same cluster (h) so that the new partition has one less cluster and the distance matrix is accordingly updated. This procedure is repeated until a partition with two clusters is obtained and the corresponding hierarchy can be provided.

Algorithm 6 SAHN algorithm (Day and Edelsbrunner 1984).

```
$S$=Training set with $N$ instances;
$S_1$=Partition with $N$ clusters one for each training each;
$M_1$=$D(S_1$ (Distance matrix of $S_1$);
for $m$ in $N$ to 2 do
    Find the nearest two clusters $(i, j)$ in $M_1$;
    Replace the two clusters ($i$ and $j$) by an agglomerated cluster $h$.
    Update $M$ by computing the distance between $h$ and the rest of the clusters $m - 1$;
end for
Output: hierarchy of clusters ($S_1$, $S_2$, ..., $S_N$);
```

In order to measure the distance between clusters different methods named linkage criteria can be used, next. In particular we used three of the most popular ones:

- Single-linkage: according to this criteria the distance between two clusters $C1 = \{c1_i, 1 \leq i \leq N\}$ and $C2 = \{c2_j, 1 \leq j \leq M\}$ is computed as the distance between the two closest instances of the two clusters (see Equation 3.1).

$$D(C1, C2) = \min_{c1_i \in C1, c2_j \in C2} d(c1_i, c2_j), \quad 1 \leq i \leq N, \quad 1 \leq j \leq M \tag{3.1}$$

- Complete-linkage: the distance between clusters $C1 = \{c1_i, 1 \leq i \leq N\}$ and $C2 = \{c2_j, 1 \leq j \leq M\}$ is computed as the distance between the two farthest instances of the two clusters (see Equation 3.2).

$$D(C1, C2) = \max_{c1_i \in C1, c2_j \in C2} d(c1_i, c2_j), \quad 1 \leq i \leq N, \quad 1 \leq j \leq M \tag{3.2}$$

- Average-linkage (Rédei 2008): according to this criteria the distance between two clusters $C1 = \{c1_i, 1 \leq i \leq N\}$ and $C2 = \{c2_j, 1 \leq j \leq M\}$ is computed as the average distance between all the instances of the first cluster ($c1_i$) and all the ones belonging to the second cluster ($c2_j$) (see Equation 3.3):

$$D(C1, C2) = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} d(c1_i, c2_j) \tag{3.3}$$

- Ward-linkage (Ward 1963): this criteria uses the sum of square errors (SSE), also known as error sum of squares (RSS), as an objective function to measure the distance between the clusters. If Ward-linkage is combined with SAHN, in each step of the procedure the pair of clusters to be merged (e.g $C1$ and $C2$ shown above) will be the one with slowest sum of square error (see Equation 3.4), that is, those with the minimum increase in total within-cluster variance after joining ($C3$):

$$\begin{gathered} SSE(C3) = \sum_{k=1}^{MN} (c3_k - \overline{c3})^2, \quad \overline{c3} = \frac{1}{NM} \sum_{k=1}^{MN} c3_k, \\ C3 = \{c3_k, 1 \leq k \leq MN\} \end{gathered} \tag{3.4}$$

The main strengths of SAHN and similar hierarchical algorithms is the easiness to interpret the results provided in a dendrogram and the fact that no information about the number of clusters is required beforehand. However, among the weaknesses we find their bad performance with large and noisy datasets and those with missing values or outliers and the difficulty to determine the right number of clusters in complex dendrograms.

3.2 Partitional clustering

Partitional clustering methods provide a single partition dividing all the instances of the training set into disjoint clusters. This kind of clustering is more suitable for large-datasets contexts where a dendrogram can be computationally very expensive.

In partitional clustering the clusters are obtained by optimising a criteria function locally (for a subset of instances) or globally (for the whole set of instances) defined (Jain et al. 1999). Usually, the employed criteria is the squared error shown in Equation 3.4. As we are describing in the following paragraphs the k-means (Lloyd 1982) popular algorithm uses in this criteria.

3.2.1 K-means

K-means (Lloyd 1982) is one of the most widely used algorithms in the literature (Wu et al. 2008) and its working principle consists of minimising the sum of square errors. As represented in Algorithm 7 the procedure of k-means starts defining the number of clusters desired, K, and continues by randomly selecting K instances as cluster centroids (geometrical centres or average between all the instances in the cluster) and assigning the instances to the nearest centroids of the clusters. Then, the centroids of the clusters are newly computed and instances are reassigned to the nearest clusters based on the new centroids. This procedure is iterated until the same instances are assigned to each cluster, that is, when the centroids are stabilised, or a particular number of iterations is achieved.

Algorithm 7 K-means algorithm.

```
Training data=$X$=$\{x_i \quad 1 \leq i \leq N\}$;
Select the number of clusters: $K \leq N$;
Randomly select $K$ centroids: $C$= $\{c_j, \quad 1 \leq j \leq K\}$;
repeat
    Assign the instances to the closest cluster centroids:
    for $i$ in 1 to $N$
    closest_c($x_i$)=$\underset{1\leq j\leq K}{\arg\min}\, d(x_i, c_j)$ ;
    end for;
    Update the $K$ cluster centroids $C$:
    for $j$ in 1 to $K$
    $c_j$=$mean(x_i|closest\_c(x_i) = j)$;
    end for;
until Cluster centroids stop changing or maximum number of iterations achieved.
```

K-means presents some advantages compared to hierarchical approaches, such as its implementation simplicity and its good and fast performance for large datasets. On the other hand, its mayor disadvantages are the difficulty of determining the number of clusters k beforehand and its sensitivity to scale and initialisation (the results for original and normalised data can totally differ).

3.2.2 K-medoids

K-medoids is a variant of k-means algorithm which instead of using the centroid as the representative instance of a cluster (C), the medoid of the cluster, $M(C)$ is employed. As shown in Equation 3.5 the medoid of a cluster is computed as the instance (x_j) with a minimum average distance to all the instances in the cluster (x_i). This is a key-difference between both algorithms because the fact that the medoid used in k-medoids is a real instance, makes the algorithm more robust against outliers which negatively affect the centroids used in k-means. In addition, the new medoids can be directly picked up from a distance matrix in contrast to new centroids which must be computed again in each step.

$$M(C) = \arg\min_{x_j \in C} \sum_{i=1}^{N} d(c_i, c_j) \quad C = \{x_i, \quad 1 \leq i \leq N\} \tag{3.5}$$

Algorithm 8 summarises the procedure of k-medoids.

Algorithm 8 K-medoids algorithm.

```
Training data=$X$=$\{x_i \quad 1 \leq i \leq N\}$;
Select the number of clusters: $K \leq N$;
Randomly select $K$ medoids: $M$= $\{m_j, \quad m_j \in X, \quad 1 \leq j \leq K\}$;
repeat
    Assign the instances to the closest cluster medoids:
    for $i$ in 1 to $N$
    closest_$m(x_i)$=$\arg\min_{1 \leq j \leq K} d(x_i, m_j)$ ;
    end for;
    Update the $K$ cluster medoids $M$:
    for $j$ in 1 to $K$
    $m_j$=$\arg\min(x_i|\text{closest\_}m(x_i) = j)$;
    end for;
until Cluster medoids stop changing or maximum number of iterations achieved.
```

The Partitioning Around Medoids (PAM) algorithm (Kaufman and P. Rousseeuw 1990) is an implementation k-medoids algorithms. This algorithm has two phases (Li et al. 2017) which are described next described:

- **Build phase**: a set of K instances are selected as medoids for an initial partition S (set of selected instances). If the sum of the distances between a particular instance and the rest of them is minimum, then that instance is selected as the first medoid, repeating the process until K medoids are obtained. In particular, for all the unselected instances i ($i \in U$) candidates to be included in the set of selected instances S a total gain is computed as shown in Equation 3.6. In the equation, each time j is the instance of the unselected set of instances without i ($j \in U - i$) and the distance between j and the closest selected instance (S) is computed (D_j). If $D_j > d(i, j)$, then the instance will increase the quality of the cluster.

$$g_i = \sum_{j \in U} \max\{D_j - d(j, i)\} \tag{3.6}$$

After computing all the total gain of the set of unselected instances (U), the instance that provides the highest gain, h, is included in the selection set and excluded from the unselected set ($S = S \cup \{h\}, U = U - \{h\}$). The process is repeated until K instances are selected.

- **Swap phase**: instances not selected as medoids ($u \in U$) are exchanged aiming to improve the quality of the cluster. In particular all possible combinations of pairs of instances selected and not selected as medoids ($s, u \in SxU$) are analysed by measuring the effect of each swap T_{su} according to Equation 3.7 and notation described below.

$$\begin{gathered} T_{su} = sum\{K_{tsu} | t \in U\} \\ K_{tsu} = \begin{cases} \min\{d(t, u) - D_t, 0\}, & \text{if } d(ts) > D_t; \\ \min\{d(t, u) - E_t\} - D_t, & \text{if } d(ts) = D_t; \end{cases} \end{gathered} \tag{3.7}$$

  - K_{tsu}: contribution of each instance t in U to the swap of s and u.
  - D_t: dissimilarity between t and the closest object in S.
  - E_t: dissimilarity between t and the second closest object in S.

  In particular, given a pair of instances (s, u) with the minimum contribution T_{su}, if its value is lower than 0, a swap will be carried out whereas the in opposite case, a halt will be carried out as no quality improvement happened. This process is repeated until the quality of the cluster is the best.

One of the biggest disadvantages of PAM is that its that is computationally expensive because each medoid is compared with the whole dataset in each iteration making difficult to deal with large datasets. On the other hand, the use of medoids allows working with many different type of distances including those employed in sequential data, e.g. edit distance Levenshtein 1966 defined as the minimum number of operations (insertion, deletion or substitution) required to transform one sequence into the other.

3.3 Clustering validation

3.3.1 Cluster Validity Indexes (CVIs)

The goal of clustering validation is to evaluate the quality of the output partition obtained in a clustering procedure. In this dissertation we focused on Cluster Validity Indexes (CVIs), which quantify the quality of a partition by measuring the compactness and separation of the clusters. In particular we focused on an extensive comparative study of CVIs performed by Arbelaitz et al. 2013b which compared a total of 30 CVIs and proposed different decision fusion strategies using them.

The reference work (Arbelaitz et al. 2013b) is focused on CVIs that can be easily evaluated by the usual methodologies and avoided those that could lead to confusion due to the need for a subjective decision by the experimenter. Most of the indices estimate the cluster cohesion (within or intra-variance) and the cluster separation (between or inter-variance) and combine them to compute a quality measure. The combination is performed by a division (ratio-type indices) or a sum (summation-type indices) (Kim and Ramakrishna 2005). For each index the authors provided an abbreviation that helps interpreting the result. In addition each an downward arrow (↓) or upward arrow (↑) is added to each abbreviation to indicate that a lower value of that index means a “better” partition or the opposite respectively. Next lines we describe the 30 CVIs used in this work (Arbelaitz et al. 2013b):

- **Dunn index** (D↑) (Dunn 1973): This index has many variants and some of them will be described next. It is a ratio-type index where the cohesion is estimated by the nearest neighbour distance and the separation by the maximum cluster diameter. The original index is defined as shown in Equation 3.8.

$$\mathrm{D}(C) = \frac{\min\limits_{c_k \in C}\{\min\limits_{c_l \in C \setminus c_k}\{\delta(c_k, c_l)\}}{\max\limits_{c_k \in C}\{\Delta(c_k)\}} \tag{3.8}$$

where

$$\delta(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_l}\{\mathrm{d_e}(x_i, x_j)\}, \tag{3.9}$$

$$\Delta(c_k) = \max_{x_i, x_j \in c_k}\{\mathrm{d_e}(x_i, x_j)\}. \tag{3.10}$$

- **Calinski-Harabasz** (CH↑) (Caliński and Harabasz 1974): This index obtained the best results in the work of Milligan and Cooper (Milligan and Cooper 1985). It is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid. The separation is based on the distance from the centroids to the global centroid, as defined in Section sec:notation. Equation 3.11 shows how CH index is computed.

$$\mathrm{CH}(C) = \frac{N-K}{K-1} \frac{\sum\limits_{c_k \in C} |c_k| \, \mathrm{d_e}(\bar{c_k}, \bar{X})}{\sum\limits_{c_k \in C} \sum\limits_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k})}. \tag{3.11}$$

- **Gamma index** (G↓) (Baker and L.J. Hubert 1975): The Gamma index is an adaptation of Goodman and Kruskal's Gamma index and can be described as shown in Equation 3.12.

$$\mathrm{G}(C) = \frac{\sum\limits_{c_k \in C} \sum\limits_{x_i, x_j \in c_k} \mathrm{dl}(x_i, x_j)}{n_w\left(\binom{N}{2} - n_w\right)} \tag{3.12}$$

where $\mathrm{dl}(x_i, x_j)$ denotes the number of all object pairs in X, namely x_k and x_l, that fulfil two conditions: (a) x_k and x_l are in different clusters, and (b) $\mathrm{d_e}(x_k, x_l) < \mathrm{d_e}(x_i, x_j)$. In this case the denominator is just a normalisation factor.

- **C-Index** (CI↓) (L.J. Hubert and Levin 1976): This index is a type of normalised cohesion estimator and its definition is provided by Equation 3.13.

$$\mathrm{CI}(C) = \frac{\mathrm{S}(C) - \mathrm{S}_{min}(C)}{\mathrm{S}_{max}(C) - \mathrm{S}_{min}(C)} \tag{3.13}$$

where

$$\mathrm{S}(C) = \sum_{c_k \in C} \sum_{x_i, x_j \in c_k} \mathrm{d_e}(x_i, x_j), \tag{3.14}$$

$$\mathrm{S}_{min}(C) = \sum \min_{x_i, x_j \in X} (n_w) \{\mathrm{d_e}(x_i, x_j)\}, \tag{3.15}$$

$$\mathrm{S}_{max}(C) = \sum \max_{x_i, x_j \in X} (n_w) \{\mathrm{d_e}(x_i, x_j)\}. \tag{3.16}$$

- **Davies-Bouldin** index (DB↓) (Davies and Bouldin 1979): This is probably one of the most used indices in CVI comparison studies. It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. DB index is computed as shown in Equation 3.17.

$$\mathrm{DB}(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{\mathrm{S}(c_k) + \mathrm{S}(c_l)}{\mathrm{d_e}(\bar{c_k}, \bar{c_l})} \right\} \tag{3.17}$$

where

$$\mathrm{S}(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k}). \tag{3.18}$$

- **Silhouette index** (Sil↑) (P.J. Rousseeuw 1987): This index is a normalised summation-type index. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbour distance. The definition of Silhouette is provided by Equation 3.19.

  $$\mathrm{Sil}(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{\mathrm{b}(x_i, c_k) - \mathrm{a}(x_i, c_k)}{\max\{\mathrm{a}(x_i, c_k), \mathrm{b}(x_i, c_k)\}} \tag{3.19}$$

  where

  $$\mathrm{a}(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \mathrm{d_e}(x_i, x_j), \tag{3.20}$$

  $$\mathrm{b}(x_i, c_k) = \min_{c_l \in C \setminus c_k} \{ \frac{1}{|c_l|} \sum_{x_j \in c_l} \mathrm{d_e}(x_i, x_j) \}. \tag{3.21}$$

- **Graph theory based Dunn and Davies-Bouldin variations** ($\mathrm{D^{MST}}$↑, $\mathrm{D^{RNG}}$↑, $\mathrm{D^{GG}}$↑, $\mathrm{DB^{MST}}$↓, $\mathrm{DB^{RNG}}$↓, $\mathrm{DB^{GG}}$↓) (Pal and Biswas 1997): These indices are variations of Dunn and Davies-Bouldin. The variation affects how the cohesion estimators are computed –$\Delta(c_k)$ for the Dunn index and $\mathrm{S}(c_k)$ for the Davies-Bouldin index.

  For each of the 3 versions –MST, RNG and GG– these 2 functions are computed in the same way. First, a particular type of graph is computed for c_k, taking the objects in the cluster as vertices and the distance between objects as the weight of each edge. Then the largest weight is taken as the value for $\Delta(c_k)$ and $\mathrm{S}(c_k)$. The difference between the 3 variants comes from the selected graph type. For MST a Minimum Spanning Tree is built, for RNG a Relative Neighbourhood Graph and for GG a Gabriel Graph.

- **Generalised Dunn indices** (gD31↑, gD41↑, gD51↑, gD33↑, gD43↑, gD53↑) (Bezdek and Pal 1998): All the variations are a combination of three variants of δ –separation estimator– and two variations of Δ –cohesion estimator. Actually, Bezdek and Pal (Bezdek and Pal 1998) proposed 6×3 variants –including the original index–, but we selected those proposals that showed the best results. Therefore we used the variants 3, 4 and 5 for δ and 1 and 3 for Δ (see Equations 3.22 to 3.26).

  $$\delta^3(c_k, c_l) = \frac{1}{|c_k||c_l|} \sum_{x_i \in c_k} \sum_{x_j \in c_l} \mathrm{d_e}\, x_i, x_j, \tag{3.22}$$

$$\delta^4(c_k, c_l) = \mathrm{d_e}(\bar{c_k}, \bar{c_l}), \tag{3.23}$$

$$\delta^5(c_k, c_l) = \frac{1}{|c_k| + |c_l|} \left(\sum_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k}) + \sum_{x_j \in c_l} \mathrm{d_e}(x_j, \bar{c_l}) \right) \tag{3.24}$$

and

$$\Delta^1(c_k) = \Delta(c_k), \tag{3.25}$$

$$\Delta^3(c_k) = \frac{2}{|c_k|} \sum_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k}). \tag{3.26}$$

- **SDbw index** (SDbw$\downarrow$) (Halkidi and Vazirgiannis 2001): This is a ratio-type index that has a more complex formulation based on the euclidean norm $||x|| = (x^T x)^{1/2}$, the standard deviation of a set of objects,

$$\sigma(X) = \frac{1}{|X|} \sum_{x_i \in X} (x_i - \bar{x})^2 \tag{3.27}$$

and the standard deviation of a partition,

$$\mathrm{stdev}(C) = \frac{1}{K} \sqrt{\sum_{c_k \in C} ||\sigma(c_k)||}. \tag{3.28}$$

The SDbw index is defined as shown in Equation 3.29.

$$\begin{aligned} \mathrm{SDbw}(C) = & \frac{1}{K} \sum_{c_k \in C} \frac{||\sigma(c_k)||}{||\sigma(X)||} \\ & + \frac{1}{K(K-1)} \sum_{c_k \in C} \sum_{c_l \in C \setminus c_k} \frac{\mathrm{den}(c_k, c_l)}{\max\{\mathrm{den}(c_k), \mathrm{den}(c_l)\}} \end{aligned} \tag{3.29}$$

where

$$\mathrm{den}(c_k) = \sum_{x_i \in c_k} \mathrm{f}(x_i, \bar{c_k}), \tag{3.30}$$

$$\mathrm{den}(c_k, c_l) = \sum_{x_i \in c_k \cup c_l} \mathrm{f}(x_i, \frac{\bar{c_k} + \bar{c_l}}{2}) \tag{3.31}$$

and

$$\mathrm{f}(x_i, c_k) = \begin{cases} 0 & \text{if } \mathrm{d_e}(x_i, \bar{c_k}) > \mathrm{stdev}(C) \\ 1 & \text{otherwise.} \end{cases} \tag{3.32}$$

- **CS index** (CS↓) (Chou et al. 2004): This index was proposed in the image compression environment, but can be extended to any other environment. It is a ratio-type index that estimates the cohesion by the cluster diameters and the separation by the nearest neighbour distance. Equation 3.33 shows how the CS index is computed.

$$\mathrm{CS}(C) = \frac{\sum_{c_k \in C} \left\{ \frac{1}{|c_k|} \sum_{x_i \in c_k} \max_{x_j \in c_k} \{\mathrm{d_e}(x_i, x_j)\} \right\}}{\sum_{c_k \in C} \min_{c_l \in C \setminus c_k} \{\mathrm{d_e}(\bar{c_k}, \bar{c_l})\}}. \tag{3.33}$$

- **Davies-Bouldin*** (DB* ↓) (Kim and Ramakrishna 2005): This variation of the Davies-Bouldin index was proposed together with an interesting discussion about different types of CVIs. The definition of this index is provided in Equation 3.34.

$$\mathrm{DB}^*(C) = \frac{1}{K} \sum_{c_k \in C} \frac{\max\limits_{c_l \in C \setminus c_k} \{\mathrm{S}(c_k) + \mathrm{S}(c_l)\}}{\min\limits_{c_l \in C \setminus c_k} \{\mathrm{d_e}(\bar{c_k}, \bar{c_l})\}}. \tag{3.34}$$

- **Score Function** (SF↑) (Saitta et al. 2007a): This is a summation-type index where the separation is measured based on the distance from the cluster centroids to the global centroid and the cohesion is based on the distance from the points in a cluster to its centroid. Equation 3.35 shows the definition of this index.

$$\mathrm{SF}(C) = 1 - \frac{1}{\mathrm{e}^{\mathrm{e}^{\mathrm{bcd}(C) - \mathrm{wcd}(C)}}} \tag{3.35}$$

where

$$\mathrm{bcd}(C) = \frac{\sum\limits_{c_k \in C} |c_k|\, \mathrm{d_e}(\bar{c_k}, \bar{X})}{N \times K}, \tag{3.36}$$

$$\mathrm{wcd}(C) = \sum_{c_k \in C} \left(\frac{1}{|c_k|} \sum_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k}) \right). \tag{3.37}$$

- **Sym-index** (Sym↑) (Bandyopadhyay and Saha 2008): This index is known as symmetry based cluster validity index and it is an adaptation of the *I* index (Maulik and Bandyopadhyay 2002) based on the Point Symmetry-Distance. The index is defined as shown in Equation 3.38.

$$\mathrm{Sym}(C) = \frac{\max\limits_{c_k, c_l \in C} \{\mathrm{d_e}(\bar{c_k}, \bar{c_l})\}}{K \sum\limits_{c_k \in C} \sum\limits_{x_i \in c_k} \mathrm{d}^*_{\mathrm{ps}}(x_i, c_k)}. \tag{3.38}$$

- **Point Symmetry-Distance based indices** (SymDB↓, SymD↑, Sym33-↑) (Saha and Bandyopadhyay 2009): These 3 indices are also based on the Point Symmetry-Distance and modify the cohesion estimator of the Davies-Bouldin, Dunn and generalized-Dunn (version 33) indices.

  The SymDB index is computed as DB, but the computation of S is redefined as described in Equation 3.39.

$$\mathrm{S}(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \mathrm{d}^*_{\mathrm{ps}}(x_i, c_k). \tag{3.39}$$

  The symD index is like D, but the Δ function is defined as

$$\Delta(c_k) = \max_{x_i \in c_k} \{\mathrm{d}^*_{\mathrm{ps}}(x_i, c_k)\}. \tag{3.40}$$

  And finally, the Sym33 index is a modification of gD33 where Δ is defined as shown in Equation 3.41.

$$\Delta(c_k) = \frac{2}{|c_k|} \sum_{x_i \in c_k} \mathrm{d}^*_{\mathrm{ps}}(x_i, c_k) \tag{3.41}$$

- **COP index** (COP↓) (Gurrutxaga et al. 2010): Although this index was first proposed to be used in conjunction with a cluster hierarchy post-processing algorithm, it can also be used as an ordinary CVI. It is a ratio-type index where the cohesion is estimated by the distance from the points in a cluster to its centroid and the separation is based on the furthest neighbour distance. Its definition is provided by Equation 3.42.

$$\mathrm{COP}(C) = \frac{1}{N} \sum_{c_k \in C} |c_k| \frac{\frac{1}{|c_k|} \sum\limits_{x_i \in c_k} \mathrm{d_e}(x_i, \bar{c_k})}{\min\limits_{x_i \notin c_k} \max\limits_{x_j \in c_k} \mathrm{d_e}(x_i, x_j)}. \tag{3.42}$$

- **Negentropy increment** (NI↓) (Lago-Fernández and Corbacho 2010): This is an index based on cluster normality estimation and, therefore, is not based on cohesion and separation estimations. Equation 3.43 shows how this index is computed.

$$\mathrm{NI}(C) = \frac{1}{2} \sum_{c_k \in C} \mathrm{p}(c_k) \log |\Sigma_{c_k}| - \frac{1}{2} \log |\Sigma_X| - \sum_{c_k \in C} \mathrm{p}(c_k) \log \mathrm{p}(c_k). \tag{3.43}$$

where $\mathrm{p}(c_k) = |c_k|/N$, Σ_{c_k} denotes the covariance matrix of cluster c_k, Σ_X denotes the covariance matrix of the whole dataset and $|\Sigma|$ denotes the determinant of a covariance matrix. Although the authors proposed the index as defined above, they later proposed a correction due to the poor results obtained. Nevertheless, we used the index in its original form since the correction does not meet the CVI selection criterion used for this work.

- **SV-Index** (SV↑) (K.R. and Žalik 2011): This ratio-type index is one of the most recent CVIs compared in this work. It estimates the separation by the nearest neighbour distance and the cohesion is based on the distance from the border points in a cluster to its centroid. It is defined as shown in Equation 3.44.

$$\mathrm{SV}(C) = \frac{\sum\limits_{c_k \in C} \min\limits_{c_l \in C \setminus c_k} \{\mathrm{d_e}(\bar{c_k}, \bar{c_l})\}}{\sum\limits_{c_k \in C} \frac{10}{|c_k|} \sum \max\limits_{x_i \in c_k} (0.1|c_k|) \{\mathrm{d_e}(x_i, \bar{c_k})\}}. \tag{3.44}$$

- **OS-Index** (OS↑) (K.R. and Žalik 2011): This is another recent ratio-type index proposed by K. R. Žalik and B. Žalik (K.R. and Žalik 2011) where a more complex separation estimator is used. In Equation 3.45 the definition of this index is given.

$$\mathrm{OS}(C) = \frac{\sum\limits_{c_k \in C} \sum\limits_{x_i \in c_k} \mathrm{ov}(x_i, c_k)}{\sum\limits_{c_k \in C} \frac{10}{|c_k|} \sum \max\limits_{x_i \in c_k} (0.1|c_k|) \{\mathrm{d_e}(x_i, \bar{c_k})\}} \tag{3.45}$$

where

$$\mathrm{ov}(x_i, c_k) = \begin{cases} \frac{\mathrm{a}(x_i, c_k)}{\mathrm{b}(x_i, c_k)} & \text{if } \frac{\mathrm{b}(x_i, c_k) - \mathrm{a}(x_i, c_k)}{\mathrm{b}(x_i, c_k) + \mathrm{a}(x_i, c_k)} < 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{3.46}$$

and

$$\mathrm{a}(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \mathrm{d_e}(x_i, x_j), \tag{3.47}$$

$$\mathrm{b}(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \notin c_k} \min(|c_k|) \{\mathrm{d_e}(x_i, x_j)\}. \tag{3.48}$$

3.4 Statistical Tests

According to Demšar (Demšar 2006), in order to compare two classifiers the non-parametric Wilcoxon-signed rank test (Wilcoxon 1945) should be used, whereas when multiple classifiers are compared the non-parametric Friedman test (Friedman 1937) with the corresponding post-hoc tests should be implemented. Following this recommendation both tests were used in order to determine whether significant differences existed between the performances of the CVIs analysed and the voting strategies we designed using them. In the next lines we summarise the working principles of both non-parametric statistical tests.

3.4.1 Wilcoxon-signed rank

The Wilcoxon-signed rank test (Wilcoxon 1945) is used to find statistically significant differences between two dependent variables. Given a sample with N pairs of instances, for pairs $\{1 \leq i \leq N\}$ with $x_{1,i}$ and $x_{2,i}$ measurements, the null hypothesis (H_0) determines that the difference between the pairs follows a symmetric distribution around zero, and the alternative hypothesis (H_1) represents the opposite case.

In the first step of the test for all the pairs (i), the absolute differences and the sign functions are computed, $|x_{2,i} - x_{1,i}|$ and $sgn(x_{2,i} - x_{1,i})$. Then after excluding all the tailed pairs, $|x_{2,i} - x_{1,i}| = 0$, the sample size is N_r and all the pairs are reordered increasingly regarding their absolute differences. At this point, the pairs are ranked accordingly so that the 1 value will be assigned to the pair with the smallest non null absolute difference; in case of ties, they will be assigned the average of the ranks of the individual ranks alternatively assigned if ties had not occurred. Then, the W test statistic will be subsequently computed as shown in Equation 3.49 with R_i representing the rank, that is, as the sum of the signed ranks.

$$W = \sum_{i=1}^{N} [sgn(x_{2,i} - x_{1,i}) R_i] \tag{3.49}$$

In case the null hypothesis is true, the distribution of the differences is expected to be approximately symmetric around zero and the distribution of positives and negatives is expected to be distributed at random among the ranks. This assumption enables to determine the probability of observing a value of W for the sample size. To do so, the sum of the positive ranks (W^+) and the absolute value of the sum of the negative ranks (W^-) are computed, keeping the lower value, $W' = min\{W^+, W^-\}$. Finally, a table of critical values for W is used to find the probability of observing a value of W for different significant levels provided for different sample sizes, $W_{critical,N_r}$. The null hypothesis at the N_r significant level will be rejected if $|W'| \leq W_{critical,N_r}$.

3.4.2 Friedman test

The procedure of Friedman test aims to determine whether the performances of different classifiers are significantly different in a group of datasets and has five steps:

Given n classifiers (rows) and k datasets (columns) first, the corresponding classifiers ranks (classifier numerical performances transformed to ranking format values) in each dataset, r_{ij}, generate a matrix of nxk dimensions.

Secondly, the average ranks of the classifiers are compared ($\mathrm{r_j}$) as shown in Equation 3.50.

$$\bar{\mathrm{r}}_j = \frac{1}{n}\sum_{i=1}^{n} r_{ij} \tag{3.50}$$

In third place, the Q test statistic is computed according to Equation 3.51.

$$Q = \frac{12N}{k(k+1)}\sum_{j=1}^{k}\left(\bar{\mathrm{r}}_j - \frac{k+1}{2}\right)^2 \tag{3.51}$$

Finally, for large values of n or k (i.e. $n > 15$ or $k > 4$), the probability distribution of Q can be approximated by that of a chi-squared distribution, that is, the distribution of a sum of the squares of k independent standard normal random variables. In this case the p-value is given by $P(\chi^2_{k-1} \geq Q)$. In the opposite case, small values of n or k, the p-values should be obtained from tables of Q specially prepared for the Friedman test.

If the p-value is significant, appropriate post-hoc test for multiple comparisons would be performed, which evaluate which pairs of classifiers have significant differences , e.g Holm's post hoc (Holm 1979). Given two classifiers i and j (k=2) with R_i and R_j average ranks computed with Friedman test for N datasets, the comparison will be carried out as shown in Equation 3.52. The value of z in the equation enables to find the p-value given in the table of normal distribution, which then is compared with a particular significance level α.

$$z = \frac{R_i - R_j}{\frac{k(k+1)}{6N}} \tag{3.52}$$

In this dissertation Holm's post hoc (Holm 1979) was used, which emulates that the tests are being carried out sequentially, using the p-values in an increasing order. Given m p-values in an increasingly ordered ($\{p_i, 1 \leq i \leq m\}$) and m corresponding hypothesis ($\{H_i, 1 \leq i \leq m\}$), the test adjusts the value of α in a step down method (García and Herrera 2008). According to the Holm's post-hoc, H_i to H_{i-1} hypothesis will be rejected if i is the smallest integer such that $p_i > \frac{\alpha}{(m-i+1)}$.

Part III

Contributions

Chapter 4

Unsupervised classification: Analysis of several decision fusion strategies for clustering validation.

4.1 Introduction

This contribution focuses on internal validation, which measures the compactness and separation of the clusters using specific indexes. For easier reading, hereinafter internal validation indexes will be denoted as Cluster Validity Indexes (CVIs). As far as we know, no research has found an "optimal" CVI able to cope successfully with all the contexts. Meanwhile, guidelines about the suitability of the indexes based on the particularity of each environment are gaining relevance. Such guidelines can be easily inferred from extensive comparative studies about the performance of the CVIs over a wide range of contexts.

Thus, the starting point of this contribution was the comparative study of internal Cluster Validity Indexes published by Arbelaitz et al. (Arbelaitz et al. 2013b). This study concluded that none of the CVIs compared showed an optimal behaviour in all the contexts, although the Silhouette index (P.J. Rousseeuw 1987) performed more robustly than the rest. Based on this work, our purpose was to obtain a more stable behaviour which would avoid the user having to select a different CVI for each particular environment. Aware of the success achieved by voting strategies in supervised learning (Schapire 1990), (Breiman 1996) we decided to export this method to our unsupervised learning scenario and to implement a decision fusion approach (Kryszczuk and Hurley 2010) for CVIs.

In our research, we analysed several decision fusion strategies for clustering validation; we implemented several voting approaches and applied them to the

CVIs used in the reference work (Arbelaitz et al. 2013b), to improve individual performances. Depending on whether the number of indexes that participated in the votes was restricted or not, our voting strategies were divided into two main types, Selective or Global Voting apiece. In addition, we used three different criteria to restrict the CVIs involved in each Selective Voting strategy: the global performance of the indexes, their factor dependent success rate and their impact on the results. Our experiments showed that most of the decision fusion approaches are more effective than using individual CVIs. Therefore, we claim that the success of these voting strategies is not limited to supervised learning, but also extends to the unsupervised learning context. This fact, leads us to believe that CVI decision fusion strategies can be a key to successfully meeting the challenges of clustering validation.

4.1.1 Related work

As mentioned before, no research to date has found a sole CVI able to cope with the variability of existing environments. Thus, we switched our attention to the extensive comparative studies of CVIs that, at least, provided some guidance to the suitability of the indexes for each situation.

Surprisingly, the main reference work in this area dates back to 1985, when Milligan and Cooper published a paper (Milligan and Cooper 1985) about internal clustering validation. They compared 30 CVIs using four hierarchical algorithms over 108 synthetic datasets. The diversity of contexts was completed using four numbers of non-overlapped clusters (2, 3, 4 or 5) and three values of either dimensionality (4, 6 or 8) or cluster sizes. Specifically, the hierarchical algorithms they used were single-linkage, complete-linkage, average-linkage and Ward's method. The results of the experiments were presented in a tabular format, showing the hit rate of the CVIs in predicting the correct number of clusters (K).

Some years later, in 1997 Bezdek et al. (Bezdek et al. 1997) also presented a comparative work of 23 CVIs but running just three times a single algorithm (EM) over 16 synthetic datasets. The experiment performed by Dimitriadou et al. (Dimitriadou et al. 2002) in 2002 was more limited in terms of number of CVIs compared (15) and besides, the 162 synthetic datasets used were of binary type. A more recent contribution provides a new perspective regarding the quality of a clustering partition (Gurrutxaga et al. 2011). In this research Gurrutxaga et al. admit that there is no single approach to defining the quality of a partition. Additionally, they support the use of Partition Similarity Measures (PSMs) used in external validation for validating the results of CVIs. Unlike their predecessors, who traditionally used the CVIs to estimate the correct number of clusters (K), they used them to predict the "best" partition, defined as the most similar to a ground truth partition of labelled data according to PSMs. Therefore, they measured the success of the CVIs based on matches between partitions predicted by the CVIs and those addressed by the PSMs. Eventually, they demonstrated that the success rate achieved by the CVIs, SR (%), was significantly higher using their approach. More exactly, they used seven

CVIs, seven synthetic and two real datasets, 10 runs of the k-means algorithm and two PSMs (Adjusted Rand (L. Hubert and Arabie 1985) and Variation of Information (Meilă 2003) (VI)).

A contemporary survey (Arbelaitz et al. 2013b) inspired by Milligan and Cooper (Milligan and Cooper 1985) provided the comparison of 30 CVIs in a wide range of environments, reaching a total of 6,480 configurations. Arbelaitz et al. 2013b used the methodological modification proposed by Gurrutxaga et al. (Gurrutxaga et al. 2011). The results are displayed in bar graph format, one per each experimental factor, showing the SR (%) of each CVI. The success rate in this case was computed in terms of matches between the partitions predicted by the CVIs and the "correct" ones identified by the Partition Similarity Measures (PSMs). Specifically, its experimental setup comprised 30 CVIs, three PSMs (Adjusted Rand (L. Hubert and Arabie 1985), Jaccard (Jaccard 1908) and VI (Meilă 2003)) and 10 runs of three clustering algorithms: k-means, Ward and average-linkage (Jain and Dubes 1988). They experimented with 20 real datasets and 720 synthetic datasets. The authors concluded that none of the CVIs compared could be considered as optimal, although indexes such as Silhouette (P.J. Rousseeuw 1987) for synthetic datasets and Score Function (SF) (Saitta et al. 2007b) for real datasets showed a relatively strong behaviour. The results showed that the overall SR (%) of the CVIs analysed was not severely affected by the experimental factors, although noise and overlap showed to be critical reducing the overall success rate up to a third. Finally, the statistical tests used Arbelaitz et al. 2013b identified three groups of CVIs with statistically significant differences in their performances, rated as best (10 CVIs), middle (14 CVIs) and worst (6 CVIs).

Concerning CVI decision fusion, some related examples can be found in the literature but they perform simple votes (equal weights for all the CVIs involved in a voting procedure). For example Sheng et al. (Sheng et al. 2005) proposed a Weighted Sum Validity Function (WSVF) where the weight assigned to the vote cast by each of the six CVIs used has the same value, assuming that the relative importance of every index is a priori the same. Conversely, we did not only test simple votes (Global Voting) but also defined some others (Selective Voting) involving just the CVIs with high relative importance and computing the corresponding weights individually according to three different criteria: the performance of the indexes, their factor dependent success rate and their impact on the results.

4.2 Experimental Setup

As the research of Arbelaitz et al. (Arbelaitz et al. 2013b) was the starting point for this contribution, we next provide some detailed information about their experimental design.

The synthetic datasets were created using three values for the numbers of clusters (K), three dimensionality values (dim), two overlap values (ov), two cluster density values (den) and two noise levels (nl). They defined the $nmin$

parameter to ensure a minimum number of objects per cluster. Table 4.1 shows the values of the parameters to design the synthetic datasets, giving way to 72 configurations which were generated 10 times each. The five variable items in Table 4.1 together with the three clustering algorithms (k-means, Ward and average-linkage (Jain and Dubes 1988)) and the three PSMs (Adjusted Rand (L. Hubert and Arabie 1985), Jaccard (Jaccard 1908) and Variation of Information (Meilă 2003) (VI)) used are controllable experimental factors (EF) in our experiments with the synthetic datasets and generate 6,480 configurations.

Experimental Factors (EF)	Values
$nmin$	100
K	2, 4, 8
dim	2, 4, 8
ov	1.5 (strict), 5 (bounded)
den	1 (not overlapped), 4 (overlapped)
nl	0 (no noise), 0.1 (noise)

Table 4.1: Values of the experimental factors (EF) used to generate the synthetic datasets.

The 20 real datasets used, drawn from the UCI repository (A. Frank and Asuncion 2010), have different characteristics: numbers of clusters, ranging from 106 to 2310, numbers of features from three to 166 and the number of classes ranging from two to 15. As these characteristics were predetermined, the number of controllable EFs for the experiments performed with the real datasets was limited to two (three clustering algorithms and three PSMs).

We finally list the 30 CVIs used in the reference work (Arbelaitz et al. 2013b) which we described in Chapter 3: Dunn index (D) (Dunn 1973), Calinski-Harabasz (CH) (Caliński and Harabasz 1974), Gamma index (G) (Baker and L.J. Hubert 1975), C-Index (CI) (L.J. Hubert and Levin 1976), Davies-Bouldin index (DB) (Davies and Bouldin 1979), Silhouette index (Sil) (P.J. Rousseeuw 1987), Graph theory based Dunn and Davies-Bouldin variations (DMST, DRNG, DGG, DBMST, DBRNG and DBGG) (Pal and Biswas 1997), Generalised Dunn indexes (gD31, gD41, gD51, gD33, gD43 and gD53) (Bezdek and Pal 1998), SDbw index (SDbw) (Halkidi and Vazirgiannis 2001), CS index (CS) (Chou et al. 2004), Davies-Bouldin* (DB*) (Kim and Ramakrishna 2005), Score function (SF) (Saitta et al. 2007b), Sym-index (Sym) (Bandyopadhyay and Saha 2008), Point Symmetry-Distance based indexes (SymDB, SymD and Sym33) (Saha and Bandyopadhyay 2009), COP index (COP) (Gurrutxaga et al. 2010), Negentropy increment (NI) (Lago-Fernández and Corbacho 2010), SV-Index (SV) (K.R. and Žalik 2011) and OS-Index (OS) (K.R. and Žalik 2011).

4.3 The proposed approach

The extensive comparative work of Arbelaitz et al. (Arbelaitz et al. 2013b) concluded that there was no optimal CVI able to cope successfully with all the factors comprising a clustering environment (i.e, the context). Therefore, our principal motivation was to devise new strategies which would provide a new method that is robust enough to face any clustering situation.

Aware of the efficiency shown by some voting techniques in the supervised learning field (Schapire 1990), (Breiman 1996) we decided to import them to our particular unsupervised learning scenario and thus to implement some CVI decision fusion strategies. Being the synthetic datasets the ones providing more controllable experimental factors (EFs) we decided to test the voting strategies first over the synthetic datasets and then use just the best ones for the real datasets.

Evaluation of the proposed approaches was carried out according to the following stages in the case of synthetic datasets: (1) Define decision fusion strategy ($\mathrm{DF_i}$); (2) calculate the success rate ($\mathrm{SR_i}$) using $\mathrm{DF_i}$ to estimate the best partition in each of the 6,480 configurations; (3) compare $\mathrm{SR_i}$ with the success rates obtained by the best CVIs presented in (Arbelaitz et al. 2013b).

In particular, we designed two main types of voting approaches according to the number of CVIs involved. The first type, named Global Voting, was implemented using the 30 CVIs mentioned before. In the second type, called Selective Voting, the number of indexes involved was restricted according to three possible criteria: the global performance of the CVIs, their factor dependent success rate and the impact they had over the results. These three criteria were defined using the overall SR (%) achieved by the CVIs for the controllable EFs of the synthetic and real datasets provided in the reference work (Arbelaitz et al. 2013b) (see Tables 4.4 and 4.5, and Table 4.8 respectively).

Additionally, some of the CVI decision fusion strategies defined are simple votes whereas in some others, the weight of vote cast by each index involved (W_{CVIk}) is individually computed as shown in Equation 4.1. Particularly, the Global Voting and the Selective Voting based on the global performance of the CVIs described in sections 5.1 and 5.2.1 are simple votes ($W_{CVIk} = 1$). Alternatively, in the Selective Voting approaches based on the factor dependent success rate and based on the impact over the results of the CVIs shown in sections 5.2.2 and 5.2.3 respectively, W_{CVIk} is individually computed for each CVI.

Let N be the number of experimental factors (EFs) available in the dataset of a voting approach, and let m be the number of values that a particular EF can get (see Table 4.1). Then, to define the weight of the vote cast by any of the CVIs involved in the voting (W_{CVIk}) we analyse the partition (k) suggested by the index in each m value of every EF. More exactly, W_{CVIk} is computed as the number of times that the index achieves the top n positions in the SR-rankings of each of the m values of the N experimental factors available in the datasets

($Top_{CVIk}^{n,EF,m}$), as denoted in Equation 4.1.

$$W_{CVIk} = \sum_{EF=1}^{N} \sum_{m=EF_{vmin}}^{EF_{vmax}} Top_{CVIk}^{n,EF,m} \tag{4.1}$$

In all the decision fusion systems after having defined W_{CVIk} for all the indexes involved, we count the total votes obtained by each partition (k). Then, the selected partition in a voting strategy (K_{max}) is computed as the one with the majority of the votes, and in the case of ties, the partition with the smallest number of clusters (k) is selected as pointed in Equation 4.2.

$$K_{max} = argmax_k \sum_k W_{CVIk} \tag{4.2}$$

Finally, to evaluate each voting strategy we compute the success rate in each value of each controllable experimental factor (EF), comparing the chosen partition with the one suggested by the Partition Similarity Measures (PSMs).

4.4 Designed Strategies and Results

We describe in this section the different decision fusion strategies analysed in the chapter.

4.4.1 Global Voting

The Global Voting approach is a simple vote that fuses the decision of 30 CVIs ($W_{CVIk} = 1$ for all the indexes involved). Table 4.2 lists the overall success rates (SRs) of this decision fusion system compared to the best individual CVIs (Arbelaitz et al. 2013b). As it can be observed, this approach cannot beat the best index for synthetic datasets (Silhouette). In particular, eight of the 30 CVIs achieved higher individual SRs than our Global Voting approach. The same rankings hold for the SRs (%) of the seven experimental factors available in the synthetic scenario.

	CVIs								Voting
	Sil	DB*	CH	gD33	gD43	gD53	SF	DB	Global Voting
Overall SR (%)	51.8	46.6	46.2	44.5	44.3	43.8	43.5	43	42.7

Table 4.2: Overall Success Rate (%) of the Global Voting approach for synthetic datasets compared to the best individual CVIs.

Considering the weakness of the results achieved by the Global Voting for synthetic datasets, we did not test this strategy on the real datasets. Aiming to achieve a better performance, we developed more sophisticated strategies denoted as Selective Voting, which are described in the next section.

4.4.2 Selective Voting

The Selective Voting strategies use a restricted group of CVIs for decision fusion. We developed three different approaches, each of which restricts the group of CVIs who vote, based on one of the following criteria: the global performance of the CVIs, their factor dependent success rate or the impact they have on the results. Next we describe the three criteria used for each Selective Voting approach and the results achieved in each case.

Global performance of the CVIs

The four Selective Voting approaches based on the global performance of the CVIs consists on simple votes ($W_{CVIk} = 1$) where we only involved one or two of the three groups of indexes with statistically significant different performances discovered by Arbelaitz et al. (Arbelaitz et al. 2013b), as shown in Table 4.3. In particular, the three voting strategies that use the *best*, *middle* and *worst* rated group of CVIs shown in Table 4.3 are denoted as *best*, *middle* and *worst vote* apiece. In addition, we developed another strategy denoted as *half vote* that uses all the indexes of the best rated group and the top five of the middle rated group.

Group	Rank$_{\text{avg}}$	CVIs
Best	9-13	Sil, DB*, CH, gD33, gD43, gD53, SDbw, DB, Sym33, COP
Middle	14-17	DMST, DRNG, DGG, SF, Sym, DBMST, DBRNG, gD41, SymDB, gD51, DBGG, gD31, SV, CS
Worst	19-23	D, SymD, G, CI, OS, NI

Table 4.3: Rated groups of CVIs with statistically significant different performances according to Arbelaitz et al. 2013b.

Regarding the results for the synthetic datasets, none of these four Selective Voting strategies based on the global performance of the CVIs was able to beat the overall SR (51.8%) achieved by the best individual index for synthetic datasets (Silhouette). The best strategy was *best vote* and achieved the second best overall SR (47.4%). The next best approach, *half vote*, achieved the third position in the overall SR-ranking with an overall SR of 46.6%. On the other hand, the strategies denoted as *middle* and *worst votes* achieved further down positions, the 15th one and the 30th one respectively, with SR values of 37.7% and of 23.8% accordingly. Finally, similar to the case of Global Voting, the overall results of these four strategies follow the same pattern for the SR (%) broken down by the seven controllable EFs of the synthetic datasets.

In conclusion, the results achieved by these four Selective Voting strategies did not meet our expectations, thus, we could claim that the simple votes (for all the participating indexes) seem not to be promising. Therefore, we did not test these approaches over the real datasets. Instead, aiming for an improvement in the results for the synthetic datasets, we computed W_{CVIk} individually for the CVIs involved in the two Selective Voting strategies described next.

Factor dependent success rates of the CVIs

Arbelaitz et al. (Arbelaitz et al. 2013b) concluded that not every experimental factor (EF) affected the same way to the performance of the CVIs. Inspired by this conclusion we designed two voting approaches involving just the CVIs with what we denoted as high and middle factor dependent success rates (SRs). In particular, the top two SR (%) of each controllable experimental factors (EFs) in Table 4.1, were classified as high factor dependent success rates and the three top ones were considered as middle factor dependent success rates. The voting strategies based on these two schemes are called the *high* and *middle factor dependent success votes*. Tables 4.4 and 4.5 show the indexes with high and middle factor dependent SRs for the seven workable EFs for the synthetic datasets: Partition Similarity Measures, number of clusters, dimensionality, overlap, density, noise and clustering algorithms.

	Partition Similarity Measures		
SR-ranking	ARand	Jaccard	VI
1	Sil	Sil	Sil
2	CH	CH	DB*
3	DB*	DB*	CH
	Number of Clusters		
SR-ranking	2	4	8
1	Sil	Sil	CH
2	gD33	CH	COP
3	SF	DB*	Sil
	Dimensionality		
SR-ranking	2	4	8
1	Sil	Sil	Sil
2	DB*	CH	CH
3	gD33	DB*	DB*
	Overlap		
SR-ranking	Yes	No	
1	Sil	Sil	
2	SF	DB	
3	DMST	DB*	

Table 4.4: Indexes with high and middle factor dependent success rates (three top ranked CVIs, $n = 3$) for four of the seven controllable experimental factors available in the synthetic datasets.

	Density		
SR-ranking	1/1	1/4	
1	Sil	Sil	
2	CH	DB*	
3	DB'	CH	
	Noise		
SR-ranking	Yes	No	
1	CH	Sil	
2	Sym	SDbw	
3	Sym33	DB*	
	Algorithms		
SR-ranking	K-means	Ward	Average Linkage
1	CH	Sil	Sil
2	Sil	DB*	CH
3	COP	DB	DMST

Table 4.5: Indexes with high and middle factor dependent success rates (three top ranked CVIs, $n = 3$) for three of the seven controllable experimental factors available in the synthetic datasets.

In these two CVI decision fusion strategies the votes cast by the indexes involved (W_{CVIk} in Equation 4.1) were individually computed according to the type of factor dependent success rates achieved (high or middle). In the *high factor dependent success vote*, W_{CVIk} is equal to the number of times it achieved the top two positions ($n = 2$ in Equation 4.1) in the SR-rankings of the manageable EFs. Similarly, in the *middle factor dependent success vote* W_{CVIk} corresponds to the number of times it achieved the top three positions in the SR-rankings ($n = 3$ in Equation 4.1). The first three columns of Table 4.6 show the set of CVIs and the weights of their votes (W_{CVIk}) in the *high* and *middle factor dependent success votes.*

Both Selective Voting approaches beat the overall SR (%) of Silhouette for the synthetic datasets as shown in Table 4.7. In particular, the improvement over this index was 1.4% for the *high factor dependent success vote* and 0.5% for the *middle factor dependent success vote*. The overall SR-rankings also remain for all the controllable EFs. Analysing the results of Table 4.7, we could claim that weighting the votes of the CVIs according to their factor dependent performance seems promising.

As the results achieved by these two voting approaches met our expectations, we also tested them on the real datasets. Since the CVIs showed to behave differently in the real datasets we calculated the new weights (W_{CVIk}) according to that behaviour. In the real context, the number of experimental factors (EFs) which could be manipulated (N in Equation 4.1) was limited to two: partition similarity measures (PSMs) and clustering algorithms. Therefore, in this context the *high* and *middle factor dependent success votes* were implemented using the CVIs with the top two and the top three success rates (SRs) for these two governable EFs.

	$\mathrm{W_{CVIk}}$ in the voting approaches					
	High factor	*Middle factor*	*Strong impact*		*Signif. impact*	
CVIs	*dep. success*	*dep. success*	n=2	n=3	n=2	n=3
Sil	16	17	34	37	25	27
CH	10	12	22	24	16	18
DB*	4	12	6	22	5	17
Sym	1	1	5	5	3	3
SDbw	1	1	5	5	3	3
gD33	1	2	3	4	2	3
COP	1	2	3	6	2	4
SF	1	2	3	6	2	4
DB	1	2	3	6	2	4
DMST	0	2	0	6	0	4
Sym33	0	1	0	5	0	3

Table 4.6: $\mathrm{W_{CVIk}}$ of the CVIs involved in the Selective Voting approaches based on the factor dependent success rate and bas.

Best CVI / Selective Voting	Overall SR (%)	Improv. on Sil (%)
High factor dependent success vote	52.5	1.4
Middle factor dependent success vote	52.1	0.5
Silhouette (Sil)	51.8	-

Table 4.7: Overall Success Rates (%) of the best CVI and the two Selective Voting factor dependent strategies for the synthetic datasets.

Table 4.8 shows the CVIs with high and middle factor dependent SRs for the two workable EFs of the real datasets, whereas the CVIs and the W_{CVIk} assigned by the *high* and *middle factor dependent success votes* in this context are displayed in the first three columns of Table 4.9.

Regarding the results (see Table 4.10), only the *high factor dependent success vote* approach was able to beat the overall results of the best CVI for the real datasets, SF. But the improvement of this voting approach over the best index in the real context (2.7%) was higher than the one observed for the synthetic context (1.4%). On the other hand, the overall SR (%) of the *middle factor dependent success vote*, 39.4%, was slightly lower than the one achieved by SF, 41.1%. For the real data, the positions achieved by the *high* and *middle factor dependent success votes* in the overall SR-rankings, first and third respectively, agree for the two governable EFs (PSMs and clustering algorithms).

	Partition Similarity Measures		
SR-ranking	ARand	Jaccard	VI
1	gD31	SF	SF
2	Sym	DGG	DGG
3	DMST	DRNG	DRNG
	Algorithms		
SR-ranking	K-means	Ward	Average Linkage
1	SF	SF	Sym
2	DGG	COP	gD51
3	DRNG	DRNG	gD31

Table 4.8: Indexes with high and middle factor dependent success (three top ranked CVIs, $n = 3$) for the two experimental factors available in the real datasets.

	Voting approaches					
	High factor	*Middle factor*	*Strong impact*		*Signif. impact*	
CVIs	*dep. success*	*dep. success*	n=2	n=3	n=2	n=3
SF	4	4	8	8	6	6
DGG	3	3	5	5	4	4
Sym	2	2	4	4	3	3
COP	1	1	3	3	2	2
gD51	1	1	3	3	2	2
gD31	1	2	1	4	1	3
DRNG	0	4	0	8	0	6
DMST	0	1	0	1	0	1

Table 4.9: $\mathrm{W_{CVIk}}$ of the CVIs involved in the Selective Votings based on the factor dependent success rates and based on the impact of the indexes over the results for real datasets.

Best CVI / Selective Voting	Overall SR (%)	Improvement on SF (%)
High factor dep. success vote	42.2	2.7
SF	41.1	-
Middle factor dep. success vote	39.4	-4.1

Table 4.10: Overall SR (%) of the best CVI and the two Selective voting based on the factor dependent success rates of the indexes for real datasets.

Considering these results, we concluded that the *high factor dependent success vote* is a good strategy, since it performed slightly more robustly than the best CVIs. However, aiming for a higher improvement over the best performances of the CVIs, we defined a new criteria for the next Selective Voting approaches.

Impact of the CVIs over the results

Arbelaitz et al. (Arbelaitz et al. 2013b) pointed out that some experimental factors (EFs) had a stronger influence on the overall success rates of the CVIs than others. Thus, we decided to quantify the impact of the EFs rating them in three different levels: *tiny*, *slight* and *great impact*. The impact levels assigned to the experimental factors were the following: *tiny impact* (Partition Similarity Measures, dimensionality and density), *slight impact* (No. clusters, overlap and clustering algorithm) and *great impact* (Noise).

This rating was the basis for the two Selective Voting approaches based on the impact of the CVIs we called *strong* and *significant impact votes* ($I = strong/significant$ in Equation 4.3). The impact of the CVIs was combined with their factor dependent success rate for these strategies. As was the case for the two previous strategies, in these ones only the decisions of those CVIs that had obtained the two and three best results (SRs) in the manageable experimental factors were fused. Consequently, the indexes involved in the *strong* and *significant impact votes* were the same used in the strategies of the previous section for the synthetic and real contexts (see Tables 4.4 and 4.5, and Table 4.7 respectively). In fact, the difference between the voting approaches mentioned is the way we computed the weight of the vote cast by each index (W_{CVIk}).

In these decision fusion schemes, the vote cast by each CVI involved (W_{CVIk}) is computed as the sum of the number of times it achieves the top two or the three positions (n) in the SR-ranking of each experimental factor ($Top^{n,EF,m}_{CVIk}$) multiplied by the impact weight assigned to that experimental factor ($WI^{EF,I}$):

$$W_{CVIk} = \sum_{EF=1}^{N} \sum_{m=EF_{vmin}}^{EF_{vmax}} Top^{n,EF,m}_{CVIk} \times WI^{EF,I} \tag{4.3}$$

Where $WI^{EF,I}$, are the weights assigned to the impact levels (I) of the experimental factors (EFs) for the *strong* and *significant impact* approaches (Table 4.11).

As showed in Table 4.11, in these two decision fusion approaches the experimental factors were rated according to three possible impact levels (*tiny*, *slight* or *strong*) and each strategy entails a particular set of weights for these levels (see in Table 4.6). In particular, the *strong impact vote* assigns a stronger set of weights to the EFs with higher impacts than the *significant impact vote*. Considering the two impact levels and the two types of ranked indexes involved, $I = strong/significant$ and $n = 2/3$ in Equation 4.3, a total of four votes based on the impact of the CVIs were designed.

In the synthetic context, W_{CVIk} was computed according to the impact defined for the seven experimental factors (EFs) available. Then, to compute W_{CVIk} in the synthetic case, we used the SR-rankings provided by the seven experimental factors (Tables 4.4 and 4.5) and the weights assigned by the *strong* and *significant impact votes* (Table 4.11). The corresponding values of W_{CVIk} for the indexes involved in these four Selective Voting approaches are shown in the last four columns of Table 4.6.

	$\mathrm{W_{CVIk}}$ in the voting approaches		
Voting	*Tiny impact*	*Slight impact*	*Great impact*
Strong impact vote	1	3	5
Significant impact vote	1	2	3

Table 4.11: Weights assigned to the impact levels, ($WI^{EF,I}$), for the *strong* and *significant impact* approaches.

As Table 4.12 illustrates, all four voting approaches based on the impact of the CVIs over the results beat the overall SR (%) of the best CVI for the synthetic datasets (Silhouette). The best overall success rate belongs to the *strong impact vote* carried out with the top two CVIs of the mentioned rankings ($n = 2$), achieving an improvement of 1.6% over Silhouette. The next best overall result corresponds to the *slight impact vote* carried out with the top two indexes of the rankings ($n = 2$), showing an improvement of 1.5% over Silhouette. On the other hand the *slight* and *strong impact vote* carried out with the three top ranked CVIs ($n = 3$), achieved lower improvements: 0.1 and 0 apiece. As before, the overall SR (%) of these four strategies follows a pattern that remains for the success rates broken down by the experimental factors.

Best CVI / Selective Voting	Overall SR (%)	Improvement on Sil (%)
Strong impact vote (n=2)	52.6	1.6
Significant impact vote (n=2)	52.6	1.5
Significant impact vote (n=3)	51.9	0.1
Strong impact vote (n=3)	51.8	0
Silhouette (Sil)	51.8	-

Table 4.12: Overall Success Rates (%) of *strong* and *significant impact votes* compared to Silhouette, the best CVI for the synthetic datasets.

As described in Table 4.12, a more efficient and stable performance is achieved with these four Selective Voting approaches for the synthetic datasets. Moreover, the improvement of the best strategy over Silhouette, (*strong impact vote* with the top two CVIs of the rankings), 1.6%, was higher than the one achieved by the best Selective Voting strategy from Section 5.2.2 (*high factor dependent success vote*), 1.4%. Hence, these four voting strategies were tested on the real data.

The two experimental factors that can be controlled for the real data are the Partition Similarity Measures and clustering algorithms. The impact levels defined for these two factors over the results were *tiny* and slight respectively.

The *tiny* and *slight* impact levels of the two experimental factors were weighted according to Table 4.11 ($WI^{EF,I}$): 1 and 3 for the *strong impact vote* and 1 and 2 for the *significant impact vote*.

As a result, the CVIs used for the real datasets were selected using the top two or top three indexes from the SR-rankings broken down by the two experimental factors mentioned (Table 4.8). Finally, the vote cast by each CVI was weighted using the same procedure described for the synthetic context (see Equation 4.3), but considering just the two impact levels of the two available experimental factors. The last four columns of Table 4.9 show W_{CVIk} for the four Selective Voting approaches based on the impact over the results for the real datasets.

Table 4.13 shows the overall SR (%) achieved by the four strategies based on the impact over the results against the one shown by the best index (SF) for the real datasets. Only the *strong* and *significant impact votes* that use the top two indexes from the success rate rankings ($n = 2$) were able to beat the overall SR of the best CVI. These two approaches improved the overall success rate of SF by 2.7%. The two Selective Voting approaches based on the impact over the results that used the best three indexes from the rankings ($n = 3$), achieved exactly the same results, which were 4.1% lower than the one achieved by SF. As before, the pattern followed by the overall results of the strategies described, continues for the SRs broken down by the EFs.

Best CVI / Selective Voting	Overall SR (%)	Improvement on SF (%)
Strong impact vote (n=2)	42.2	2.7
Significant impact vote (n=2)	42.2	2.7
SF	41.1	-
Significant impact vote (n=3)	39.4	-4.1
Strong impact vote (n=3)	39.4	-4.1

Table 4.13: Overall Success Rates (%) of the best CVI and the *strong* and *significant impact votes* for the real datasets.

Unlike for synthetic datasets, the highest improvement rate over the best CVI for the real datasets is the same registered by the best Selective Voting strategy based on the factor dependent success rates of the indexes (*high factor dependent success vote*), 2.7%. Considering the results for both types of datasets, the best performance belongs to the *strong impact vote* that uses the two top ranked CVIs from the SR rankings provided by the experimental factors. In fact, the *strong impact vote* (with the top two CVIs) achieves the highest SR (%) for both synthetic or real datasets, beating the former best approach named *high factor dependent success vote* for the real case. Thus, weighting the votes of CVIs correctly seems important for decision fusion strategies.

4.4.3 Statistical Tests

Although, the use of statistical tests is not very usual in the unsupervised context, Arbelaitz et al. (Arbelaitz et al. 2013b) adapted a methodology used in the supervised scenario where several classification algorithms are compared by running them on several datasets and computing a "quality" estimate, such as the accuracy or the AUC value, for each algorithm and dataset pair. In this context Demšar (Demšar 2006) proposed to use a single test over all the algorithms and all the dataset and Arbelaitz et al. (Arbelaitz et al. 2013b) adapted this particular proposal to the unsupervised learning context replacing the classification algorithms by CVIs. However, this was not enough, since in the experiments a Boolean value (success / not success) for each CVI-configuration pair is obtained instead of a "quality" estimate and the configurations obtained by varying the clustering algorithm and PSM cannot be considered independent because they are based on the same dataset. The proposed solution, and the one adopted in this chapter, was to add for each dataset the number of successes each CVI obtained for each clustering algorithm–partition similarity measure pair. Moreover, in order to obtain a more precise estimate, the number of successes obtained in every run was also added —remember that 10 datasets were created for each combination of dataset characteristics. We thus obtained 72 values (one for each of the 72 combinations of dataset characteristics obtained varying the values of five of the seven controllable experimental factors) ranging from 0 to 90 for each CVI or decision fusion strategy, that gave us a "quality" estimate for independent datasets.

We compared the performance of our best voting strategy (*strong impact vote*) with the 10 best CVIs according to the reference work using the Friedman Aligned test (Friedman 1937) to check the existence of statistical differences and the Holm's post hoc (García and Herrera 2008) for pairwise index comparisons, with confidence levels of 5%. We finally used the Wilcoxon-signed rank test to compare the two best options; the *strong impact vote* approach with the best CVI in (Arbelaitz et al. 2013b), Silhouette.

According to the Friedman-Aligned test there were significant differences in a one-to all way, computing a p-value of the order of 4.2-10, which did not exceed the threshold defined in this configuration. In this case the Holm's post hoc test established that there were statistically significant differences between our best voting approach and nine of the best indexes compared; all except Silhouette. Finally, applying the Wilcoxon-signed rank test we confirmed the existence of statistical differences between our best voting approach and Silhouette, computing an asymptotic p-value of about 0.027, which was within the range defined for the existence of statistical differences for this scheme.

4.4.4 Summary

In the unsupervised learning environment, the correct partition of data is not available, making it difficult to evaluate the performance of clustering algorithms. Therefore, one of the biggest challenges in this area is the validation of the results obtained by the algorithms. Amongst the various proposals currently under discussion, one of the most popular approaches is the one based on internal Cluster Validity Indexes (CVIs). Comparative studies of such indexes show that there is no optimal CVI able to cope successfully with all the contexts.

The aim of this contribution was to implement and analyse several decision fusion strategies over the CVIs studied in an extensive comparative work published in the bibliography (Arbelaitz et al. 2013b), motivated by the success achieved by voting strategies in supervised learning. Thus, this experimental contribution consists of designing and implementing different CVI decision fusion strategies and then evaluating their performance in order to discover which of them are promising and eventually select the best one. Experiments with several strategies showed that the majority of the decision fusion approaches designed cope with the diversity of contexts more effectively than single CVIs.

Specifically, we designed two main types of decision fusion approaches depending on the number of CVIs participating in the voting, a Global Voting using all them (30) and three different groups of Selective Voting approaches where the indexes involved were selected considering three characteristics: their global performance (best/middle/worst), their factor dependent success rate (high/middle) and the impact they had on the results (strong/significant). In the last two Selective Voting strategies the vote cast by each CVI was weighted according to the characteristic used (factor dependent success rate/impact), whereas in the remaining one as well as in the Global Voting, equal weights were used.

Regarding the results, on the one hand, we observed that the Selective Voting strategies performed better than the Global Voting and on the other hand, we found that weighting the votes according to a particular criteria was more effective than using equal weights. More concretely, the decision fusion which selects the CVIs according to their impact on the results and strongly weights their votes, *strong impact vote*, was found to be the best approach. Furthermore, this best voting strategy was proven to be significantly better than the top 10 ranked indexes of the reference work (Arbelaitz et al. 2013b) according to the Friedman Aligned test (Friedman 1937) carried out.

Chapter 5

Modelling the interaction of users with disabilities

5.1 Introduction

In recent decades, there has been a trend towards a dramatic increase in the amount of information stored on the web and its subsequent use. Website access has become an important tool for information seeking, communication and participation processes in our society, and consequently, digital competence is considered basic nowadays. This makes it important to familiarise and enable people with disabilities in the use of digital devices and applications, and, to adapt site interaction to their needs.

Unfortunately, a theoretically accessible design might not be enough to ensure that people with disabilities enjoy smooth access to a website (Arbelaitz et al. 2016). In this context, the adaptation of the site to the users becomes crucial. Adaptations could be determined according to the results of specific questionnaires but this would be limited to the users participating in the questionnaire. Moreover, in general web applications it is all too easy to fail to recognize the full range of types of users who might be interested in using them or who might need to navigate in them (Dillon 2001). Another option for gathering adaptation proposals is from the analysis of the interaction of the users with the website. For example building adaptive systems able to generate models based solely on in-use information. This option will be more general and applicable to new users accessing to the site.

The specific adaptations required will depend on the user characteristics, the problems the user is having while navigating, etc. In this context, the detection in use time of the navigation problems or the type of device being used, will be a compulsory initial stage to then be able automatically adapt the site to the user and thus, to improve the user experience.

The use of web mining (Liu 2006) for these objectives has many advantages. It is not disruptive, it is based on statistical data obtained through real navigation data (decreasing the possibility of false assumptions) and is, itself, adaptive (when the characteristics of the user change, the collected data allows the automatic change of the interaction schema). When the user is a person with physical, sensory or cognitive restrictions, data mining is the easiest (and frequently almost the only) way to model user habits or characteristics.

In this chapter we present a system with a two-step architecture to detect user navigation problems while the users are interacting with a website. The first step is dedicated to detect automatically the device being used to interact with the computer while the user is navigating on the web. The second step tackles the issue of detecting the problem the user is having while navigating on the site. The system is based on data collected by RemoTest (Valencia et al. 2015), a tool to collect the complete user interaction data. The complete data mining process includes some specific steps as some meetings with accessibility experts to define some of the features to be used in the system.

The results showed that the application of a complete data mining process to the data collected by RemoTest is a promising strategy for automatically detecting user problems and affords the opportunity to provide specific adaptations in the future.

5.1.1 Related work

The application of machine learning techniques requires large amounts of data to be collected. Data collection in the context of users with disabilities is not an easy task, and this probably limits the number of works carried out in the area.

When machine learning techniques are applied to user interaction data the features extracted from the interaction are critical. Depending on the extracted features the machine learning algorithms will be able to solve the problem or not. Almanji et al. (Almanji et al. 2014) present a review of features extracted from the client-side interaction data of users with pointing devices who suffer from upper limb impairments due to cerebral palsy. They propose a model that measures the influence of the MACS (Manual Ability Classification System) level of each user and the characteristics of the analysed features. Among the analysed features, movement time, acceleration-deceleration cycles and average speed are the most significant. Authors claim that for individuals with cerebral palsy, it is more important to focus on methods to increase speed because they already appear to have enough accuracy.

Hurst et al. (Hurst et al. 2008) propose systems to automatically detect pointing performance with the aim of learning how to deploy adaptations at the correct time without prior knowledge of the participants' ability. To this regard, able bodied participants were also included in the experiment. They use client-side interaction data to build several systems to: (a) distinguish pointing behaviours of individuals without problems from individuals with motor impairments, (b) distinguish pointing behaviours of young people from people with Parkinson or older adults, and (c) determine the need for the Steady Click adaptation that was designed to minimise pointer slips during clicking. All the systems are built over labelled datasets including features related to clicking, features related to movement, pause features and task specific features. They used wrapper methods to select the features through C4.5 classifier. Concerning the results, the systems presented achieved high accuracy values: (a) 92.7%, (b) 91.6% and (c) 94.4%.

De Santana and Baranauskas (Santana and Baranauskas 2015) propose a remote evaluation tool, named WELFIT, for identifying web usage patterns through client-side interaction data (event streams). They provide insights into differences in the composition of event streams generated by people with and without disabilities. The tool uses SAM (Sequence Alignment Method) for measuring the distances between event streams and uses a previously proposed heuristic (Santana and Baranauskas 2010) to point out usage incidents. They label the groups built in the clustering procedure as AT (users using assistive technologies), or non-AT, according to the majority in each group. With the aim of identifying web usage patterns within the discovered groups, they found significant differences in the distribution of several features between AT and non-AT users.

Augstein et al. (Augstein et al. 2017) present a personalised interaction approach where a set of metrics are computed based on different interaction tests performed by 22 users (four of them with cognitive impairments) and then used to recommend the so-called 'interaction device setting'(IDS) that best fits the user needs. A ranked list of all IDSs and an overall suitability value for each is provided and the user can select the desired setting to work with two real-world interaction tasks, scrolling in larger documents and navigation through the windows start menu. In particular, the interaction tests were performed using three different IDSs: physical pressure based on a smart phone vibration absorption, physical pressure based on a smart phone magnetic field manipulation and hand or arm shaking using a smart watch or an armband with integrated position/acceleration sensor. According to their results, in more than 95% of the cases the recommended IDS was the one the user had expected.

We stand that to our knowledge, no work has analysed a set of users (with and without physical impairments) interacting with their preferred device (keyboard, trackball, joystick or mouse) and tried to find out any problematic pattern that could happen in any of the two types of tasks defined: mechanical task or mechanical and cognitive task requiring some cognitive effort. This is important since previously unknown problems can be detected this way and this is what we have done in this chapter.

5.2 Description of the platform used for the experiments: RemoTest

The RemoTest platform (Valencia et al. 2015) provides the necessary functionalities to assist researchers to define web-based user experiments, manage experimental remote/in-situ sessions and analyse the gathered interaction data.

This platform admits a wide range of experiments. The architecture of the platform consists of a hybrid architecture model that includes some functions in a client-side module and the other ones in some server-side modules. Figure 5.1 below shows the general architecture of RemoTest platform and the interaction between its four modules (Arrue et al. 2018):

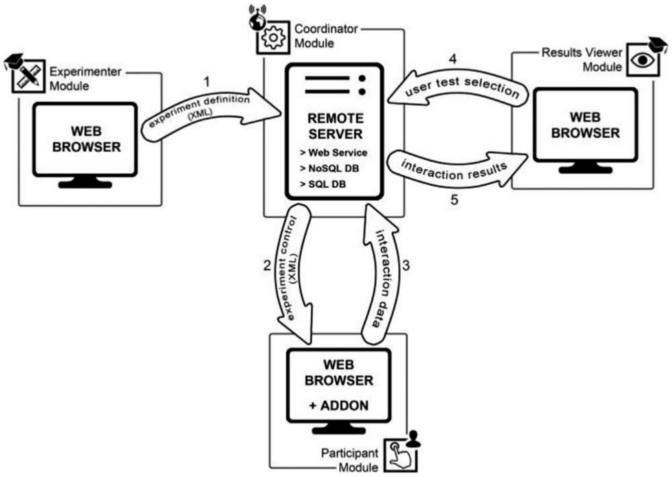


Figure 5.1: Description of RemoTest (Arrue et al. 2018).

- Experimenter Module (EXm): this module provides assistance to the researchers during the experiment definition process and stores it in an XML file based on specific vocabulary created to describe the experiments, comprehensive enough to detail the main elements, e.j objectives, stimuli to be presented, task time limits, questionnaires to be filled in by participants etc.

- Coordinator Module (COm): this module creates personalised experimental sessions for each participant using the information of the XML file generated in the EXm module.

- Participant Module (PAm): this module uses the experimental sessions transferred by the COm module, allowing the participants to visualise them and storing their corresponding interaction data.

- Results Viewer Module (RVm): this module organisers and presents the interaction data gathered in the experiments.

In order to build the system proposed in this chapter we did not work with the information stored in the RVm module but directly with the interaction units or events gathered and interaction information stored in PAm, such as cursor movements, key presses, scrolls, clicks, etc.

5.3 Experiments with users

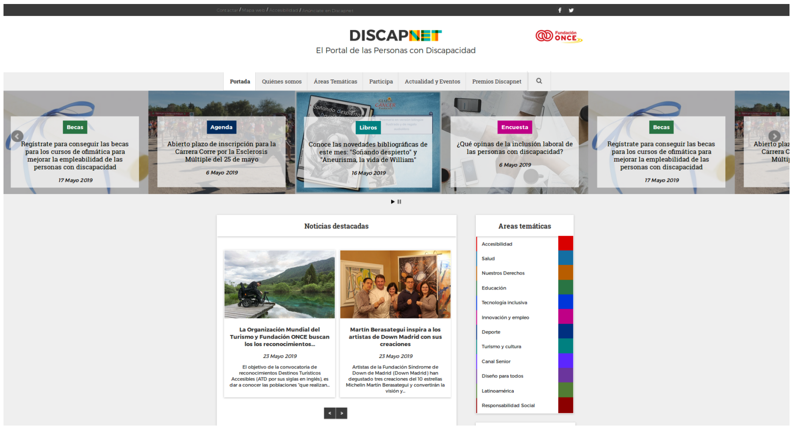
Fifteen subjects took part in the study which were divided into five groups based on the input device used for pointing and clicking actions: two keyboard users, two keyboard users using a headpointer to interact with the keyboard (keyboard+headpointer users), one trackball user, four joystick users and six mouse users.

All subjects from the first four groups were participants with motor-impairments most of them with over seven years of experience and using daily the computer. The subjects in the last group only included users without disabilities who had more than seven years of daily use of the mouse as an input device.

The same Dell Precision M6700 laptop running a 64 bit version of Windows 7 was utilised in all sessions. An additional widescreen LCD monitor (aspect ratio 16:10) with a diagonal size of 24 inches and display resolution set to 1920x1200 pixels was used to present stimuli to participants. Firefox add-ons implementing the virtual aids for the cursor were installed in this computer.

Before starting the study, participants were encouraged to adjust the pointer motion behaviour to meet their preferences. Subjects with disabilities used their own personal input devices to complete the study. All non disabled participants used the same optical USB mouse (Dell M- UVDEL1).

Two different websites were selected as stimuli for the experiment: the Discapnet website, `http://www.discapnet.com/`, which provides information to people with disabilities (see Figure 5.2) and the institutional website from the Council of Gipuzkoa, `http://www.gipuzkoa.eus/` (see Figure 5.3).



El Portal de las Personas con Discapacidad

Regístrate para conseguir las becas para los cursos de ofimática para mejorar la empleabilidad de las personas con discapacidad

17 Mayo 2019

Abierto plazo de inscripción para la Carrera Corre por la Esclerosis Múltiple del 25 de mayo

6 Mayo 2019

Conoce las novedades bibliográficas de este mes: "Soñando despierto" y "Aneurisma, la vida de William"

16 Mayo 2019

¿Qué opinas de la inclusión laboral de las personas con discapacidad?

6 Mayo 2019

Regístrate para conseguir las becas para los cursos de ofimática para mejorar la empleabilidad de las personas con discapacidad

17 Mayo 2019

Abierto plaz
Carrera C
Múltij

Noticias destacadas

La Organización Mundial del Turismo y Fundación ONCE buscan los los reconocimientos...

23 Mayo 2019

El objetivo de la convocatoria de reconocimientos Destinos Turísticos Accesibles (ATD por sus siglas en inglés), es dar a conocer las poblaciones "que realizan...

Martín Berasategui inspira a los artistas de Down Madrid con sus creaciones

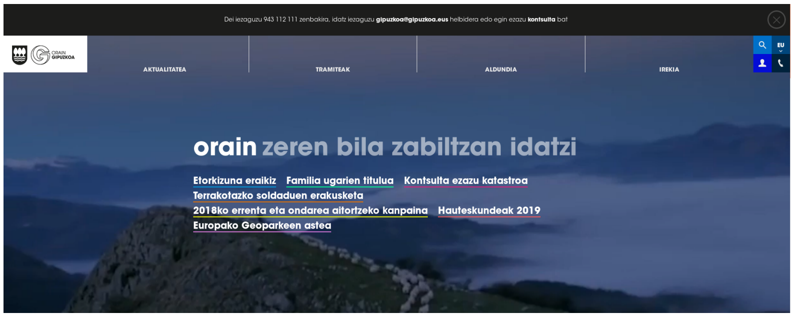
23 Mayo 2019

Artistas de la Fundación Síndrome de Down de Madrid (Down Madrid) han degustado tres creaciones del 10 estrellas Michelin Martín Berasategui y convertirán la visión y...

Areas temáticas

Figure 5.2: Home page of Discapnet website: `http://www.discapnet.com`

A third website, that provides information about the Bidasoa local area (Bidasoa Turismo), `http://www.bidasoaturismo.com`, was used for training purposes, so participants could learn how to use the new cursor virtual enhancements (see Figure 5.4). All three websites claim, within their accessibility



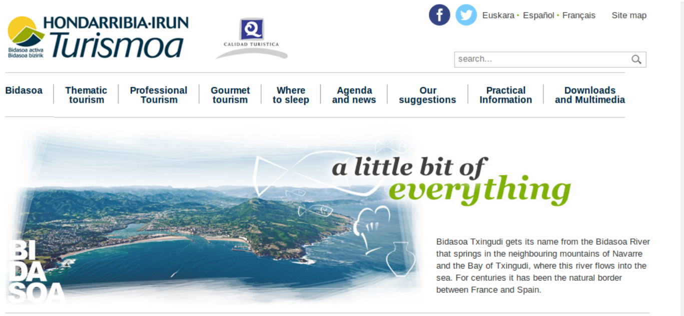
Dei iezaguzu 943 112 111 zenbakira, idatz iezaguzu **gipuzkoa@gipuzkoa.eus** helbidera edo egin ezazu **kontsulta** bat

orain zeren bila zabiltzan idatzi

Etorkizuna eraikiz Familia ugarien titulua Kontsulta ezazu katastroa
Terrakotazko soldaduen erakusketa
2018ko errenta eta ondarea aitortzeko kanpaina Hauteskundeak 2019
Europako Geoparkeen astea

Figure 5.3: Home page of the website of the Council of Gipuzkoa: `http://www.gipuzkoa.eus/`

sections, to conform to certain level of the WCAG 1.0 guidelines (Discapnet to Level AA, Gipuzkoa and Bidasoa to the Level A).



Bidasoa Txingudi gets its name from the Bidasoa River that springs in the neighbouring mountains of Navarre and the Bay of Txingudi, where this river flows into the sea. For centuries it has been the natural border between France and Spain.

Figure 5.4: Home page of Bidasoa Turismo website: `http://www.bidasoaturismo.com`

5.4 Tasks' characteristics

The users carried out two types of tasks:

- **MiniTask**: each MiniTask consisted of clicking on one highlighted target. After each target-clicking the position of the cursor was reset to the center of the screen. These types of tasks do not have any cognitive phase that requires thinking about the requested information and looking for it in the website, therefore these tasks measure exclusively the users' motor skills also named mechanical tasks. They are designed to be straightforward and short navigations which are very interesting to study the intentioned navigation of the user.
- **SearchingTask**: each task consisted of searching for a precise web page in a website after having been given its title. The target web pages were

between two and four steps away in the website. These tasks require a cognitive phase where the user spends time thinking and searching for the requested information also named cognitive tasks. The intentioned navigation starts when the user identifies the target.

In total each user carried out 156 tasks (144 MiniTasks and 12 SearchingTasks carried out in two websites) whose corresponding interaction information was used to detect first the device being used and then the possible navigation problems experienced.

5.5 Feature extraction

The interaction of the users with the website was collected with RemoTest and converted to a labelled dataset to be used in a supervised classification environment. We supposed that each of the users interacted similarly in every page visited during the experiments carried out, and that the interaction somehow depended on the type of device the user was using. So, the device being used was used to label the dataset examples, generating a dataset with five classes: keyboard, keyboard + headpointer, trackball, joystick and mouse.

In order to build an automatic detection system, first for devices and then for problems, we agree with accessibility experts to extract 25 features. However, based on the experts' suggestions, the set of features used for device (D) detection and problem (P) detection were different: 19 features were used for device detection (DB_1) and 11 features for problem detection ($\mathrm{DB}_{2\mathrm{M}}$ and $\mathrm{DB}_{2\mathrm{S}}$), which are shown in Table 5.1 marked as as D and P in column Use respectively.

For the extraction of the features in each page, we divided the space into eight quadrants (1-8) as shown in Figure 5.5, that are divided by the horizontal (H_i), vertical (V_i) and diagonal (D_i) axes. Then, we computed the direction of the cursor movements (slope) inside the defined quadrants and axes.

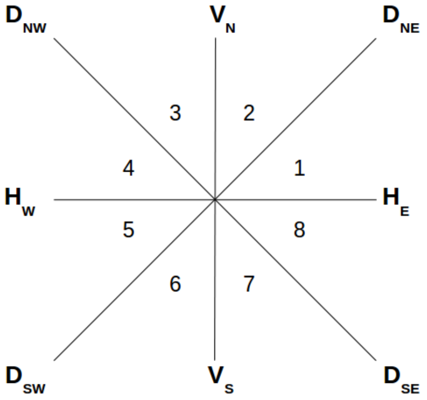


Figure 5.5: Space division to extract some features: horizontal (H_i), vertical (V_i) and diagonal (D_i) axes and eight quadrants (1-8).

Id	**Feature**	**Use**	**$\mathbf{P_r}$**	**Unit**	**Description**
1	NEvent	D	1	#	No. events (cursor-move, click, hover...)
2	NSpKeysPress	D	1	#	No. special keys pressed (no letter/digit).
3	NWheel	D	1	#	No. times the wheel has been used.
4	NHVMov	D	1	#	No. movements aligned with the horizontal or vertical axes.
5	NDMov	D	1	#	No. movements aligned with diagonals axes.
6	MedGapTime	D	1	*ms*	Median of the time intervals without cursor movements (gaps).
7	MedSpeed	D	1	px/s	Median of the cursor movements speeds.
8	MedAcc	D	1	px/s^2	Median of the cursor accelerations.
9	NKeyPress	D	2	#	No. keys pressed.
10	CurDist	D	2	*px*	Total distance traveled with the cursor.
11	RCurDistOpt	D/P	2	*ratio*	Ratio between CurDist and the distance between the initial position of the cursor and the target (optimal distance).
12	RStrQuadCh	D/P	2	*ratio*	NStrQuadCh / NQuadCh
13	NClick	D/P	3	#	No. clicks.
14	NScroll	D	3	#	No. times the scroll has been used.
15	RHVDmov	D	3	*ratio*	(NCross + NDMov) / (NCross + NDMov + NnotHVDMov).
16	NQuadCh	D/P	3	#	No. quadrant changes in the direction of movements.
17	NStrQuadCh	D	3	#	No. strong changes ($\geq$ 2 quadrants) in the direction of movements.
18	NnotHVDMov	D	3	#	No. movements not aligned with horizontal, vertical or diagonal axes.
19	TotTime	D/P	3	*ms*	Total time spent in the page.
20	ClickDist	P		*px*	Average distance between click-down and click-up actions.
21	ClickTime	P		*ms*	Average time between click-down and click-up actions.
22	DiagCurArea	P		*px*	Length of the diagonal of the rectangle that circumscribes the area traversed by the cursor.
23	MedJerk	P		px/s^3	Median of the changes of accelerations of the cursor movements.
24	NCross	P		#	No. times the cursor crosses the clickable area limits.
25	NGap	P		#	No. gaps.

Table 5.1: Description of the 25 features extracted in each visited page, 19 for device detection and 11 for problem detection.

5.5.1 Feature extraction for device detection

Each of the entries of the generated dataset contains the summary of the interaction of a user with a visited page. To summarise this interaction, we extracted a total of 19 features considered meaningful by the accessibility experts for the device (D) detection task, who in addition, grouped them into three priority levels (P_r) according to their usefulness: 1-high (eight features), 2-middle (four features), 3-low (seven features).

Table 5.1 summarises the 19 features extracted (marked as D and D/P in column Use) and their priority levels (column P_r). In the table, the abbreviations for features' units are pixels (px) for distances; seconds (s) and milliseconds (ms) for times; and appearance times (#) for counting.

5.5.2 Feature extraction for problem detection

Following the recommendations of the accessibility experts a total of 11 features were finally used for the problem (P) detection task, which are shown in Table 5.1 marked as P in column Use. As it can be observed, five of these features were also used for the device detection task (marked as D/P in column Use of Table 5.1): number of clicks (NClick), number of quadrant changes in the direction of movements (NQuadCh), ratio between the number of quadrant changes and the number of strong changes in the direction of movements (RStrQuadCh/NQuadCh), ratio between the CurDist feature and the distance between the initial position of the cursor and the target (RCurDistOpt), and total time spent in the page (TotTime).

In the recording process some errors or fluctuations may appear in the interaction data that lead to outlier navigation behaviours or impossible behaviours. In order to reduce the effect of these behaviours and obtain a smoother signal, a Butterworth filter (Proakis and Manolakis 1992) configured as a low-pass filter was applied to the following features: cursor speeds (MedSpeed), cursor accelerations (MedAcc), cursor jerks (MedJerk) and the angles of the cursor direction (NnotHVDMov, NHVMov, NQuadCh, NStrQuadCh, RHVDMov, RStrQuadCh). In the case of the nominal variables the smoothing was carried out based on a Simple Moving Mode method and using a subset of 11 elements (the five previous elements, the current element and the five subsequent elements).

Table 5.2 shows the average values obtained for the features used for problem detection for each of the devices, which was our final goal (data is split by devices and task times). The average values of the variables for each device and task type shown in Table 5.2 suggest that the overall behaviour is affected by both the type of device used and the type of task. For instance, the MiniTask part of the table shows that the highest values of MedJerk can be found for the mouse, in contrast, the smallest ones for the keyboard. The trackball is prone to register more NQuadCh. When it comes to TotTime, keyboard users spend the most time in completing the task, while the mouse users are the ones spending the least. For NGap, keyboard users make the most stops in navigation. As regards

NCross, mouse and keyboard users are the most skilled controlling the cursor around the target. The ClickTime with the joystick is much higher.

Devices	**Features**										
	MedJerk	***RCurDisOpt***	***RStrQuadCh***	***NClick***	***NQuadCh***	***TotTime (s)***	***NGap***	***NCross***	***ClickTime (ms)***	***ClickDist***	***DiagCurArea***
						MiniTask					
Keyboard	208	1.1	0.3	0.1	1.3	1.1	4.4	5.1	9	0.3	418
Joystick	868	1.4	0.5	0.1	4.8	6.3	1.6	16.0	258	0.8	476
Trackball	2,039	1.6	0.5	0.1	8.8	5.9	1.4	29.5	86	0.0	465
Mouse	3,330	1.2	0.2	0.1	1.4	1.6	0.4	7.4	91	0.3	315
						SearchingTask					
Keyboard	234	2.2	0.5	0.8	2.9	46.9	8.3	5.4	311	1.0	575
Joystick	707	5.9	0.5	0.8	18.1	40.4	8.9	31.3	322	4.0	941
Trackball	1,917	6.2	0.5	0.8	40.5	39.6	9.0	59.7	117	0.1	1,117
Mouse	3,758	14.3	0.5	0.9	10.3	13.1	5.8	14.3	109	5.4	781

Table 5.2: Average values (centroids) obtained for each feature used for problem detection and for each device, keyboard, joystick, trackball and mouse, and for each type of task, MiniTask and SearchingTask.

The trends of the features across devices in the SearchingTask part are similar to the ones described in the MiniTask part but not the values, which are higher. Some of the differences to highlight appear in the features NGap, ClickTime and ClickDist. The former becomes equal for all devices when the searching process is added, with trackball, joystick and keyboard being the devices with more stops. High values of ClickTime and ClickDist denote that users need more time and distance to make the click, keyboard users need more time (ClickTime) whereas mouse users travel more distance (ClickDist). In addition, the increase of the value of RCurDisOpt, the ratio between the traveled distance and the optimal distance, or TotTime, denote that SearchingTasks are, as expected, longer and more difficult.

In conclusion, the observable differences between devices and tasks suggest that the identification of the device will be important to later identify problems. This confirms our hypothesis about the need of a two phase system where the device is detected first, and the dataset is divided by device for the second phase.

5.6 Automatic device detection: first approach

5.6.1 Description of the dataset

In this first step of the architecture the entire dataset was given to the machine learning algorithms, without distinguishing between devices and MiniTask or SearchingTask URL-navigations, because, from the device point of view the task type should be irrelevant.

In order to compute all the features proposed by the experts a minimum activity within each URL navigation was established. As a result only the interactions fulfilling the five following conditions were added to the dataset:

- The number of MouseMove events has to be greater than or equal to five. Cursor-move events are recorded more or less every 10ms and offer information about cursor position.
- The distance traveled by the cursor has to be greater than zero.
- The information about the dimensions of the clicked target area has to be recorded by RemoTest.
- The click-up and click-down events need to appear among the stored data.
- The task duration (elapsed time) must be greater or equal than 4 seconds (s).

The reasoning behind the last condition, elapsed time $\geq$ 4s, was to be able to analyse the influence of the different intervals of the interaction (last 25% ...). Thus, we just considered tasks that could be split in smaller segments with sufficient activity to compute the features needed. This criteria was not considered in the adopted approach first because it enabled the selection of shorter navigations, mainly MiniTasks, and second because the use of a particular segment of the navigation (extracted splitting long sessions) was not found to be effective in improving the results.

After applying the above mentioned requirements, a 5-class unbalanced dataset with 20 (19 + class) features was generated (see Table 5.3). All the features were standardised (standard score was calculated) so that their differences in ranges did not affect to the performance of the built classifiers.

As this first approach was an intermediate step to achieve the final goal to detect navigation problems experienced by the users, features were extracted thinking in both objectives. To confirm our decision, we experimented with the complete set of features extracted (25) shown in Table 5.1, eight features considered by accessibility experts as significant to detect the used device (priority 1 features in Table 5.1) and features selected by two of the most used automatic feature selection algorithms (García et al. 2015): the Correlation-based Feature Subset Selection (M.A. Hall 1998) and a Wrapper (R. Kohavi and G.H John 1997) feature selection option which optimises the features for a given classifier (J48 in our case).

	class	**Number of examples**
	Joystic	584
	Keyboard	347
	Keyboard+headpointer	338
	Trackball	171
	Mouse	235
Total		1,675

Table 5.3: Class distribution of the generated dataset.

5.6.2 Results and analysis

The calculated features were used to build classifiers to classify user interaction data according to the device used for navigation. We built classifiers with the complete set of features extracted, the features considered to be the most important by the accessibility experts and the features selected by some automatic feature selection processes.

Experiments were run in Weka (M. Hall et al. 2009) with 4 basic classifiers, Naïve Bayes (NB) (G.H. John and Langley 1995), IBK (a k-NN implementation) (Aha et al. 1991), SVM (J.C. Platt 1999) and J48 (J.R Quinlan 1993) with default parameters and two decision tree (J48) based multiple classifiers, bagging (Bagg.) (Breiman 1996) and boosting (Boos.)(Freund and Schapire 1996), with 25 iterations. A five fold cross-validation (5 fold-CV) strategy was used for validation (80% for training and 20% for testing). As it can be observed in Table 5.4, four datasets differing in the contained features were evaluated:

- The most important features according to the experts (Priority 1 features)
- All the extracted features (All features)
- The features selected by the Correlation-based Feature Subset Selection method (CF Subset Eval)
- The features selected by the wrapper selection method with J48 as classifier and Genetic Search as search algorithm (Wrapper J48)

The values in Table 5.4 show that classification rates were not as high as expected. Generally the best rates were obtained with the most complex classifiers or multiple classifiers: bagging and boosting.

Focusing the analysis on how the sets of features affect to the performance of the system, it seems that the set proposed based on the experience of the accessibility experts (Priority 1 features) is only the best option in the case of Naïve Bayes (NB) classifiers. The rest of the classifiers behave better using the complete set of features or automatically selected sets of features. These two outcomes lead us to analyse confusion matrices in order to discover the source of the error.

Used features	**Accuracy (%) of the classifiers**					
	NB	IBK	SVM	J48	Bagg.	Boos.
Priority 1 features	66.09	67.46	62.09	71.88	74.75	75.52
All features	57.85	67.82	67.10	74.45	79.34	79.76
CF Subset Eval	59.64	68.96	64.12	74.87	77.97	77.25
Wrapper J48	57.85	67.88	66.57	75.82	79.7	80.78

Table 5.4: Classification results for different feature sets and classifiers in the 5-class dataset.

Analysis of the source of the error

To analyse the source of the error, we selected one of the best classifiers, the outcome of a J48 based boosting process applied to a dataset generated using the features selected with the Wrapper Feature selection process, and, studied its confusion matrix (see Table 5.5).

classified as	a	b	c	d	e	Fm (%)
Joystick = a	114	0	0	1	2	95.00
Keyboard = b	0	49	20	1	0	72.59
Keyboard+headpointer = c	0	16	51	0	0	73.91
Mouse = d	8	0	0	35	4	80.46
Trackball = e	1	0	0	3	30	85.71

Table 5.5: Confusion matrix + F-measure, Fm (%). Boosting with a Wrapper feature selection for the 5-class problem.

The values clearly show that the main source of error comes from not being able to differentiate classes keyboard and keyboard+headpointer. This was to be expected somehow, since although managing it differently, in both cases the finally used device is the keyboard. On the other edge, the joystick users are very accurately classified obtaining a F-measure (%) value of 95% and the mistakes done with mouse and trackball users are also few.

In this sense, we decided to simplify the problem to four classes, that is, we joined into the same class keyboard users and keyboard+headpointer users. The new dataset had still 1,675 examples but distributed now in the following way: Joystick (584), Keyboard (685), Trackball (171) and Mouse (235). In this regard, as we will explain later, accessibility experts supported this decision for the future goal of problem detection, suggesting differentiated virtual aids to improve the accessibility of the cursor for each group of devices (J.E. Pérez et al. 2016).

Solving the 4-class problem

The same experiments described in the previous sections were repeated in Weka for the new 4-class dataset; NB, IBK, SVM and J48, bagging and boosting models were built and evaluated based on a 5 fold-CV strategy.

As it could be expected, the values in Table 5.6 show that classification rates increased for all classifiers. This means that the systems built combining the features extracted from the experiments carried out with RemoTest with machine learning algorithms are able to differentiate the used device accurately.

Used features	**Accuracy (%) of the classifiers**					
	NB	IBK	SVM	J48	Bagg.	Boos.
Priority 1 features	82.99	83.52	80.48	87.22	89.19	90.39
All features	77.31	83.16	84.54	88.9	92.42	93.07
CF Subset Eval	81.19	82.75	80.54	87.34	90.69	91.22
Wrapper J48	77.49	84.12	83.52	89.67	92.48	92.66

Table 5.6: Classification results for different feature sets and classifiers in the 4-class dataset.

Comparing classifiers' performance, the same trends observed in the 5-class dataset were repeated; the best rates were obtained with the most complex classifiers or multiple classifiers bagging and boosting. With regard to the sets of features seeming to work better, they are again the automatically selected ones or the complete set of features. In particular, the best results (93%) were achieved when using boosting with the complete set of features extracted.

classified as	a	b	c	d	Fm (%)
Joystick = a	108	2	5	2	91.91
Keyboard = b	1	136	0	0	98.91
Mouse = c	6	0	40	1	84.21
Trackball = d	3	0	3	28	86.15

Table 5.7: Confusion matrix + F-measure, Fm (%). Boosting with a Wrapper feature selection for the 4-class problem.

If we further analyse the confusion matrices, see Table 5.7 for an example, we realise that the classifier was able to nearly perfectly differentiate the keyboard users from the rest, maintaining the general ability to differentiate devices for the three remaining options. This suggest the idea of building a hierarchical system to differentiate the different types of devices, where first keyboard is differentiated from the rest, and then remaining devices, joystick/trackball/mouse, are distinguished.

Analysis of the importance of the features

As in any data mining process, the features used in the classification process affected the efficiency of the classifiers. As stated before, it seems that the features considered to be the most important by the accessibility experts were not the best to use for classification. Therefore, we considered that the comparison of the features selected by the two feature selection processes applied to the two datasets (5-class and 4-class) and the experts could give us some clues about the importance of the extracted features. Table 5.8 contrasts the selection done by the experts and the one done by automatic algorithms. Each of the features could have been selected at most four times.

		Automatically selected	
		Very popular	Less popular

Experts	**Priority=1**	Nevents(4)
		Nwheel(4)
		MedGapTime(4)
		MedAcc(4)
		MedSpeed(3)
			NHVMov(2)
			NspKeyPress(2)
			NDMov(1)

	Priority≠1	RHVDMov (4)
		NnotHVDMov(4)
		NkeyPress(3)
			RcurDistOpt(2)
			NquadCh(2)
			TotTime(2)
			Nclick(2)
			CurDist(1)
			NStrQuadCh(1)
			Nscroll(2)
			RstrQuadCh(0)

Table 5.8: Use of the features.

In conclusion, only five out of the eight features considered very important by the experts where considered effective for the classification process by most of the four automatic feature selection processes carried out. However, there were three features, RHVDMov, NnotHVDMov and NkeyPress, not considered by the experts that are important for classification and will probably need to be taken into account by the experts in the future.

Although as stated before some features seemed not to be determinant intuitively for device detection, there was a single one, RstrQuadCh, not selected by any of the feature selection processes executed. However, this feature will probably be informative for problem detection.

5.7 Automatic device detection: adopted approach

From the previous analysis we learned that the type of task was irrelevant for the device detection goal and thus, in the adopted approach we also used the complete dataset (DB_1) without distinguishing between devices and MiniTask or SearchingTask URL-navigations.

Concerning future adaptations, some accessibility experts suggest that rather than adapting the website to people with disabilities, they should be provided with virtual aids to enhance their cursor accessibility as an intermediate solution. They point out that the solutions for the users should be different according to two main groups of devices: devices with restricted movements (RestrictedMov) and devices with free movements (FreeMov)(J.E. Pérez et al. 2016). This is in line with the outcome of the previous approach, where the two meta-devices are better distinguished, keyboard (with or without headpointer) and the group with the rest of devices. Accordingly, we modified the previous classifier system where all the devices were classified in a single step to better fit the new user interaction adaptation context, proposing a two level hierarchical approach to discriminate between devices. In the first level, two meta-classes were defined for classification, placing together the devices with similar behaviour: discrete input devices or devices with restricted movements, RestrictedMov (keyboard), and devices with non-restricted movements or FreeMov (joystick, trackball and mouse). In the second level, devices grouped into FreeMov meta-class were modelled and classified, i.e., the joystick, trackball and mouse classes.

Experiments were run in Weka with the same algorithms used in the previous approach: the basic classifiers, Naïve Bayes (NB), IBK, SVM and J48, with default parameters and the two decision tree (J48) based multiple classifiers, bagging (Bagg.) and boosting, with 25 iterations. Accordingly, a five-fold cross validation (5-fold CV) strategy was used for validation (80% for training and 20% for testing). Regarding the set of features the best one from the previous approach was selected, that is, the one made up of the 19 features marked as D and D/P in column Use of Table 5.1).

5.7.1 Description of the new dataset

In the adopted solution the requirement of minimum elapsed time value ($\geq$ 4s) was not considered, what increased a 22% the size of dataset described in Section 5.6 (from 1,656 to 2,148 examples). In particular, just the following four requirements were established: No. MouseMove events ≥ 5, distance travelled by the cursor ≥ 0, dimensions of the clicked target area are recorded = true, No. click-up/click-down events $\neq 0$.

Theoretically, in MiniTasks there were 2,160 entries (15 users x 144 MiniTasks) and in SearchingTasks there should be, approximately, 540 examples (15 users x 12 SearchingTasks x 2-4 clicks). However, the requirements established to ensure a minimum activity, diminished that theoretical number of recorded entries to 1,713 for MiniTasks (DB_{2M}) and to 435 for SearchingTasks (DB_{2S}). For the device detection phase, the examples (DB_1) were first divided into two datasets with 464 and 1,686 examples, resulting from the two meta-devices, devices with restricted movements (RestrictedMov) and those with free-movements (FreeMov), and the two types of tasks joined (MiniTasks and SearchingTasks). Then the second meta-device was divided into three datasets with 130, 628 and 938 examples respectively arising from the three devices with free-movements, trackball, joystick and mouse. See Table 5.9.

Meta-Device	**Device**	**Problem detection**		**Device det.**	**N.users**
		MiniTask	SearchingTask	Task	
RestrictedMov	Keyboard	388	76	464	4
FreeMov	Trackball	98	32	130	1
	Joystick	491	127	618	4
	Mouse	736	200	938	6
	Total	1,325	359	1,686	11
RestrictedMov +FreeMov	All	1,713 (DB_{2M})	435 (DB_{2S})	2,148 (DB_1)	15

Table 5.9: Number of examples in the dataset for different types of devices and tasks.

5.7.2 New results and analysis

As suggested by the accessibility experts, we considered as a critical error the misclassification of devices of different meta-classes (first level) and as a non-critical error the misclassification of devices of the same meta-class (second level). Hence, in the first level of the approach the classes that must be perfectly discriminated are addressed to minimise the critical error, whereas in the second level, where interactions are more similar, the classes with smaller differences are tackled.

Table 5.10 makes clear that the classifiers are able to discriminate the two meta-classes with high accuracy. Although all but NB classifiers obtain high accuracy, the best results are obtained with meta-classifiers, especially with boosting J48, with accuracy values of 99.26. As a result, the critical error is very small.

Set of features	**Accuracy (%) of the classifiers**					
	NB	IBK	SVM	J48	Bagg.	Boos.
All (19 features)	90.33	96.06	96.30	98.57	99.04	99.26

Table 5.10: Accuracy values obtained when discriminating the two meta-classes for different classifiers (critical error). The best value is shaded.

With the intention of analysing the origin of the critical error in the first level a confusion matrix is presented in Table 5.11. The results in the table show that the critical error of the first level is 0.74% (see Table 5.10) and that the F-measure (%) in the RestrictedMov group is slightly lower than in the FreeMov group (98.29% < 99.53%). Note that the four navigations made with devices with non restricted movements (FreeMov) misclassified as navigations made by devices with restricted movements (RestrictedMov) have been computed as critical errors in the first level thus they will not be considered as non-critical errors in the second level.

Assigned class ⇒	a	b	Fm (%)
RestrictedMov = a	460	4	98.29
FreeMov = b	12	1,672	99.52

Table 5.11: Confusion matrix and F-measure, Fm (%), values generated applying boosting J48 to the two class dataset.

In the second level we tried to discriminate the devices that convey the analogue movement: joystick, trackball and mouse. Table 5.12 shows that the best value was obtained again with boosting J48 with 90.13% of accuracy so the two levels of the hierarchy will be built using the boosting J48 classifier.

Set of features	**Accuracy (%) of the classifiers**					
	NB	IBK	SVM	J48	Bagg.	Boos.
All (19 features)	73.33	80.10	80.10	84.24	87.17	90.13

Table 5.12: Accuracy values obtained when discriminating the three no restricted movement classes for different classifiers (non-critical error).

Assigned class ⇒	a	b	c	Fm (%)
Trackball = a	122	1	3	73.72
Joystick = b	38	536	40	90.16
Mouse = c	45	38	849	93.09

Table 5.13: Confusion matrix and F-measure, Fm (%), values generated applying boosting J48 to the dataset with all features and with examples of corresponding class.

In order to analyse the origin of the non-critical error in the second level Table 5.13 is presented. Focusing on the F-measure values, it can be seen that the majority of errors are made misclassifying the mouse as a trackball, although joystick has been misclassified more times percentage-wise. In particular the non-critical error of the second level is 9.87 (see Table 5.12). Therefore, the global error of the two level automatic device classifier system is 8.43% and since the critical error (0.74%) is lower than the global smallest error of the approach described in Section 5.6 (6.93%, see Table 5.6) where all the devices were classified in a single step, we consider this hierarchical solution better for adapting the interaction of future users.

5.8 Automatic problem detection

5.8.1 Clustering for pattern discovery

Clustering algorithms, a type of unsupervised learning algorithm, can be used to discover behavioural patterns within the data when no prior knowledge about its structure or class exists. Based on the premise of devices having different values for features ($feat$), the idea is to first perform the clustering for each device (dev) and type of task, MiniTasks (M) and SearchingTasks (S), and then, automatically select the clusters (i, j) showing anomalous behaviour; the ones with higher deviation in the 11 features selected by the experts for this task. With this aim, the average behaviour of the examples grouped within a cluster, cluster centroid (MiniTasks $MC^{dev}_{i,feat}$ and SearchingTasks $SC^{dev}_{j,feat}$ respectively), is compared to the overall behaviour, see global centroid (MGC^{dev}_{feat}, SGC^{dev}_{feat}) in Table 5.2 where values for each type of device and task appear.

One of the main parameters of the clustering algorithm is the number of clusters (k) generated to obtain the best partition. According to Chapter 4, the use of decision fusion strategies between several CVIs was proven to be more effective than the use of a single CVI in the selection of the most suitable k. Accordingly, we detected navigation problems by executing k-means algorithm (Lloyd 1982) with different k values and then, selecting the best k according to a decision fusion strategy carried out with eight Cluster Validity Indexes (CVIs): Silhouette (P.J. Rousseeuw 1987), Davies-Bouldin variation (Kim and Ramakrishna 2005), Caliński-Harabasz (Caliński and Harabasz 1974), Davies-Bouldin (Davies and Bouldin 1979), COP (Gurrutxaga et al. 2010) and the Generalised Dunn indexes GD33, GD43 and GD53 (Bezdek and Pal 1998).

Pattern discovery in MiniTasks

Compared to SearchingTasks ($\mathrm{DB}_{2\mathrm{S}}$), the MiniTasks ($\mathrm{DB}_{2\mathrm{M}}$) are more straightforward, they do not have a thinking period where the user does not have any fixed direction. With regard to the variables, the biggest difference between the two types of tasks is the average time needed to complete the task (shown in Table 5.14). Focusing on the values, for the second group (SearchingTasks) the cognitive component clearly takes effect and, consequently average, median and maximum times increase considerably for all devices but their rank and differences are maintained in both types of tasks. The values for the rest of features also vary considerably according to the device.

To compare the average behaviour of a cluster (i) to the overall behaviour of all the navigations in the corresponding device (dev), all the values of the features (feat) were previously normalised (normal distribution) by device. As shown in Equation 5.1, cluster centroids ($MC_{i,feat}^{dev}$) deviating more than or equal to two standard deviations ($2stdev_{feat}^{dev}$) of the global centroid (MGC_{feat}^{dev}) were considered to be good candidates to identify navigation problems.

$$MC_{i,feat}^{dev} \geq MGC_{feat}^{dev} + 2stdev_{feat}^{dev}, dev \in \{Ke, Jo, Tr, Mo\}, i \in \mathbb{N}, i \leq 15 \tag{5.1}$$

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			MiniTask			
Keyboard	3.9	7.5	10.1	10.6	12.7	38.9
Joystick	2.0	4.1	5.3	6.3	7.1	42.3
Trackball	2.5	4.4	5.4	5.9	6.5	19.4
Mouse	0.6	1.0	1.2	1.6	1.6	22.0
			SearchingTask			
Keyboard	3.7	21.6	38.7	46.9	61.9	176.1
Joystick	5.3	14.9	27.1	40.4	58.2	164.8
Trackball	7.5	15.1	28.7	39.6	49.4	162.5
Mouse	1.4	5.4	8.7	13.1	15.9	81.4

Table 5.14: Time needed in completing each type of task by each device (in seconds).

Considering the sizes of the datasets for MiniTasks (one per device), the different k values we tested for k-means algorithm (Lloyd 1982), were 10, 15, 20 and 25. After using the eight Cluster Validity Indexes (CVIs) to evaluate the best partitions for each device, we computed the average k selected by the CVIs for the four devices (14.53). Then, we selected accordingly the nearest k (k=15) for the k-means used in the final problem detection process carried out in the four MiniTask datasets.

Problematic patterns and indicators

Clusters where the values of the features were deviated were automatically selected and then we inferred some problematic patterns and their meanings. The problematic patterns inferred and their meanings are described in the following paragraphs.

- **Pattern 1: too much distance.** Its indicator is the feature RCurDistOpt. It is a good predictor of a target selection problem; the cursor has travelled a much longer distance than the expected one. However, the causes of the problem are unknown. In order to explore the cause of the problem it is necessary to analyse the other features triggered with RCurDistOpt.

- **Pattern 2: too much time.** Its indicator is the feature TotTime. It is good predictor of a task completion difficulty; the task has taken more time than expected. However, in order to explore the cause of the problem it is necessary to analyse the other features triggered with TotTime. Thus, this feature is normally triggered by numerous different problems.

- **Pattern 3: rectifications in direction.** Its indicators are the features NQuadCh or RStrQuadCh that are triggered with the feature NGap because the adjustment of directions normally requires making short stops.
- **Pattern 4: unnecessary clicks.** Its indicator is the feature NClick. In a straightforward task it is an unexpected behaviour, more than that, it seems to be an indicator of ungainliness.
- **Pattern 5: difficulties around the target.** Its indicator is the feature NCross. Specifically, it indicates lack of control and precision in landing on the target. Whenever the target is passed over, there is a need to adjust the direction so it is usual to see it in combination with Pattern 3. In this context, the user may miss clicking on the target, so it will also frequently appear in combination with Pattern 4.
- **Pattern 6: long clicks.** Its indicator is the feature ClickTime. In a straightforward task it is an unexpected behaviour, indeed it seems to be more an indicator of indecisive behaviour.
- **Pattern 7: too many stops.** Its indicator is the feature NGap. Mostly it is related with the action of rectification (Pattern 3), but there are cases when the user makes no rectification and continues with the same direction, giving the idea that she/he is retaking control of the cursor.

Table 5.15 summarises the problematic patterns (Patt) inferred from the clusters selected automatically and the deviated features in each of the patterns. Where x indicates a significant deviation of a feature and [x] indicates a deviation of, at least, one of the marked features (OR condition).

From the table it follows that although experts initially marked 11 features as suitable indicators of navigation problems, the features MedJerk, ClickDist and DiagCurArea were not matched with any of the seven problematic patterns inferred but they may be connected to other types of problems not identified in this contribution. Consequently, their values are not shown neither in the table that describe the problematic patterns (Table 5.15) nor in the tables that show examples of clusters with problematic patterns (Tables 5.16 and 5.17).

Table 5.16 shows one of the clusters identified as problematic for each device. The table includes the standard deviation of the centroids of the clusters and the problematic patterns related (column Pattern), marking in bold the deviated features, that is, those which deviate more than two standard deviations (stdev). The table also includes the reference to a figure where one of the navigations of the cluster is illustrated (column Figure). Note that Pattern 6 (long clicks) is difficult to visualise, as shown in the navigation pattern represented in Figure 5.10. It is clear that each of the navigations represented is linked to several of the identified patterns, meaning that users having problems are probably finding difficulties in many aspects.

	Features							
Patt	***RCurDistOpt***	***RStrQuadCh***	***NClick***	***NQuadCh***	***TotTime (s)***	***NGap***	***NCross***	***ClickTime (ms)***
P1	x							
P2					x			
P3		[x]		[x]		x		
P4			x					
P5							x	
P6								x
P7						x		

Table 5.15: Description of the seven problematic patterns inferred and the related features for MiniTasks.

			Centroid: $MC_{i,feat}^{dev}$							
Figure	**Device**	**Pattern**	***RCurDistOpt***	***RStrQuadCh***	***NClick***	***NQuadCh***	***TotTime (s)***	***NGap***	***NCross***	***ClickTime (ms)***
5.6	Keyboard	1, 2, 3, 7	**2.21**	1.48	0.87	**2.06**	**2.21**	**2.24**	1.78	-0.03
5.7	Joystick	1, 2, 4, 5	**4.48**	0.94	**4.52**	**2.35**	**6.77**	5.61	**2.58**	0.58
5.8	Trackball	6	-0.13	0.41	-0.18	0.15	0.56	1.73	1.00	**2.49**
5.9	Mouse	1, 2, 3, 4	**6.51**	1.03	**6.33**	**5.82**	**7.24**	**7.94**	0.77	1.08

Table 5.16: Problematic patterns in the MiniTasks navigation traces presented and their corresponding centroids with deviated features according to Equation 5.1 in bold.

In Figures 5.6 to 5.9, an example per device and cluster is represented graphically where most of the patterns detected in the cluster are visible: keyboard in Figure 5.6, joystick in Figure 5.7, trackball in Figure 5.8 and mouse in Figure 5.9). Next we describe the symbology used in the figures:

- Big square: it represents the starting point.
- Medium square: it indicates a cursor stop.
- Little circumference: it indicates the cursor's position every 10ms.

- Medium circumference: it shows where a scroll was made.
- Medium cross: it means that a click was made.
- Big cross: it indicates the target position.

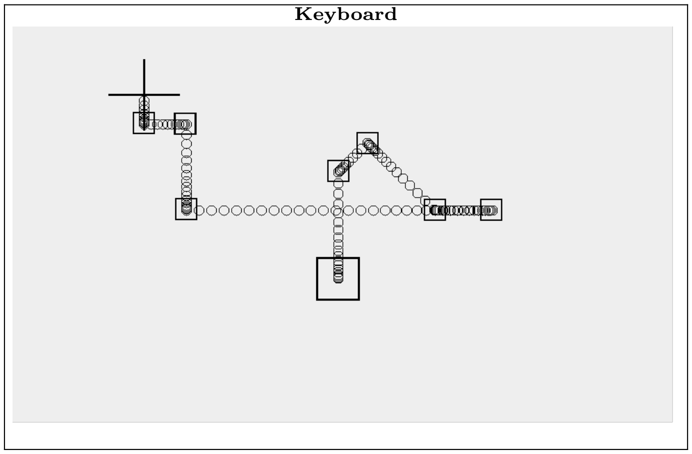


Figure 5.6: Example of problematic keyboard navigation extracted from cluster 7 with evident rectifications in the direction (P3) and excess of stops (P7).

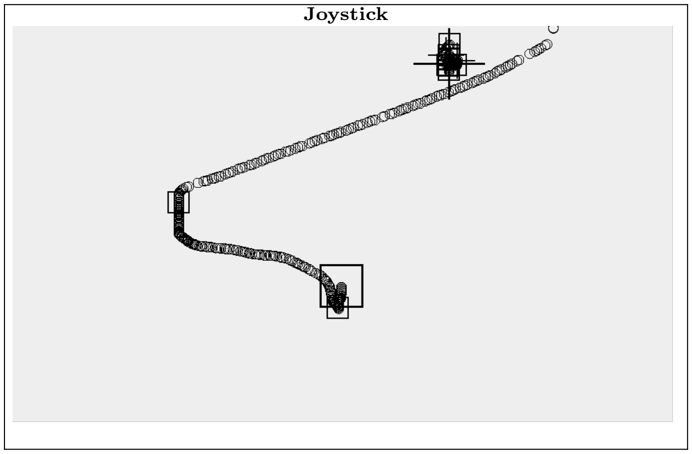


Figure 5.7: Example of problematic joystick navigation extracted from cluster 12, where time excess (P2) and difficulties around the target (P5) are particularly noticeable.

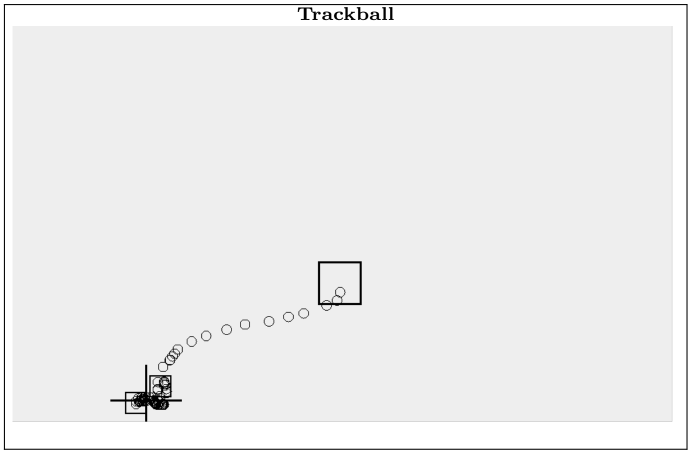


Figure 5.8: Example of problematic trackball navigation extracted from cluster 10 with long clicks (P6).

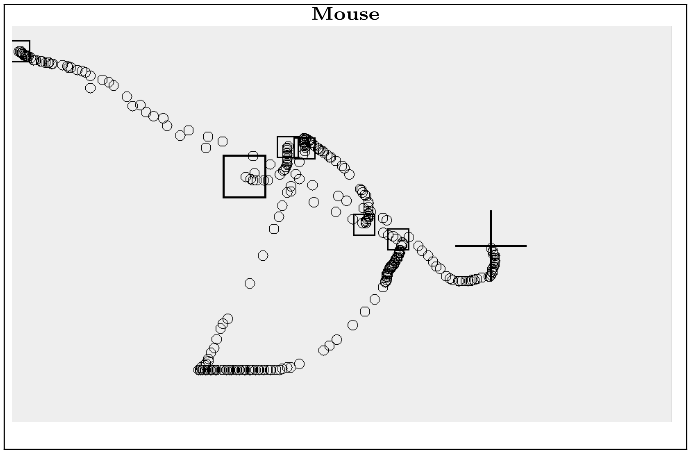


Figure 5.9: Example of problematic mouse navigation extracted from cluster 7 where too much distance (P1) is clearly manifest.

Pattern discovery in SearchingTasks

To compare the average behaviour of a cluster (j) to the overall behaviour of all the navigations in the corresponding device (dev), all the values of the features (feat) were previously normalised (normal distribution) by device. As shown in Equation 5.2, in this case, cluster centroids ($SC^{dev}_{j,feat}$) deviating more than or equal to one standard deviations ($1stdev^{dev}_{feat}$) of the global centroid (SGC^{dev}_{feat}) were considered to be good candidates for identifying navigation problems. The higher averages and standard deviations of the features in the SearchingTasks justify this new threshold, as the type of tasks here are more complex than in the previous case.

$$SC^{dev}_{j,feat} \geq SGC^{dev}_{feat} + 1stdev^{dev}_{feat}, dev \in \{Ke, Jo, Tr, Mo\}, j \in \mathbb{N}, j \leq 4 \tag{5.2}$$

Considering the sizes of the datasets for the SearchingTask (one per device), the different k values tested for k-means algorithm (Lloyd 1982), were set to K ϵ $\mathbb{N}$, k $\leq$ 12. After using the eight Cluster Validity Indexes (CVIs) to evaluate the best partitions for each device, we computed the average k selected by the CVIs for the four devices (3.72) and selected accordingly the nearest k (k=4) for the k-means used in the final problem detection process carried out in the four SearchingTask datasets. Although the selection was made based on the CVIs, the outcome was a set of partitions with similar cluster sizes in both cases. For the MiniTask navigations (partitioned with k=15) on average 25.9, 32.7, 6.5 and 49.1 navigations per cluster were obtained (keyboard, joystick, trackball and mouse respectively) whereas in the case of SearchingTask navigations similar cluster sizes were obtained using k=4: 19.0, 31.8, 8.0 and 50.5 (keyboard, joystick, trackball and mouse respectively).

The patterns arising in the deviated clusters where mostly identical to the ones discovered for MiniTasks (see section 5.8.1). The only difference appeared with the feature NCross which never deviated for SearchingTasks and thus, Pattern 5 (the difficulties around the target) was not inferred for this context. This is comprehensible, as in the MiniTasks the users must reach particular targets, whereas in the SearchingTasks they freely decide the targets they want to reach and so may be more precise in this exercise.

			Centroid: $SC^{dev}_{j,feat}$							
Figure	**Device**	**Pattern**	***RCurDistOpt***	***RStrQuadCh***	***NClick***	***NQuadCh***	***TotTime (s)***	***NGap***	***NCross***	***ClickTime (ms)***
5.10	Keyboard	1, 2, 3, 7	0.20	0.30	**1.07**	**1.27**	**1.47**	**1.55**	0.61	0.51
5.11	Joystick	2, 3, 4	0.75	-0.18	0.19	**1.32**	**1.33**	**1.17**	0.38	0.17
5.12	Trackball	4, 6	-0.23	-0.13	**1.56**	0.21	0.49	0.91	0.22	**1.49**
5.13	Mouse	2, 3, 4, 7	0.83	-0.04	**1.18**	**2.11**	**1.39**	**1.96**	0.44	0.19

Table 5.17: Problematic patterns in the SearchingTasks navigation traces shown and their corresponding centroids with deviated features according to Equation 5.2 in bold.

Table 5.17 shows one of the clusters identified as problematic for each device. The table includes the standard deviation of the centroids of the clusters and the related problematic patterns (column Pattern), marking in bold the deviated features (> 1 stdev). The table also includes the reference to a figure where one of the navigations of the cluster is illustrated (column Figure).

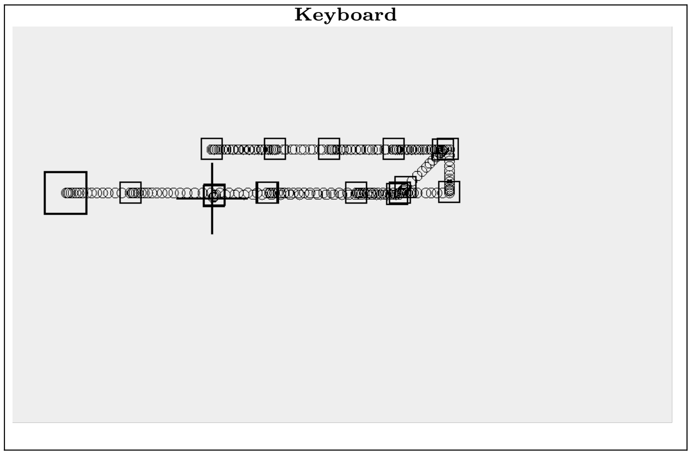


Figure 5.10: Example of problematic keyboard navigation extracted from cluster 3, where too much distance (P1) can be easily recognised.

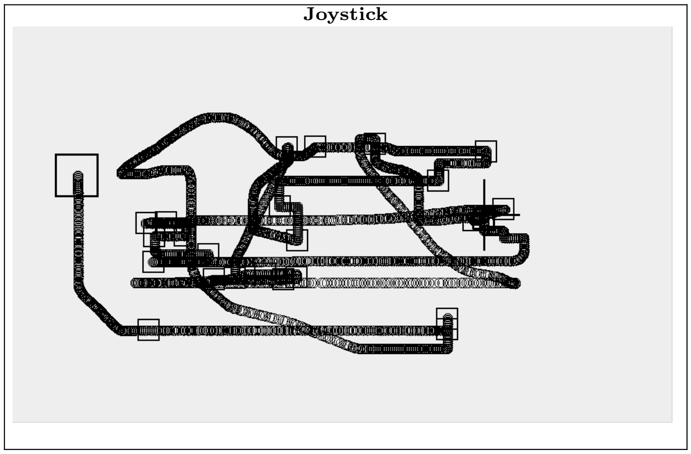


Figure 5.11: Example of problematic joystick navigation extracted from cluster 1, where time excess (P2) and rectifications in the direction (P3) are very perceptible.

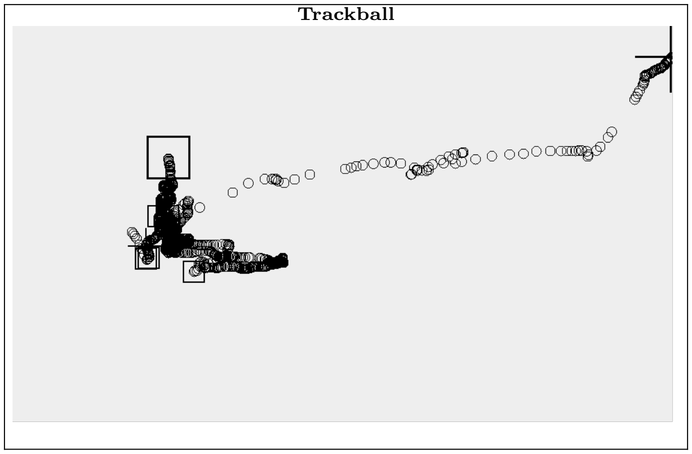


Figure 5.12: Example of problematic trackball navigation extracted from cluster 2 that clearly reveals unnecessary clicks (P4).

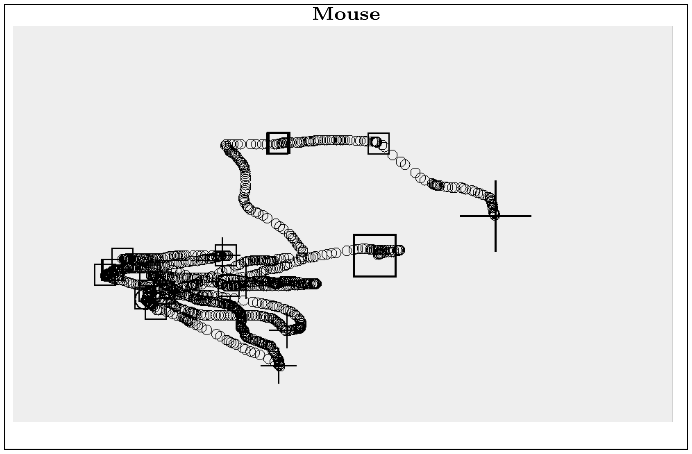


Figure 5.13: Example of problematic mouse navigation extracted from cluster 4, where too many stops (P7) can be easily identified.

5.8.2 Use of the detected patterns to facilitate navigation

The designed system can be used to detect navigation problems in real time. When a user is interacting with a web platform, 25 features can be extracted from the data recorded with the RemoTest platform, the corresponding 19 used in a first stage to automatically detect the used device based on the model described in Section 5.7. Then, according to the type of device and task, the nearest cluster (IBK or k-NN applied in the centroids of the clusters) of the adequate partition can be used to determine if the user is having problems or not and which problems she/he is having.

The set of problematic patterns detected has practical applications. Problematic patterns do not only describe the accessibility barriers experienced by the users, but also enable the definition of the most adequate adaptation techniques in order to avoid them. With this aim, in this section potential sources of problematic patterns are analysed and a number of feasible solutions are proposed.

Pattern 1 (too much distance) and Pattern 2 (too much time) occur as a consequence of other patterns found in the study. For example, these patterns have a direct connection to Pattern 5 (difficulties around the target), implying that the user requires increased distance and time to select the desired target. Similarly, Pattern 3 (rectifications in direction), can be a consequence of either the handling of the input device used or the specific characteristics of the user. For example, discrete input devices (devices with restricted movements such as keyboards) allow the user to move the cursor only in predefined directions; horizontal H_i (where i can be E or W), vertical V_i (where i can be N or S) and diagonal D_i (where i can be NE, SE, SW, NW). When targets cannot be reached directly following one of these predefined angles, the user must rectify the trajectory (selecting different angles) to reach the target.

People having uncontrolled movements, e.g. people with cerebral palsy, can experience difficulties in maintaining the position of the hand while they are moving the cursor, bringing about rectifications in the cursor path (Pattern 3) and also produces more stops during the target selection (Pattern 7). In addition, when the target size is not adequate, uncontrolled movements can make target selection difficult (Pattern 5). A lack of control can also provoke involuntary movements during the clicking process, moving the cursor away from the target and therefore generating “unnecessary clicks” (Pattern 4). In addition, these users might also have difficulties to press the buttons of the joystick, trackball or mouse to perform a click action if they have difficulty in stopping the ongoing action, producing “long clicks” (Pattern 6). This can cause clicks outside the target (Pattern 4) or also move the cursor during the click, preventing the click event to be triggered.

Different strategies can be followed to allow the user to make effective target selections. For instance, the *bubble cursor* technique increases the cursor selection area, depending on the number of selectable targets within reach, (Grossman and Balakrishnan 2005). An alternative solution is a *magnetic* target that attracts the cursor (Park et al. 2006). In this case, when the cursor is near a target, the cursor is pulled towards the target center making its activation easier. Another possible adaptation is the so called *goal crossing* (Wobbrock and Gajos 2008),which activates the target when the cursor crosses it. This adaptation makes the clicking action unnecessary and therefore the selection of small targets becomes easier. The *cross cursor* technique (J.E. Pérez et al. 2016) divides the screen in zones by means of a crossbar that allows a remote target to be selected by typing one letter, indicating the coordinates of the zone. This procedure reduces the required number of corrections and stops. The *goal crossing* technique appears to be an appropriate technique for people who use trackballs since it does not require handling precision and minimises the use of buttons. On the other hand, joystick users can benefit from both *magnetic* and *bubble cursor* techniques and keyboard users usually prefer the *cross cursor* technique since it helps to reduce the effort to select the target. Therefore adequate techniques to help to reduce the occurrence of the patterns 3, 4, 5, 7 can be selected depending on the used device.

With regard to problems relating to Pattern 6, diverse techniques can be used. For instance, *click on down*, *click on up*, *steady click* or *goal crossing.* Using *click on down* the target is selected just when the user presses the button. Similarly, *click on up* selects the target when the user releases the button. Instead, *steady click* "freezes" the cursor during the click enabling the target selection even if the user moves the cursor out away from the target (Trewin et al. 2006).

To select the most appropriate technique for each user, it is necessary to observe how they select the objectives. If the user tends to put the cursor on the target at the beginning of the click but moves the cursor during the click, the *steady click* or *click on down* techniques can be used. On the other hand, if the user moves the cursor during the click and tends to put the cursor on the target at the end, the technique to use would be the *click on up.*

Since each pattern has diverse associated techniques to avoid the related accessibility barriers, the selection of the most appropriate one for a specific user depends on his or her particular characteristics. Some of them can be selected by the users themselves when presetting their interfaces but they can also be selected by adaptive systems depending on the detected device and problem.

5.8.3 Summary

The importance of digital competence nowadays makes it essential to enable people with disabilities in the use of digital devices and applications, and moreover, to be able to automatically adapt site interaction to their necessities. Most of the current adaptable systems are linked to predefined user profiles. However, the automatic detection of user characteristics allows adaptive systems to be built, that is, to provide automatic adaptations to suit user characteristics.

In this chapter we made a contribution on adaptive systems by proposing a system with a two-step architecture that detects the web navigation problems of users with physical disabilities (see Figure 5.14).

The first step is to detect automatically the device being used to interact with the computer (joystick, keyboard, keyboard+headpointer, trackball or mouse). The second step is to detect the problems the user may be having while she/he is interacting with the computer. Knowing the device being used and the problems being encountered will allow the most adequate adaptation to be deployed. The system proposed in this chapter is based on web user interaction data collected by the RemoTest platform, and a complete data mining process applied to the data. In particular, 25 features were computed based on the interaction data gathered for each type of task and used them in the two stages of the system. Specifically for the first stage we built a hierarchical classifier with the best set of features (19) able first to discriminate between the two main groups of devices and then the specific device within each of them. From the adaptations point of view, we consider that the classifier built was effective, as the critical error occurred when missclassifying restricted movements and free movements devices was very small, 0.74%. In the second stage of the system for each type of task and each device k-means clustering algorithm was ran and then, clusters with high probabilities of containing problematic navigation patterns were automatically filtered based on particular standard deviation thresholds for a set of 11 meaningful features. By means of a visual analysis of the navigation traces grouped in the clusters we observed a total of seven problematic patterns: (P1) too much distance, (P2) too much time,(P3) rectifications in directions, (P4) unnecessary clicks, (P5) difficulties around the target, (P6) long clicks and (P7) too many stops. We closed this contribution by discussing the hypothetical reasons behind the detected patterns and by suggesting, according to these patterns and the devices used, the most suitable adaptation technique in some cases.

Limitations

The proposed system, due to its nature has some limitations. First, as the system has been built based on client-side data, the system can only be extended to new users if they allow the corresponding Firefox add-ons implementing the virtual aids for the cursor to be installed in their computers. The use of client-side data also limits the amount of data that could be used to build the system as it has had to be collected in controlled experiments.

The system is able to detect the used device and problems the users are having and we suggest adequate adaptations but the adaptations have not yet been activated.

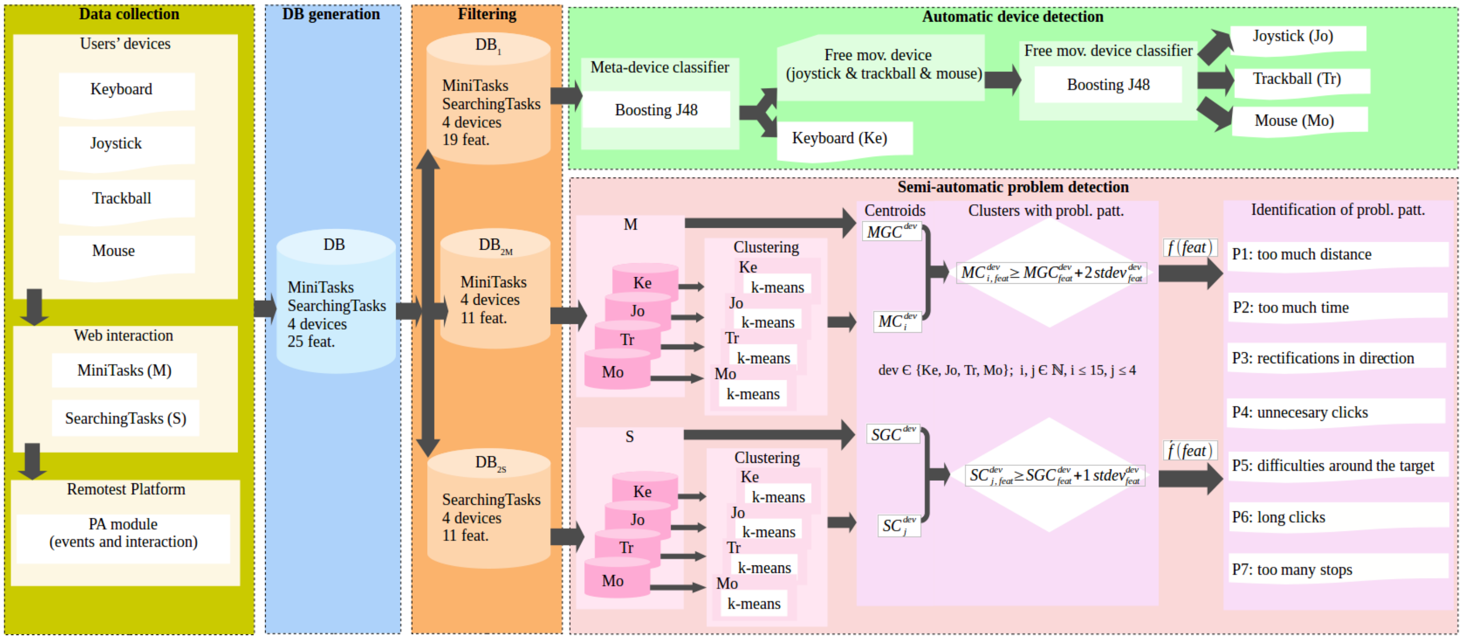


Figure 5.14: Description of the designed system.

Chapter 6

Modelling the interaction with specific web platforms

6.1 Introduction

The responsibility of health care professionals (especially General Practitioners) is shifting from a reactive patient-by-patient role to a proactive manager of population health. This shift requires the availability of health data and information tools that give a population-level view of such data, which allow the identification of individual patients that require intervention. Consequently, the use of medical dashboards is becoming increasingly important in using this data to improve healthcare. While the current wealth of clinical data satisfies the availability premise, it becomes, at the same time, a double-edged sword in that medical dashboards suffer from information overload. What is more, clinicians have varying levels of practical clinical experience, different problem-solving skills, and vary considerably in their IT skills. As the information density in the clinical environment is increasing rapidly and the role of medical dashboards is still at an early stage, it is of paramount importance to build smart adaptive systems that cater for the needs of clinicians and support them in the transition towards a proactive management of population health.

Medical dashboards display population data and are being used to monitor the health of communities and support clinicians in decision-making activities (Dowding et al. 2015). A few examples highlight the benefits of the visual nature of dashboards including successful interventions for diabetes care (Dagliati et al. 2018) and management of alerts triggered by drug-drug interactions (Simpao et al. 2014). Typically, medical dashboards display data in a tabular fashion and contain images, charts, numeric and textual information that may tax the perception and cognition of their user. While dashboards may help to alleviate information overload, paradoxically, they may also contribute to this problem by cluttering the screen with information and widgets – some have coined this phenomenon as the “blizzard of dashboards” (Kalra et al. 2016).

Information overload and substandard usability are well-known problems for electronic health record (EHR) systems (Middleton et al. 2013; Ratwani et al. 2015). To address this problem, usability guidelines that are sensitive to specific clinical settings and their typical tasks have been derived from general-purpose guidelines (Kushniruk and Patel 2004; Zhang and Walji 2011). Similar usability guidelines have also been formulated for medical dashboards (Brown et al. 2016). While implementing usability guidelines may address some of the most prominent and critical usability issues, users still feel overwhelmed by the amount of information on screen.

Consequently, it has been suggested that personalised and adaptive user interface capabilities should be implemented in order to: mitigate the complexity of audit and feedback interventions (Landis-Lewis et al. 2015); address information overload in electronic medical records (Zahabi et al. 2015); and improve the effectiveness of clinical tools for decision making (Brehaut et al. 2016). When it comes to medical dashboards, this is not without difficulties in that adapting the user interface to the user's needs calls not only for eliciting such needs, but also taking their skills and expertise into consideration (Dowding et al. 2017). Yet, detecting the skills and expertise of users (i.e. competence) that will inform the adaptations is particularly challenging due to the evolving nature of knowledge acquisition. This suggests that systems that adapt to the skills and expertise of the users should track competence automatically and unobtrusively.

In this context, in the first contribution we address the following research question: can we determine the users' visual behaviour based on their exhibited interactive behaviour? Preliminary work has analysed the relationship between gaze, which is a good indicator of interest (Ehmke and Wilson 2007), and interactive behaviour of users, finding out that mouse and gaze are strongly related: dwell times on specific regions are correlated with the likelihood of visiting that region with the mouse (Chen et al. 2001). In gaze prediction models for search engine results pages (SERPs), the inclusion of the mouse coordinates, the velocity and direction of the cursor, and the time elapsed since starting to view the results achieves an accuracy of 77% (Huang et al. 2012). However, unlike websites and SERPs, medical dashboards are constrained by grid layouts where data is displayed in a tabular fashion, which determines the variability of behaviours that can be exhibited.

On the other hand, we make a second contribution by computing proxies of competence on two cohorts of users of a medication safety intervention: a group of pharmacist who led the intervention (primary users) and a group of non-pharmacist who engaged less (secondary users).

Both analysis were carried out on the Salford Medication Safety Dashboard (SMASH), one of two vital components of a pharmacist-led information technology intervention for safe prescribing of medications in primary care, from which we collected user interaction data and whose purpose and functionalities we describe later in section 6.2. In particular we used the gaze and interaction data collected in a user study with six clinicians in the lab and the interaction data logged from a ten-month observational study with 35 clinicians in SMASH.

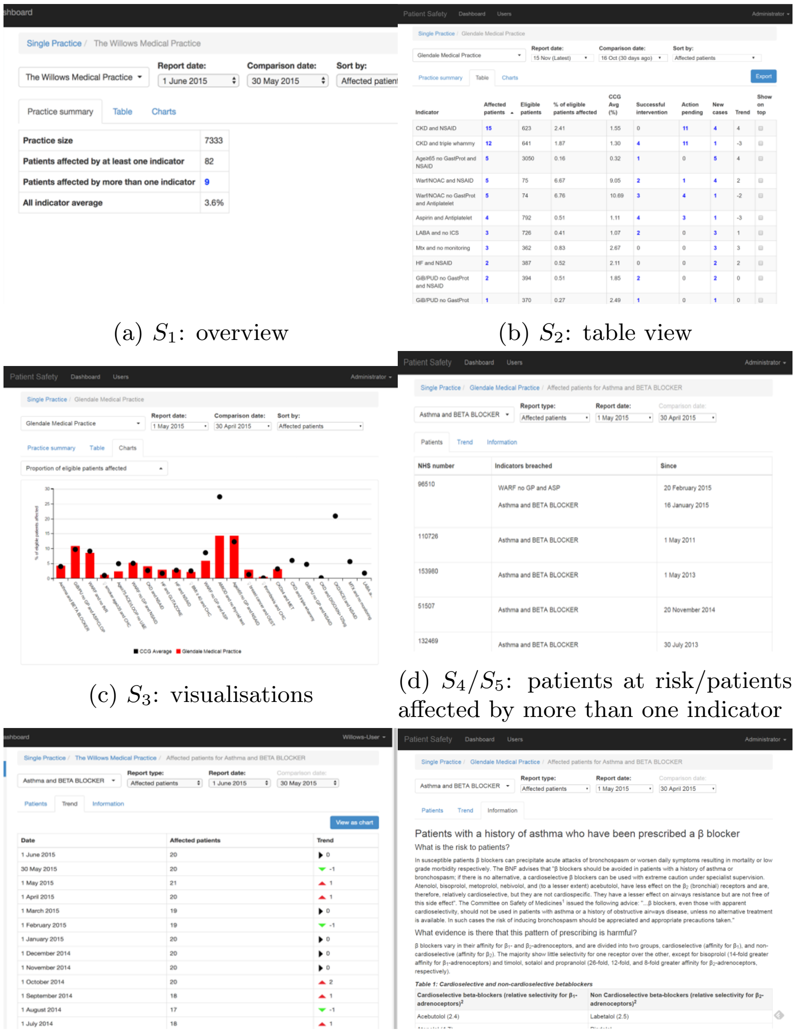
More concretely, in the first contribution, using exploratory unsupervised learning procedures we clustered the user study participants based on the collected interactive behaviour and we employed inferential statistics to find relationships between their visual behaviour. Then, we applied the same clustering analysis on the interaction data by adding the interaction of the 35 clinicians of the observational study and analysed whether the lab findings could reliably be extrapolated to a setting where no eye-tracking device is deployed. In the second contribution, using supervised learning techniques on the interaction data of the observational study, we were able to characterise and automatically distinguish the interactive behaviour of primary users who were leading the intervention and secondary users who used the dashboard to engage in safe prescribing practices.

6.2 Context: The SMASH Intervention

The SMASH intervention aims to determine whether implementation of a pharm-acist-led complex intervention reduces the incidence of potentially hazardous prescribing and medication monitoring practices in primary care across Salford, UK (Williams et al. 2018). The SMASH dashboard was implemented in 2016 and the quantitative evaluation of the impact on rates of potentially hazardous prescribing is ongoing with results expected in 2019. A concurrent qualitative process evaluation of the SMASH intervention has also been published (Jeffries et al. 2018).

The SMASH intervention is comprised of two main components: a web-based interactive dashboard that highlights patients exposed to potentially hazardous prescribing in general practices, and dedicated clinical pharmacist support involving collaborative working with practice staff to resolve hazardous prescribing cases and prevent their future occurrence using root cause analysis. The SMASH intervention follows that of the landmark pharmacist-led information-technology based intervention (PINCER) trial (Avery et al. 2012) but the incorporation of the interactive dashboard is novel.

The SMASH dashboard was co-designed with key stakeholders (Keers et al. 2015) and incorporates a refined set of 13 prescribing safety indicators which have previously been applied to measure the rate of potentially hazardous prescribing and medication monitoring (Akbarov et al. 2015; Stocks et al. 2015). For instance, the dashboard identifies all patients' with a history of peptic ulcer who have been prescribed a non-steroidal anti-inflammatory drug (NSAID; e.g. ibuprofen) without co-prescription of gastro-protective medication, which places them at risk of gastro-intestinal bleeding, a major adverse event with high mortality rates. The dashboard displays summary statistics for each of the indicators, counting how many patients are currently at risk in a given practice and relating those numbers to previous episodes and other practices. In addition, pharmacists and general practitioners (GPs) can view which patients are currently at risk for each indicator. The dashboard is deployed in Salford, a city in the Greater Manchester conurbation, comprising a population of 270,000 served by primary care with additional linkage to secondary care records.



The Willows Medical Practice

Report date: 1 June 2015

Comparison date: 30 May 2015

Sort by: Affected patien

Practice size	7333
Patients affected by at least one indicator	82
Patients affected by more than one indicator	9
All indicator average	3.6%

(a) S_1: overview

Glendale Medical Practice

Report date: 15 Nov (Latest)

Comparison date: 16 Oct (30 days ago)

Sort by: Affected patients

Indicator	Affected patients	Eligible patients	% of eligible patients affected	CCG Avg (%)	Successful intervention	Action pending	New cases	Trend	Show on top
CKD and NSAID	15	623	2.41	1.55	0	11	4	4	□
CKD and triple whammy	12	641	1.87	1.30	4	11	1	-3	□
Age≥65 no GastProt and NSAID	5	3050	0.16	0.32	1	0	5	4	□
Warf/NOAC and NSAID	5	75	6.67	9.05	2	1	4	2	□
Warf/NOAC no GastProt and Antiplatelet	5	74	6.76	10.69	3	4	1	-2	□
Aspirin and Antiplatelet	4	792	0.51	1.11	4	3	1	-3	□
LABA and no ICS	3	726	0.41	1.07	2	0	3	1	□
Mtx and no monitoring	3	362	0.83	2.67	0	0	3	3	□
HF and NSAID	2	387	0.52	2.11	0	0	2	2	□
GiB/PUD no GastProt and NSAID	2	394	0.51	1.85	2	0	2	0	□
GiB/PUD no GastProt	1	370	0.27	2.49	1	0	1	0	□

(b) S_2: table view

Glendale Medical Practice

Report date: 1 May 2015

Comparison date: 30 April 2015

Sort by: Affected patients

(c) S_3: visualisations

Patient Safety
Dashboard
Users
Administrator
Single Practice / Glendale Medical Practice / Affected patients for Asthma and BETA BLOCKER
Patients
Trend
Information

Asthma and BETA BLOCKER

Report type: Affected patients

Report date: 1 May 2015

Comparison date: 30 April 2015

NHS number	Indicators breached	Since
96510	WARF no GP and ASP Asthma and BETA BLOCKER	20 February 2015 16 January 2015
110726	Asthma and BETA BLOCKER	1 May 2011
153980	Asthma and BETA BLOCKER	1 May 2013
51507	Asthma and BETA BLOCKER	20 November 2014
132469	Asthma and BETA BLOCKER	30 July 2013

(d) S_4/S_5: patients at risk/patients affected by more than one indicator

ashboard
Willows-User
Single Practice / The Willows Medical Practice / Affected patients for Asthma and BETA BLOCKER
Patients
Trend
Information
View as chart

Asthma and BETA BLOCKER

Report type: Affected patients

Report date: 1 June 2015

Comparison date: 30 May 2015

Date	Affected patients	Trend
1 June 2015	20	▶ 0
30 May 2015	20	▼ -1
1 May 2015	21	▲ 1
1 April 2015	20	▲ 1
1 March 2015	19	▶ 0
1 February 2015	19	▼ -1
1 January 2015	20	▶ 0
1 December 2014	20	▶ 0
1 November 2014	20	▶ 0
1 October 2014	20	▲ 2
1 September 2014	18	▲ 1
1 August 2014	17	▼ -1
1 July 2014	18	▲ 1

(e) S_6: trends

Patient Safety
Dashboard
Users
Administrator
Single Practice / Glendale Medical Practice / Affected patients for Asthma and BETA BLOCKER
Patients
Trend
Information

Asthma and BETA BLOCKER

Report type: Affected patients

Report date: 1 May 2015

Comparison date: 30 April 2015

Patients with a history of asthma who have been prescribed a β blocker

What is the risk to patients?

In susceptible patients β blockers can precipitate acute attacks of bronchospasm or worsen daily symptoms resulting in mortality or low grade morbidity respectively. The BNF advises that "β blockers should be avoided in patients with a history of asthma or bronchospasm; if there is no alternative, a cardioselective β blockers can be used with extreme caution under specialist supervision. Atenolol, bisoprolol, metoprolol, nebivolol, and (to a lesser extent) acebutolol, have less effect on the $β_2$ (bronchial) receptors and are, therefore, relatively cardioselective, but they are not cardiospecific. They have a lesser effect on airways resistance but are not free of this side effect". The Committee on Safety of Medicines[1] issued the following advice: "...β blockers, even those with apparent cardioselectivity, should not be used in patients with asthma or a history of obstructive airways disease, unless no alternative treatment is available. In such cases the risk of inducing bronchospasm should be appreciated and appropriate precautions taken."

What evidence is there that this pattern of prescribing is harmful?

β blockers vary in their affinity for $β_1$- and $β_2$-adrenoceptors, and are divided into two groups, cardioselective (affinity for $β_1$), and non-cardioselective (affinity for $β_2$). The majority show little selectivity for one receptor over the other, except for bisoprolol (14-fold greater affinity for $β_1$-adrenoceptors) and timolol, sotalol and propranolol (26-fold, 12-fold, and 8-fold greater affinity for $β_2$-adrenoceptors, respectively).

Table 1: Cardioselective and non-cardioselective betablockers

Cardioselective beta-blockers (relative selectivity for $β_1$-adrenoceptors)[2]	Non Cardioselective beta-blockers (relative selectivity for $β_2$-adrenoceptors)[2]
Acebutolol (2.4)	Labetalol (2.5)

(f) S_7: indicator information

Figure 6.1: Screenshots of the SMASH dashboard (from the top-left to bottom right): (a) S_1: practice overview, (b) S_2: tabular view of the safety indicators; (c) S_3: the visualisation of the safety indicators; (d) S_4/S_5: the list of patients at risk; (e) S_6: indicator trends and (f) S_7: screen containing evidence about why an indicator is a safety hazard.

As shown in Figure 6.1 the user interface of SMASH is divided into seven screens or views: (a) S_1: a landing page containing a tabular overview of a given practice including the size of the practice and the number of patients affected by more than one indicator; (b) S_2: a table view displaying the number of patients who are affected by the indicators, their severity, the number of eligible patients and the percentage of patients who are affected. Indicators can be contraindications between drugs and conditions (e.g. chronic kidney disease and NSAIDs) or between drugs, habits and demographics; (c) S_3: graph-based visualisations displaying the incidence of indicators as time-series. Clicking on the number of patients at risk on the table view S_2 leads to (d) S_4, a list of patients at risk for a specific indicator, while clicking on the link'Patients affected by more than one indicator' on the overview page (a) S_1 leads to (d) S_5, which is a patient list similar to S_4 but containing only those patients that are affected by more than one indicator. (e) S_6 displays the trends for a given indicator over time and (f) S_7 contains information and pointers to the medical literature about why a certain indicator is considered a risk.

The SMASH dashboard logs the user interface events triggered by the users in a dataset on the server. Because SMASH is a mouse-driven application the collected events are mostly mouse clicks and mouse hovers. A third event logged is the page load event, which signals navigation to a different view (e.g. from the data table S_2 to the data visualisation S_3) that does not necessarily entail an update in the URL and is triggered by clicking on the 'Selection menu'. For each event, SMASH collects the user id, the identifier of the session (i.e. every time a user logs in, a new session is established), the timestamp, the URL where the event took place and the specific element on the user interface where the event occurred indicated by an XPATH statement.

6.3 User studies and metrics

As mentioned in the introduction, in our contributions we used the gaze and interaction data from two different user studies, lab and observational, which we describe in this section.

6.3.1 Lab study

In the first contribution we analysed fixation data collected by an eye-tracker and the user interaction data collected by SMASH for participants of the lab study, as well as the interaction data of the participants from the observational study which is described hereunder.

Six participants (four male) took part in the lab study, five General Practitioners (GPs) and one pharmacist, with an average age of 38 (stdev = 10 and age range = 30–56). In particular, the Tobii X2-60 eye-tracker was employed in the laboratory study to log gaze information including fixation coordinates on the screen, duration of the fixations and the saccades (movement of eyes between the fixations).

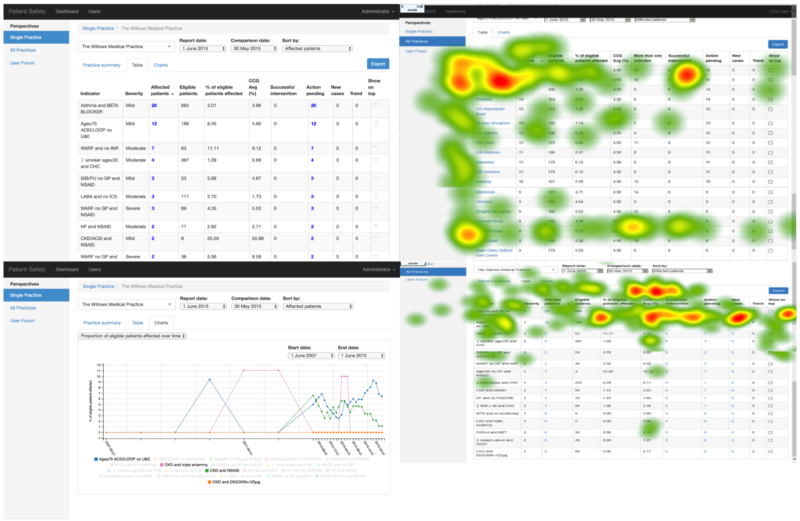
In this study participants were asked to complete nine tasks classified in three ways: a) Identification of patients at risk: i.e. ''List up to three patients at risk for indicator X"; b) Identification of problems in the practice and their evolution over time: i.e. ''Identify the three indicators with the largest number of patients affected"; c) Comparison of problems between practices: i.e. ''Identify three indicators in which your practice performs worse than others"

The main nine areas of interest (AOIs) of the SMASH user interface (see Figure 6.3) considered in the design of the lab study we analysed, were defined based on the gaze patterns observed in a previous pilot study described in the next section.

Pilot study

Five participants (3 female) who were 39 years old on average (stdev = 13.5 and age range = 27–62) took part in the pilot study. All of them were computer savvy and familiar with the domain and terminology of the medication safety dashboard. Two of them were members of the Research User Group, a pool of users who frequently take part in e-Health studies, and of the remaining three: one had a degree in nursing, another one was doing a PhD in nursing and one was a medical microbiologist. In this study participants were asked to complete the same nine tasks described in the laboratory study.

A qualitative analysis of the heatmaps of the pilot study yielded some interesting insights: the visual search strategies on the dashboard followed particular patterns. Figure 6.2 shows some examples of the heatmap patterns generated in different screens of SMASH.

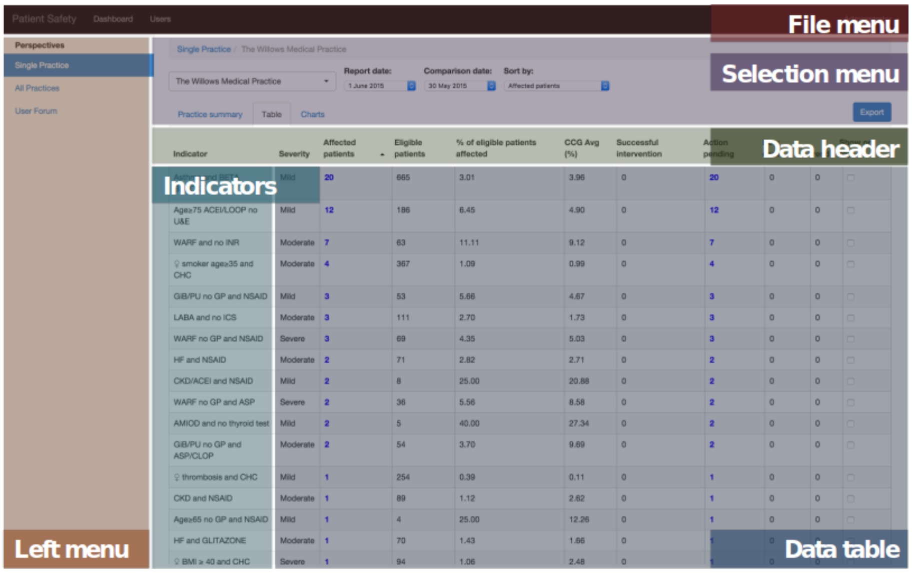


Indicator	Severity	Affected patients	Eligible patients	% of eligible patients affected	CCG Avg (%)	Successful intervention	Action pending	New cases	Trend	Show on top
Asthma and BETA BLOCKER	Mild	20	665	3.01	3.96	0	20	0	0	□
Age≥75 ACEI/LOOP no U&E	Mild	12	186	6.45	4.90	0	12	0	0	□
WARF and no INR	Moderate	7	63	11.11	9.12	0	7	0	0	□
♀ smoker age≥35 and CHC	Moderate	4	367	1.09	0.99	0	4	0	0	□
GiB/PU no GP and NSAID	Mild	3	53	5.66	4.67	0	3	0	0	□
LABA and no ICS	Moderate	3	111	2.70	1.73	0	3	0	0	□
WARF no GP and NSAID	Severe	3	69	4.35	5.03	0	3	0	0	□
HF and NSAID	Moderate	2	71	2.82	2.71	0	2	0	0	□
CKD/ACEI and NSAID	Mild	2	8	25.00	20.88	0	2	0	0	□
WARF no GP and	Severe	2	36	5.56	8.58	0	2	0	0	□

Figure 6.2: Heatmap patterns (right) generated on the pilot study for the table view (top-left) and visualisations (bottom-left) of SMASH.

The C-shaped behaviour in Figure 6.2 (top-right) suggests that users look at the data header, the list of indicators on the left and the values in a row belonging to a particular indicator. On the other hand, the paint drop pattern in Figure 6.2 (bottom-right) indicates that users look at the header and the top rows and visual search is restricted to a few columns. This strategy can be explained by the fact that some users discovered that clicking on a header sorted the indicators based on the values of the corresponding column/variable, which was an effective strategy for completing many tasks and reduced the need for visual exploration. While the boundaries between the components of the dashboard are clear, it is difficult to establish the AOIs in a tabular environment.

The gaze patterns found as well as the demarcation of existing user interface elements informed the design of the areas of interest (AOIs), accounting for nine of them: (1) file menu, (2) selection menu, (3) left menu, (4) practice summary, (5) data header, (6) indicators, (7) data table, (8) chart (visualisation of data) and (9) drop down menu (any drop-down menu folding down after clicking). As depicted by Figure 6.3 the findings suggest that the column containing the safety indicators, the table header and the remaining rows should constitute independent areas of interest.



Single Practice / The Willows Medical Practice

The Willows Medical Practice

Report date: 1 June 2015

Comparison date: 30 May 2015

Sort by: Affected patients

Indicator	Severity	Affected patients	Eligible patients	% of eligible patients affected	CCG Avg (%)	Successful intervention	Action pending			
Asthma and BETA	Mild	20	665	3.01	3.96	0	20	0	0	□
Age≥75 ACEI/LOOP no U&E	Mild	12	186	6.45	4.90	0	12	0	0	□
WARF and no INR	Moderate	7	63	11.11	9.12	0	7	0	0	□
♀ smoker age≥35 and CHC	Moderate	4	367	1.09	0.99	0	4	0	0	□
GIB/PU no GP and NSAID	Mild	3	53	5.66	4.67	0	3	0	0	□
LABA and no ICS	Moderate	3	111	2.70	1.73	0	3	0	0	□
WARF no GP and NSAID	Severe	3	69	4.35	5.03	0	3	0	0	□
HF and NSAID	Moderate	2	71	2.82	2.71	0	2	0	0	□
CKD/ACEI and NSAID	Mild	2	8	25.00	20.88	0	2	0	0	□
WARF no GP and ASP	Severe	2	36	5.56	8.58	0	2	0	0	□
AMIOD and no thyroid test	Mild	2	5	40.00	27.34	0	2	0	0	□
GIB/PU no GP and ASP/CLOP	Moderate	2	54	3.70	9.69	0	2	0	0	□
♀ thrombosis and CHC	Mild	1	254	0.39	0.11	0	1	0	0	□
CKD and NSAID	Moderate	1	89	1.12	2.62	0	1	0	0	□
Age≥65 no GP and NSAID	Mild	1	4	25.00	12.26	0	1	0	0	□
HF and GLITAZONE	Moderate	1	70	1.43	1.66	0	1	0		
♀ BMI ≥ 40 and CHC	Severe	1	94	1.06	2.48	0	1	0		

Figure 6.3: Six of the nine AOIs defined in SMASH; the remaining three AOIs correspond to pop-up dialogues (9) and charts (8).

6.3.2 Observational study

35 participants, 10 pharmacists and 25 non-pharmacists, took part in the observational study that ran for a period of 10 months and where user interaction data was collected by the SMASH dashboard. In this study no tasks were given to the participants since the dashboard was used for the purpose it was intended: the identification of those patients at risk and the promotion of good prescrip-

tion practices. We expected that the participants in this study would carry out tasks of a higher ecological validity than those given in the lab setting.

The group of pharmacists named primary user, leaded the intervention and received 2–3 hours of face-to-face formal structured training based on the training principles of the PINCER trial (Sadler et al. 2014) including: an interactive seminar covering the background and rationale of the SMASH project, a guided tour of the SMASH dashboard, and the principles of root cause analysis to identify the cause of problems. The group of non-pharmacists, named secondary users, consisted of eight members of the Clinical Commissioning Group (CCG), eight GPs, five managers, and four other including nurses and pharmacy technicians who were trained by the primary users following similar procedures.

In the qualitative evaluation of the intervention, primary users indicated that the dashboard added value to their work, while secondary users reported some resistance to engage with it, as some perceived the dashboard was owned by primary users. These attitudes had implications for engagement in that primary users engaged more with the intervention than their colleagues (Jeffries et al. 2018). This is important because the literature concerning engagement at the workplace suggests that those individuals who are engaged are more competent and perform better (Rich et al. 2010; Christian et al. 2011). This finding has been confirmed in a variety of settings including healthcare (Laschinger et al. 2009).

Therefore, the main goal of this study, which we analysed in our second contribution, was to explore whether we could model the interactive behaviour in terms of competence of two groups of electronic dashboard users who reported different levels of engagement in the qualitative evaluation of the SMASH intervention. This would allow us to examine how the SMASH dashboard is being used and identify distinctive interactive behaviours, and how this may then inform our understanding of the use of medical dashboards in general (Dowding et al. 2015).

6.3.3 Computed metrics

Gaze metric

In particular, as shown in Equation 6.1 the gaze activity of the lab study (G) was computed using the average fixation duration (henceforth fd) feature gathered by the eye tracker in the nine Areas Of Interest (AOI_j) that were defined as a consequence of the pilot study:

$$G = fd_{AOI_j}, j \in \mathbb{N}, j \leq 9 \tag{6.1}$$

Fixation duration is known to be a proxy for cognitive load (Ehmke and Wilson 2007) so our premise is that, if we want to relate visual behaviour to interactive behaviour on SMASH, cognitive load might well be an indicator to profile participants.

Interaction metrics

Based on the information of the logs gathered by SMASH (user and session IDs, events, timestamps and URLs), we computed features for exploration and dwell time for the lab and observational studies. In both features clicks events are used as reference, knowing that a click triggers an update of the current view by filtering information or leads to another screen of the dashboard.

- **Exploration (e):** median of the number of mouse hovers between two consecutive mouse clicks. This is based on the fact that since mouse location on screen is a proxy of gaze location (Guo and Agichtein 2010), it can be used to quantify visual exploration. Higher exploration values suggest more visual search activities.
- **dwell time (d):** median of the elapsed time between two consecutive mouse clicks. This is supported by the fact that the time spent on a screen is an indicator of how effective users are processing information and solving problems. A study found that, on information seeking tasks, longer times were correlated with lower cognitive ability (Chin et al. 2009). Lower dwell time conveys higher efficiency accomplishing tasks.

Using these features, we created two interaction metrics (vectors) per participant, considering their global interaction in all of the screens ($V1$ global perspective) or their interaction in each of the screens ($V2$ screen perspective).

The first metric $V1$ defined Equation 6.2 describes user interaction on all of the screens available and is computed as a vector of two features, the global exploration (e) and dwell time (d) on SMASH.

$$V1 = (e, d) \tag{6.2}$$

The second metric $V2$ defined in Equation 6.3, takes into consideration the above features (e_{S_i}, d_{S_i}) in each of the seven screens of SMASH ($S_i, i \in \mathbb{N}$, $i \leq 7$) and is represented as a vector of 14 features per participant.

$$V2 = (e_{S_i}, d_{S_i}), i \in \mathbb{N}, i \leq 7 \tag{6.3}$$

6.4 Inferring visual behaviour from interaction data

6.4.1 Description of the datasets

In this section we describe the three datasets built in this first contribution with the metrics computed based on the data gathered during the lab an observational studies, one using the gaze data of the lab study and two using the interaction data of both studies.

Gaze dataset: lab study

In particular for each of the six participants of the lab study we computed the gaze activity (G) described in Equation 6.1, which measures the average fixation duration (seconds), fd_{AOI_j}, in each of the nine Areas of Interest defined ($AOI_j, j \in \mathbb{N}, j \leq 9$). Table 6.1 shows the values of G for all the lab participants.

Part	M	Feature	Areas of Interest (AOI_j)								
			j=1	j=2	j=3	j=4	j=5	j=6	j=7	j=8	j=9
P1	G	fd_{AOI_j}	0.0	173.3	241.9	167.0	254.5	240.1	168.1	164.3	140.6
P2	G	fd_{AOI_j}	230.2	223	213.2	158.5	200.8	222.9	228	207.2	148.5
P3	G	fd_{AOI_j}	212.9	220.5	268.7	295.0	294.0	248.2	244.7	263.2	235.5
P4	G	fd_{AOI_j}	223.2	250	299.3	289.4	260.4	234.5	275.4	203.1	79.0
P5	G	fd_{AOI_j}	201.6	151.4	201.2	189.0	264.8	163.0	243.3	278.9	91.5
P6	G	fd_{AOI_j}	193.5	173.4	248.8	0.0	254.9	204.2	151.2	176.1	157.5

Table 6.1: Values of the gaze activity (G) metric (M) for the lab participants (P), computed based on the fixation duration on average (seconds) in each of the nine AOIs (fd_{AOI_j}).

Interaction datasets: lab and observational studies

On the other hand, we computed two interaction metrics for the lab participants, $V1$ and $V2$, based on the global and screen divided ($Si, i \in \mathbb{N}, i \leq 7$) explorations ($e/e_{S_i}$) and dwell times ($d/d_{S_i}$) (see Equations 6.2 and 6.3). Table 6.2 shows the values of $V1$ and $V2$ for all the lab participants.

Finally, $V1$ and $V2$ were also computed for the 35 participants of the observational study based on all the interaction they carried out in SMASH, this is, the user perspective. In Section 6.5 we specify the values on average of these metrics for all of the participants and for the two main groups of participants, primary and secondary users. In this case, we built a dataset with the interaction data ($V1$ and $V2$) of 41 participants, six from the lab study and 35 from the observational study (user perspective).

Participant	Metric	Feature	Global	Screens (Si)						
				i=1	i=2	i=3	i=4	i=5	i=6	i=7
P1	$V1$	e	4.00							
		d	4.28							
	$V2$	e_{S_i}		3.00	4.00	3.00	6.00	0.00	4.00	0.00
		d_{S_i}		3.90	6.50	3.75	6.95	0.00	2.80	0.00
P2	$V1$	e	5.00							
		d	6.54							
	$V2$	e_{S_i}		5.50	5.00	3.00	4.00	0.00	3.00	0.00
		d_{S_i}		6.03	9.23	2.99	3.35	0.00	1.22	0.00
P3	$V1$	e	4.00							
		d	3.22							
	$V2$	e_{S_i}		3.50	4.00	3.00	3.50	0.00	3.00	4.00
		d_{S_i}		2.37	4.49	4.88	3.01	0.00	4.22	2.41
P4	$V1$	e	3.00							
		d	3.31							
	$V2$	e_{S_i}		4.00	3.00	3.50	2.00	0.00	2.00	0.00
		d_{S_i}		5.36	3.82	4.11	0.95	0.00	0.64	0.00
P5	$V1$	e	3.00							
		d	2.26							
	$V2$	e_{S_i}		3.00	3.00	3.00	3.00	5.00	3.00	0.00
		d_{S_i}		2.77	1.71	1.90	2.18	3.60	5.76	0.00
P6	$V1$	e	4.00							
		d	9.43							
	$V2$	e_{S_i}		9.00	4.00	3.00	6.00	4.00	0.00	0.00
		d_{S_i}		36.39	6.20	6.59	7.13	5.40	0.00	0.00

Table 6.2: Values of the interaction metrics $V1$ and $V2$ for the lab participants, computed based on the global and screen divided exploration (e/e_{S_i}) and dwell time (d/d_{S_i}).

6.4.2 Results and analysis

Gaze data analysis in the lab

In order to find participants with similar fixation durations across the different AOIs (fd_{AOI_j}) we run Pearson correlation analysis between the G vectors. Consequently, a positive correlation between any two participants would entail similar visual behaviours in terms of cognitive load. In particular, we paired those participants with similar visual behaviour using the highest value for the Pearson correlation computed in each case – note that data was normally distributed according to the Shapiro-Wilk test ($p > 0.05$). Table 6.3 shows the results of the Pearson correlation coefficient (r) for the gaze metric G computed for the six participants of the lab study.

Pearson correlation (r) computed for G

	P1	P2	P3	P4	P5	P6
P1	1.00	-0.05	**0.63**	0.33	0.15	0.26
P2	-0.05	1.00	-0.40	0.47	0.42	**0.55**
P3	**0.63**	-0.40	1.00	0.41	0.45	-0.18
P4	0.33	0.47	0.41	1.00	**0.53**	-0.03
P5	0.15	0.42	0.45	**0.53**	1.00	0.20
P6	0.26	**0.55**	-0.18	-0.03	0.20	1.00

Groups: {P1,P3}, {P2,P6} and {P4,P5}

Table 6.3: Pearson correlation coefficient (r) computed for the gaze data, G, of the six lab participants.

Analysing Table 6.3, we identified three groups, which paired P1 and P3 (r = 0.63, p-value = 0.06), another one pairing P2 and P6 (r = 0.55, p-value = 0.11) and a last one pairing P4 and P5 (r = 0.53, p-value = 0.13). It is well known that p-values are sensitive to the sample size. Since the G vectors contain nine items, an alpha value < 0.95 is justifiable so we can say that the moderate-high correlations found show a clear tendency towards significance.

Interaction data analysis in the Lab

For the sake of identifying those users from the lab study who exhibited similar interactive behaviours, we applied different clustering algorithms including k-means and single-linkage method (Jain and Dubes 1988) to the two interaction metrics we defined, the global one $V1$ and the screen divided one $V2$.

Specifically, the k-means algorithm (k=3 and Euclidean distance) clustered clinicians in three pair wise groups for $V1$: P4-P5, P1-P3 and P2-P6. To this regard, a Silhouette analysis on Cluster Validity Indexes (Arbelaitz et al. 2013b) indicated that for $V1$ k=3 was the most appropriate cluster configuration when compared to $k = 4$ and $k = 5$ obtaining scores of 0.51, 0.20 and 0.003 respectively.

To better understand the structure of these groups we carried out a second clustering procedure including the centroids of the clusters generated by the k-means algorithm. We then computed the distance matrix for $V1$ and then calculated the centroids using the Euclidean distance again. The resulting distance matrix can be visualised using hierarchical clustering techniques (Jain and Dubes 1988). Thus, we applied the single-linkage and Ward hierarchical clustering algorithms using the Euclidean distance. Figure 6.4 shows the resulting dendrogram for the single-linkage clustering procedure, where the height at which participants are grouped represents the distance between clusters. The arrangement of clusters in the dendrogram shows three main branches that group the participants and the centroids together: P4-P5-C1, P1-P3-C2 and P2-P6-C3 which indicates that the groups discovered by k-means applied to the global interactive behaviour of the participants remain stable.

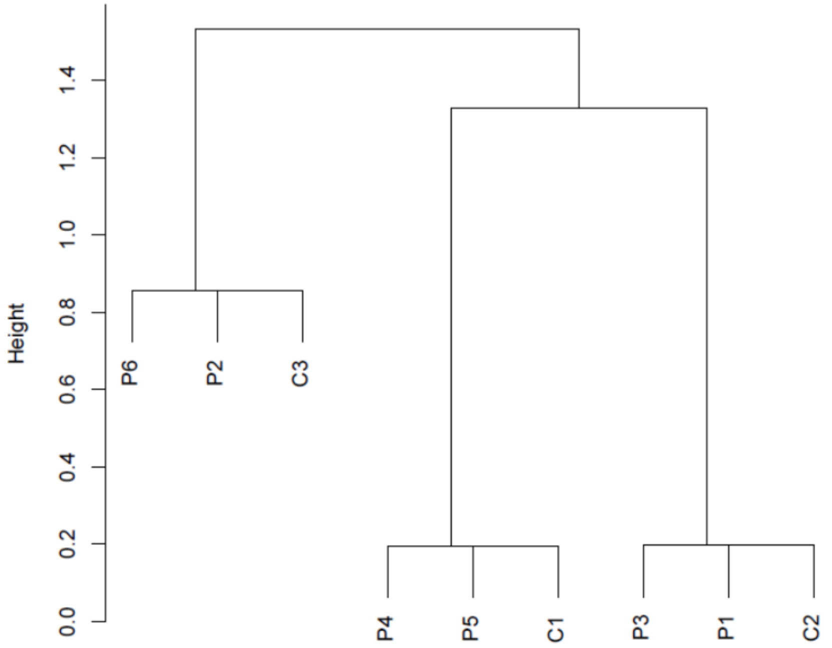


Figure 6.4: Single-linkage algorithm dendrogram for the distance matrix and the computed centroids of $V1$.

Regarding the closeness or similarity of the patterns, it can be observed that the groups of P1-P3 and P4-P5 are more compact since they are placed at the bottom of the dendrogram. To better show the proximity of the six participants, we performed a more exhaustive study of the global interaction analysis and used the neighbour-joining tree estimation of Saitou and Nei (Saitou and Nei 1987) over the distance matrix of $V1$, excluding the centroids. In the resulting tree shown in Figure 6.5, it can be seen that P2 and P6 are at some distance from the remaining participants, and P6 in particular is further than any other. This suggests that the cluster of groups P2 and P6 was not as compact as the other ones.

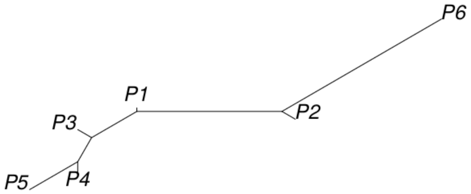


Figure 6.5: Neighbour joining method for the distance matrix of $V1$.

We applied the same pattern discovery method on $V2$, that is, we ran the k-means algorithm (k=3) using the Euclidean distance, then computed the centroids of the resulting clusters. As done with in the previous procedure ($V1$), the k value was selected according to Silhouette, which indicated that for $V2$, among the values tested (k={3,4,5}), k=3 was the most appropriate cluster configuration, achieving scores of 0.002, -0.011 and -0.046 respectively. Again, we computed the distance matrix for $V2$ and the centroids calculated in the previous step, using the Euclidean distance. Accordingly, the size of the distance matrix was of 9 x 9 measuring the distance based similarity of the six participants and the three centroids (C1-C3). Finally, we studied the proximity of the patterns discovered in $V2$, using the single-linkage algorithm (Jain and Dubes 1988) in the computed distance matrix. Figure 6.6 shows the visual output of the single-linkage algorithm and illustrates how participants were distributed in the same form for the two clustering procedures used in $V2$, k-means and hierarchical clustering. These patterns matched the ones found for $V1$ when using the same clustering procedures. When we analysed the proximity of the patterns represented in Figure 6.7, we found again that P2 and P6 are quite far from the other two groups.

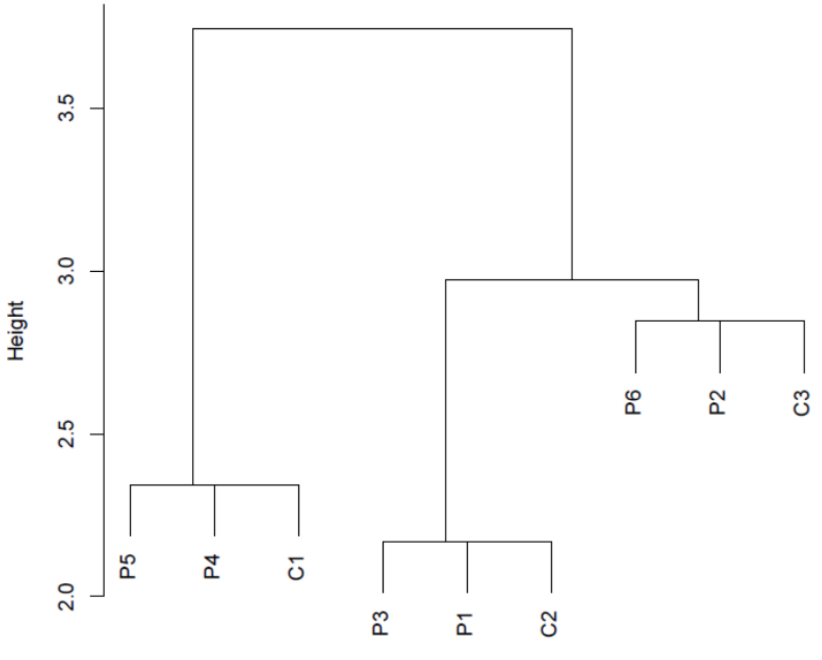


Figure 6.6: Single-linkage algorithm dendrogram for the distance matrix and the computed centroids of $V2$.

The patterns discovered on user interaction data (see Figures 6.4 and 6.6) and the emerging gaze patterns generate the same groupings. That is, those individuals having similar interactive behaviour happen to have related visual behaviour. Specifically, individuals with a similar cognitive load (as indicated by fixation durations) exhibit similar mouse use as captured by the exploration and dwell time features on SMASH ($V1$) and on its seven views ($V2$).

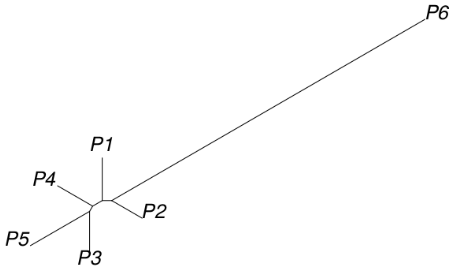


Figure 6.7: Neighbour joining method for the distance matrix of $V2$.

Interaction data analysis in the observational study

We carried out the analysis of user interaction data including the data of the lab participants and that of those who took part in the observational study. The purpose of analysing the two datasets together was to ascertain whether the emerging clusters would include the six laboratory participants in the same pairs. If the pairs of users fell again in the same clusters we could speculate that those participants belonging to the same cluster would have similar search behaviour to their lab counterparts. We therefore re-run our analysis (i.e. k-means and Euclidean distance) on $V1$, which this time accounted for 41 participants (i.e. 35 from the observational study + six from the lab).

A Silhouette analysis on Cluster Validity Indexes indicated, again, that k = 3 is the most appropriate cluster configuration for all ks, where $3 \leq k \leq 10$. Table 6.4 below shows the distribution of the lab study participants in the generated clusters.

	Clusters		
	1	2	3
Lab participants	P6	P4, P5	P1, P2, P3
No. participants from the observational study	2	19	14

Table 6.4: Results of k-means (k=3, d = Euclidean) for $V1$ when merging the participants of the lab and the observational study

The results indicate that the six laboratory participants are grouped in a similar way. P2 was the only participant falling in a different group, as instead of belonging to the same pair as P6, it was a member of P1 and P3's group. Taking into account the proximity analysis carried out for $V1$ (see Figure 6.5) this finding was not surprising given that P2 and P6's clustering was unstable. Hence, the fact that P2 switched groups would be understandable.

6.5 Characterizing different types of users

6.5.1 Description of the datasets

In this section we describe the two datasets built in this second contribution with the metrics computed based on the data collected in the observational study, one using the global perspective ($V1$) and the other using the screens perspective ($V2$).

The interaction data generated by users of the observational study was retrieved from the dataset and cleansed, which involved identifying users accessing the platform with different credentials, removing variables that were not necessary for this study and eliminating entries corresponding to software engineers and people testing the platform. We analysed the data from two perspectives: users and sessions. In the users approach all the events of each user are compiled in one record irrespective of the sessions (accounting for 35 records, split in two cohorts of primary and secondary users). In the sessions approach, one record contains all the events corresponding to a single session (accounting for 564 records).

Since users have to be logged into the SMASH dashboard in order to access the platform we could have computed session durations by using the total period during which users are logged in. However, this method overestimates the length of the session as long periods of inactivity would be included. Hence, we took a more granular approach by which 20 minutes of inactivity would indicate that the session was finished and another session would start as soon as the activity resumed. This approach is in line with the literature on identifying user sessions (Heer and Chi 2002). By applying this method we record a total of 564 sessions: 419 corresponding to primary users (74%) and 145 to secondary users (26%) distributed as 64 exhibited by CCG staff, 27 by GPs, 32 by general practice managers and 22 by other. The unequal number of sessions confirms what qualitative studies (Jeffries et al. 2018) reported on higher engagement of primary users: 29% of the users (primary) generated 74% of the sessions.

Both interaction representations (global and screens) were computed for each user and for each session. In both types of analysis the records are labelled according to the user group of the user (i.e. primary and secondary). Table 6.5 shows the distribution of users and sessions per user group.

User group	N (%)	Sessions (%)
Primary users	10 (29)	419 (74)
Secondary users	25 (71)	145 (26)
GPs	8 (23)	27 (5)
CCG staff	8 (23)	64 (11)
Managers	5 (14)	32 (6)
Other	4 (11)	22 (4)

Table 6.5: Number of users and number of sessions per group.

Tables 6.6 and 6.7 show values of exploration (e, e_{S_i}) and dwell time (d, d_{S_i}) on the global ($V1$) and screens ($V2$) representation from the user and session perspectives respectively. According to the tables in general, secondary users exhibit higher values for both features, although, this behaviour is not consistent across all the screens.

User group	Metric	Feature	Global	Screens (Si)						
				i=1	i=2	i=3	i=4	i=5	i=6	i=7
All	$V1$	e	3.40							
		d	3.47							
	$V2$	e_{S_i}		3.84	3.21	3.29	2.83	2.34	1.87	1.60
		d_{S_i}		8.93	2.75	16.02	3.68	6.28	2.15	7.93
Primary	$V1$	e	2.90							
		d	2.23							
	$V2$	e_{S_i}		3.70	3.30	2.75	2.50	3.50	2.70	3.40
		d_{S_i}		16.88	2.82	2.54	2.10	5.88	1.59	3.27
Secondary	$V1$	e	3.60							
		d	3.97							
	$V2$	e_{S_i}		3.90	3.18	3.50	2.96	1.88	1.54	0.90
		d_{S_i}		5.75	2.72	21.41	4.31	6.44	2.38	9.81

Table 6.6: Values of the interaction metrics $V1$ and $V2$ computed based on the global and screen divided exploration (e/e_{S_i}) and dwell time (d/d_{S_i}) average on observational study participants from the user perspective.

User group	Metric	Feature	Global	Screens (Si)						
				i=1	i=2	i=3	i=4	i=5	i=6	i=7
All	$V1$	e	3.79							
		d	3.28							
	$V2$	e_{S_i}		4.18	3.43	1.05	1.94	0.79	0.62	0.47
		d_{S_i}		6.81	10.01	4.6	3.03	3.4	1.05	3.35
Primary	$V1$	e	3.65							
		d	2.85							
	$V2$	e_{S_i}		4.08	3.52	0.81	1.82	0.87	0.71	0.53
		d_{S_i}		5.53	6.97	3.59	3.1	3.84	1.23	3.55
Secondary	$V1$	e	4.19							
		d	4.51							
	$V2$	e_{S_i}		4.45	3.14	1.74	2.29	0.58	0.36	0.22
		d_{S_i}		10.48	11.01	7.48	2.84	2.14	0.53	2.75

Table 6.7: Values of the interaction metrics $V1$ and $V2$ computed based on the global and screen divided exploration (e/e_{S_i}) and dwell time (d/d_{S_i}) average on observational study participants from the session perspective.

6.5.2 Results and analysis

The user perspective

In our analysis we used the Weka machine learning software (M. Hall et al. 2009) to classify the different types of users – primary and secondary – using machine learning algorithms, and used 10-fold cross validation (CV) to evaluate the performance of the algorithms. In particular, among the 10-top algorithms (Wu et al. 2008), we selected those applicable to our problem (AdaBoost, IBK, J48, NB and SMO) and we included some other very extended algorithms such as Bagging and MLP.

We calculated precision (ratio of users that were correctly classified, out of those who were predicted to belong to a particular class: i.e. primary or secondary), recall (ratio of users that were correctly classified, out of those who belong to a particular class) and F-measure (the harmonic mean of precision and recall).

The dataset was found to be unbalanced as 10 primary users accounted for 29% of users out of 35 individuals. Table 6.8 shows the performance of the algorithms where IBK, J48 and Naïve Bayes (NB) produce scores for precision, recall and F-measure above 0.80 for the screens representation. Scores were lower when not taking into consideration the particular screens (i.e. global) although the MLP algorithm achieved values above 0.80 for both representations.

Algorithm	Representation	Precision	Recall	F-measure
AdaBoost	global	0.73	0.71	0.72
	screens	0.77	0.77	0.77
Bagging	global	0.67	0.71	0.67
	screens	0.74	0.74	0.74
IBK	global	0.71	0.71	0.71
	screens	0.84	0.83	0.83
J48	global	0.60	0.60	0.60
	screens	0.84	0.83	0.83
MLP	global	0.83	0.83	0.83
	screens	0.86	0.86	0.86
NB	global	0.77	0.74	0.75
	screens	0.90	0.89	0.89
SMO	global	0.51	0.71	0.60
	screens	0.72	0.74	0.71

Table 6.8: Precision, recall and F-measure on users per algorithm.

The scores in Table 6.8 were useful to the extent that false positives (computed for precision) and false negatives (computed for recall) can be tolerated, which is ultimately dependent on the purpose of the classifiers. While discriminating users (primary vs secondary) is certainly necessary for user modelling purposes, the extraction of the characteristics of each user group provided valuable information that could inform future adaptations. Among the selected classifiers the classifier with the clearest explaining capacity is J48 (Witten et al. 2016), where the explanation is given by the path between the root node and the leaves where the examples have fallen in the classification process. We analysed the structure of the J48 classifier generated with the complete sample for the global and screens representation, using all the available information, because this would be the tree that would be deployed in a real system.

Despite J48 having lower scores than other algorithms for the global representation ($V1$), dwell time (d) was able to discriminate primary from secondary users in the SMASH dashboard platform. According to the structure of J48's classification in Figure 6.8, if dwell time was smaller or equal to 2.5 seconds (i.e. the user spends 2.5 seconds or less between mouse clicks), primary users accounted for 62% (8/13) of the users (note that initially primary users comprised 29% of users), whereas secondary users constituted 91% (19/22) of the individuals (up from the original 71%) when dwell time was restricted to observations greater than 2.5 seconds. Note that the average dwell time score when using the SMASH dashboard was 3.47 (stdev = 1.87), which revealed that primary users spent less time between clicks (mean 2.23 seconds) compared to their counterparts (mean 3.97 seconds).

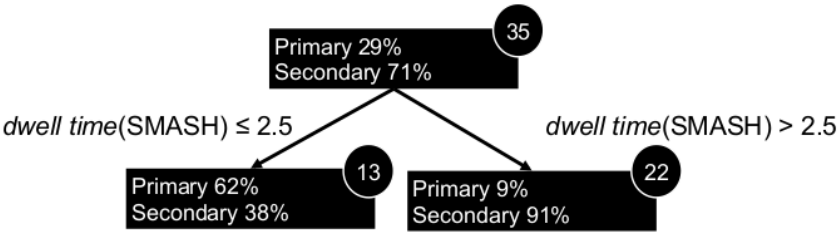


Figure 6.8: Graphical representation of a J48 pruned tree for global ($V1$) resulting in 0.6 for precision, 0.6 for recall and 0.6 for F-measure. The circles on the top-right part of each square indicate the number of individuals who fell in this condition while the numbers in the square convey the distribution of these individuals by group.

The structure of a J48 pruned tree is illustrated in Figure 6.9 for the screens representation. The average value for the exploration metric on the indicators information screen (e_{S_7}) was 1.6 mouse hovers (stdev = 2.31). However, primary users' exploration was higher than the average (3.35 mouse hovers) and secondary users exploration (0.9 mouse hovers). If exploration was restricted to users with values of 0 – meaning there was no activity on the indicators information screen – secondary users constituted 95% (19/20) from an initial 71% of the users in this node of the classification tree. The remaining 15 users (53% primary and 47% secondary) who exhibited some activity in the indicators information screen ($e_{S_7} > 0$) fell into three conditions.

The users whose exploration activity was more than 3 mouse hovers on the screen listing patients affected by potentially hazardous prescribing indicators (e_{S_4}) were exclusively secondary (four individuals, accounting for 100% of users in this node). Note that the average is 2.83 mouse hovers, 2.50 for primary users vs. 2.96 for secondary users. If the activity was equal to or less than 3 mouse hovers and, if time spent on the trends screen (d_{S_6}) was less than or equal to 1.84 seconds, all the users in this node were primary users (accounting for eight individuals), while when dwell time was greater than 1.84 seconds, one was a primary user and the remaining two were not – note that the average dwell time is 2.15 seconds, 1.59 seconds for primary users (stdev = 0.92) and 2.38 seconds (stdev = 4.91) for secondary users.

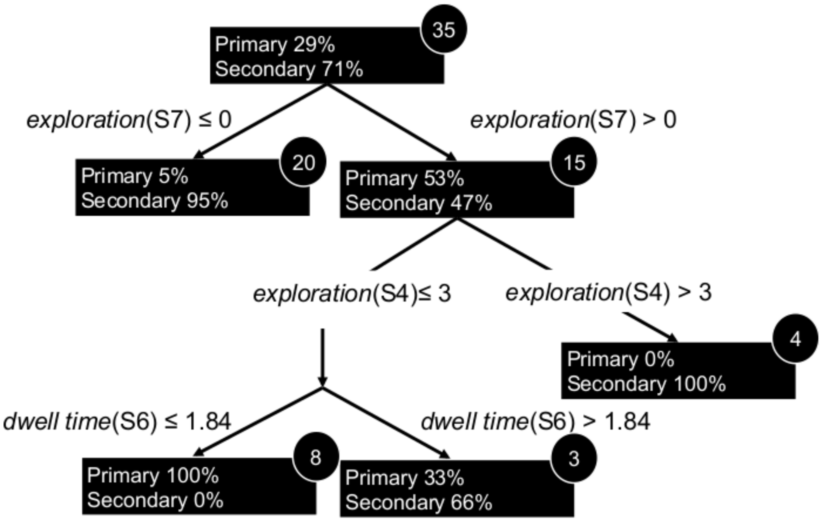


Figure 6.9: Graphical representation of a J48 pruned tree for screens ($V2$) resulting in 0.84 for precision, 0.83 for recall and 0.83 for F-measure.

The session perspective

In the sessions approach, since different session observations for the same user are not independent, in addition to the 10-fold CV we also performed an approximately stratified 3-fold CV, splitting the dataset in three folds of similar size and similar primary/secondary proportion but keeping all the sessions of each user in the same fold. This way, we were able to estimate the performance metrics for unseen users. In particular, we used the same seven algorithms described in the user perspective: AdaBoost, IBK, J48, NB, SMO, Bagging and

MLP. Table 6.9 shows the scores obtained when the data was analysed from a session perspective.

Algorithm	CV	Analysis	Precision	Recall	F-measure
AdaBoost	3-fold	global	0.65	0.73	0.68
		screens	0.73	0.71	0.67
	10-fold	global	0.68	0.73	0.69
		screens	0.73	0.76	0.72
Bagging	3-fold	global	0.71	0.72	0.69
		screens	0.68	0.70	0.68
	10-fold	global	0.70	0.74	0.7
		screens	0.77	0.78	0.77
IBK	3-fold	global	0.63	0.64	0.63
		screens	0.68	0.69	0.68
	10-fold	global	0.65	0.65	0.65
		screens	0.72	0.73	0.73
J48	3-fold	global	0.59	0.74	0.63
		screens	0.69	0.69	0.68
	10-fold	global	0.67	0.72	0.68
		screens	0.73	0.76	0.73
MLP	3-fold	global	0.70	0.73	0.69
		screens	0.70	0.74	0.7
	10-fold	global	0.67	0.72	0.68
		screens	0.71	0.75	0.71
NB	3-fold	global	0.64	0.71	0.64
		screens	0.64	0.48	0.47
	10-fold	global	0.61	0.71	0.64
		screens	0.69	0.59	0.62
SMO	3-fold	global	0.55	0.74	0.63
		screens	0.58	0.73	0.63
	10-fold	global	0.55	0.74	0.63
		screens	0.55	0.74	0.63

Table 6.9: Precision, recall and F-measure on sessions per algorithm.

We analysed the structure of the J48 classifier generated with the complete sample for the screens representation. As mentioned above, the performance of the tree would be the one estimated with the 10-fold CV for regular users and 3-fold CV for new users. The rightmost branch of the J48 classification structure in Figure 6.10 indicates that 81% of the sessions belonged to secondary users when there were more than 3.5 mouse hovers on the visualisations screen (1.74 mouse hovers by secondary users vs. 0.81 by primary users on S_3), 0 mouse hovers on the screen containing information about prescribing safety indicators (0.22 mouse hovers by secondary users vs. 0.53 by primary users on S_7) and users spent less than 4.47 seconds between clicks on the screen showing patients at risk (S_5), where 3.40 seconds were spent on average.

92% of the sessions belonged to primary users when the number of mouse hovers between clicks on visualisations (S_3) was less than or equal to 3.5 mouse hovers and the time spent on the landing page (S_1) was less than or equal to 2.8 seconds (average time was 6.8 seconds). When users spent more than 2.8 seconds on the landing page, the screen containing information about patients at risk (S_5) was decisive to classify the users: when they spent less than or equal to the cut-off value(0.04 seconds, average was 3.4 seconds), 53% of the sessions were exhibited by secondary users, whereas 85% of the sessions belonged to primary users otherwise.

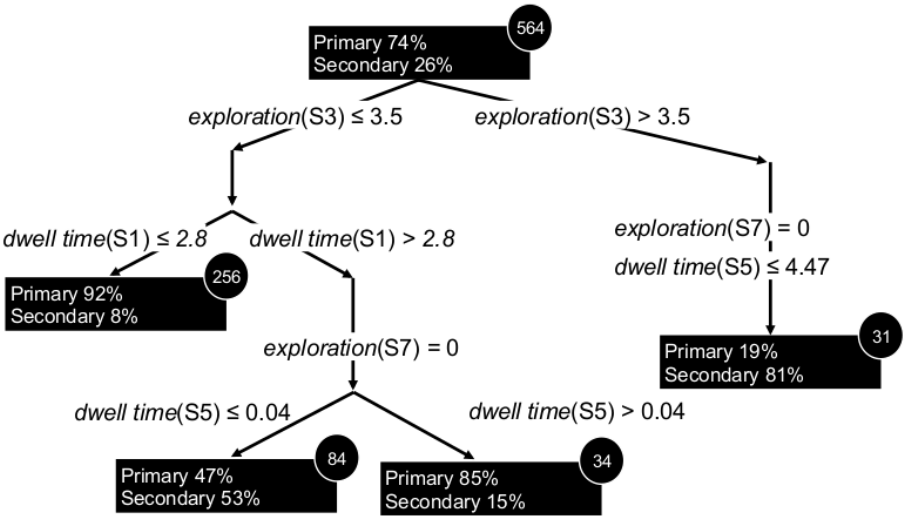


Figure 6.10: Graphical representation of the most relevant nodes of the J48 pruned tree for screens ($V2$) resulting in 0.73 for precision, 0.76 for recall and 0.73 for F-measure.

6.5.3 Discussion

We found that our initial expectations that primary SMASH dashboard users would exhibit lower values for dwell time and exploration were met in both representations (i.e. global and screens) and perspectives – see users in Table 6.6 and sessions in Table 6.7. When considering the classification algorithms, these expectations were also met as exhibited by the values of cut-off points in the classification trees in Figures 6.8–6.10. In general, descriptive statistics and classification algorithms confirm that lower dwell time and exploration was characteristic of those who engaged more with the intervention (i.e. primary users).

Lower values of dwell time were key to distinguish primary from secondary dashboard users in the global representation (Figure 6.8), in the screen that displayed the prescribing safety indicator trends of a given practice (S_6 in Figure 6.9) and on the landing page containing the overview of the practice (S_1 in Figure 6.10). The behaviour of secondary users was characterised by higher dwell time on the screen that displays patients at risk (S_4 in Figure 6.9) and higher values of exploration on the visualisations screen (S_3 in Figure 6.10). It should be noted that the usage of the dashboard is not exclusive to the user groups but what characterises the user groups is how this usage is exhibited. Specifically, these findings suggest that secondary users exhibit characteristic behaviours on screens showing a detailed breakdown of the safety of patients (patients at risk and visualisations), while primary users are characterised by their use of the SMASH dashboard to monitor population health on screens showing the overview of the practice and trends.

Yet, there were exceptions: higher dwell time was characteristic of primary users on the screen showing patients affected by more than one indicator (see S_5 in 6.7 and Figure 6.10). We know from an ongoing study (Jeffries et al. 2019) that primary dashboard users spend most of their time on S_4/S_5 (i.e. patients at risk) because they would check these patients' electronic health records and perhaps make some phone calls. Additionally, lower values of exploration corresponded to secondary users on the screen describing particular indicators (S_7 in Table 6.6 and Table 6.7, and Figure 6.9 and Figure 6.10). It is worth noting that in these instances the cut-off values were close to 0 suggesting that secondary users did not exhibit low dwell time and exploration values because they were more effective, but because they did not access those screens. This implies that, when the functionalities are accessed, dwell time and exploration serve as proxies that characterise the interactive behaviour in the SMASH dashboard in that they discriminate primary users with a high accuracy. For each algorithm the screens representation performs better than the global one, which means that including the screens in the modelling has added value.

These findings suggest that the two user groups (primary and secondary users) have different characteristic behaviours when interacting with the SMASH dashboard. These interactive behaviours, which are modelled using features that are proxies of competence, make the two user groups distinguishable. Lower values of dwell time are indicators of users being more effective in processing information and solving problems (Chin et al. 2009), which suggests that primary users were more competent carrying out overseeing tasks in the SMASH dashboard. Since longer visual activity conveyed by exploration is known to be an indicator of less efficient search (Ehmke and Wilson 2007), we attribute higher exploration values observed in secondary users to lower levels of engagement and, consequently, lower performance. This is in line with the literature that indicates that those who are more engaged perform better (Rich et al. 2010; Christian et al. 2011).

Two design recommendations emerge from these outcomes. Since the perceived lack of competence is a barrier to use this kind of interventions (Jeffries et al. 2017), we could monitor the competence of SMASH dashboard users (and similar interventions) and intervene if needed. Individuals belonging to groups of users who are less engaged (and are less competent) could be given support using tailored educational nudges to encourage their learning. They could be provided with personalised messages about their performance with respect to their peers, which could help to challenge their perceptions if they underestimated themselves and increase self-efficacy. These nudges could potentially be delivered by retrieving the current URL and keeping track of mouse events to compute dwell time and exploration, which can be carried out in real time in the browser. Since this method does not require to remote storage of interaction or personal data, user confidentiality and privacy are respected, removing potential barriers for acceptance of such systems by prospective users (Angulo and Ortlieb 2015). The second recommendation is about adapting the workflows in the SMASH dashboard according to the characteristic use of the two groups. Informed by the stereotypical uses of the dashboard, SMASH should facilitate workflows for (a) monitoring population health and (b) for a more detailed analysis of individuals at risk, by grouping the screens accordingly. Transitions between these two workflows should also be possible by mapping an analogy of the information visualisation mantra (i.e. overview, filtering, details-on-demand) (Shneiderman 1996) into dashboards for managing population health through progressive disclosure principles: monitoring population health, filtering, breakdown-of-data on demand.

Methodological Considerations

Primary users comprised 29% of the users in the study while their sessions accounted for 74% of the total sessions. This means that since we had more recurring sessions from primary users we might have confounded learning effects in the user perspective. While our conclusion still holds (primary users are more efficient carrying out overseeing activities in the SMASH dashboard) the reason they are more confident could be a result not only of their engagement with the intervention, but also to their expertise due to the fact they were pharmacists. Nevertheless, we found that while having this expertise is beneficial, it is not essential to engage with the intervention (Jeffries et al. 2019).

6.6 Summary

Making medical software easy to use and actionable is challenging due to the characteristics of the data (its size and complexity) and its context of use. This results in user interfaces with a high-density of data that do not support optimal decision-making by clinicians. Anecdotal evidence indicates that clinicians demand the right amount of information to carry out their tasks. This suggests that adaptive user interfaces could be employed in order to cater for the information needs of the users and tackle information overload. Yet, since these information needs may vary, it is necessary first to identify and prioritise them, before implementing adaptations to the user interface. As gaze has long been known to be an indicator of interest, eye tracking allows us to unobtrusively observe where the users are looking, but it is not practical to use in a deployed system.

In the first contribution, we address the question of whether we can infer visual behaviour on a medication safety dashboard (SMASH) through user interaction data. Towards that goal, we first analysed the gaze (fixation duration on the Areas of Interest defined) and interaction data (global and screen divided exploration and dwell time) collected in a lab study with six participants which completed nine tasks. Using Pearson correlation on gaze data, lab participants with similar gaze behaviours were paired, obtaining a total of three groups. The same pairings were obtained from the k-means procedure used on the interaction data of these participants, which implies a connection between the gaze and interaction behaviours. What is more, the six lab study participants were similarly paired when applying the k-means algorithm on their global exploration together with that of the observational study participants (35). Therefore, the gaze behaviour of the observational study participants may have been similar to the one showed by lab study participants grouped in the same clusters.

In the second contribution we aimed to characterise the use of SMASH by exploring and contrasting interactions from primary users who were leading the intervention and secondary users who used the dashboard to engage in safe prescribing practices. To that end, we analysed the global and screen divided interaction data of the observational study and applied supervised learning algorithms to classify primary against secondary users. Regarding the results, we observed values for accuracy above 0.8, indicating that 80% of the time we were able to distinguish a primary user from a secondary user. In particular, the Multilayer Perceptron (MLP) yielded the highest values of precision (0.88), recall (0.86) and F-measure (0.86). The behaviour of primary users was distinctive in that they spent less time between mouse clicks (lower dwell time) on the screens showing the overview of the practice and trends. Secondary users exhibited a higher dwell time and more visual search activity (higher exploration) on the screens displaying patients at risk and visualisations. In other words, primary users were more competent on population health monitoring activities, while secondary users struggled on activities involving a detailed breakdown of the safety of patients. Informed by these findings, we propose workflows that group these activities and adaptive nudges to increase user engagement.

Chapter 7

Modelling the interaction and use of e-Services

7.1 Introduction

According to the eGovernment Benchmark 2018 report (Tinholt et al. 2018), 66% of the services delivered by public administrations in Europe were fully available online, which represents an increase of 17% since 2012 (Tinholt et al. 2015). The present high online availability is not surprising, since citizens can benefit from public e-Services which are delivered at any time (during 24 hours, seven days a week) and provided in a personalised way (different languages, adaptations for disabled users, etc.) (González et al. 2007). To this regard, designing accessible and personalised e-Services is crucial so that they can be fully inclusive for the wide variability of citizens who use them. However, the extraction of the user profiles needed for personalising such services is a difficult task, since frequently they do not require registration or if so, no sensible information about the user (e.g., about disabilities or limitations) is collected (Abascal et al. 2019). In this scenario, web usage mining techniques can be used for modelling users from e-Services by gathering their interaction data unobtrusively from Web server logs (Abascal et al. 2013).

Unfortunately, due to privacy concerns public institutions often do not facilitate access to the navigation data of the services they deliver. Proof of this were the collaborations we carried out with the Gipuzkoa Provincial Council (GPC) and the University of the Basque Country (UPV/EHU) who allowed us to analyse the navigation on their websites, `www.gipuzkoa.eus` and `www.ehu.eus` respectively, but not any particular service. Indeed, in the analysis performed in the website of the Gipuzkoa Provincial Council (Yera et al. 2016a) we concluded that modelling the users' web interaction was an extremely difficult task because added to the lack of user registration, the goal of the users was previously unknown and the updating process of the website hindered to reproduce their navigations. Thus, we pointed out some essential technical requirements that e-Services should fulfil to enable the modelling of the interaction of its users: a minimum number of users (100-1000), a final goal, several steps to achieve the final goal where users must select different options and availability of the URLs requested by the users during their interaction to reconstruct the navigation. In turn, we concluded that for each transaction the following data should be gathered in the logs of the e-Services if machine learning techniques want to be used for the analysis: the user identification (if registered) or IP, the timestamp, the step of the process or the URL, the options selected by the users in that step / URL, the achievement (success/ failure) in case it is the goal.

In absence of interaction data of a particular e-Service, in this chapter, we present two contributions made in this area using machine learning procedures: the modelling of the interaction on the enrolment web information area of the University of the Basque Country (UPV/EHU) as an e-Service and an empirical analysis of the use of e-Services in Europe based on surveys provided by Eurostat. In the first contribution different systems were built to automatically classify users reading enrolment information of the UPV/EHU and those carrying out searching type tasks, which in addition enabled their characterisation. In contrast, in the second contribution based on survey data supplied by Eurostat we defined two indexes to quantify the use of e-Government services (EGUI/EGUI$^+$), and using supervised learning procedures we characterised the null and total levels of use of these e-Services.

7.2 Modelling the interaction with e-Services

This contribution presents a research result of the collaboration with the University of the Basque Country (UPV/EHU). Since February to the middle of March 2016 the university provided us with access to the navigation data of its whole website. In order to provide clues for future service improvements, our main goal has been to model the university enrolment web information area (`www.ehu.eus/web/sarrera-acceso`) as an e-Service and to extract as much knowledge as possible from it through data mining processes.

Initially, we analysed the structure and content of the whole website of the university, in order to identify the parts related to the enrolment area. Then, we studied the usage of this particular area extracting the navigation sessions of the users from the log files stored in the servers, and labelling them as success or fail based on the end of the navigation. Finally, we used supervised and unsupervised learning algorithms to answer three meaningful questions: whether the sequence of URLs visited by the user and the way the user navigates (the two types of information used to represent user sessions) affected the success or failure of her/his navigation, if both sources, the navigation sequence and the navigation style, were closely related and if it is possible to foresee if new sessions will be successful or not, just analysing the beginning of the navigation.

7.2.1 Context: The website of the University of the Basque Country

The UPV/EHU is the public University of the Basque Country with campuses over the three provinces of this region: Biscay, Gipuzkoa and Álava. This institution was established in 1980 and it has around 45,000 students and a staff of around 3,500 workers.

In this research we analysed the usage of the web page of the UPV/EHU (`www.ehu.eus`) and more specifically we were interested in the enrolment e-Service. This university has an online enrolment procedure that can be completed using an IT application called GAUR. However, this process requires to be logged and the institutions have difficulties to provide such data due to privacy issues. Thus, we focused in the navigations of the enrolment area whose main domain is `wwww.ehu.eus/web/sarrera-acceso`.

The enrolment area can be accessed using the top menu (University access option) displayed in all URLs of the site. This area provides information about the university (staff, contact and location), the access to the university (types of access, academic calendar, admission and enrolment procedure, degree offer...) and scholarships. The main web page of the enrolment area is shown in Figure 7.1 below.

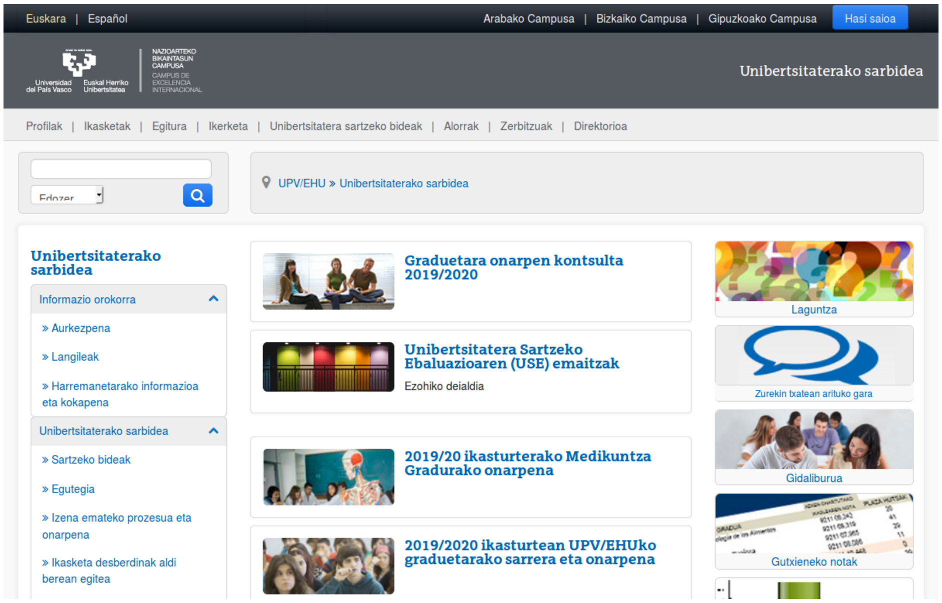

Euskara | Español
Arabako Campusa | Bizkaiko Campusa | Gipuzkoako Campusa
Hasi saioa
Universidad del País Vasco
Euskal Herriko Unibertsitatea
NAZIOARTEKO BIKAINTASUN CAMPUSA
CAMPUS DE EXCELENCIA INTERNACIONAL
Unibertsitaterako sarbidea
Profilak | Ikasketak | Egitura | Ikerketa | Unibertsitatera sartzeko bideak | Alorrak | Zerbitzuak | Direktorioa
Edozer
UPV/EHU » Unibertsitaterako sarbidea
Unibertsitaterako sarbidea
Informazio orokorra
» Aurkezpena
» Langileak
» Harremanetarako informazioa eta kokapena
Unibertsitaterako sarbidea
» Sartzeko bideak
» Egutegia
» Izena emateko prozesua eta onarpena
» Ikasketa desberdinak aldi berean egitea
Graduetara onarpen kontsulta 2019/2020
Unibertsitatera Sartzeko Ebaluazioaren (USE) emaitzak
Ezohiko deialdia
2019/20 ikasturterako Medikuntza Gradurako onarpena
2019/2020 ikasturtean UPV/EHUko graduetarako sarrera eta onarpena
Laguntza
Zurekin txatean arituko gara
Gidaliburua
Gutxieneko notak

Figure 7.1: The main web page of the enrolment area of the UPV/EHU (`www.ehu.eus/web/sarrera-acceso`).

7.2.2 Preprocessing, session classification and description of the datasets

In this section we first describe the preprocessing and the session classification carried out with the navigation data of the University of the Basque Country and then, we explain how we generated the navigation sequence and navigation style datasets (DBs) for the analysis. For this process a java project was implemented in the Eclipse platform V3.8.1 and finally executed as a jar file in the terminal of Ubuntu 14.04 LTS.

The UPV/EHU provided us access to their logs from the end of February to the middle of March 2016 and we focused the analysis on the navigation in the area that supplies enrolment information. These months partly cover the pre-enrolment dates for the secondary school students, from the middle of January to the middle of March (2016).

The preprocessing consisted on the one hand on removing the unsuitable URL request from the user sessions which did not satisfy the following criteria: Request method $\in$ {GET, POST}, URL extensions $\in$ {aspx, .htm, .html, .pdf, .doc, .xml}, URL content $\neq$ {admin, error, rss, piwik, wposta} and server's answer = 2XX (no errors). On the other hand, only user sessions with a meaningful navigation in the area of interest were considered for the analysis, those meeting the next requirements: session length $\geq$ 3 clicks, number of URLs within the enrolment area $\geq$ 1 and inactivity period (session gap) $\leq$ 10 minutes. On average after the preprocessing the number of sessions was diminished by 94%.

In our contribution we assumed that all users aim to obtain information about the different enrolment options and we accordingly classified user sessions, concretely, based on the kind of web page they last visited. Specifically, the URLs requested in the sessions were characterised considering two criteria: the content of the URL (whether the text or the links were dominant) and the area of the URL (whether it corresponded to the enrolment area or not).

Regarding the content, those URLs with text format (.pdf/.doc/.docx) were classified as of content type. The remaining URLs were classified as content or scatter type (when links were dominant) using the LCIndex (Link Content index) (Arbelaitz et al. 2016) described in Equation 7.1, where: $Nlinks$ is the number of links in the webpage and $Nwords$ is the number of words appearing in the webpage and $NwordsLinks$ is the number of words used in the links of the page.

$$LCIndex = \frac{Nwords - NwordsLinks}{Nlinks}; URLtype = \begin{cases} scatter, & LCIndex \leq 10 \\ content, & LCIndex > 10 \end{cases} \tag{7.1}$$

Finally, we considered to be successful (classified as success) the user sessions finished in a URL with enrolment information (content type and within the enrolment area) because this will probably help in the enrolment process. On the contrary, the user sessions ended in web pages with little information (scatter type and any area) were considered to be of failure type, as they probably were carrying searching type tasks. Since we are focusing in users that want to enrol, the sessions ended in a content type URL outside the enrolment area were considered out of our scope (possible success in another area). Table 7.1 describes the session classification defined.

	Last URL of the session	
Type of session	Area	Type
Success	Enrolment	Content
Failure	Enrolment Not Enrolment	Scatter

Table 7.1: User session classification of the enrolment web information area of the University of the Basque Country (UPV/EHU).

The use of the last URL for the session classification was supported by the statistical analysis carried out for the complete navigation sequences, revealing that the nature of the last URL (type/area) determined the nature of the majority of URLs visited in the session. In this way in the failure type sessions, last URL of scatter type and from any area, the proportion of URLs of failure type on average was a majority, 53% for those ended in URL from the enrolment area and 52% for those ending in a URL outside the enrolment area. Similarly, in the success type sessions, last URL of content type and from the enrolment area, the proportion of URLs of this type was higher than that of failure type, 53%.

After the log was preprocessed and the user sessions were classified, data was prepared for analysis. As stated in the introduction, user sessions were analysed from two points of view: navigation sequence and navigation style. Accordingly, we created two datasets with the selected sessions: the first one containing the sequence of URLs visited in each user session, and the second one, representing the user sessions with a vector of interaction features calculated from the information contained in the log files combined with the content and the structure of the site. The features that represent the user sessions in the second dataset were computed according to the time, the URL classification (content/area) or the number of clicks. As shown in Table 7.2 a total of 18 interaction features were computed, which are classified in three categories: seven click related features (session length, proportion of scatter/content type or enrolment area/outside the enrolment area URLs requested...), seven time related features (session duration, click duration on average in the two types of URLs and two different areas...) and four transition related features (number of transitions based on the type and area of the URL).

To better understand the datasets, in Tables 7.3 and 7.4 we show a particular user session from the dataset with the sequence of URLs visited and the values (not normalised) of the interaction features computed for the same session according to the information of the log files respectively. The session example shown in these tables is of success type, as the last URL visited is of content type and inside the enrolment web information area of the UPV/EHU.

Int. feature	Description
No. click	Number of clicks (length of the session).
No. scat. %	Number of scatter type URLs / length of the session
No. cont. %	Number of content type URLs / length of the session
No. enr. %	Number URLs of the enrolment area / length of the session
No. not-enr. %	Number URLs from outside the enrolment area / length of the session
No. ind. %	Number of times the start page of the enrolment area (index) is visited / length of the session
No. ref-sear. %	Number of URLs that a have web search engine as reference
T-ses	Duration of the session (s)
T-click_avg	Average duration of a click (s)
T-scat._avg	Average duration of a click on a scatter type URL (s)
T-cont._avg	Average duration of a click on a content type URL (s)
T-enr._avg	Average duration of a click on a URL of the enrolment area (s)
T-not-enr._avg	Average duration of a click on a URL outside the enrolment area (s)
T-ind.-avg	Average duration of a click on the start page of the enrolment area (s)
No. cont.-scat.	Number of transitions content-scatter type URLs
No. scat.-cont.	Number of transitions scatter-content type URLs
No. enr.-not-enr.	Number of transitions inside-outside enrolment area URLs
No. not-enr.-enr.	Number of transition outside-inside enrolment area URLs

Table 7.2: Interaction features used to represent the user sessions.

N	URL	Type	Area
1.	`http://www.ehu.eus/es/web/medikuntza-odontologia/medikuntza-14-15`	Scat.	Not-Enr.
2.	`http://www.ehu.eus/es/web/medikuntza-odontologia/gasteiz`	Scat.	Not-Enr.
3.	`http://www.ehu.eus/documents/1546271/2600354-/horario6_vitoria_castellano_2014-2015.pdf`	Cont.	Not-Enr.
4.	`http://www.ehu.eus/es/web/medikuntza-odontologia/medikuntza-plana`	Cont.	Not-Enr.
5.	`http://www.ehu.eus/es/web/medikuntza-odontologia/tramiteak`	Scat.	Not-Enr.
6.	`http://www.ehu.eus/es/web/medikuntza-odontologia`	Cont.	Not-Enr.
7.	`http://www.ehu.eus/eu/web/sarrera-acceso/gutxieneko-notak`	Cont.	Enr.

Table 7.3: An example of a navigation session expressed as sequence of URLs.

Interaction features		
No. click = 7	No. scat. = 43%	No. cont. = 57%
No. enr. = 14%	No. not-enr. = 86%	No. ind. = 0
No. ref-sear. = 29%	T-ses. = 98 s	T-click_avg = 14 s
T-scat._avg = 14 s	T-cont._avg = 14 s	T-enr._avg = 21 s
T-not-enr._avg = 13 s	T-ind._avg = 0	No. cont.-scat. = 2
No. scat.-cont. = 2	No.enr.-not-enr. = 0	No. not-enr.-enr. = 1

Table 7.4: Interaction features computed for the navigation session shown in Table 7.3.

7.2.3 Automatic classifier system based on supervised learning techniques

For this approach we analysed the navigation data of 49 days, from 23/02/2016 to 12/04/2016. After the preprocessing described in Section 7.2.2, a total of 25,467 sessions were obtained (around 6% of the 416,354 sessions available), 10,734 of them of success type (42.1%) and 14,733 (57.9%) of them of failure type.

Considering the future goal of building a system able to classify new user sessions of the enrolment e-Service, in this first approach based on supervised learning techniques we explored two options to automatically classify the sessions using the dataset with the interaction features: one building 10 C4.5 (J.R Quinlan 1993) and another one building 10 CTC (Consolidated Tree Construction) (J.M. Pérez et al. 2007) trees. In fact, these two supervised learning approaches will provide us not only the specific discrimination capacity of the system to classify new sessions as success or failure, but also a concrete description of the interaction features to be used in the process.

For this task, the sessions of the dataset were chronologically ordered and the first 25,000 were selected for experiments. Then, as shown in Figure 7.2 below the new dataset was divided in 10 parts of 2,500 sessions($Fold_i|i \in \mathbb{N}, i \leq 10$) and each part was again divided into 10 segments of 250 sessions ($F_{ij}|i, j \in \mathbb{N}, i, j \leq 10$). Every split respects the chronological order as it would happen in exploitation in a real system. As in an ordinary 10 fold-cv procedure, the first nine parts of this dataset ($Fold_i|i \in \mathbb{N}, i \leq 9$) were used for training whereas the last one with the newest sessions ($Fold_{10}$) was kept for test. To build each of the 10 trees a particular segment from the 10 parts of the dataset available was used (first= F_{i1}, second= F_{i2}, third= F_{i3} etc.) but always using these concrete segments from the nine first parts ($F_{ij}|i, j \in \mathbb{N}, i \leq 9, j \leq 10$) as training (2,250 sessions) and the newest segment ($F_{10j}, j \in \mathbb{N}, i \leq 10$) for test (250 sessions). This way we ensured that the data were equally distributed on time among the trees so that they had similar learning processes and that the newest sessions ($Fold_{10}$) were used for test.

The algorithms were implemented in Visual C++ although they are also available as official Weka (M. Hall et al. 2009) packages (C4.5 and J48Consolidated[1]).

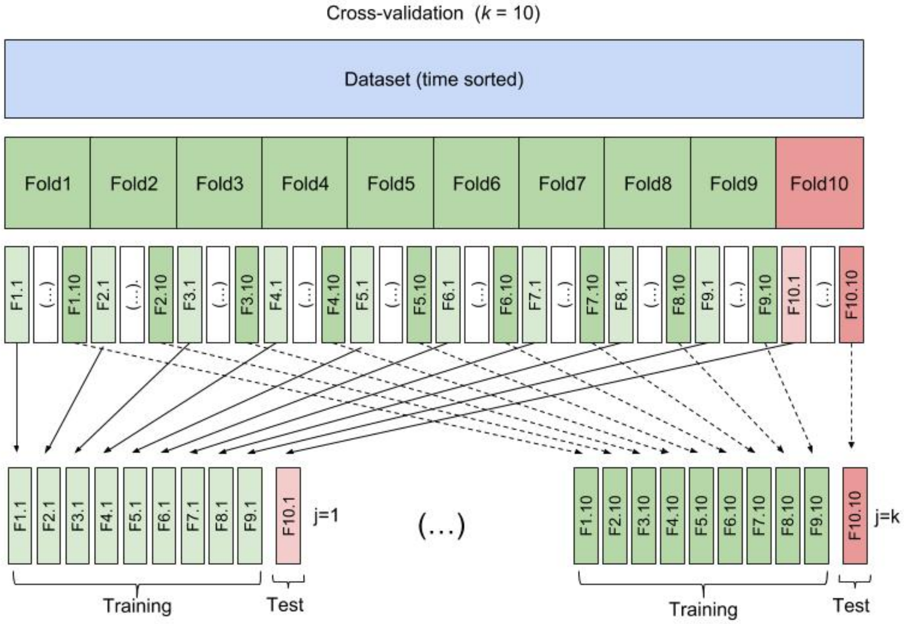


Figure 7.2: Procedure to build 10 C4.5 and 10 CTC trees using the dataset with the interaction features

Ideally, the best classifier will be the one with better classification performance, that is, the one with a low classification error and a simple and stable explanation. Regarding the average error the C4.5 decision trees achieved a lower value (0.0500) than the CTC ones (0.0626). Conversely, the CTC trees achieved a higher average value (0.9828) for the Area Under the ROC (AUC) than the one achieved by the C4.5 (0.9665) ones. These low classification errors and high AUC values show that both types of trees are able to discriminate between the success and failure navigations defined, however, in order to better compare them a paired t-test was carried out using both metrics. The t-test revealed on the one hand that the AUC of the CTC approach was significantly better than the AUC of the C4.5 approach (significance level 0.05) and on the other hand, that there were not statistically significant differences between the classification errors of the two options (significance level 0.05). Concerning the explaining capacity of the trees, we noticed that the structures of the CTC trees were simpler, with values on average for the number leaves and for the number of internal nodes, 14.4 and 13.4 respectively, lower than the ones achieved by the C4.5 trees, 44.3 and 43.3. Accordingly, the explanation provided by the CTC approach was proof to be more stable, achieving an average value of common

[1] `http://www.sc.ehu.es/aldapa/weka-ctc/`

nodes among the different trees (21.81%) higher than the one obtained in the C4.5 approach (6.02%).

Therefore, the CTC approach was found to be the best option as it achieved a significantly better AUC value and it provided more stable and simple explanations. According to the structure of the CTC trees these are the main interaction features to differentiate each type of session (success / failure): the average duration of a click on a URL of content or scatter type (T-cont._avg and T-scat._avg), the proportion of content type URLs in the session (No. cont. %) and the number of content type URL– scatter type URL transitions (No. cont.-scat.). More specifically, the CTC highlighted the following rules to discriminate each type of navigation:

Main rules to detect failure type sessions:

- (T-cont._avg $\leq$ 13.95 s) AND (T-scat._avg $>$ 14.06 s)
- (T-cont._avg $\leq$ 13.95 s) AND (T-scat._avg $\leq$ 14.06 s) AND (No. cont. % $\leq$15%)

Main rules to detect success type sessions:

- T-cont._avg $>$ 13.95 s
- (T-cont._avg $\leq$ 13.95 s) AND (T-scat._avg $\leq$ 14.06 s) AND (No. cont. % $>$ 15%) AND (No. cont.-scat. $\leq$ 0.01)

Hence, failure type sessions are closely linked with short times in content type URLs (T-cont_avg $\leq$ 13.95 s) and long times in scatter type URLs (T-scat._avg $>$ 14.06 s). This could represent the navigation of those users who are not able to find certain information. In addition, the sessions with small values for the two interaction features mentioned T-cont._avg / T-scat._avg, and low proportion of content type URLs (No. cont. % $\leq$ 15%) have more probabilities to be of failure type. Conversely, success type sessions have a close relation with longer times on average in content type URLs (T-cont._avg $>$ 13.95 s). To a lesser extent, short times in content and link type URLs, high proportions of content type URLs and an absence of content-scatter type URL transitions (No. cont.-scat. $\leq$ 0.01) lead more easily to success type sessions.

Finally, it should be remarked that the rules mentioned above, are very similar to the ones provided by the C4.5 trees. This fact reinforces the validity of the interaction features noted to discriminate the types of sessions defined. Thus, we think that the conclusions achieved will be very effective to classify the new user sessions and improve the enrolment e-Service of the UPV/EHU in the future.

Additionally, we decided to compare the effectiveness of using supervised and unsupervised learning techniques to build classifier systems in terms of accuracy to classify success and failure type sessions. Thereby, in the next section we describe several classifier systems built based on unsupervised learning procedures using each of the two session representations available (navigation sequence / navigation style) and a combination of both.

7.2.4 Automatic classifier system based on unsupervised learning techniques

The goal of this contribution was to model the enrolment web information area so that it could be improved in the future. We mainly focused on characterising successful and failure sessions and detecting failure sessions thereby actions can be taken. With that aim in the first approach just described we built an automatic classifier system based on supervised learning procedures using the interaction features computed from user navigation sessions. In contrast, in this second approach we explore the potential of unsupervised learning techniques to automatically classify new user sessions but using the two representations proposed, the set of interaction features or the sequences of URLs visited.

To that end, we first analyse whether the two aspects studied, navigation sequence and navigation style were meaningful to decide if a user session will be of success or failure type, whether they can be used to foresee the type of new sessions, and if they are complementary or not. Unsupervised learning techniques used allow on the one hand, characterising success and failure patterns. On the other hand, the centroids we computed in the resulting clusters of the two session representations, provide stable patterns that enable to tune the success and failure session classification system in order to control its level of precision. The next sections describe the analysis carried out.

Analysis of the discriminating capacity of the navigation sequence and navigation style

In this approach we analysed the navigation data of 53 days, from 23/02/2016 to 16/04/2016, obtaining after the preprocessing a total of 26,467 sessions. The 25,467 sessions of the first 49 days (same period analysed in the previous approach) were used to model the enrolment web information area of the UPV/EHU and build automatic classifier systems based on unsupervised learning techniques, whereas the last four days (1,000 sessions) were kept to validate these classifiers. The class distribution in the validation dataset was similar to the modelling dataset with 37.9% ($\sim$ 42.1%) of sessions of success type and 62.1% ($\sim$ 57.9%) of them of failure type.

The dataset built with sequences of URLs was used to analyse how the navigation sequence is interrelated with the type of sessions defined. We used PAM (k-medoids) (Kaufman and P. Rousseeuw 1990) clustering algorithm that allows to group sequences into high quality clusters (Barioni et al. 2008) with edit distance. Broadly, in a clustering procedure the number of clusters selected is ideally high, as it contributes to create clusters with as many cases as possible of the same type. We selected the k value according to the Silhouette Cluster Validity Index (Arbelaitz et al. 2013b) which indicated that k=50 was the most appropriate cluster configuration when compared to k=25, k=75 and k=100. The scores for each k in ascending order were 0.046, 0.069, 0.055 and 0.047 respectively.

To evaluate the discernment power of the approach for the two types of navigation sessions, we focused on the clusters where the superiority of success or failure cases was over 74%. In the selected clusters, the total number of sessions grouped was significant (42% of the whole dataset), and a suitable representation of each type of session can be found (12% of success and 29% of failure). Table 7.5 summarises the types of examples grouped in the eight success type clusters and the 17 clusters of failure type.

Table 7.5 shows that half of the clusters (25) have a proportion of one class or the other one above 74%. The whole dataset contains 14,733 user sessions of failure type and the clusters labelled as failure, 6,669. Consequently, although the percentage of failure type sessions in the dataset is 58%, within the failure type clusters that probability raises to 89%. Similarly, being the success type user sessions 42% in the whole dataset, in the selected success type clusters this percentage increases up to 82%.

Feature	Clusters with a no. success-sessions $\geq$ 74%	Clusters with a no. failure-sessions $\geq$ 74%
No. clusters	8 (16%)	17 (34%)
No. success-sessions	2,551 (81.9%)	842 (11.2%)
No. failure-sessions	564 (18.1%)	6,669 (88.8%)
No. sessions-clusters	3,115	7,511
No. sessions-DB (%)	12.20%	29.50%

Table 7.5: Results of PAM (k=50) in the dataset built with sequences of URLs.

These results suggest that the navigation sequence (URLs visited) and the success/failure of a user session are connected. Hence, we can gather that there is a chance to automatically classify the navigation of new users of the UPV/EHU enrolment e-Service, based on the navigation sequence.

The dataset with the interaction features of the sessions (Table 7.2) was used to determine whether unsupervised learning procedures support the findings of the supervised learning approach, that is, that the navigation style is tightly interrelated with the success/failure of a session. We ran the k-means algorithm (Lloyd 1982) with Euclidean distance using k=50 in the previously normalised (normal distribution) dataset. Then, we selected the clusters with a superiority of success or failure cases over 74%, which grouped 43% of the total number of sessions of the DB (14% of the success and 29% of failure). Results are shown in Table 7.6.

As it happened with the navigation sequence dataset, Table 7.6 shows that in more than half of the clusters (27/50), the proportion of one of the types of sessions defined or the other is higher than 74%. In the case of the failure clusters there are 6,842 sessions of failure type what raises form being 58% of the sessions in the complete dataset, to 93% within those clusters. Likewise, being the success type sessions 42% of the dataset, in the success clusters this number raises to 88%. Thus, the results show that the navigation style is discriminant for the two types of navigations defined, success and failure.

Feature	Clusters with a no. success-sessions ≥ 74%	Clusters with a no. failure-sessions ≥ 74%
No. clusters	6 (12%)	21 (42%)
No. success-sessions	3,212 (87.7%)	544 (7.4%)
No. failure-sessions	451 (12.3%)	6,842 (92.6%)
No. sessions-clusters	3,663	7,386
No. sessions-DB (%)	14.40%	29.00%

Table 7.6: Results of k-means (k=50) in the interaction features dataset.

Comparison of the navigation sequence and navigation style

According to the results shown in the previous paragraphs, we can state that whether a session will be of success or failure type depends on both the navigation sequence and the navigation style. But it would be interesting to know if both perspectives are closely related or not. With this aim we compared the partitions of the two clustering procedures using the Jaccard index (Jaccard 1908) which provided a very low value (0.04) in the comparison, showing that both results are quite different. This suggests that in the navigation of the UPV/EHU enrolment area, the navigation sequences (URLs visited) and the navigation style described by the interaction features are independent, and thus the design of each concrete URL does not affect much to how the user navigates.

Hence, in principle to classify the navigation of new users in the enrolment area of the website, both view points, could be useful and might be complementary. To this regard, in the following paragraphs, Section Description of the automatic classifier systems based on unsupervised learning techniques, we describe the validation process performed to test these hypotheses.

Characterisation of the types of session based on the navigation style

Additionally, we extracted the main characteristics of the six success type clusters and the 21 failure type clusters to model both types of navigation, which are summarised in Table 7.7. According to the table, these are the main characteristics for the failure type sessions compared to the success type sessions: the click-streams on average tend to be larger (No. click = 13.1 vs. 6.2); they are more focused on scatter type URLs (No. scat. % = 80% vs. 30%); the internal and external navigations are more balanced (No. enr. % = 48% vs. 92%); the total duration of these sessions is higher (T-ses = 184.6 s vs. 106.3 s); the duration on average of a click on a scatter type URL is longer (T-scat._avg = 27.8 s vs. 11.7 s); the duration on average of a click on a content type URL is shorter (T-cont._avg = 7.4 s vs. 20.9 s); there is at least one transition from enrolment area to outside this area (No. enr.-not-enr. = 0.8 vs. 0); there is almost no transition from outside the enrolment area to the enrolment area (No. not-enr.-enr = 0.4 vs. 1.1).

Interaction features	Failure type sessions (21 clusters)	Success type sessions (6 clusters)
No. click	13.1	6.2
No. scat. %	**80**	**30**
No. cont. %	**20**	**70**
No. enr. %	48	92
No. not-enr. %	52	8
T-ses (s)	184.6	106.3
T-scat._avg	**27.8**	**11.7**
T-cont._avg	**7.4**	**20.9**
No. scat.-cont	**1.95**	**1.19**
No. cont.-scat	**0.91**	**2.04**
No. enr.-not-enr.	0.8	0
No. not-enr.-enr	0.4	1.1

Table 7.7: Main interaction features of the sessions inside the six success type clusters and the 21 failure type clusters obtained from the navigation style data.

This is partially in line with the main rules provided by the CTC trees used in the supervised learning system described in Section 7.2.3, where the interaction features marked in bold in Table 7.7 were found to be decisive for the session classification: the time on average on content/scatter type URLs, the proportion of content type URLs and the number of transitions from content to scatter type URLs. In particular, according to the decision trees, the duration on average shorter than 13.95 seconds on content type URLs (together with other rules) were related with failure type sessions, whereas the opposite case was found to be related with success type sessions.

Description of the automatic classifier systems based on unsupervised learning techniques

As mentioned before, a total of 1,000 sessions of the initial dataset were kept for validating the system (validation dataset). These sessions were also represented as sequences of URLs visited (navigation sequence perspective) and as vectors of the interaction features described in Table 7.2 (navigation style perspective). The next paragraphs describe the three automatic classifiers built using clustering, one based on the navigation sequence, another one based on the navigation style and a last one based on both representations.

In the system based on the navigation sequence we first computed the medoids of the 25 clusters with a wide percentage of sessions ($\geq 74\%$) of success or failure type (Table 7.5) and labelled them with the majority class of their corresponding cluster. Then, for each new session of the navigation sequence dataset, we calculated the 10 nearest medoids (*10-NM*) using the edit distance and finally labelled each new session with the most voted type according to the simple voting performed using different numbers (km) of nearest medoids: *km-NM*; $km \in \{1, 3, 4, 7, 9, 10\}$. We measured the accuracy (%) in terms of number of examples where the type of sessions is guessed in the validation set (see Table 7.8). The voting approach that involved the 10 nearest medoids was found to be the best one, reaching an accuracy of 61.90% thus the results were not very good.

	Voting approaches performed with medoids (*km-NM*)					
	1-NM	*3-NM*	*5-NM*	*7-NM*	*9-NM*	*10-NM*
Accuracy (%)	55.70	56.10	55.90	61.80	59.90	61.90

Table 7.8: Accuracy (%) of the system built with the *km-NM* of the 25 clusters obtained with PAM (k=50) procedure used in the navigation sequence dataset.

Similarly, in the system based on the navigation style, we computed the centroids (average) of the 27 clusters with a wide majority of sessions ($\geq$ 74%) of success or failure type (Table 7.6) and labelled them with the majority class of their corresponding cluster. Then, each session of the new navigation style dataset was labelled with the most voted type according to the different numbers (kc) of nearest centroids (*NC*) computed with the Euclidean distance: *kc-NC*; $kc \in \{1, 3, 5, 7, 9, 10\}$. The results of this system (see Table 7.9) were better than the previous ones. In this case, the voting that involved the five nearest centroids (*5-NC*) was found to be the best, reaching an accuracy of 78.2%.

	Voting approaches performed with centroids (*kc-NC*)					
	1-NC	*3-NC*	*5-NC*	*7-NC*	*9-NC*	*10-NC*
Accuracy (%)	75.50	76.70	78.20	77.50	76.90	74.80

Table 7.9: Accuracy (%) of the system built with the *kc-NC* of the 27 clusters obtained with k-means (k=50) procedure used in the navigation style dataset.

In order to analyse if both systems were complementary or not, we built a third system that combined the votes of the system based on the navigation sequence dataset, *km-NM*; $km \in \{2, 3, 4, 5, 6\}$, and the votes of the best approach of the system based on the navigation style dataset (*5-NC*). Accordingly, each new session was classified using the most voted type among the km nearest medoids and kc nearest centroids involved. The best voting was the one that involved the nine nearest neighbours (*9-NN*), using the four nearest medoids (*4-NM*) for the navigation sequence dataset and the five nearest centroids (*5-NC*) for the navigation style dataset. Table 7.10 shows the accuracy obtained for different configurations. Although no weighting of other more complex strategies where used in this first approach, it seems that both options could be complementary.

	Voting approaches performed with medoids and centroids (*k-NN* = *km-NM* + *kc-NC*)				
	7-NN *2-NM* *5-NC*	*8-NN* *3-NM* *5-NC*	*9-NN* *4-NM* *5-NC*	*10-NN* *5-NM* *5-NC*	*11-NN* *6-NM* *5-NC*
Accuracy (%)	78.20	77.90	78.70	76.90	77.40

Table 7.10: Accuracy (%) of the system built combining the votes of the systems built with the *km-NM* and the *kc-NC* of the navigation sequence and navigation style datasets respectively.

In summary, we can state that the automatic classifier system based on supervised learning techniques described in the Section 7.2.3 (first approach) performed better than the systems based on unsupervised learning techniques described in this section (second approach). Indeed, the accuracy obtained obtained in the first approach (98%) was higher than the one achieved by the best classifier of the second approach (78.70%). In addition, the navigation style perspective seems to be more effective to classify success and failure type sessions than the navigation sequence perspective.

Failure detection subsystem based on the navigation style

Finally, we built a subsystem to detect users that were having problems (failure type sessions) enabling to adapt the restriction level (minimising the false positives).

In this subsystem new sessions were classified based exclusively on the 21 clusters of the navigation style dataset that detected failure type sessions (Table 7.6). Specifically, for each new session we computed the nearest centroid (*1-NC*) of the selected clusters and then, we reordered the validation dataset according to the distance (ascendant order). The smaller the distance is, the higher probability will the pattern have to be of failure type.

This allowed us to segment the new users and only work with those who were more similar to the failure patterns detected with the following accuracy values: 100%, 99% and 91% for the 10%, 15% and 25% nearest new sessions to the computed centroid respectively. This method also enabled us to define a distance threshold (1.657) to classify new users individually with a high certainty to be of failure type. Table 7.11 shows the accuracy of the failure detection subsystem for different number of nearest new sessions (%) based on the nearest centroid (*1-NC*).

	Number of nearest new sessions (%)						
	100%	75%	50%	25%	10%	5%	2.5%
Accuracy (%)	62.10	71.87	77.40	91.20	100.00	100.00	100.00

Table 7.11: Accuracy (%) of the failure subsystem built for different number of nearest new sessions (%) based on the nearest centroid (*1-NC*).

7.3 Modelling the use of e-Services

In this section we describe the contribution made to modelling the use of e-Services, where an empirical analysis of the use of e-Services in European countries based on survey data provided by Eurostat (Eurostat 2004) was carried out.

7.3.1 Introduction

Despite the efforts made by European governments and administrations in recent decades, the digital divide in the old continent still persists. The Digital Economy and Society Index (DESI) (European Commission 2018), a composite index which has been published annually since 2014 by the European Commission to measure the progress of the 28 European Union (EU) countries towards a digital economy and society, provides an idea of the digital divide in Europe. In particular, DESI regroups 34 indicators in five principal policy areas weighted as follows: 25% connectivity, 25% human capital, 15% use of Internet services, 20% integration of digital technology and 15% digital public services. From 2014 to 2018 the highest digital divide between the 28 EU countries (max-min) was reduced 11%, from 58% to 47%. Nevertheless, analysing the DESI of 2018, we observe that only four countries achieved high DESI values ($\geq$70%), whereas around half of the countries (15) scored medium values ($\in [50\%, 70\%)$) and about a third of the countries (9) achieved low values ($< 50\%$). Although these statistics indicate a slight improvement, it is clear that more forceful actions are called for.

One of the aspects affected by the consequences of the digital divide and also analysed within DESI (European Commission 2018) is the e-Government use or adoption. e-Government has several aspects, including social, technical, economic, political, and public administrative but most works define the mission of e-Government as systems that use Information and Communication Technology (ICT) to provide citizens with a better service (Shareef et al. 2011; Layne and Lee 2001). E-Government has been defined as the use of digital technology, especially Web-based applications, to enhance access to -and efficiently deliver of- government information and services. Although it has featured a substantial growth, development and diffusion, citizens in all developed and developing countries may not be willing to adopt such services (Carter and Bélanger 2005).

According to Shareef et al. (Shareef et al. 2011) e-Government has several aspects, including social, technical, economic, political, and public administrative. Most dominating concepts of e-Government arise from the technical perspective and a combination of the socio-economic and public administrative perspectives but in the academic literature, the adoption models offered so far are mainly conceptual. For instance, inside DESI e-Government is represented by the E-Government Development Index (EGDI) a composite index which has been published biannually in the UN E-Government Survey since 2010 (UN 2010) and which considers three aspects: telecommunications infrastructure, human capital and online services. In 2018 the average value of the EGDI for the 28 EU countries (80%) was rated as Very-High ($> 75\%$) and in line with DESI, the average EGDI score for the period 2010-2018 improved by 14% (UN 2010; UN 2012; UN 2014; UN 2016; UN 2018). Although some digital divide related themes are still mentioned in the 2018 UN E-Government Survey (UN 2018): access, affordability, age, bandwidth, content, disability, education, gender, migration, location, mobile, speed and useful usage, it seems that the e-Government situation in Europe is promising, at least from a theoretical point of view.

Another conceptual analysis is provided by the World Economic Forum's Networked Readiness Index (NRI), also referred to as Technology Readiness, which measures the propensity for countries to exploit the opportunities offered by Information and Communication Technology (ICT). It is published in collaboration with INSEAD (a graduate business school with campuses in Europe, Asia, and the Middle East), as part of their annual Global Information Technology Report (GITR) (Dutta et al. 2015). The report is regarded as the most authoritative and comprehensive assessment of how ICT impacts the competitiveness and well-being of nations. The index is a composite of three components: the environment for ICT offered by a given country or community (market, political, regulatory, and infrastructure environment), the readiness of the country's key stakeholders (individuals, businesses, and governments) to use ICT and the usage of ICT among these stakeholders.

Beyond the conceptual models, some limited empirical studies exist. For instance, the study carried out by Schwester (Schwester 2009) in US municipalities, concludes that e-Government adoption is a function of financial, technical, and human resources. Holding all other factors constant, municipalities with higher operating budgets, more full-time IT staff, and technical resources are more likely to implement a comprehensive e-Government platform. However, extensive empirical studies among the actual users to validate and generalise the models are absent (Shareef et al. 2011).

In this context, we consider the Eurostat Community Statistics on Information Society (CSIS) (Eurostat 2004) (Eurostat CSIS Surveys) an opportunity to carry out an extensive empirical study in European countries. They stated that in 2018 the average e-Government Use (EGU) in the 28 EU countries reached just 52%. The EGU is computed as the percentage of individuals who used the Internet to interact with public authorities, for example by obtaining information from public websites and downloading or submitting official forms.

The aim of this contribution is to offer some insights into the empirical e-Government adoption across Europe (26 EU countries). Inspired on the e-Government adoption options described by different authors (Bélanger and Carter 2008; Nam 2014; Thompson et al. 2005) and based on the Eurostat CSIS Surveys' (Eurostat 2004) e-Government Use (EGU) question, which asks about the contact of respondents with public authorities or public services, we defined two indexes: the E-Government Use Index (EGUI) and an extreme version of it ($EGUI^+$). With regard to $EGUI^+$ we defined four ranges: very high, high, low and very low and characterised the extreme levels of e-Government practical use (null and complete) by applying supervised learning procedures to the Eurostat data of two countries from each of the four $EGUI^+$ ranges. In addition, the ranking comparison carried out between $EGUI^+$ and four composite indexes measuring the level of e-readiness of a country provided by United Nations (UN 2010; UN 2012; UN 2014; UN 2016) and The World Economic Forum (Dutta and Mia 2010; Dutta and Mia 2011; Dutta and B. Bilbao-Osorio 2012; B. Bilbao-Osorio et al. 2013; Bilbao-Osorio et al. 2014; Dutta et al. 2015) determined that the index we defined is highly correlated with them.

7.3.2 Eurostat CSIS Surveys

In this section we first describe the data on e-Government extracted from the Eurostat's Community Statistics on Information Society (CSIS) 2009-2015 (Eurostat CSIS Surveys). Then we show the two indexes defined to characterise the e-Government practical use, EGUI and $EGUI^+$, based on the information extracted from the Eurostat CSIS Surveys.

Description of the Eurostat CSIS Surveys

From 2002 to the present, Eurostat CSIS Surveys have been annually conducted in all Member States, in two countries of the European Free Trade Association (EFTA), as well as in the candidate countries for future membership of the EU, and those in the process of accession to the EU. The data collection is based on Regulation (EC) 808/2004 (European Parliament and Council of the European Union 2004) of the European Parliament and the Council of the European Union and since 2011 the transmission of microdata to Eurostat is mandatory.

The Eurostat CSIS Surveys collect data on access and use of information and communication technologies (ICT) from households and individuals. The survey covers households with at least one member aged between 16 and 74 and individuals in this age range. Information on access to ICT, e.g. Internet connection, is collected at household level while statistics on the use of ICT, mainly on the use of the Internet, is gathered for individuals. Annual core subjects (included every year) and episodic topics on various ICT phenomena (changing for different years) are distinguished in the survey. There are six annual core subjects: access to ICT, use of computers, use of the Internet, e-Government, e-Commerce and e-Skills. To analyse variables of access and use of ICT in relation to household or individual characteristics, a number of so called social background variables, $b.v_i$, are collected (see Table 7.12). These include composition, income and regional location of the household as well as the age, gender, educational attainment and employment situation of individuals.

Code	Description	Type	Value	Description
HH_CHILD	No. children	$\mathrm{b.v_i}$	[1-4]	From one to four or more
HH_IQ	Income quartile	$\mathrm{b.v_i}$	[1-4]	Lowest / Second lowest / Second highest / Highest
AGECLS	Age range	$\mathrm{b.v_i}$	[1-8]	≤ 15 / $\in$ {[16-24], [25-34], [35-44], [45-54], [55-64], [65-74]} / $\geq$ [75]
SEX	Gender	$\mathrm{b.v_i}$	[1-2]	Male / Female
ISCED	Education level	$\mathrm{b.v_i}$	1	Primary/lower secondary
			2	Upper secondary
			3	Tertiary
EMPST	Employment situation	$\mathrm{b.v_i}$	1	Employee/self-employed
			2	Unemployed
			3	Student
			4	Not in the labour force
OCC_ICT	ICT occupation	$\mathrm{b.v_i}$	[0-1]	Non ICT / ICT professional
OCC_MAN	Manual occupation	$\mathrm{b.v_i}$	[0-1]	Non manual / Manual worker
IACC	Internet access	$\mathrm{q_i}$	[0-1]	No / Yes
CU	Computer use	$\mathrm{q_i}$	1	>a year ago/never
			2	$\in$ (3 months-a year) ago
			3	< 3 months ago
CFU	Computer freq. of use	$\mathrm{q_i}$	1	$\leq$ once a month/year
			2	$\leq$ once a week
			3	(Almost) every day
IU	Internet Use	$\mathrm{q_i}$	1	$\geq$ a year ago/never
			2	$\in$ (3 months-a year) ago
			3	< 3 months ago
IFU	Internet freq. of use	$\mathrm{q_i}$	1	$\leq$ once a month/year
			2	$\leq$ once a week
			3	(Almost) every day
IBUY	Buy goods over the Internet	$\mathrm{q_i}$	1	$\geq$ year ago/never
			2	$\in$ (3 months-a year) ago
			3	< 3 months ago
EGU	**E-Government use**	$\mathbf{q_i/ d.v_i}$	**1**	**Null**
			2	**OI**
			3	**OI & DF**
			4	**OI & DF & SF**

OI = obtain information, DF = download forms, SF = send filled forms

Table 7.12: Questions ($\mathrm{q_i}$) and background variables ($\mathrm{b.v_i}$) for the characterisation of the dependent variable ($\mathrm{d.v_i}$) e-Government practical use (EGU).

We were given access to the annual micro-datasets for the period 2009-2015, which we used for the analysis on the practical use of e-Government. Some questions in the micro-datasets varied from year to year, thus, for the analysis we used just the seven questions about ICT (q_i) common to all the years and the eight background variables ($b.v_i$), which are shown in Table 7.12 above.

Among the seven questions selected, one is at a household level and related to Internet access (IACC) whereas the remaining six are at individual level, two related to computer use (CU / CFU) and four related to the use of Internet (IU, IFU, IBUY and EGU). In the last row of Table 7.12 we show in bold the question selected as dependent variable ($d.v_i$) to measure e-Government practical use, EGU, which was obtained by coding a question about the activities related to interaction with public services or administrations through the Internet for private purposes, providing four possible values: 1 if none of the three possible activities was carried out, 2 if the obtaining information activity (OI) was carried out, 3 if the OI and the downloading official forms (DF) activities were completed and 4 if OI, DF and sending filled in forms activities were carried out.

Analysing the micro-datasets, we realised that six countries were missing data for the year 2008 and thus, we focused our analysis on the period 2009-2015. United Kingdom and Croatia were removed from our analysis because they were missing data for two of the years (2009 and 2010) of the period of time of our scope. Therefore our analysis comprises a total of 767,691 surveys from 26 different EU countries. Table 7.13 illustrates the ample variability in the number of surveys for the countries selected over the years. As shown in the table, Italy is the country with the biggest total number of CSIS Surveys (133,698) which is 25 times higher than that of the country with the smallest number, Malta (5,327), although in 2015 this country had 134 times lower population.

Literature analysis suggest that the information provided in the surveys to be promising for the empirical study proposed in this contribution. In the study conducted by Carter and Bélanger (Carter and Bélanger 2005), perceived ease of use, compatibility and trustworthiness appear to be significant predictors of citizens' intention to use an e-government service. In an empirical study conducted by Shareef et al. (Shareef et al. 2011), the authors observed that, e-Government adoption behaviour differs based on service maturity levels, i.e., when functional characteristics of organisational, technological, economical, and social perspectives of e-Government differ. A user will not arrive at an intention to use an e-Government system, which requires computer knowledge to get a competitive advantage, unless the user has competence from experience in the use of modern ICT. From technological, behavioural, economic, and organisational perspectives, it is anticipated that failing to get hands-on experience of technology will not create in the user an attitude favorable to adopting the system. Therefore, from organisational perspectives, computer self-efficacy is an important predictor of whether a user will adopt an e-Government system instead of using traditional government services. Bélanger and Carter (Bélanger and Carter 2008) propose a model of e-Government trust composed of disposition to trust, trust of the Internet (TOI), trust of the government (TOG) and perceived risk. Results from a citizen survey (214 responses) indicate that disposition to trust positively affects TOI and TOG, which in turn affects intentions to use an e-Government service. According to Nam (Nam 2014) the degree of e-Government use for a specific purpose is predicted by five sets of determinants: psychological factors of technology adoption, civic mindedness, information channels, trust in government, and socio-demographic and personal characteristics. Socio-demographic conditions influence usage level of various transactional services provided by e-Government. Perceived ease of use facilitates the acquisition of general information through e-Government.

		Number of Eurostat CSIS Surveys							
Country	Co	2009	2010	2011	2012	2013	2014	2015	Total
Austria	AT	4,634	4,620	3,178	3,454	3,371	3,291	3,455	26,003
Belgium	BE	4,049	4,109	3,872	3,899	4,000	3,794	0	23,723
Bulgaria	BG	2,832	3,325	4,876	4,064	4,682	5,167	4,847	29,793
Cyprus	CY	1,562	1,601	1,879	2,350	2,234	2,677	2,609	14,912
Czech R.	CZ	4,233	4,682	4,119	5,514	5,606	5,265	5,439	34,858
Denmark	DK	3,399	3,100	2,942	2,974	3,071	3,128	3,044	21,658
Estonia	EE	2,751	3,043	2,946	3,604	3,792	2,763	1,919	20,818
Greece	EL	1,538	1,568	1,865	1,482	1,813	2,080	2,522	12,868
Spain	ES	8,586	9,268	9,295	8,312	8,509	8,837	9,076	61,883
Finland	FI	1,989	2,053	2,164	2,141	2,107	1,967	2,072	14,493
France	FR	2,180	3,323	4,819	6,517	5,675	4,831	6,711	34,056
Hungary	HU	4,092	4,373	4,793	4,811	4,656	4,844	4,593	32,162
Ireland	IE	4,321	4,520	3,683	6,653	6,815	6,054	5,401	37,447
Italy	IT	18,133	18,461	19,143	18,611	19,229	19,539	20,582	133,698
Lithuania	LT	6,551	6,484	6,150	5,931	5,947	6,450	4,262	41,775
Luxemb.	LU	1,126	1,204	1,060	1,297	1,134	1,072	1,132	8,025
Latvia	LV	0	4,252	4,742	4,043	4,264	3,533	4,306	25,140
Malta	MT	583	634	812	709	852	881	856	5,327
Netherl.	NL	3,304	3,323	3,392	3,563	3,459	2,954	3,435	23,430
Norway	NO	878	803	856	778	842	854	902	5,913
Poland	PL	5,746	6,568	6,341	6,080	5,285	10,642	4,844	45,506
Portugal	PT	2,578	2,745	2,799	3,126	3,415	3,689	3,992	22,344
Romania	RO	4,731	5,688	6,154	6,216	7,819	8,570	9,405	48,583
Sweden	SE	3,207	2,976	2,124	1,033	1,110	1,067	966	12,483
Slovenia	SI	1,136	1,213	1,235	1,210	1,384	1,318	1,157	8,653
Slovakia	SK	2,682	3,025	2,930	3,357	3,593	3,320	3,233	22,140

Table 7.13: Number of Eurostat CSIS Surveys analysed for the period 2009-2015 in each country.

E-Government Use Indexes: EGUI / EGUI$^+$

The literature identifies different e-Government adoption levels. Bélanger and Carter (Bélanger and Carter 2008) differentiated the dependent variable ”Adoption” into two sub-groups:

- Adoption 1: Decision to accept and use an e-Government system to view, collect information, and/or download forms for different government services as the user requires with the positive perception of receiving a competitive advantage.

- Adoption 2: Decision to accept and use an e-Government system to interact with, and seek government services, and/or search for queries for different government services as the user requires with the positive perception of receiving a competitive advantage.

On the other hand, Nam (Nam 2014) and Thompson et al. (Thompson et al. 2005) identified three main purposes of e-Government use: information use, service use or engaging in electronic transactions with government and policy research or to participate in government decision making. The first two, could be equivalent to the Adoption 1 and 2 defined in Bélanger and Carter 2008.

Bearing these definitions in mind, in order to quantify e-Government practical use we defined two indexes, EGUI and EGUI$^+$, which are computed as ratios between the number of answers (#) to the question on EGU (EGU$_\mathrm{i}$) that reveal some level of e-Government use (i $\in$ {2,3,4}) and the ones that indicate no use (i=1). Equation 7.2 specifies how the two defined e-Government Use Indexes are computed. As it can be observed, EGUI takes into account the Adoption 1 or information use idea and EGUI$^+$ is an extreme version of EGUI that only involves the users engaged in electronic transactions, null use against complete use (#EGU$_\mathrm{i}$, i $\in$ {1,4}).

$$EGUI = \frac{\sum_{i=2}^{4} \#EGU_i}{\#EGU_1} \quad ; \quad EGUI^+ = \frac{\#EGU_4}{\#EGU_1} \tag{7.2}$$

In Table 7.14 we provide the list of countries ordered according to the EGUI$^+$ ranking, the total number of possible answers gathered for the EGU question (#EGU$_\mathrm{i}$, i $\in \{1, 2, 3, 4\}$), and the EGUI and EGUI$^+$ values.

Based on the EGUI$^+$ values we were able to rate the countries into four different e-Government use levels: very high ($\geq$ 2.0), high ($\in [1.0, 2.0)$), low ($\in [0.5, 1.0)$) and very low ($<$ 0.5). As a result, two countries were rated as having a very high level (DK, NO), six as having a high level (FI, NL, SE, FR, IE, EE), eight with a low level (AT, LU, ES, PT, SI, HU, LT, LV) and 10 with a very low level (BE, MT, EL, CY, SK, IT, BG, CZ, PL, RO).

	$\#EGU_i$				Value		
Country	i=1	i=2	i=3	i=4	EGUI	$EGUI^+$	$EGUI^+$ level
DK	2,955	3,718	1,701	13,284	6.33	4.50	Very High
NO	1,263	1,003	771	2,876	3.68	2.28	
FI	3,564	2,656	1,656	6,617	3.07	1.86	High
NL	7,231	3,986	1,278	10,935	2.24	1.51	
SE	3,285	2,573	2,085	4,540	2.80	1.38	
FR	11,844	5,263	4,497	12,452	1.88	1.05	
IE	16,406	2,677	1,647	16,717	1.28	1.02	
EE	8,202	4,251	394	7,971	1.54	0.97	
AT	9,747	4,993	4,451	6,812	1.67	0.70	Low
LU	2,948	1,083	1,982	2,012	1.72	0.68	
ES	26,320	11,815	7,409	16,339	1.35	0.62	
PT	11,689	2,784	921	6,950	0.91	0.59	
SI	3,222	1,690	1,896	1,845	1.69	0.57	
HU	14,802	5,803	3,127	8,430	1.17	0.57	
LT	23,661	4,532	329	13,253	0.77	0.56	
LV	10,589	7,706	1,263	5,582	1.37	0.53	
BE	11,525	4,528	2,303	5,367	1.06	0.47	Very Low
MT	2,636	723	776	1,192	1.02	0.45	
EL	6,375	2,842	949	2,702	1.02	0.42	
CY	7,678	1,799	2,247	3,188	0.94	0.42	
SK	10,210	5,250	2,769	3,911	1.17	0.38	
IT	88,551	13,377	13,555	18,215	0.51	0.21	
BG	18,996	5,231	1,809	3,757	0.57	0.20	
CZ	21,731	6,969	2,169	3,989	0.60	0.18	
PL	29,955	6,473	3,708	5,370	0.52	0.18	
RO	39,003	5,751	1,268	2,561	0.25	0.07	

Table 7.14: Average values of EGUI and $EGUI^+$ (2009-2015) in the 26 EU countries analysed.

7.3.3 Characterisation of extreme values of E-Government Use (EGU)

Aiming to obtain a greater understanding of e-Government practical use, we characterised the factors involved in the $EGUI^+$ index. As a preliminary study we computed the Pearson correlation for the 26 countries to get the correlation of the 14 independent variables with the two values of the dependent variable EGU: EGU_1 and EGU_4. This provided us with a global picture of factors which most influenced the extreme values of e-Government use in Europe. To facilitate the interpretation of the correlation results, the irrelevant answers (9=no answer

/ don't know) were removed for this analysis.

According to Pearson, a high frequency to buy goods over the Internet (IBUY) was the variable with the highest correlation coefficient ($|r| = 0.43$) with e-Government use (EGU), which according to Cohen (Cohen 1988) suggests a medium strength correlation ($0.3 < |r| < 0.5$). In addition, a medium strength correlation ($|r| = 0.34$) was also found between education level (ISCED) and e-Government use. Finally, manual occupation (OCC_MAN) and Internet frequency of use (IFU) were found to be inversely and positively correlated with EGU respectively ($|r| = 0.27$), which is considered nearly medium strength correlations. In all the cases the p-value of the test was lower than the significance level alpha, 0.05 and thus, the correlations found are significant although the majority of the values are of small strength ($0.1 < |r| < 0.3$).

To find more specific characteristics of e-Government use, we used the supervised learning algorithm Consolidated Tree Construction (CTC) (J.M. Pérez et al. 2007), which beyond a specific discriminating capacity to distinguish between the two extreme levels of EGU, provided a particular and stable description of the most influential variables for each EGU value. For the analysis we selected two countries from each of the four EGUI$^+$ levels defined, very high, high, low and very low.

In particular an experiment was run in Weka (M. Hall et al. 2009) with CTC for the eight countries selected, using the 14 independent variables and the dependent variable EGU with two possible values, null (EGU$_1$) and complete (EGU$_4$). A ten-fold cross-validation (10-fold CV) strategy was used for validation. Table 7.15 shows the characteristics of the datasets and the obtained classification rates. As can be observed the datasets are quite unbalanced in the majority of countries selected: columns #EGU$_\text{i}$ and #EGU$_\text{i}$ (%). Thus, in order to obtain a better characterisation of the minority EGU class in each country, CTC was run using a distribution of the minority class of 50% and 2% of each dataset as the minimum number of instances per leaf, which limits the minimum size of any decision node to the specified value.

		#EGU$_\text{i}$		#EGU$_\text{i}$ (%)		CTC average results			
Co.	EGUI$^+$ level	i=1	i=4	i=1	i=4	Pr	Re	Fm	Acc
DK	Very High	2,955	13,284	18	82	0.85	0.83	0.84	0.83
NO	Very High	1,263	2,876	31	69	0.76	0.74	0.74	0.74
IE	High	16,406	16,717	50	50	0.73	0.73	0.73	0.73
EE	High	8,202	7,971	51	49	0.73	0.73	0.73	0.73
LV	Low	10,589	5,582	65	35	0.77	0.74	0.74	0.73
BE	Low	11,525	5,367	68	32	0.77	0.71	0.72	0.71
PL	Very Low	29,955	5,370	85	15	0.86	0.74	0.77	0.74
RO	Very Low	39,003	2,561	94	6	0.94	0.67	0.75	0.67

Table 7.15: Number of answers for extreme EGU levels (#EGU$_\text{i}$, i $\in$\{1,4\}) and CTC average results in eight countries with four EGUI$^+$ levels.

According to Table 7.15 the average results achieved by the CTC trees in terms of precision (Pr), recall (Re), F-measure (Fm) and accuracy (Acc) were good with values over 0.71 in the four groups, except in Romania where recall and accuracy scored 0.67. This is not surprising since Romania has a very unbalanced dataset, with 94% of the surveys being of null e-Government use type (#EGU$_1$), which reduces the recall and accuracy of the minority class.

The structures of the classification trees provide an explanation of the classification. In Figures 7.3, 7.4 and 7.5 CTC trees obtained for Denmark, Belgium and Poland are shown by way of example. In the CTC trees displayed, in the leaf nodes the first number (0/1) represents the class given to the leaf node (EGU$_1$/EGU$_4$), whereas inside the parenthesis, the numbers before and after the slash represent the number of examples involved and the number of misclassified examples respectively.

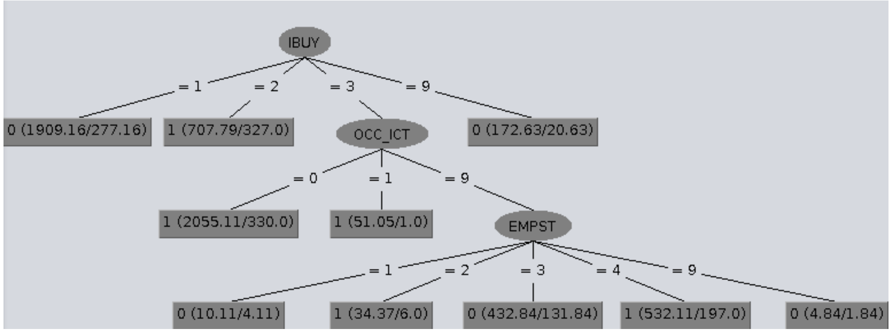


Figure 7.3: CTC tree obtained for Denmark which has a very high EGUI$^+$ level.

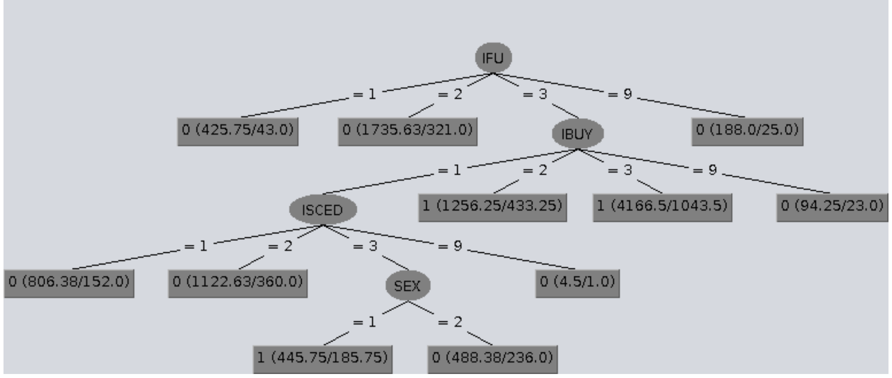


Figure 7.4: CTC tree obtained for Belgium which has a low EGUI$^+$ level.

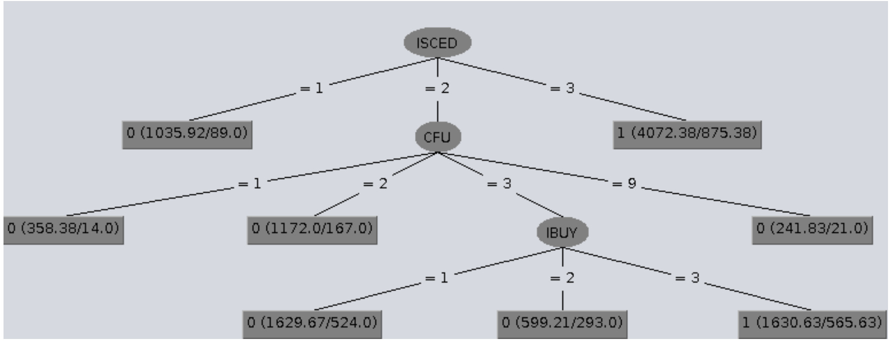


Figure 7.5: CTC tree obtained for Poland which has a very low EGUI$^+$ level.

	Main rules for extreme levels of e-Government use	
Country	Null: EGU$_1$	Complete: EGU$_4$

DK	IBUY=1
	IBUY=3 & EMPST=1/3
		IBUY=3 & OCC_ICT=0
		IBUY=2

NO	IFU=3 & IBUY=1
	IFU≠3
		IFU=3 & IBUY=3
		IFU=3 & IBUY=2 & ISCED≠1

IE	IBUY=1 & CFU≠3
	IBUY=1 & CFU=3 & ISDEC≠3
		IBUY=3 & IFU=3 & ISCED=3
		IBUY=3 & IFU=3 & ISCED=2
		& AGECLS≠2

EE	IBUY=1 & OCC_MAN≠0
	IBUY=1 & OCC_MAN=0
	& CFU≠3
		IBUY=3 & ISCED≠1
		IBUY=1 & OCC_MAN=0
		& CFU=3

LV	ISCED=1
	ISCED=2 & IBUY=1
		ISCED=3 & EMPST≠4
		ISCED=2 & IBUY≠1

BE	IFU≠3 IFU=3 & IBUY=1 & ISCED≠3	IFU=3 & IBUY≠1

PL	ISCED=1
	ISCED=2 & CFU≠3
		ISCED=3
		ISCED=2 & CFU=3 & IBUY=3

RO	OCC_MAN≠0	OCC_MAN=0

Table 7.16: Main rules provided by the CTC trees for complete an null e-Government use.

Globally analysing the structures of the classification trees, we concluded that excluding the countries with a very low EGUI$^+$ level, complete e-Government use (EGU$_4$) was closely related to recent online shopping (IBUY=3), whereas the same action carried out more long time ago (IBUY=1) seemed to be connected to null e-Government use (EGU_1). Table 7.16 summarises the main rules provided by the CTC trees for each country and their descriptions are given in the next paragraphs. According to the table, the following variables were found to have a close connection with extreme e-Government use levels, listed in a descending number of appearances in the main 29 rules presented for the eight countries: Buy goods over the Internet (IBUY) = 21/29, Education level (ISCED) = 14/29, Internet frequency of use (IFU) = 8/29, Computer frequency of use (CFU) = 6/29, Manual occupation (OCC_MAN) = 5/29, Employment situation (EMPST) = 2/29 and ICT occupation (OCC_ICT) = 1/29.

Countries with a very high $EGUI^+$ level: Denmark and Norway

In Denmark the citizens who rarely bought goods over the Internet (IBUY=1) and those who had done online shopping recently and were employees/self-employees or students (IBUY=3 & EMPST= 1/3) did not use e-Government tools (EGU_1). On the other hand, Danish citizens who did use e-Government tools (EGU_4) had bought online quite recently and were not ICT professionals (IBUY=3 & OCC_ICT=0) or had done online shopping between three months and a year previously (IBUY=2). These main CTC rules found for EGU in Denmark are also available in the most representative nodes of each class (0:EGU_1/1:EGU_4) in Figure 7.3.

In Norway the people who used the Internet almost everyday but had not bought goods over the Internet for a time (IFU=3 & IBUY=1) and the ones who did not daily use the Internet (IFU≠3) did not use e-Government tools. Conversely, Norwegian people who used the Internet almost everyday and had recently bought goods over the Internet (IFU=3 & IBUY=3) did use e-Government. In addition, Norwegians who used the Internet almost daily, had bought goods over the Internet quite recently and did not have a low education level (IFU=3 & IBUY=2 & ISCED≠1) also used e-Government tools.

Countries with a high $EGUI^+$ level: Ireland and Estonia

In Ireland the citizens not using e-Government tools (EGU_1) were those who hardly ever bought goods over the Internet and who did not use the computer daily (IBUY=1 & CFU≠3) together with those who hardly ever did online shopping, used the computer almost daily but did not have a high education level (IBUY=1 & CFU=3 & ISCED≠3). On the other hand, the Irish citizens who recorded a complete use of e-Government (EGU_4) were those who had bought goods over the Internet recently, used the Internet daily and had a high education level (IBUY=3 & IFU=3 & ISCED=3). Additionally, Irish people who had shopped online recently, used the Internet almost daily, had a medium education level and were not in the age range of 16-24 years also used e-Government tools (IBUY=3 & IFU=3 & ISCED=2 & AGECLS≠2).

In Estonia citizens who had not bought online for long time and had a manual occupation (IBUY=1 & OCC_MAN≠0), as well as those who had not shopped online for a long time, but did not have a manual occupation and did not use the computer daily (IBUY=1 & OCC_MAN=0 & CFU≠3) did not use e-Government tools. However, Estonian citizens who had bought goods over the Internet recently and did not have a low education level (IBUY=3 & ISCED≠1) or those who had bought online long time ago, did not have a manual occupation but used the computer almost every day (IBUY=1 & OCC_MAN=0 & CFU=3), did use e-Government tools.

Countries with a low EGUI$^+$ level: Latvia and Belgium

In Latvia, people with a low education level (ISCED=1) and those with a medium education level who had bought goods over the Internet only a long time ago (ISCED=2 & IBUY=1) had no inclination to use the e-Government tools (EGU_1). On the other hand, Latvians with a high education level who were not retired (ISCED=3 & EMPST ≠ 4) and those with a medium education level who had bought online in the previous 12 months (ISCED=2 & IBUY≠1) did use such tools (EGU_4).

In Belgium the null use of e-Government is related to citizens who did not use the Internet daily (IFU≠3) along with the ones who used it almost daily but had not bought goods over the Internet for a long time and did not have a high education level (IFU=3 & IBUY=1 & ISCED≠3). On the other hand, Belgian citizens making a complete use of e-Government, used the Internet daily and had shopped online within the previous 12 months (IFU=3 & IBUY≠1). The most representative nodes of each class ($0{:}EGU_1/1{:}EGU_4$) in the CTC tree shown in Figure 7.4 also exhibit the main CTC rules found for EGU in Belgium.

Countries with a very low EGUI$^+$ level: Poland and Romania

In Poland citizens with a low education level (ISCED=1) together with those with a medium education level who did not use the computer daily (ISCED=2 & CFU≠3) showed a null trend towards the use of e-Government (EGU_1). On the other hand, Polish who used e-Government tools (EGU_4) had a high (ISCED=3) or a medium education level, used the computer almost daily and had bought goods over the Internet quite recently (ISCED=2 & CFU=3 & BUY=3). The most representative nodes of each class ($0{:}EGU_1/1{:}EGU_4$) shown in Figure 7.5 also illustrate the main CTC rules found for EGU in Poland.

Romania was the only country where the two extreme e-Government use levels were characterised by two rules that involved a single factor, manual occupation (OCC_MAN): manual workers (OCC_MAN≠0) did not use the e-Government tools, whereas non manual workers did use them.

7.3.4 Comparison between EGUI$^+$ and other indexes

Aiming to analyse if the performance of the index we defined to measure the practical e-Government use is similar to other conceptual indexes broadly used as indicators of related features such as e-Readiness of a country, we selected four indexes and compared them to EGUI$^+$: E-Government Development Index (EGDI) and its Online Service Index (OSI) component, and the Networked Readiness Index (NRI) and its Government Use (GU) component. Next we describe the indexes mentioned and the comparison carried out.

7.3.4.1 Description of the indexes

From 2001 to the present, The United Nations Department of Economic and Social Affairs (UNDESA) has published the UN E-Government Survey (UN 2018). In 2003 this survey began to provide an analysis of the progress in using e-government via the E-Government Development Index (EGDI), a composite index based on the weighted average of three normalised (norm.) indices, assigning one third weight to each of them (see Equation 7.3): the Telecommunications Infrastructure Index (TII), the Human Capital Index (HCI) and the Online Service Index (OSI). As a composite indicator, the EGDI is used to measure the readiness and capacity of national institutions to use ICTs to deliver public services (UN 2018). Prior to the normalisation of the three component indicators, the Z-score standardisation procedure is implemented for each component indicator to ensure that the overall EGDI is decided equally by the three component indexes.

$$EGDI = \frac{1}{3}(TII_{norm.} + HCI_{norm.} + OSI_{norm.}) \tag{7.3}$$

The OSI index, one of the three components of the EGDI index described in Equation 7.3, is a composite normalised score based on an independent survey questionnaire, conducted by UNDESA, which assesses the national online presence of all 193 United Nations Member States. The survey questionnaire computes several features related to online service delivery, including whole-of-government approaches,open government data, e-participation, multi-channel service delivery, mobile services, usage up-take, digital divide as well as innovative partnerships through the use of ICTs. (UN 2018)

The World Economic Forum has been annually publishing The Global Information Technology Report (Dutta et al. 2015) since 2001, where the Networked Readiness Index (NRI) is provided. As shown in Equation 7.4, the NRI is a composite index computed as the weighted average of four main subindexes (subind.), being all the weights a quarter: Environment subindex, Readiness subindex, Usage subindex and Impact subindex.

$$NRI = \frac{1}{4}(Enviroment_{subind.} + Readiness_{subind.} + Usage_{subind.} + Impact_{subind.}) \tag{7.4}$$

The Usage subindex of NRI assesses the level of ICT adoption by a society's main stakeholders: government, businesses and individuals (Dutta et al. 2015). In particular, the Usage subindex is computed as the weighted average of three pillars (using weights of one third), the Individual usage, Business usage and the Government usage. In this case we focused on the Government usage pillar, which assesses the leadership and success of the government in developing and implementing strategies for ICT development, as well as in using ICTs, as measured by the availability and quality of government online services (Dutta et al. 2015). The Government usage pillar is computed as the average of the importance of ICTs to government vision, the Government Online Service Index and the Government success in ICT promotion.

7.3.4.2 Ranking comparison

In order to study the relationship between e-Government adoption (EGUI$^+$) and the level of e-Government readiness (EGDI), network readiness (NRI), national online presence (OSI) and ICT adoption by government (GU), we compared their rankings for the 26 countries analysed. For the comparison we tried to use similar time periods, 2009-2015 period for the annual indexes or indicators, EGUI$^+$, NRI and GU (Dutta and Mia 2010; Dutta and Mia 2011; Dutta and B. Bilbao-Osorio 2012; B. Bilbao-Osorio et al. 2013; Bilbao-Osorio et al. 2014; Dutta et al. 2015), and 2010-2016 period for the biannual indexes or indicators, EGDI and OSI (UN 2010; UN 2012; UN 2014; UN 2016).

Specifically we computed the number of positions won or lost (positive or negative value) by the countries from the EGDI, OSI, NRI and GU rankings to the EGUI$^+$ ranking, which in general terms is low (see Table 7.17). As shown in Table 7.17, we grouped the countries into three different sets using ±5 positions as a threshold for the ranking differences appreciated (nearly a 20% of the ranking) represented by the following codes: blue-bold if they drop more than five positions, green-roman if they drop or gain fewer than five positions (stable countries) and red-italic if they gain more than five positions.

According to Table 7.17, for a great majority of the countries involved in the analysis, 74% on average (green-roman ones) the practical e-Government use does match the features measured by the conceptual the indexes, EGDI, OSI, NRI and GU. To this regard, we found twelve countries (46%) appearing in all the groups with small ranking differences (stable countries): Austria (AT), Bulgaria (BG), Cyprus (CY), Denmark (DK), Hungary (HU), Lithuania (LT), Luxembourg (LU), Latvia (LV), Netherlands (NL), Norway (NO), Portugal (PT) and Sweden (SE). In addition, we observed that EGDI is the most similar index to EGUI$^+$, since 88% of countries (23/26) are of stable type.

On the other hand, only 16% of the countries (blue-bold ones) on average showed higher positions in the rankings provided by the rest of the indexes than for that of EGUI$^+$ ($<$ -5 positions). Analysing all the groups with high negative ranking differences with EGUI$^+$ (blue-bold ones), we did not find any country common to all of them but Czech Republic (CZ) and Belgium (BE) could be considered as common since they are nearly in the blue-bold groups of EGDI and GU indexes respectively. In addition, the negative ranking differences were lower for EGDI than for the rest of indexes, where we observed that Czech Republic (CZ) was the country with higher drops, falling from 7th, 4th and 3rd positions in the OSI, NRI and GU rankings to the 26th one in the EGUI$^+$

ranking.

Finally 9% of the countries on average achieved lower positions (red-italic) in the rankings of the four conceptual indexes than in the ones provided by EGUI$^+$ ($>$ 5 positions), although the EGDI one does not have any country with such ranking rises. In the rest of indexes, we found that Estonia (EE) is the only country common to all the groups with high positive ranking differences. In addition, Estonia also was the country with the highest ranking rise for the index we defined, rising from 22nd, 24th and 26th positions in the OSI, NRI and GU rankings to 8th one in that of EGUI$^+$.

	Ranking dif.		Ranking dif.		Ranking dif.		Ranking dif.
Co.	EGDI-EGUI$^+$	Co.	OSI-EGUI$^+$	Co.	NRI-EGUI$^+$	Co.	GU-EGUI$^+$
IT	**-10**	**CZ**	**-19**	**CZ**	**-22**	**CZ**	**-23**
BE	**-7**	**EL**	**-16**	**ES**	**-9**	**MT**	**-12**
PL	**-6**	**IT**	**-12**	**BE**	**-6**	**SK**	**-10**
CZ	-4	**BE**	**-6**	MT	-5	**EL**	**-7**
ES	-4	**ES**	**-6**	PL	-5	**RO**	**-6**
FR	-4	**PL**	**-6**	SE	-4	BE	-4
NL	-3	LT	-5	IT	-3	BG	-4
LT	-2	RO	-4	LU	-3	ES	-3
MT	-2	FI	-2	RO	-3	SE	-3
RO	-2	NL	-2	CY	-2	PT	-2
AT	0	HU	-1	EL	-2	IT	-1
BG	0	AT	0	HU	-2	LT	-1
EE	0	MT	0	AT	-1	LU	-1
EL	0	BG	1	IE	-1	PL	-1
SE	0	LV	1	NL	-1	CY	0
LU	1	NO	2	BG	1	DK	0
DK	2	PT	2	LT	1	HU	1
NO	2	CY	3	PT	2	NL	1
SI	2	SE	3	SI	2	AT	2
CY	3	SI	3	SK	2	NO	2
FI	3	SK	4	LV	3	LV	3
SK	3	DK	5	NO	3	FI	4
HU	4	LU	5	*FI*	*7*	SI	4
LV	4	*FR*	*13*	*DK*	*8*	*IE*	*9*
IE	5	*IE*	*13*	*FR*	*14*	*FR*	*12*
PT	5	*EE*	*14*	*EE*	*16*	*EE*	*16*

Table 7.17: EGUI$^+$, EGDI, OSI, NRI and GU ranking comparison for the 26 countries analysed.

For a deeper analysis of the similarity between the performance of the four conceptual indexes and that of EGUI$^+$ we computed four pairwise comparisons based on Kendall correlation (Kendall 1938) using the rankings provided by each index. Table 7.18 shows the results of Kendall pairwise tests between EGUI$^+$ and EGDI, OSI, NRI and GU, in terms correlation values ($\mathcal{T}$) and significance (p-values). The second and third columns of the table show the performance of the stable countries (green-roman ones in Table 7.17) suggesting that the four indexes are highly correlated with EGUI$^+$ at 0.05 significance level, p-value $< \alpha$, with correlation values ($\mathcal{T}$) on average of 0.8. In addition, in the in the fourth and fifth columns of Table 7.18 we also show the results of the Kendall tests carried out analysing the complete set of countries. In this case, the values of $\mathcal{T}$ decreased down to 0.5 on average, being EGDI the index which scores the highest correlation value ($\mathcal{T}$=0.72) with the index we defined. In the global comparison, we observed higher correlation values for the indexes measuring the e-readiness of the countries (EGDI and NRI) than for the indicators of features related with e-Government (OSI and GU).

	Stable* countries		26 countries	
Index	p-value	$\mathcal{T}$	p-value	$\mathcal{T}$
EGDI	3.87×10^{-06}	0.72	8.90×10^{-10}	0.78
OSI	2.35×10^{-02}	0.37	5.51×10^{-05}	0.82
NRI	4.04×10^{-03}	0.46	2.38×10^{-08}	0.81
GU	4.69×10^{-03}	0.43	2.29×10^{-07}	0.79

* Ranking differences with EGUI$^+$ $\leq \pm 5$ positions.

Table 7.18: Results of the Kendall pairwise correlation tests between EGUI$^+$ and EGDI, OSI, NRI and GU.

Considering all the above, we can state that the empirical analysis carried out on e-Government adoption across Europe through EGUI$^+$ index concur to a large extent with the theoretical studies which measure the level of e-readiness of European countries through different indexes (EGDI, OSI, NRI and GU).

7.3.5 Discussion

On the one hand the digital divide makes the task of providing universally accessible online government services challenging (Schwester 2009) and on the other hand, citizen confidence in the ability of an agency to provide online services is imperative for the widespread adoption of e-government initiatives (Bélanger and Carter 2008). According to Shareef et al. (Shareef et al. 2011), e-Government adoption behaviour differs when functional characteristics of organisational, technological, economical, and social perspectives of e-Government differ. The first part of the empirical study carried out based on Eurostat CSIS surveys, the classification of countries in different e-Government use levels (see

7.14) is in concordance with the statement since, first of all, no all countries have the same e-Government use level, and, although with exceptions, more developed and wealthier countries seem to have higher levels of e-Government use.

With regard to the factors affecting the e-Government use, buying goods in the Internet could be expected to be one of the factors directly related to e-Government use due to the similarities existing between e-commerce and e-government. According to Schwester (Schwester 2009) the same way factors from Technology Acceptance, Diffusion of Innovation and trustworthiness models play a role in user acceptance of e-commerce, it is expected that they will also affect e-Government adoption. The outcome of our study shows that in countries with higher e-Government adoption according to Eursotat CSIS surveys, IBUY, the variable related to e-commerce is the one conditioning most of the times the use or not use of e-Government services.

But, this is not always the case, there are differences between commercial businesses and government agencies (Bélanger and Carter 2008). E-commerce and e-Government differ in their reasons for existence (profit vs. service) and constituents served (target market vs. population at-large). Businesses can choose their customers; however, in e-Government, agencies are responsible for providing access to the entire eligible population, including individuals with lower incomes and disabilities (Schwester 2009). Mandatory relationships exist only in e-Government. Citizens perceive businesses differently than government. In addition, the structure of businesses is different from the structure of agencies in the public sector. Decision-making authority is less centralised in government agencies than in businesses. This dispersion of authority impedes the development and implementation of new government services. The third difference is accountability. In a democratic government, public sector agencies are constrained by the requirement to allocate resources and provide services 'in the best interest of the public'. The political nature of government agencies is also a feature that makes e-Government and e-commerce different. These factors could be related with the fact that in countries with lower adoption level, other factors such as education level and occupancy appear to be related to the e-Government adoption.

On the other hand, some authors, Afyonluoglu and Alkar (Afyonluoglu and Alkar 2017) for instance, compared 16 international e-government benchmarking studies completed between 2001 - 2016 by five active organisations including UN and WEF and identified the common points and the differences with respect to 22 different criteria including indexes such as EGDI and NRI. They pointed out that none of studies compared measures the "usage of e-services by citizens", "governance model of e-Government", "benefits of e-services" and "satisfaction", suggesting that they should be considered for future e-Government framework improvements. Similarly, Jadi and Jie (Jadi and Jie 2017) use the EPI E-participating index, a supplementary indicator designed by the UN, as an output of government effort to evaluate the performance of e-Government systems. The authors state, that although the EGDI is used as a benchmark to provide a numerical ranking of e-Government development, building websites, infrastructures and providing online services only shows how the readiness of the government to exploit the facilities. However, in addition to those indicators the performance of e-Government systems can be analysed by measuring to what extent citizens are using these facilities. The second part of this work is in line with the previous lines, where we have compared the e-Government adoption and the e-readiness of 26 EU countries, based on the EGUI$^+$ index empirically computed from the 2009-2015 Eurostat CSIS Surveys, the EGDI index and its OSI component published in the 2010-2016 UN e-Government Surveys, and the NRI and its GU component provided by the 2010-2015 World Economic Forum's Global Information Technology Reports. According to our analysis, it seems that in the majority of the countries the situation of the e-Government does not differ substantially despite using different calculation methods. To this regard we think that the compute of e-readiness of countries (EGDI, NRI) could be improved by including the real use of e-Services quantified in the index we define (EGUI$^+$).

7.4 Summary

In this chapter we presented two contributions made on e-Services, the first one focused on modelling the interaction of users in the enrolment web information area of a university and the second one empirically analysing the use of e-Government services in Europe.

In particular, in the first contribution we analysed the navigation in the enrolment area of the University of the Basque Country, carrying out a complete data mining process which showed that successful and failure navigation behaviours can be automatically modelled using data mining techniques. To that end, two domains were defined to represent the navigations of the users: 28 interaction features extracted from the recorded click-streams (navigation style) and URLs visited by the users in each session (navigation sequence). On the one hand, using supervised learning, CTC trees (J.M. Pérez et al. 2007) over the first domain (navigation style), we are able to automatically distinguish the two navigation types with an accuracy rate of 98%. On the other hand, using unsupervised learning techniques, in each of the two domains and in a combination of both, we were also able to automatically detect the success and failure navigation sessions but achieving a lower accuracy, around 78%. Besides, an additional subsystem to detect failure type sessions was built based on the navigation style using unsupervised procedures, which enabled to tune the accuracy in order to achieve higher values, 100% for a 10% of the new nearest sessions. The two main systems built based on the navigation style to automatically classify success and failure type navigations, were also able to identify the main rules (supervised) and characteristics (unsupervised) of each type of session, e.g. the time spent on average on content (text is dominant) or scatter (links are dominant) type URLs. Thus, we think that this contribution is a suitable basis to improve the e-Service analysed in a near future.

In the second contribution, we analysed the practical use of e-Services supplied by Governments across Europe (e-Government adoption for 26 EU countries) based on the empirical data provided by Eurostat (Eurostat 2004). Based on the data obtained we first quantified this factor by defining two indexes, the E-Government Use Index (EGUI) and an extreme version of it (EGUI^+). Then, using CTC trees (J.M. Pérez et al. 2007), we characterised the use/non use of e-Government services in a selection of countries with different EGUI^+ levels. These procedures achieved an average accuracy of 73% and determined the main factors related to the practical use of e-Government in each of the countries, e.g. the frequency of buying goods over the Internet or the education level. In addition, we compared one of the proposed index (EGUI^+) to other indexes measuring the level of e-readiness of a country such as the E-Government Development Index (EGDI) its Online Service Index (OSI) component, the Networked Readiness Index (NRI) and its Government Use component. The ranking comparison found that EGUI^+ was correlated with the four indexes mentioned at 0.05 significance level. The outcomes contribute to gaining an understanding of what are some of the factors characterising the practical use of e-Government in Europe. Thus, our findings can provide some guidelines with which to improve the interaction of citizens with web services and information offered by institutions depending on their e-Government use level (EGUI^+).

Part IV

Conclusions

Chapter 8

Conclusions

This dissertation presents contributions to improve HCI systems based on machine learning techniques in different contexts. Therefore, there are contributions in machine learning and HCI contexts:

In the context of machine learning, the work deals with one of the main difficulties that the use of clustering procedures presents, which is evaluating the quality of partitions in all type of contexts. Cluster Validity Indexes (CVIs) enable this evaluation but none of them has proven to be the best in all situations thus, in the first contribution several decision fusion approaches using different indexes are proposed as an effective alternative.

In HCI, the contributions are framed in three areas. The first one belongs to the accessibility context and presents a system to automatically detect navigation problems of users with and without disabilities.

The medical informatics area is analysed in the next one by first, connecting visual and interaction behaviours on a medical dashboard used to support the decision-making of clinicians (SMASH) and second, by automatically detecting and characterising two main cohorts of users based on their interaction behaviour: primary (pharmacists) vs secondary (non-pharmacists).

Finally, we also contributed to the area of e-Services modelling their interaction and use. On the one hand, based on the interaction data gathered in the website of a university (UPV/EHU) potential students aiming to enrol the university were modelled. On the other hand, based on survey data provided by Eurostat a quantification and characterisation of the e-Government adoption was accomplished.

Our contributions in such different HCI environments prove the importance of machine learning to generate better HCI systems in the future. In the following sections the main conclusions aroused in each of the contexts will be break down.

8.1 Machine learning - Clustering validation.

The analysis carried out shows that the design of CVI decision fusion strategies similar to the ones used in multiple classifier systems for clustering validation requires weighting the votes according to the characteristics of the CVIs involved and the experimental factors available (e.g noise level). None of the strategies with unweighted votes, showed any improvement in performance compared to the best CVI (Silhouette), whereas nearly every voting strategy weighting the CVIs according to their performance showed to behave better than single CVIs.

In particular, the best voting strategy for real and synthetic datasets uses the two CVIs with the highest success rates in each controllable experimental factor and weights each vote according to the importance of each factor defined beforehand. In this regard, the Friedman and Wilcoxon tests performed indicated that the results of the best voting strategy were significantly better than the ones given by the 10 top ranked indexes of the reference work (Arbelaitz et al. 2013b).

In light of the results achieved, we think that decision fusion strategies are a successful path to determine which is the best partition for each context, which is the key issue in the unsupervised learning field. Thus, we believe that new contributions on decision fusion strategies for clustering validation can help reducing the uncertainty about the suitability of the partitions generated by the algorithms.

8.2 HCI - Accessibility

A system built carrying out a complete data mining process on the data collected by RemoTest platform showed to be a promising strategy to automatically detect the web navigation problems that users with and without disabilities may be experiencing.

The first step, a hierarchical two-level approach based on supervised learning procedures to automatically discriminate four different devices first in two groups (keyboard and others) and second, trackball, joystick and mouse obtained high accuracy: 99.26 in the first level and 89.84 in the second level. As future adaptations discussed for each type of devices differ significantly, minimizing the critical error occurred when discriminating both groups of devices was vital.

The second step of the system detects automatically problems that each user may be experiencing while carrying out two types of tasks using unsupervised learning procedures. On the one hand MiniTasks are directed navigations in which users are asked to click highlighted targets and on the other hand, SearchingTasks are searching and directed navigations where users are asked to search a particular web page. Concretely, for each task and each device, the navigation traces grouped in clusters with automatically detected deviated features were visually analysed to find problematic patterns. A total of seven problematic patterns including too much distance, too much time, rectifications in directions, unnecessary clicks, difficulties around the target, long clicks, and too many stops were found in MiniTasks, and, all of these, except difficulties around the target, were also observed in SearchingTasks, which is understandable given that in the first scenario the users had to reach the targets appearing in the screen.

In addition we were able to point out the reasons behind the detected patterns and suggest suitable adaptations according to the patterns and the devices used.

8.3 HCI - Medical informatics

In this contribution for which two analysis were accomplished using interaction and visual data gathered in two studies (lab and observational) carried out with clinicians using a medical dashboard (SMASH), conclusions on two matters were obtained: synergies between visual and interactive behaviours and user modelling based of interactive behaviour.

8.3.1 Synergies between visual and interactive behaviours

SMASH has seven screens and we analysed the screen divided interaction as well as the global interaction of lab study participants, who had to complete specific tasks. The clustering analysis of the interaction data showed that the users' interactive behaviour was similar across all screens of SMASH. Alternatively, if the behaviour was different (i.e. a specific screen led to a different behaviour) this also changed similarly across the participants within the clusters. This finding suggests that incorporating the specific screen in the user model may not make any difference to the way dwell time and exploration metrics, elapsed time and number of mouse hovers between two consecutive clicks respectively, are used in the model.

On the other hand, the SMASH Interface is divided in nine different areas of interest (AOIs) considered relevant for the gaze activity and we analysed the AOIs divided gaze activity of lab study participants, in terms of fixation duration. The analysis of the gaze activity yields the same groups as the ones generated by the analysis of user interaction. This indicates that those participants who exhibited a particular interactive behaviour in terms of dwell time and exploration had a similar visual behaviour in terms of fixation duration on the AOIs. Since gaze is a proxy of attention and, at the same time, attention precedes action (Huang et al. 2012), we can say that these groupings are not incidental and the exhibition of particular interactive behaviours might be determined, to some extent, by the duration of fixations on specific areas of interest.

The inclusion in the analysis of the interaction data of the participants from the observational study who used SMASH as part of their daily activities, showed that the resulting behaviours are stable across the two settings despite the fact that different tasks were conducted.

This has promising consequences in that visual behaviour could be inferred using interactive data alone. Some interaction data analysis can be carried out in real time in the browser. The processing and analysis of interaction data is straightforward –provided that the right metrics are being monitored– and the deployment of eye-tracking devices beyond laboratory settings is not to be expected in the medium term. The computed metrics to measure the interaction and the gaze activity can be used 1) to infer usability problems in real-time and 2) to inform adaptations on the user interface. Without having gaze data from the participants of the observational study, we can hypothesise that their visual behaviour might be similar to that of the laboratory participants who were grouped in their respective clusters. Future work will certainly pursue this lead.

8.3.2 User modelling based on interactive behaviour

This contribution successfully modelled the interactive behaviour of two different cohorts of electronic dashboard users of the observational study. In addition to the explicit differences derived from the descriptive analysis, we identified the differences that characterised the two groups in terms of their interactive behaviours. These differences are important to understand everyday use of the SMASH dashboard where primary pharmacist users are more competent on screens that provide summary and trends information, and secondary general practice staff users are less competent on screens containing a detailed breakdown of the data. We propose workflows that encompass these activities in a coherent sequence and personalised educational nudges to foster engagement.

The contributions of the work are twofold. On the one hand, a methodological contribution suggests that it is feasible to characterise the interactive behaviour of users in a medication safety dashboard using user interaction events such as mouse hovers and elapsed time between two consecutive clicks. On the other hand, an empirical contribution advances into our understanding of how medical dashboards are used by health care stakeholders, an area which remains largely unexplored and is key to perform adaptations that cater for the users' ability to perceive, process and make data actionable (Dowding et al. 2015).

8.4 HCI - E-Services

The last two contributions lead us to conclude that modelling both, the interaction and the use of e-Service, is possible.

8.4.1 Modelling the interaction with e-Services

This contribution presents the modelling of the enrolment web information area of the UPV/EHU using web mining techniques. The navigation sessions, were classified in two types based on the last URL visited: success (users interested in enrolment information) and failure (users not reaching enrolment information).

In the first approach, using two supervised algorithms, C4.5 and CTC, we were able to characterise failure and success type sessions based on the set of interaction features computed for them. According to the paired t-test carried out the CTC approach provided a significantly better AUC value (0.9828>0.9665), as well as more simple and stable explanation. CTC indicated that failure type sessions spent shorter times on average URLs with text dominance and longer times on average URLs with links dominance. In contrast, success type URLs were closely linked to long times in URLs with text dominance. Other alternative rules indicated that low proportions of such type of URLs are related with failure and high ones with success. This characterisation agreed with the one carried out with unsupervised classification where similar features were found to be decisive for session classification.

On the other hand, unsupervised learning algorithms applied to both session representations used, set of interaction features and sequence of URLs visited, produced partitions where half of the clusters had a high proportion (>74%) of one of the navigation types defined, thereby showing that the two perspectives give rise to automatically detect success and failure type sessions. However, when comparing both partitions with Jaccard index we concluded that both perspectives were not connected and thus, may be complementary. Later, the combination of the selected clusters with a majority of success or failure type sessions in three different voting systems, showed that the session representation using a set of interaction parameters was more effective than the one using the sequence of URLs visited to classify new sessions (accuracy of 78.2%). The combination of both representations could be complemented successfully, performing slightly better than the previous strategies. Finally, a failure detection subsystem was built based on the set of interaction parameters allowed defining a distance threshold to classify new sessions as failure type with high probability.

Thanks to both approaches, we have taken the first step to model the enrolment web information area of the UPV/EHU and considering the results, we can state that either supervised and unsupervised learning techniques are useful in that process, although the first one has shown to be more accurate.

8.4.2 Modelling the use of e-Services

In this contribution we analysed the e-Government adoption, the practical use of e-Services supplied by Governments, across Europe (for 26 EU countries) based on the empirical data provided by Eurostat 2004. The outcomes contribute to gaining insight on some of the factors influencing the e-Government adoption in Europe and can provide some guidelines to improve the interaction of citizens with web services and information offered by institutions depending on their e-Government use level.

The data provided information to quantify the adoption level and classify countries into four e-Government use levels (very high, high, low and very low). The characterisation of two countries from each level using supervised learning procedures, CTC trees, could differentiated individuals doing null e-Goverment use from those doing complete use of it with an average accuracy of 73% in the eight countries selected: Denmark and Norway, Ireland and Estonia, Latvia and Belgium and Poland and Romania (e-Government use level in descendent order).

In addition, Pearson correlation analysis revealed that European citizens who had bought goods quite recently over the Internet and those with high education levels did a complete use e-Government tools. These findings are in line with the ones aroused from the CTC structures of the majority of countries analysed.

Finally, for the 26 EU countries analysed, the comparison of the rankings provided by the index measuring the e-Government adoption we defined and other conceptual indexes measuring the level of e-readiness of a country showed that they all were correlated at 0.05 significance level. Therefore, we can state that the adoption levels extracted from the empirical analysis are in general aligned with more conceptual and theoretical values.

In summary, we think that our research results provide some key-aspects that could be considered for future strategic decisions on the improvement of e-Government adoption in different European countries, in terms of knowledge about the most influential factors on null and complete e-Government use and also in terms of proposing complementary indexes based on empirical data.

8.5 Further work

The four contributions presented in this dissertation leave room for different future lines of work.

In the context of clustering validation, further work could include aspects such as: testing the decision fusion approaches proposed over a large number of synthetic and real datasets, designing new voting strategies using other CVIs than those used herein, computing more precisely the impact of the experimental factors used to define the weights assigned to the CVIs in order to improve the results or designing voting approaches that are specialised in particular environments (noisy, overlapped, high dimensionality. . .).

Regarding the accessibility context, the main future line of action would be the implementation of suitable adaptations in the system proposed. Therefore, as both the device used and the problematic patterns are automatically discovered, the corresponding adaptation techniques should be activated accordingly. To this regard the system could be improved by comparing the interaction features of new navigations where adaptations have been activated with the ones computed in previous navigations where no adaptation was activated.

In the first analysis done in the medical informatics context, new experiments with remote users could be carried out as further work, where the gaze information of such participants will need to be gathered together with the interaction information in order to validate the relationships found between visual and interaction behaviours. In addition the visual information gathered in a video format could be used to enrich the analysis by detecting for example problems that users may have experienced while interacting with the dashboard. Regarding the second analysis, new experiments could be designed providing the same training to all participants so that other type of user profiles could be automatically detected and characterised. In addition, involving the designers of the medical dashboard in the future, could enable the design of the medical dashboard to be automatically adapted to the profiles extracted in order to enhance the user experience. Finally, more participants should be engaged in both analysis thereby the conclusions extracted can be generalised.

In the context of e-Services, regarding the analysis of the University of the Basque Country, we would like to do an in depth analysis of the complementary system, considering more complex voting criteria e.g using weighted votes. In addition, we will like to review the features selected to represent the sessions and find out if the use of a subset of these features leads to better results. Additionally, we wish to do an in-depth analysis of the models created in order to anticipate to the future users, and to identify and improve those elements having a negative influence on the usability. Regarding the analysis of e-Government adoption in Europe, our study could be extended by including new Eurostat (2016-2020) and UN (2018-2020) data and by involving other e-Government indicators suggested by various authors (Seri et al. 2014; Kabbar and Dell 2013; Jadi and Jie 2017). Finally, the research could be enriched by extending the geographical area of interest or focusing more closely on an smaller area.

8.6 Related publications

Throughout this dissertation different type of publications, international journal papers, international and national conference papers, book chapter and internal research reports, have been made in four different contexts. Below the related publications are summarised for each category and context.

- International journals:
  - Clustering validation (Yera et al. 2017a): Ainhoa Yera, Olatz Arbelaitz, Jose Luis Jodra, Ibai Gurrutxaga, Jose María Pérez, and Javier Muguerza. Analysis of several decision fusion strategies for clustering validation. Strategy definition, experiments and validation. *Pattern Recognition Letters*, 85, 42-48, 2017.
  - Accesibility (Yera et al. 2019b), submitted on November 2019: Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, Javier Muguerza, Juan Eduardo Pérez, and Xabier Valencia. Automatic web navigation problem detection based on client-side interaction data. *Data Mining and Knowledge Discovery.*
  - Medical informatics (Yera et al. 2019c): Yera, Ainhoa, Javier Muguerza, Olatz Arbelaitz, Iñigo Perona, Richard N. Keers, Darren M. Ashcroft, Richard Williams, Niels Peek, Caroline Jay, and Markel Vigo. Modelling the interactive behaviour of users with a medication safety dashboard in a primary care setting. *International journal of medical informatics*, 129, 395-403, 2019.

  - E-Services (Yera et al. 2019a), submitted on October 2019: Ainhoa Yera, Olatz Arbelaitz, Oier Jauregi, and Javier Muguerza. Characterisation of e-Government adoption in Europe. *PLOS ONE.*

- International Conferences:

  - Accesibility (Perona et al. 2019): Iñigo Perona, Ainhoa Yera, Olatz Arbelaitz, Javier Muguerza, Juan Eduardo Pérez, and Xabier Valencia. Towards automatic problem detection in Web navigation based on client-side interaction data. *In Proceedings of the XX International Conference on Human Computer Interaction*, 41:1-41:4, ACM, 2019.
  - Medical informatics (Yera et al. 2018a): Ainhoa Yera, Javier Muguerza, Olatz Arbelaitz, Iñigo Perona, Richard N. Keers, Darren M. Ashcroft, Richard Williams, Niels Peek, Caroline Jay, and Markel Vigo. Inferring Visual Behaviour from User Interaction Data on a Medical Dashboard. *In Proceedings of the 2018 International Conference on Digital Health*, 55-59. 2019.
  - E-Services (Yera et al. 2018c): Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, and Javier Muguerza. Modelling the enrolment eService of a university using machine learning techniques. *In Proceedings of the XVI International Conference e-Society*, 83-91, 2018.

- National Conferences:

  - Accesibility (Perona et al. 2016): Iñigo Perona, Ainhoa Yera, Olatz Arbelaitz, Javier Muguerza, Nikolaos Ragkousis, Myriam Arrue, Juan Eduardo Pérez, and Xabier Valencia. Automatic device detection in web interaction. *In Procedings of the XVII Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2016)*, 835-844, 2016.
  - Accesibility (Perona et al. 2017): Iñigo Perona, Ainhoa Yera, Olatz Arbelaitz, Javier Muguerza, Juan Eduardo Pérez, and Xabier Valencia. Web elkarrekintzan erabilitako gailuen detekzio automatikoa. *II. IKERGAZTE NAZIOARTEKO IKERKETA EUSKARAZ. Kongresuko artikulu-bilduma Ingeniaritza eta Arkitektura*, 22-29, 2017.
  - E-Services (Yera et al. 2017b): Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, and Javier Muguerza. UPV/EHUko eZerbitzu baten modelatzea ikasketa automatikoaren bidez*II. IKERGAZTE NAZIOARTEKO IKERKETA EUSKARAZ. Kongresuko artikulu-bilduma Ingeniaritza eta Arkitektura*, 111-118, 2017.
  - E-Services (Yera et al. 2018b): Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, and Javier Muguerza. Modeling the navigation on enrolment web information area of a university using machine learning techniques. *Advances in Artificial Intelligence. 18th Conference of the*

Spanish Association for Artificial Intelligence, CAEPIA 2018. Lecture Notes in Artificial Intelligence 11160 (LNAI 11160), 307-316, 2018.

- Book chapter:
  - Accesibility (Abascal et al. 2019): Julio Abascal, Xabier Gardeazabal, Juan Eduardo Pérez, Xabier Valencia, Olatz Arbelaitz, Javier Muguerza and Ainhoa Yera. Personalizing the user interface for people with disabilities. *Personalized Human-Computer Interaction*, part III, chapter 10, 254-282, 2019.
- Internal research reports:
  - E-Services (Yera et al. 2016a): Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, and Javier Muguerza. Análisis de la estructura, contenido y uso del sitio web de la Diputación Foral de Gipuzkoa - Gipuzkoako Foru Aldundiaren webgunearen egitura, eduki eta erabilera analisia. *Internal research report*, 2016.
  - E-Services (Yera et al. 2016b): Ainhoa Yera, Iñigo Perona, Olatz Arbelaitz, and Javier Muguerza. Análisis de la navegación en la web de la UPV/EHU - UPV/EHUko webgunearen nabigazioaren azterketa. *Internal research report*, 2017.

Bibliography

Abascal, J., O. Arbelaitz, M. Arrue, A. Lojo, J. Muguerza, J.E. Pérez, I. Perona, and X. Valencia (2013). "Enhancing Web Accessibility through User Modelling and Adaption Techniques". In: vol. 33. DOI: 10.3233/978-1-61499-304-9-427.

Abascal, J., O. Arbelaitz, X. Gardeazabal, J. Muguerza, J.E. Pérez, X. Valencia, and A. Yera (2019). "Personalizing the User Interface for People with Disabilities". In: *Personalized Human-Computer Interaction.* Ed. by Augstein Mirjam, Herder Eelco, and Wörndl Wolfgang, pp. 253–282. ISBN: 978-3-11-055248-5.

Afyonluoglu, M. and A.Z. Alkar (2017). "Comparison and Evaluation of International e-government Benchmarking Studies". In: *European Conference on e-Government.* Academic Conferences International Limited, pp. 283–293.

Aha, D.W., D. Kibler, and M.K. Albert (1991). "Instance-based learning algorithms". In: *Machine Learning* 6.1, pp. 37–66. ISSN: 1573-0565. DOI: 10.1007/BF00153759. URL: http://dx.doi.org/10.1007/BF00153759.

Akbarov, A., E. Kontopantelis, and Sperrin M. et al (2015). "Primary care medication safety surveillance with integrated primary and secondary care electronic health records: a cross-sectional study." In: *Drug Saf* 38, pp. 671–682.

Almanji, A., T.C. Davies, and N.S. Stott (2014). "Using cursor measures to investigate the effects of impairment severity on cursor control for youths with cerebral palsy". In: *International Journal of Human-Computer Studies* 72.3, pp. 349–357. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2013.12.003. URL: http://www.sciencedirect.com/science/article/pii/S1071581913001985.

Angulo, J. and M. Ortlieb (2015). ""WTH..!?!" Experiences, Reactions, and Expectations Related to Online Privacy Panic Situations". In: *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015).* Ottawa: USENIX Association, pp. 19–38. ISBN: 978-1-931971-249. URL: https://www.usenix.org/conference/soups2015/proceedings/presentation/angulo.

Arbelaitz, O., I. Gurrutxaga, J. Infante, J. Muguerza, and J.M. Pérez (2013a). "CTC: Competitive in an Analysis of Genetic based Algorithms for Rule Induction in Imbalanced Datasets". In: *Actas de la XV Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2013)*, pp. 19–28.

Arbelaitz, O., I. Gurrutxaga, J. Muguerza, J.M. Pérez, and I. Perona (2013b). "An extensive comparative study of cluster validity indices". In: *Pattern Recognition* 46.1, pp. 243–256. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2012.07.021. URL: http://www.sciencedirect.com/science/article/pii/S003132031200338X.

Arbelaitz, O., A. Lojo, J. Muguerza, and I. Perona (2016). "Web mining for navigation problem detection and diagnosis in Discapnet: A website aimed at disabled people". In: *Journal of the Association for Information Science and Technology* 67.8, pp. 1916–1927. DOI: 10.1002/asi.23506. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23506.

Arrue, M., X. Valencia, J.E. Pérez, L. Moreno, and J. Abascal (2018). "Inclusive Web Empirical Studies in Remote and In-Situ Settings: A User Evaluation of the RemoTest Platform". In: *International Journal of Human-Computer Interaction*, pp. 1–16. DOI: 10.1080/10447318.2018.1473941.

Augstein, M., T. Neumayr, W. Kurschl, D. Kern, T. Burger, and J. Altmann (2017). "A Personalized Interaction Approach: Motivation and Use Case". In: *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization.* ACM, pp. 221–226.

Avery, T. et al. (2012). "A pharmacist-led information technology intervention for medication errors (PINCER): A multicentre, cluster randomised, controlled trial and cost-effectiveness analysis". In: *Lancet* 379, pp. 1310–1319. DOI: 10.1016/S0140-6736(11)61817-5.

Baker, F.B. and L.J. Hubert (1975). "Measuring the Power of Hierarchical Cluster Analysis". English. In: *Journal of the American Statistical Association* 70.349, pp. 31–38. ISSN: 01621459.

Bandyopadhyay, S. and S. Saha (2008). "A Point Symmetry-Based Clustering Technique for Automatic Evolution of Clusters". In: *IEEE Transactions on Knowledge and Data Engineering* 20 (Issue: 11), pp. 1441–1457. DOI: 10.1109/TKDE.2008.79.

Barioni, M.C.N., H.L. Razente, A.J.M. Traina, and C. Traina (2008). "Accelerating k-medoid-based algorithms through metric access methods". In: *Journal of Systems and Software* 81.3. Selected Papers from the 2006 Brazilian Symposia on Databases and on Software Engineering, pp. 343–355. ISSN: 0164-1212. DOI: https://doi.org/10.1016/j.jss.2007.06.019. URL: http://www.sciencedirect.com/science/article/pii/S0164121207001768.

Bélanger, F. and L. Carter (2008). "Trust and risk in e-government adoption". In: *The Journal of Strategic Information Systems* 17.2. eGovernment Strategies: ICT innovation in international public sector contexts, pp. 165–176. ISSN: 0963-8687. DOI: https://doi.org/10.1016/j.jsis.2007.12.002. URL: http://www.sciencedirect.com/science/article/pii/S0963868707000637.

Bezdek, J.C., W.Q. Li, Y. Attikiouzel, and M. Windham (1997). "A geometric approach to cluster validity for normal mixtures". In: *Soft Computing* 1.4, pp. 166–179. ISSN: 1432-7643. DOI: 10.1007/s005000050019. URL: https://doi.org/10.1007/s005000050019.

Bezdek, J.C. and N.R. Pal (1998). "Some new indexes of cluster validity". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.3, pp. 301–315. ISSN: 1083-4419. DOI: 10.1109/3477.678624.

Bilbao-Osorio, B., S. Dutta, and B. Lanvin (2013). "The global information technology report 2013". In: *World Economic Forum*, pp. 1–383. URL: http://www3.weforum.org/docs/WEF_GITR_Report_2013.pdf.

Bilbao-Osorio, S. B. Dutta, and B. Lanvin (2014). "The global information technology report 2014". In: *World Economic Forum*, pp. 1–343. URL: http://www3.weforum.org/docs/WEF_GlobalInformationTechnology_Report_2014.pdf.

Brehaut, J.C., H.L. Colquhoun, K.W. Eva, K. Carroll, A. Sales, S. Michie, N.M. Ivers, and J.M. Grimshaw (2016). "Practice Feedback Interventions: 15 Suggestions for Optimizing Effectiveness." In: *Ann Inter Med* 164 6, pp. 435–441.

Breiman, L. (1996). "Bagging Predictors". In: *Machine Learning* 24.2, pp. 123–140. ISSN: 1573-0565. DOI: 10.1023/A:1018054314350. URL: https://doi.org/10.1023/A:1018054314350.

Brown, B., P. Balatsoukas, R. Williams, M. Sperrin, and I. Buchan (2016). "Interface design recommendations for computerised clinical audit and feedback: Hybrid usability evidence from a research-led system". In: *Int J Med Inform* 94, pp. 191–206. DOI: 10.1016/j.ijmedinf.2016.07.010.

Caliński, T. and J. Harabasz (1974). "A dendrite method for cluster analysis". In: *Communications in Statistics-Simulation and Computation* 3.1, pp. 1–27.

Carter, L. and F. Bélanger (2005). "The utilization of e-government services: citizen trust, innovation and acceptance factors*". In: *Information Systems Journal* 15.1, pp. 5–25. DOI: 10.1111/j.1365-2575.2005.00183.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2575.2005.00183.x.

Card, S.K., A. Newell, and T.P. Moran (1983). *The Psychology of Human-Computer Interaction.* Hillsdale, NJ, USA: L. Erlbaum Associates Inc. ISBN: 0898592437.

Chen, M.C., J.R. Anderson, and M.H Sohn (2001). "What Can a Mouse Cursor Tell Us More?: Correlation of Eye/Mouse Movements on Web Browsing". In: *CHI '01 Extended Abstracts on Human Factors in Computing Systems.* CHI EA '01. Seattle, Washington: ACM, pp. 281–282. ISBN: 1-58113-340-5. DOI: 10.1145/634067.634234. URL: http://doi.acm.org/10.1145/634067.634234.

Chin, J., W.T. Fu, and T. Kannampallil (2009). "Adaptive information search: Age-dependent interactions between cognitive profiles and strategies". In: pp. 1683–1692. DOI: 10.1145/1518701.1518961.

Chou, C.H., M.C. Su, and E. Lai (2004). "A new cluster validity measure and its application to image compression". In: *Pattern Analysis and Applications* 7.2, pp. 205–220. ISSN: 1433-7541. DOI: 10.1007/s10044-004-0218-1. URL: http://dx.doi.org/10.1007/s10044-004-0218-1.

Christian, M.S., A.S. Garza, and J. Slaughter (2011). “Work Engagement: A Quantitative Review and Test of Its Relations with Task and Contextual Performance”. In: *Pers Psychol* 64, pp. 89–136. DOI: 10.1111/j.1744-6570.2010.01203.x.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J. : L. Erlbaum Associates. ISBN: 0805802835. DOI: https://doi.org/10.4324/9780203771587. URL: http://www.zentralblatt-math.org/zmath/en/search/?an=0747.62110.

Cortes, C. and V. Vapnik (1995). “Support-Vector Networks”. In: *Machine Learning* 20.3, pp. 273–297. ISSN: 1573-0565. DOI: 10.1023/A:1022627411411. URL: https://doi.org/10.1023/A:1022627411411.

Dagliati, A. et al. (2018). “A dashboard-based system for supporting diabetes care”. In: *J Am Med Inform Assoc* 25 5, pp. 538–547.

Davies, D.L and D.W. Bouldin (1979). “A Cluster Separation Measure”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 224–227. ISSN: 0162-8828. DOI: 10.1109/TPAMI.1979.4766909.

Day, W.H.E. and H. Edelsbrunner (1984). “Efficient algorithms for agglomerative hierarchical clustering methods”. In: *Journal of Classification* 1.1, pp. 7–24. ISSN: 1432-1343. DOI: 10.1007/BF01890115. URL: https://doi.org/10.1007/BF01890115.

Demšar, J. (2006). “Statistical Comparisons of Classifiers over Multiple Data Sets”. In: *J. Mach. Learn. Res.* 7, pp. 1–30. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1248547.1248548.

Dillon, A. (2001). “Beyond usability: process, outcome and affect in human computer interactions”. In: *Canadian Journal of Information Science* 26.4, pp. 57–69. URL: http://hdl.handle.net/10150/106391.

Dimitriadou, E., S. Dolničar, and A. Weingessel (2002). “An examination of indexes for determining the number of clusters in binary data sets”. In: *Psychometrika* 67.1, pp. 137–159. ISSN: 1860-0980. DOI: 10.1007/BF02294713. URL: https://doi.org/10.1007/BF02294713.

Dowding, D., R. Randell, P. Gardner, G. Fitzpatrick, P. Dykes, J. Favela, S. Hamer, Z. Whitewood-Moores, N. Hardiker, E. Borycki, and L. Currie (2015). “Dashboards for improving patient care: Review of the literature”. In: *Int J Med Inform* 84.2, pp. 87–100. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2014.10.001. URL: http://www.sciencedirect.com/science/article/pii/S1386505614001890.

Dowding, D., J.A. Merrill, N. Onorato, Y. Barrón, R. Rosati, and D. Russell (2017). “The impact of home care nurses’ numeracy and graph literacy on comprehension of visual display information: Implications for dashboard design”. In: *J Am Med Inform Assoc* 25, pp. 175–182. DOI: 10.1093/jamia/ocx042.

Dunn, J.C. (1973). “A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters”. In: *Journal of Cybernetics* 3.3, pp. 32–57. DOI: 10.1080/01969727308546046. eprint: http://dx.doi.org/10.1080/01969727308546046. URL: http://dx.doi.org/10.1080/01969727308546046.

Dutta, S. and I. Mia (2010). "The global information technology report 2009–2010". In: *World Economic Forum and INSEAD, SRO-Kundig Geneva, Switzerland*, pp. 1–415. URL: https://www.itu.int/net/wsis/implementation/2010/forum/geneva/docs/publications/GITR%5C%202009-2010_Full_Report_final.pdf.

Dutta, S. and I. Mia (2011). "The global information technology report 2010–2011". In: *World Economic Forum and INSEAD, SRO-Kundig Geneva, Switzerland*, pp. 1–411. URL: http://reports.weforum.org/wp-content/pdf/gitr-2011/wef-gitr-2010-2011.pdf.

Dutta, S. and B. Bilbao-Osorio (2012). "The global information technology report 2012". In: *World Economic Forum and INSEAD, SRO-Kundig Geneva, Switzerland*, pp. 1–413. URL: http://www3.weforum.org/docs/Global_IT_Report_2012.pdf.

Dutta, S., T. Geiger, and B. Lanvin (2015). "The global information technology report 2015". In: *World Economic Forum*. Vol. 1. 1. Citeseer, pp. 1–357. URL: http://www3.weforum.org/docs/WEF_Global_IT_Report_2015.pdf.

Ehmke, C. and S. Wilson (2007). "Identifying Web Usability Problems from Eye-tracking Data". In: *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...But Not As We Know It - Volume 1*. BCS-HCI '07. University of Lancaster, United Kingdom: British Computer Society, pp. 119–128. ISBN: 978-1-902505-94-7. URL: http://dl.acm.org/citation.cfm?id=1531294.1531311.

European Parliament and Council of the European Union (2004). *"Regulation (EC) No 808/2004 of the European Parliament and of the Council"*.

Eurostat (2004). *Individuals using the internet for interaction with public authorities, Dataset code: tin00012*. URL: http://ec.europa.eu/eurostat/web/products-datasets/-/tin00012.

European Commission (2018). *The Digital Economy and Society Index (DESI)*. https:/ec.europa.eu/digital-single-market/desi.

Faghfouri, A. and M. Frish (2011). "Robust discrimination of human footsteps using seismic signals". In: *Proc SPIE*. DOI: 10.1117/12.882726.

Frank, A. and A. Asuncion (2010). *UCI Machine Learning Repository*. URL: http://archive.ics.uci.edu/ml/.

Freund, Y. and R.E. Schapire (1996). *Experiments with a New Boosting Algorithm*. Bari, Italy. URL: http://dl.acm.org/citation.cfm?id=3091696.3091715.

Friedman, M. (1937). "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". In: *Journal of the American Statistical Association* 32.200, pp. 675–701. DOI: 10.1080/01621459.1937.10503522. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1937.10503522. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522.

García, S. and F. Herrera (2008). "An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons". In: *Journal of Machine Learning Research* 9.Dec, pp. 2677–2694.

García, S., J. Luengo, and F. Herrera (2015). *Data Preprocessing in Data Mining*. Springer International Publishing Switzerland.

González, R., J. Gasco, and J. Llopis (2007). "E-government success: some principles from a Spanish case study". In: *Industrial Management & Data Systems* 107.6, pp. 845–861.

Gosset, W.S. (1908). "The probable error of a mean". In: *Biometrika* 6, pp. 1–26.

Grossman, T. and R. Balakrishnan (2005). "The Bubble Cursor: Enhancing Target Acquisition by Dynamic Resizing of the Cursor's Activation Area". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '05. Portland, Oregon, USA: ACM, pp. 281–290. ISBN: 1-58113-998-5. DOI: 10.1145/1054972.1055012. URL: http://doi.acm.org/10.1145/1054972.1055012.

Guo, Q. and E. Agichtein (2010). "Towards Predicting Web Searcher Gaze Position from Mouse Movements". In: *CHI '10 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '10. Atlanta, Georgia, USA: ACM, pp. 3601–3606. ISBN: 978-1-60558-930-5. DOI: 10.1145/1753846.1754025. URL: http://doi.acm.org/10.1145/1753846.1754025.

Gurrutxaga, I., I. Albisua, O. Arbelaitz, J.I Martín, J. Muguerza, J.M Pérez, and I. Perona (2010). "SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index". In: *Pattern Recognition* 43.10, pp. 3364–3373. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2010.04.021. URL: http://www.sciencedirect.com/science/article/pii/S0031320310001974.

Gurrutxaga, I., J. Muguerza, O. Arbelaitz, J.M Pérez, and J.I Martín (2011). "Towards a standard methodology to evaluate internal cluster validity indices". In: *Pattern Recognition Letters* 32.3, pp. 505–515. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2010.11.006. URL: http://www.sciencedirect.com/science/article/pii/S0167865510003636.

Halkidi, M. and M. Vazirgiannis (2001). "Clustering validity assessment: finding the optimal partitioning of a data set". In: *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pp. 187–194. DOI: 10.1109/ICDM.2001.989517.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten (2009). "The WEKA Data Mining Software: An Update". In: *SIGKDD Explor. Newsl.* 11.1, pp. 10–18. ISSN: 1931-0145. DOI: 10.1145/1656274.1656278. URL: http://doi.acm.org/10.1145/1656274.1656278.

Hall, M.A. (1998). "Correlation-based Feature Subset Selection for Machine Learning". PhD thesis. Hamilton, New Zealand: University of Waikato.

Han, J., J. Pei, and M. Kamber (2011). *Data mining: concepts and techniques*. Elsevier.

Hastie, T., R. Tibshirani, and J.H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer. ISBN: 9780387848846. URL: https://books.google.es/books?id=eBSgoAEACAAJ.

Heer, J. and E.H. Chi (2002). “Separating the Swarm: Categorization Methods for User Sessions on the Web”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’02. Minneapolis, Minnesota, USA: ACM, pp. 243–250. ISBN: 1-58113-453-3. DOI: 10.1145/503376.503420. URL: http://doi.acm.org/10.1145/503376.503420.

Holm, S. (1979). “A simple sequentially rejective multiple test procedure”. In: *Scandinavian Journal of Statistics* 6, pp. 65–70.

Huang, J., R. White, and G. Buscher (2012). “User See, User Point: Gaze and Cursor Alignment in Web Search”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’12. Austin, Texas, USA: ACM, pp. 1341–1350. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2208591. URL: http://doi.acm.org/10.1145/2207676.2208591.

Hubert, L.J. and J.R. Levin (1976). “A general statistical framework for assessing categorical clustering in free recall”. In: *Psychological Bulletin* 83, pp. 1072–1080.

Hubert, L. and P. Arabie (1985). “Comparing partitions”. In: *Journal of Classification* 2.1, pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075. URL: https://doi.org/10.1007/BF01908075.

Hurst, A., S.E. Hudson, J. Mankoff, and S. Trewin (2008). “Automatically Detecting Pointing Performance”. In: *Proceedings of the 13th International Conference on Intelligent User Interfaces*. IUI ’08. Gran Canaria, Spain: ACM, pp. 11–19. ISBN: 978-1-59593-987-6. DOI: 10.1145/1378773.1378776. URL: http://doi.acm.org/10.1145/1378773.1378776.

ISO (2018). “9241-11: Usability: definitions and concepts”. In: *Ergonomics of human-system interaction*.

ISO (2019). “9241-210: Human-centred design for interactive systems”. In: *Ergonomics of human-system interaction*.

Jaccard, P. (1908). “Nouvelles recherches sur la distribution florale”. In: *Bull. Soc. Vaud. Sci. Nat.* 44, pp. 223–270.

Jadi, Y. and L. Jie (2017). “An Efficiency Measurement of E-Government Performance for United Nation Ranking Index”. In: *World Academy of Science, Engineering and Technology, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering* 11.1, pp. 279–282. ISSN: eISSN:1307-6892. URL: http://waset.org/Publications?p=121.

Jain, A.K. and R.C. Dubes (1988). *Algorithms for Clustering Data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN: 0-13-022278-X.

Jain, A.K., M.N. Murty, and P.J. Flynn (1999). “Data Clustering: A Review”. In: *ACM Comput. Surv.* 31.3, pp. 264–323. ISSN: 0360-0300. DOI: 10.1145/331499.331504. URL: http://doi.acm.org/10.1145/331499.331504.

Jeffries, M., D. Phipps, R.L. Howard, A. Avery, S. Rodgers, and D. Ashcroft (2017). “Understanding the implementation and adoption of an information technology intervention to support medicine optimisation in primary care: qualitative study using strong structuration theory”. In: *BMJ Open* 7.5. ISSN: 2044-6055. DOI: 10.1136/bmjopen-2016-014810. eprint: https://bmjopen.bmj.com/content/7/5/e014810.full.pdf. URL: https://bmjopen.bmj.com/content/7/5/e014810.

Jeffries, M., R.N. Keers, D.L. Phipps, R. Williams, B. Brown, A.J. Avery, N. Peek, and D.M. Ashcroft (2018). "Developing a learning health system: Insights from a qualitative process evaluation of a pharmacist-led electronic audit and feedback intervention to improve medication safety in primary care". In: *PLOS ONE* 13.10, pp. 1–16. DOI: 10.1371/journal.pone.0205419. URL: https://doi.org/10.1371/journal.pone.0205419.

Jeffries, M., W.T. Gude, and R.N. Keers (2019). *Understanding the utilisation of a novel interactive electronic medication safety dashboard by pharmacists and clinicians in general practice: a mixed methods study (Paper under review).*

John, G.H. and P. Langley (1995). "Estimating Continuous Distributions in Bayesian Classifiers". In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. UAI'95. Canada: Morgan Kaufmann Publishers Inc., pp. 338–345. ISBN: 1-55860-385-9. URL: http://dl.acm.org/citation.cfm?id=2074158.2074196.

Kabbar, E. and P. Dell (2013). *"Weaknesses of the E-Government Development Index"*. Ed. by Shiro Uesugi. Vienna: Springer Vienna, pp. 111–124. ISBN: 978-3-7091-1425-4. URL: https://doi.org/10.1007/978-3-7091-1425-4%5C_7.

Kalra, D., I. Buchan, and N. Paton (2016). *Three Gurus of Big Data.* Ed. by The Translational Scientist. https://thetranslationalscientist.com/issues/0816/three-gurus-of-big-data/. Accessed: 2019-10-01.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction To Cluster Analysis.* Vol. 344. John Wiley & Sons. ISBN: 0-471-87876-6. DOI: 10.2307/2532178.

Keers, R.N., R. Williams, C. Davies, N. Peek, and D.M. Ashcroft (2015). *Improving medication safety in primary care: developing a stakeholder-centred electronic prescribing safety indicator dashboard.* DOI: https://doi.org/10.1002/pds.3812.

Kendall, M.G. (1938). "A new measure of rank correlation". In: *Biometrika* 30.1-2, pp. 81–93. ISSN: 0006-3444. DOI: 10.1093/biomet/30.1-2.81. eprint: http://oup.prod.sis.lan/biomet/article-pdf/30/1-2/81/423380/30-1-2-81.pdf. URL: https://doi.org/10.1093/biomet/30.1-2.81.

Kim, M. and R.S. Ramakrishna (2005). "New indices for cluster validity assessment". In: *Pattern Recognition Letters* 26.15, pp. 2353–2363. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.04.007. URL: http://www.sciencedirect.com/science/article/pii/S016786550500125X.

Kohavi, R. and G.H John (1997). "Wrappers for feature subset selection". In: *Artificial Intelligence* 97.1. Relevance, pp. 273–324. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(97)00043-X. URL: http://www.sciencedirect.com/science/article/pii/S000437029700043X.

Kohavi, Ron and Foster Provost (1998). "Glossary of Terms". In: *Machine Learning - Special issue on applications of machine learning and the knowledge discovery process* 30.2-3, pp. 271–274. ISSN: 0885-6125.

K.R., Žalik and B. Žalik (2011). "Validity index for clusters of different sizes and densities". In: *Pattern Recognition Letters* 32.2, pp. 221–234. ISSN:

0167-8655. DOI: http://dx.doi.org/10.1016/j.patrec.2010.08.007. URL: http://www.sciencedirect.com/science/article/pii/S0167865510002928.

Kryszczuk, K. and P. Hurley (2010). "Estimation of the Number of Clusters Using Multiple Clustering Validity Indices". In: *Multiple Classifier Systems*. Ed. by Neamat El Gayar, Josef Kittler, and Fabio Roli. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 114–123. ISBN: 978-3-642-12127-2.

Kuhn, M. and K. Johnson (2013). *Applied predictive modeling.* URL: http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/.

Kushniruk, A.W. and V.L. Patel (2004). "Cognitive and usability engineering methods for the evaluation of clinical information systems". In: *J Biomed Inform* 37.1, pp. 56–76. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2004.01.003. URL: http://www.sciencedirect.com/science/article/pii/S1532046404000206.

Lago-Fernández, L.F. and F. Corbacho (2010). "Normality-based validation for crisp clustering". In: *Pattern Recognition* 43.3, pp. 782–795. ISSN: 0031-3203. DOI: http://dx.doi.org/10.1016/j.patcog.2009.09.018. URL: http://www.sciencedirect.com/science/article/pii/S0031320309003628.

Landis-Lewis, Z., J. C Brehaut, H. Hochheiser, G. Douglas, and R. S Jacobson (2015). "Computer-supported feedback message tailoring: Theory-informed adaptation of clinical audit and feedback for learning and behavior change". In: *Implement Sci* 10, p. 12. DOI: 10.1186/s13012-014-0203-z.

Laschinger, H., P. Wilk, J. Cho, and P. Greco (2009). "Empowerment, engagement and perceived effectiveness in nursing work environments: Does experience matter?" In: *J Nurs Manag* 17, pp. 636–646. DOI: 10.1111/j.1365-2834.2008.00907.x.

Layne, K. and J. Lee (2001). "Developing fully functional E-government: A four stage model". In: *Government Information Quarterly* 18.2, pp. 122–136. ISSN: 0740-624X. URL: https://doi.org/10.1016/S0740-624X(01)00066-1.

Levenshtein, V.I (1966). "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10, p. 707.

Li, Z., G. Wang, and G. He (2017). "Milling tool wear state recognition based on partitioning around medoids (PAM) clustering". In: *The International Journal of Advanced Manufacturing Technology* 88. DOI: 10.1007/s00170-016-8848-1.

Liu, B. (2006). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer-Verlag New York, Inc.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. ISSN: 0018-9448. DOI: 10.1109/TIT.1982.1056489.

Maulik, U. and S. Bandyopadhyay (2002). "Performance Evaluation of Some Clustering Algorithms and Validity Indices". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.12, pp. 1650–1654. ISSN: 0162-

8828. DOI: 10.1109/TPAMI.2002.1114856. URL: http://dx.doi.org/10.1109/TPAMI.2002.1114856.

Meilă, M. (2003). “Comparing Clusterings by the Variation of Information”. In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 173–187. ISBN: 978-3-540-45167-9.

Middleton, B., M. Bloomrosen, M.A. Dente, B. Hashmat, R. Koppel, J.M. Overhage, T.H. Payne, S.T. Rosenbloom, C. Weaver, and J. Zhang (2013). “Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA”. In: *J Am Med Inform Assoc* 20.e1, e2–e8.

Milligan, G.W. and M.C. Cooper (1985). “An examination of procedures for determining the number of clusters in a data set”. In: *Psychometrika* 50.2, pp. 159–179. ISSN: 1860-0980. DOI: 10.1007/BF02294245. URL: https://doi.org/10.1007/BF02294245.

Nam, T. (2014). “Determining the type of e-government use”. In: *Government Information Quarterly* 31.2, pp. 211–220. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2013.09.006. URL: http://www.sciencedirect.com/science/article/pii/S0740624X14000483.

Pal, N.R. and J. Biswas (1997). “Cluster validation using graph theoretic concepts”. In: *Pattern Recognition* 30.6, pp. 847–857. ISSN: 0031-3203. DOI: http://dx.doi.org/10.1016/S0031-3203(96)00127-6. URL: http://www.sciencedirect.com/science/article/pii/S0031320396001276.

Park, J., S.H. Han, and H. Yang (2006). “Evaluation of cursor capturing functions in a target positioning task”. In: *International Journal of Industrial Ergonomics* 36.8, pp. 721–730. ISSN: 0169-8141. DOI: https://doi.org/10.1016/j.ergon.2006.05.004. URL: http://www.sciencedirect.com/science/article/pii/S0169814106001041.

Pérez, J.M., J. Muguerza, O. Arbelaitz, I. Gurrutxaga, and J.I. Martín (2007). “Combining multiple class distribution modified subsamples in a single tree”. In: *Pattern Recognition Letters* 28.4, pp. 414–422. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2006.08.013.

Pérez, J.M., I. Albisua, O. Arbelaitz, I. Gurrutxaga, J.I. Martín, J. Muguerza, and I. Perona (2010). “Consolidated trees versus bagging when explanation is required”. In: *Computing* 89.3, pp. 113–145. ISSN: 1436-5057. DOI: 10.1007/s00607-010-0094-z. URL: https://doi.org/10.1007/s00607-010-0094-z.

Perona, I., A. Yera, O. Arbelaitz, J. Muguerza, N. Ragkousis, M. Arrue, J.E Pérez, and X. Valencia (2016). “Automatic device detection in web interaction”. In: *VIII Workshop on Theory and Applications of Data Mining (TAMIDA 2016); In Procedings of the XVII Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2016).* Salamanca (Spain), pp. 835–844.

Pérez, J.E., X. Valencia, M. Arrue, and J. Abascal (2016). “A Usability Evaluation of Two Virtual Aids to Enhance Cursor Accessibility for People with Motor Impairments”. In: *Proceedings of the 13th Web for All Conference.*

W4A '16. Montreal, Canada: ACM, 20:1–20:4. ISBN: 978-1-4503-4138-7. DOI: 10.1145/2899475.2899489. URL: http://doi.acm.org/10.1145/2899475.2899489.

Perona, I., A. Yera, O. Arbelaitz, J. Muguerza, J.E. Pérez, and X. Valencia (2017). "Web elkarrekintzan erabilitako gailuen detekzio automatikoa". In: *II. Ikergazte Nazioarteko Ikerketa Euskaraz. Kongresuko artikulu-bilduma Ingeniaritza eta Arkitektura*. Ed. by Iñaki Alegria, Ainhoa Latatu, Miren Josu Omaetxebarria, and Patxi Salaberri. Ingeniaritza eta arkitektura. Udako Euskal Unibertsitatea (UEU). Iruñea, Euskal Herria, pp. 22–29.

Perona, I., A. Yera, O. Arbelaitz, J. Muguerza, J.E. Pérez, and X. Valencia (2019). "Towards Automatic Problem Detection in Web Navigation Based on Client-side Interaction Data". In: *Proceedings of the XX International Conference on Human Computer Interaction*. Interacción '19. Donostia, Gipuzkoa, Spain: ACM, 41:1–41:4. ISBN: 9781450371766. DOI: 10.1145/3335595.3335642. URL: http://doi.acm.org/10.1145/3335595.3335642.

Platt, J. (1998). "Fast Training of Support Vector Machines using Sequential Minimal Optimization". In: *Advances in Kernel Methods - Support Vector Learning*. Ed. by B. Schoelkopf, C. Burges, and A. Smola. MIT Press. URL: http://research.microsoft.com/%5C~jplatt/smo.html.

Platt, J.C. (1999). "Advances in Kernel Methods". In: ed. by Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola. Cambridge, MA, USA: MIT Press. Chap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. ISBN: 0-262-19416-3.

Preece, J., Y. Rogers, and H. Sharp (2001). *Beyond Interaction Design: Beyond Human-Computer Interaction*. New York, NY, USA: John Wiley & Sons, Inc. ISBN: 0471402494.

Proakis, J.G. and D.G. Manolakis (1992). *Digital Signal Processing (2Nd Ed.): Principles, Algorithms, and Applications*. Indianapolis, IN, USA: Macmillan Publishing Co., Inc. ISBN: 0-02-396815-X.

Quinlan, J.R. (1986). "Induction of Decision Trees". In: *Mach. Learn.* 1.1, pp. 81–106. ISSN: 0885-6125. DOI: 10.1023/A:1022643204877. URL: http://dx.doi.org/10.1023/A:1022643204877.

Quinlan, J.R (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 1558602402.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ratwani, R.M., R.J. Fairbanks, A.Z. Hettinger, and N.C. Benda (2015). "Electronic health record usability: analysis of the user-centered design processes of eleven electronic health record vendors". In: *J Am Med Inform Assoc* 22.6, pp. 1179–1182. ISSN: 1527-974X. DOI: 10.1093/jamia/ocv050. eprint: http://oup.prod.sis.lan/jamia/article-pdf/22/6/1179/6956965/ocv050.pdf. URL: https://doi.org/10.1093/jamia/ocv050.

Rédei, G.P. (2008). "UPGMA (unweighted pair group method with arithmetic means)". In: *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*. Dordrecht: Springer Netherlands, pp. 2068–2068. ISBN: 978-1-4020-6754-

9. DOI: 10.1007/978-1-4020-6754-9_17806. URL: https://doi.org/10.1007/978-1-4020-6754-9_17806.

Rich, B.L., J.A. Lepine, and E. Crawford (2010). "Job Engagement: Antecedents and Effects on Job Performance". In: *Acad Manage J* 53, pp. 617–635. DOI: 10.5465/AMJ.2010.51468988.

Rousseeuw, P.J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1". In: ed. by David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group. Cambridge, MA, USA: MIT Press. Chap. Learning Internal Representations by Error Propagation, pp. 318–362. ISBN: 0-262-68053-X. URL: http://dl.acm.org/citation.cfm?id=104279.104293.

Sadler, S., S. Rodgers, R. Howard, C. Morris, and T. Avery (2014). "Training pharmacists to deliver a complex information technology intervention (PINCER) using the principles of educational outreach and root cause analysis". In: *Int J Pharm Pract* 22, pp. 47–58. DOI: 10.1111/ijpp.12032.

Saha, S. and S. Bandyopadhyay (2009). "Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* 39 (Issue: 4), pp. 420–425. DOI: 10.1109/TSMCC.2009.2013335.

Saitta, S., B. Raphael, and I.F.C. Smith (2007a). "A Bounded Index for Cluster Validity". In: *Machine Learning and Data Mining in Pattern Recognition.* Ed. by Petra Perner. Vol. 4571. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 174–187. ISBN: 978-3-540-73498-7. DOI: 10.1007/978-3-540-73499-4_14. URL: http://dx.doi.org/10.1007/978-3-540-73499-4_14.

Saitta, S., R Raphael B., and I.F.C. Smith (2007b). "A Bounded Index for Cluster Validity". In: *Machine Learning and Data Mining in Pattern Recognition.* Ed. by Petra Perner. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 174–187. ISBN: 978-3-540-73499-4.

Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." In: *Molecular biology and evolution* 4.4, pp. 406–425. DOI: 10.1093/oxfordjournals.molbev.a040454.

Santana, V. de and M.C. Baranauskas (2010). "Summarizing Observational Client-side Data to Reveal Web Usage Patterns". In: *Proceedings of the 2010 ACM Symposium on Applied Computing.* SAC '10. Sierre, Switzerland: ACM, pp. 1219–1223. ISBN: 978-1-60558-639-7. DOI: 10.1145/1774088.1774344. URL: http://doi.acm.org/10.1145/1774088.1774344.

Santana, V. de and M.C. Baranauskas (2015). "WELFIT: A remote evaluation tool for identifying Web usage patterns through client-side logging". In: *International Journal of Human-Computer Studies* 76, pp. 40–49. ISSN: 1071-

5819. DOI: https://doi.org/10.1016/j.ijhcs.2014.12.005. URL: http://www.sciencedirect.com/science/article/pii/S1071581914001682.

Schwester, R. (2009). "Examining the Barriers to e-Government Adoption". In: *Electronic Journal of e-Government* 7.1, pp. 113–122. ISSN: 1479-436-9X. URL: www.ejeg.com.

Schapire, R.E. (1990). "The strength of weak learnability". In: *Machine Learning* 5.2, pp. 197–227. ISSN: 1573-0565. DOI: 10.1007/BF00116037. URL: https://doi.org/10.1007/BF00116037.

Schapire, R.E. (1999). "A Brief Introduction to Boosting". In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'99. Stockholm, Sweden: Morgan Kaufmann Publishers Inc., pp. 1401–1406. URL: http://dl.acm.org/citation.cfm?id=1624312.1624417.

Seri, P., A. Bianchi, and P. Matteucci (2014). "Diffusion and usage of public e-services in Europe: An assessment of country level indicators and drivers". In: *Telecommunications Policy* 38.5. Special issue on : Selected papers from the 10th Conference in Telecommunications, Media and Internet Techno-economics Special issue on : The development of public e-services: Empirical analysis and policy issues., pp. 496–513. ISSN: 0308-5961. DOI: https://doi.org/10.1016/j.telpol.2014.03.004.

Shareef, M.A., V. Kumar, U. Kumar, and Y.K. Dwivedi (2011). "e-Government Adoption Model (GAM): Differing service maturity levels". In: *Government Information Quarterly* 28.1, pp. 17–35. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2010.05.006. URL: http://www.sciencedirect.com/science/article/pii/S0740624X10000985.

Shannon, C.E. (1948). "A Mathematical Theory of Communication". In: *Bell System Technical Journal* 27.3, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

Sheng, W., S. Swift, L. Zhang, and X. Liu (2005). "A Weighted Sum Validity Function for Clustering with a Hybrid Niching Genetic Algorithm". In: *Trans. Sys. Man Cyber. Part B* 35.6, pp. 1156–1167. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2005.850173. URL: https://doi.org/10.1109/TSMCB.2005.850173.

Shneiderman, B. (1996). "The eyes have it: a task by data type taxonomy for information visualizations". In: *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343. DOI: 10.1109/VL.1996.545307.

Shneiderman, B. (1997). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* 3rd. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc. ISBN: 0201694972.

Simpao, A.F., L.M. Ahumada, and B.R. et al Desai (2014). "Optimization of drug–drug interaction alert rules in a pediatric hospital's electronic health record system using a visual analytics dashboard". In: *J Am Med Inform Assoc* 22, pp. 361–369.

Sneath, P.H.A. and R.R. Sokal (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* Freeman.

Stocks, S.J., E. Kontopantelis, A. Akbarov, S. Rodgers, A.J. Avery, and D.M. Ashcroft (2015). “Examining variations in prescribing safety in UK general practice: cross sectional study using the Clinical Practice Research Datalink”. In: *BMJ* 351. DOI: 10.1136/bmj.h5501. eprint: https://www.bmj.com/content/351/bmj.h5501.full.pdf. URL: https://www.bmj.com/content/351/bmj.h5501.

Thompson, D.V., R.T. Rust, and J. Rhoda (2005). “The business value of e-government for small firms”. In: *International Journal of Service Industry Management* 16.4, pp. 385–407. ISSN: 0956-4233. DOI: https://doi.org/10.1108/09564230510614022 6.

Tinholt, D., N. Van der Linden, M. Ehrismann, G. Cattaneo, S. Aguzzi, L. Jacquet, S. Vanmarcke, G. Noci, M. Benedetti, and G. Marchio (2015). *Future-proofing eGovernment for the Digital Single Market. Final insight report, June 2015 - Study.* DOI: 10.2759/32843. URL: https://ec.europa.eu/futurium/en/system/files/ged/egovernmentbenchmarkinsightreport.pdf.

Tinholt, D., Van der Linden N., S. Enzerink, R. Geilleit, A. Groeneveld, G. Cattaneo, S. Aguzzi, F. Pallaro, G. Noci, M. Benedetti, G. Marchio, and A. Salvadori (2018). *eGovernment Benchmark 2018: Securing eGovernment for all.* DOI: 10.2759/371003. URL: https://publications.europa.eu/en/publication-detail/-/publication/82749e75-f389-11e8-9982-01aa75ed71a1/language-en.

Trewin, S., S. Keates, and K. Moffatt (2006). “Developing Steady Clicks:: A Method of Cursor Assistance for People with Motor Impairments”. In: *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility.* Assets ’06. Portland, Oregon, USA: ACM, pp. 26–33. ISBN: 1-59593-290-9. DOI: 10.1145/1168987.1168993. URL: http://doi.acm.org/10.1145/1168987.1168993.

UN (2010). *United Nations E-Government Survey 2010: Leveraging e-government at a time of financial and economic crisis.* URL: https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2010-Survey/Complete-survey.pdf.

UN (2012). *United Nations E-Government Survey 2012: E-Government for the People.* URL: https://publicadministration.un.org/egovkb/Portals/egovkb/Documents/un/2012-Survey/Complete-Survey.pdf.

UN (2014). *United Nations E-Government Survey 2014: E-Government for the future we want.* URL: https://doi.org/10.18356/73688f37-en.

UN (2016). *United Nations E-Government Survey 2016:E-Government in Support of Sustainable Development.* DOI: https://doi.org/10.18356/d719b252-en.

UN (2018). *United Nations E-Government Survey 2018: Gearing E-Government to Support Transformation Towards Sustainable and Resilient Societies.* DOI: https://doi.org/10.18356/d54b9179-en.

Valencia, X., J.E. Pérez, U. Muñoz, M. Arrue, and J. Abascal (2015). “Human-Computer Interaction – INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part

I". In: ed. by Julio Abascal, Simone Barbosa, Mirko Fetter, Tom Gross, Philippe Palanque, and Marco Winckler. Springer International Publishing. Chap. Assisted Interaction Data Analysis of Web-Based User Studies, pp. 1–19.

Ward, J.H. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244. URL: http://www.jstor.org/stable/2282967.

Williams, R., R. Keers, W. Gude, M. Jeffries, C. Davies, B. Brown, E. Kontopantelis, J.A. Avery, M.D. Ashcroft, and N. Peek (2018). "SMASH! The Salford medication safety dashboard". In: *J Innov Health Inform* 25, pp. 183–193. DOI: 10.14236/jhi.v25i3.1015.

Wilcoxon, F. (1945). "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6, pp. 80–83. ISSN: 00994987. URL: http://www.jstor.org/stable/3001968.

Witten, I.H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems).* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 0120884070.

Witten, I.H., E. Frank, M.A. Hall, and C.J. Pal (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques.* 4th. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 0128042915.

Wobbrock, J.O. and K.Z. Gajos (2008). "Goal Crossing with Mice and Trackballs for People with Motor Impairments: Performance, Submovements, and Design Directions". In: *ACM Trans. Access. Comput.* 1.1, 4:1–4:37. ISSN: 1936-7228. DOI: 10.1145/1361203.1361207. URL: http://doi.acm.org/10.1145/1361203.1361207.

Wu, X. et al. (2008). "Top 10 algorithms in data mining". In: *Knowledge and Information Systems* 14.1, pp. 1–37. ISSN: 0219-3116. DOI: 10.1007/s10115-007-0114-2. URL: https://doi.org/10.1007/s10115-007-0114-2.

Xu, R. and D. Wunsch (2008). *Clustering.* Vol. 10. John Wiley & Sons.

Yang, Q., N. Banovic, and J. Zimmerman (2018). "Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* CHI '18. Montreal QC, Canada: ACM, 130:1–130:11. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173704. URL: http://doi.acm.org/10.1145/3173574.3173704.

Yera, A., O. Arbelaitz, J. Muguerza, and I. Perona (2016a). *Análisis de la estructura, contenido y uso del sitio web de la Diputación Foral de Gipuzkoa.* unpublished EHU-KAT-IK-NN-16. University of the Basque Country UPV/EHU.

Yera, A., O. Arbelaitz, J. Muguerza, and I. Perona (2016b). *Análisis de la navegación en la web de la UPV/EHU.* unpublished EHU-KAT-IK-01-17. University of the Basque Country UPV/EHU.

Yera, A., O. Arbelaitz, J.L. Jodra, I. Gurrutxaga, J.M. Pérez, and J Muguerza (2017a). "Analysis of several decision fusion strategies for clustering validation. Strategy definition, experiments and validation". In: *Pattern Recogni-*

tion Letters 85, pp. 42–48. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2016.11.009. URL: http://www.sciencedirect.com/science/article/pii/S0167865516303324.

Yera, A., I. Perona, O. Arbelaitz, J. Muguerza, J.E. Pérez, and X. Valencia (2017b). "UPV/EHUko eZerbitzu baten modelatzea ikasketa automatikoaren bidez". In: *II. Ikergazte Nazioarteko Ikerketa Euskaraz. Kongresuko artikulu-bilduma Ingeniaritza eta Arkitektura.* Ed. by Iñaki Alegria, Ainhoa Latatu, Miren Josu Omaetxebarria, and Patxi Salaberri. Ingeniaritza eta arkitektura. Udako Euskal Unibertsitatea (UEU). Iruñea, Euskal Herria, pp. 111–118.

Yera, A., J. Muguerza, O. Arbelaitz, I. Perona, R. Keers, D. Ashcroft, R. Williams, N. Peek, C. Jay, and M. Vigo (2018a). "Inferring Visual Behaviour from User Interaction Data on a Medical Dashboard". In: *Proceedings of the 2018 International Conference on Digital Health.* DH '18. Lyon, France: ACM, pp. 55–59. ISBN: 9781450364935. DOI: 10.1145/3194658.3194676. URL: http://doi.acm.org/10.1145/3194658.3194676.

Yera, A., I. Perona, O. Arbelaitz, and J. Muguerza (2018b). "Modeling the Navigation on Enrolment Web Information Area of a University Using Machine Learning Techniques". In: *Advances in Artificial Intelligence.* Ed. by Francisco Herrera, Sergio Damas, Rosana Montes, Sergio Alonso, Óscar Cordón, Antonio González, and Alicia Troncoso. Cham: Springer International Publishing, pp. 307–316. ISBN: 9783030003746.

Yera, A., I. Perona, O. Arbelaitz, and J. Muguerza (2018c). "Modelling the enrolment eService of a university using machine learning techniques". In: *Proceedings of the 2018 International Conference e-Society 2018 (ES'18).* Ed. by Piet Kommers and Pedro Isaías. Lisbon, Portugal, pp. 83–91. ISBN: 9789898533753. URL: http://www.iadisportal.org/digital-library/modelling-the-enrolment-eservice-of-a-university-using-machine-learning-techniques.

Yera, A., O. Arbelaitz, O. Jauregi, and J. Muguerza (2019a). "Automatic web navigation problem detection based on client-side interaction data (paper submitted)". In: *Data Mining and Knowledge Discovery.*

Yera, A., O. Arbelaitz, O. Jauregi, and J. Muguerza (2019b). "Characterization of e-Government adoption in Europe (paper submitted)". In: *PLOS ONE.*

Yera, A., J. Muguerza, O. Arbelaitz, I. Perona, R.N Keers, Ashcroft D.M., R. Williams, N. Peek, C. Jay, and M. Vigo (2019c). "Modelling the interactive behaviour of users with a medication safety dashboard in a primary care setting". In: *International Journal of Medical Informatics* 129, pp. 395–403. ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2019.07.014. URL: http://www.sciencedirect.com/science/article/pii/S1386505619301662.

Zahabi, M., D.B. Kaber, and M. Swangnetr (2015). "Usability and Safety in Electronic Medical Records Interface Design: A Review of Recent Literature and Guideline Formulation". In: *Hum Factors* 57.5. PMID: 25850118, pp. 805–834. DOI: 10.1177/0018720815576827. eprint: https://doi.

org/10.1177/0018720815576827. URL: https://doi.org/10.1177/0018720815576827.

Zhang, J. and M.F. Walji (2011). “TURF: Toward a unified framework of EHR usability”. In: *Journal of Biomedical Informatics* 44.6, pp. 1056–1067. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2011.08.005. URL: http://www.sciencedirect.com/science/article/pii/S1532046411001328.

Zhu, Y., L. Zhou, C. Xie, G.J. Wang, and T. Nguyen (2019). “Forecasting SMEs’ credit risk in supply chain finance with an enhanced hybrid ensemble machine learning approach”. In: *International Journal of Production Economics* 211. DOI: 10.1016/j.ijpe.2019.01.032.