UNIVERSITAT POLITÈCNICA DE CATALUNYA

Doctoral programme:

AUTOMATIC CONTROL, ROBOTICS AND COMPUTER VISION

Doctoral Dissertation

A CONTRIBUTION TO THE RANKING AND DESCRIPTION OF CLASSIFICATIONS

Germán Sánchez-Hernández

Co-Advisors:
Juan Carlos Aguado Chao
Núria Agell Jané

June 2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Juan Carlos Aguado Chao
(Principal Co-Advisor)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Núria Agell Jané
(Principal Co-Advisor)

Approved for the University Committee on Graduate Studies:

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola de Doctorat

Curso académico: 2012-2013

Acta de calificación de tesis doctoral

Nombre y apellidos
Germán Sánchez Hernández

DNI / NIE / Pasaporte
53075559T

Programa de doctorado
Automática, Robótica y Visión (ARV)

Unidad estructural responsable del programa
Instituto de Organización y Control de Sistemas Industriales (IOC), Departamento de Ingeniería de Sistemas, Automática e Informática Industrial (ESAII)

Resolución del Tribunal

Reunido el Tribunal designado a tal efecto, el doctorando / la doctoranda expone el tema de la su tesis doctoral titulada _ **A contribution to the ranking and description of classifications** _____

_____.

Acabada la lectura y después de dar respuesta a las cuestiones formuladas por los miembros titulares del tribunal, éste otorga la calificación:

☐ APTA/O ☐ NO APTA/O

(Nombre, apellidos y firma) Presidente/a	(Nombre, apellidos y firma) Secretario/a	
(Nombre, apellidos y firma) Vocal	(Nombre, apellidos y firma) Vocal	(Nombre, apellidos y firma) Vocal

_____, _____ de _____ de _____

El resultado del escrutinio de los votos emitidos por los miembros titulares del tribunal, efectuado por la Escuela de Doctorado, a instancia de la Comisión de Doctorado de la UPC, otorga la MENCIÓN CUM LAUDE:

☐ SÍ ☐ NO

(Nombre, apellidos y firma) Presidenta de la Comisión de Doctorado	(Nombre, apellidos y firma) Secretaria de la Comisión de Doctorado

Barcelona a _____ de _____ de _____

Acknowledgments/Agradecimientos

Agradecer a las empresas e instituciones académicos...

Fundació ESADE, AIS, UPC

This work is partially supported by the SENSORIAL Research Project (TIN2010-20966-C02-01, 02), funded by the Spanish Ministry of Science and Information Technology. This research work has been partially conducted during a three-months visiting period to the Centre for Computational Intelligence (CCI) at De Montfort University in Leicester (UK). The participation and interest of the firm Textil Seu SA in the SENSORIAL project is also gratefully acknowledged.

I have many people to thank for their support during the preparation and writing of this dissertation. As with most of them I communicate in Spanish (or Catalan), that is the language I have used in these acknowledgments.

Este doctorado ha supuesto un recorrido a lo largo de un camino desafiante que no habría sido capaz de recorrer sin la ayuda de mi familia, amigos y colegas. Entre ellos, el agradecimiento más especial es para mis padres Germán e Inés per haberme educado con rectitud y amor. No solo me han apoyado incondicionalmente estos últimos años de doctorado, sino que lo han hecho desde mis primeros garabatos a lápiz. Esperemos que tengan razón en eso de "ya llegarán los frutos". También tengo que agradecer a mis hermanos Darío y Almudena por el ejemplo a seguir que me han proporcionado desde bien pequeño.

A mis co-directores, Juan Carlos Aguado y Núria Agell, por su comprensión, entusiasmo y dedicación en la dirección de este trabajo. Buena parte de esta tesis les corresponde, especialmente a Núria, que me ha llevado de la mano como si de una madre se tratara desde aquel día que bajó al CPD de ESADE preguntando por alguien con ganas de trabajar con ella. ¡Gracias, Núria!

A los miembros del grupo GREC por toda la ayuda ofrecida en mis momentos de ofuscación.

Cecilio, Francisco, Mònica, Mónica, Quico, Xari... aún sigo aprendiendo de vosotros (¡y lo que me queda!). A la gente del JARCA y CCIA, por hacerme disfrutar de los congresos tanto a nivel científico como social. Y a Paco Chiclana, por su inmejorable acogida durante mi estancia en Leicester y su excelencia a la hora de trabajar.

A mis colegas, amigos y compañeros de penas y alegrías Albert Samà, Ricardo Campos y Josean Sanabria. El primero está también a punto de leer la tesis, mientras que los otros dos lo hicieron hace no muchos meses. Ha sido un placer trabajar con vosotros.

En general, a todos los amigos con los que he compartido mi vida en estos últimos años. A los Esade Crew, Pichotes, colegas de AIS, gente del CPD, FIBers, badalonins, hattrickeros... al fin puedo responder a vuestra reiterativa pregunta: "*Muy bien pero... ¿cuándo lees la tesis?*". Mención especial para Julio, por entenderme como nadie y por estar siempre ahí, sin importar la distancia física.

Y finalmente, a mi mujer Verónica (no, no estamos casados pero sí, es mi mujer) por su comprensión y apoyo no solo durante estos últimos años sino desde la primera época de exámenes en la FIB. Gran parte de la motivación para acabar esta tesis es debida a las ganas de darle una alegría, así que todo esto va por ti. Porque pronto seremos tres. Te quiero, bonita. ¡Gracias, Vero!

Abstract

This thesis presents a novel and complete fuzzy multi-criteria decision making (MCDM) methodology. This methodology is specifically designed for selecting classifications in the framework of unsupervised learning systems. The main results obtained are twofold. On the one hand, the definition of fuzzy criteria to be used to assess the suitability of a set of given classifications and, on the other hand, the design and development of a natural language generation (NLG) system to qualitatively describe them.

Unsupervised learning systems often produce a large number of possible classifications. In order to select the most suitable one, a set of criteria is usually defined and applied sequentially to assess and filter the obtained classifications. This is done, in general, by using a true-false decision in the application of each criterion. This approach could result in classifications being discarded and not taken into account when they marginally fail to meet one particular criterion even though they meet other criteria with a high score. An alternative solution to this sequential approach has been introduced in this thesis. It consists of evaluating the degree up to which each fuzzy criterion is met by each classification and, only after this, aggregating for each classification the individual assessments. This overall value reflects the degree up to which the set of criteria is globally satisfied by each classification.

Five fuzzy criteria are defined and analysed to be used collectively to evaluate classifications. The corresponding single evaluations are then proposed to be aggregated into a collective one by means of an Ordered Weighted Averaging (OWA) operator guided by a fuzzy linguistic quantifier, which is used to implement the concept of fuzzy majority in the selection process. In addition, a NLG system to qualitatively describe the most important characteristics of the best classification is designed and developed in order to fully understand the chosen classification. Finally, this new

methodology is applied to a real business problem in a marketing context. The main purpose of this application is to show how the proposed methodology can help marketing experts in the design of specific-oriented marketing strategies by means of an automatic and interpretable segmentation system.

Resumen

En esta tesis se presenta una novedosa y completa metodología difusa y multicriterio (MCDM). Esta metodología está específicamente diseñada para la selección de clasificaciones en el marco de los sistemas de aprendizaje no supervisado. Los principales resultados obtenidos son de dos tipos. Por un lado, la definición de criterios difusos que se utilizan para evaluar la idoneidad de un conjunto de clasificaciones dadas y, por otro lado, el diseño y desarrollo de un sistema de generación de lenguaje natural (NLG) para describirlos cualitativamente.

Los sistemas de aprendizaje no supervisado producen a menudo un gran número de posibles clasificaciones. Con el fin de seleccionar la más adecuada, se suele definir y aplicar de forma secuencial un conjunto de criterios para evaluar y filtrar las clasificaciones obtenidas. Esto se hace, en general, mediante el uso de una decisión de verdadero o falso en la aplicación de cada criterio. Este enfoque podría dar lugar al descarte de clasificaciones cuando marginalmente no cumplen con algún criterio particular, a pesar de que cumplen con otros criterios incluso con una puntuación más alta. En esta tesis se introduce una solución alternativa a este enfoque secuencial. Esta alternativa consiste en evaluar el grado hasta el cual cada criterio difuso es cumplido por cada clasificación, y sólo después de esto, agregar para cada clasificación las evaluaciones individuales. Este valor general refleja el grado hasta el cual el conjunto de criterios es globalmente satisfecho por cada clasificación.

Se definen y analizan cinco criterios difusos para ser usados de forma colectiva para evaluar clasificaciones. Se propone a continuación la agregación de las correspondientes evaluaciones individuales en una colectiva por medio de un operador OWA guiado por un cuantificador lingüístico difuso, que se utiliza para poner en práctica el concepto de mayoría difusa en el proceso de selección. Además, un sistema NLG es diseñado y desarrollado para describir cualitativamente las características más

importantes de la mejor clasificación con el fin de entender plenamente la clasificación elegida. Por último, esta nueva metodología es aplicada a un problema empresarial real en un contexto de marketing. El propósito principal de esta aplicación es mostrar cómo la metodología propuesta puede ayudar a los expertos de marketing en el diseño de estrategias de marketing específicas y orientadas por medio de un sistema de segmentación automática e interpretable.

Resum

En aquesta tesi es presenta una nova i completa metodologia difusa i multicriteri (MCDM). Aquesta metodologia està especificament disenyada per a la selecció de classificacions en el marc dels sistemes d'aprenentatge no supervisat. Els principals resultats obtinguts són de dos tipus. D'una banda, la definició de criteris difusos que s'utilitzen per avaluar la idoneïtat d'un conjunt de classificacions donades i, d'altra banda, el disseny i desenvolupament d'un sistema de generació de llenguatge natural (NLG) per descriure'ls qualitativament.

Els sistemes d'aprenentatge no supervisat produeixen sovint un gran nombre de possibles classificacions. Amb la finalitat de seleccionar la més adequada, se sol definir i aplicar de forma seqüencial un conjunt de criteris per avaluar i filtrar les classificacions obtingudes. Això es fa, en general, mitjançant l'ús d'una decisió de cert o fals en l'aplicació de cada criteri. Aquest enfocament podria donar lloc al descartament de classificacions quan marginalment no compleixen amb algun criteri particular, a pesar que compleixen amb altres criteris fins i tot amb una puntuació més alta. En aquesta tesi s'introdueix una solució alternativa a aquest enfocament seqüencial. Aquesta alternativa consisteix en avaluar el grau fins al qual cada criteri difús és complert per cada classificació, i només després d'això, agregar per a cada classificació les avaluacions individuals. Aquest valor general reflecteix el grau fins al qual el conjunt de criteris és globalment satisfet per cada classificació.

Es defineixen i analitzen cinc criteris difusos per ser usats de forma col·lectiva per avaluar classificacions. Es proposa a continuació l'agregació de les corresponents avaluacions individuals en una de col·lectiva per mitjà d'un operador OWA guiat per un quantificador lingüístic difús, que s'utilitza per posar en pràctica el concepte de majoria difusa en el procés de selecció. A més, un sistema de NLG és dissenyat i desenvolupat per descriure qualitativament les característiques més importants de la millor classificació amb la finalitat d'entendre plenament la classificació triada.

Finalment, aquesta nova metodologia és aplicada a un problema empresarial real en un context de màrqueting. El propòsit principal d'aquesta aplicació és mostrar com la metodologia proposada pot ajudar als experts de màrqueting en el disseny d'estratègies de màrqueting específiques i orientades per mitjà d'un sistema de segmentació automàtica i interpretable.

Contents

1 Introduction **1**

1.1 Motivation and framework 2

1.2 Objectives 3

1.3 Theoretical background 3

1.4 Structure of the doctoral thesis 5

1.5 Publications derived from this thesis 6

2 Literature review **9**

2.1 Criteria for selecting classifications 10

2.2 Aggregation functions based on OWA operators 14

2.3 Data-to-text systems for generating natural language 18

3 Fuzzy criteria for selecting classifications **25**

3.1 First criterion: useful number of classes 27

3.2 Second criterion: balanced classes 30

3.3 Third criterion: coherent classification 38

3.4 Fourth criterion: dependency on external variables 43

3.5 Fifth criterion: accuracy of the predictive model 45

3.6 Conclusions 46

4 Natural language-based automatic qualitative description of clusters **49**

4.1 Signal analysis 51

4.2 Data interpretation 64

4.3 Document planning 70
4.4 Microplanning and realisation 85
4.5 Conclusions 92

5 Application to market segmentation **95**
5.1 Dataset 96
5.2 Obtaining segmentations 99
5.3 Ranking and selecting segmentations 101
5.4 Qualitative class description 106
5.5 Discussion and managerial implications 128
5.6 Conclusions 132

6 Limitations and future research **133**

A Learning Algorithm for Multivariate Data Analysis (LAMDA) **137**

B Tables of the case study **141**

List of Tables

2.1 Clustering validation criteria (Part I) . . . 12
2.2 Clustering validation criteria (Part II) . . . 13
2.3 Preference results in Reiter et al. (2005) where statisticallly significant results are in **bold**. . . . 20
2.4 Data-to-text systems . . . 23

3.1 Marginal Adequacy Degrees (MADs) of individuals in classification $\mathcal{C}_1$ . . . 42
3.2 MADs of individuals in classification $\mathcal{C}_2$ . . . 43
3.3 Differences between maximum and minimum MAD . . . 43
3.4 Contingency table . . . 44

4.1 Contingency table of the example . . . 55
4.2 Expected values in the example . . . 56
4.3 Example: variables independent on class but dependent when combining them . . . 56
4.4 Example: combination of independent variables becoming dependent on class . . . 57
4.5 Example: variables independent on class and also independent when combining them 57
4.6 Example: combination of independent variables remaining independent on class . . . 58
4.7 Values of importance in the example . . . 59
4.8 Conditional and joint distributions in the example . . . 62
4.9 Summary of type-A rules, showing which attributes are checked by each rule . . . 64
4.10 Summary of the activation of type-B rules, showing which attributes are checked by each rule . . . 71
4.11 Example of messages ordered according to the defined criteria and the resulting groups 88

4.12 Compatibility between type-B rules 89
4.13 Summary of type-B rules 94

5.1 Description of variables 98
5.2 Number of segmentations obtained by hybrid connective 102
5.3 Distribution of variable *PromosSensit* 104
5.4 Extract of the best segmentations using fuzzy selection criteria OWA methodology . 105
5.5 Frequencies of each obtained interval in the discretisation process 106
5.6 Chosen modalities for the discretised variables 107
5.7 Contingency table of variable *Competition* 107
5.8 Expected frequencies of variable *Competition* 109
5.9 values of importance (VoIs) associated with variable *Competition* 109
5.10 Cut points obtained for each variable and number of relevant VoIs detected 110
5.11 Conditional frequencies of variable *Competition* 111
5.12 Percentiles of conditional frequencies to detect extreme frequencies (EFs) 112
5.13 Number of obtained extreme frequencies 112
5.14 Initial messages detected for variable *Competition*. The rest of messages are included in Table B.10 113
5.15 Groups of messages affected by rule A.1 114
5.16 Groups of messages of variable *Competition* affected by rule A.2. The rest of groups are shown in Table B.12 115
5.17 Messages of variable *PromosSensit* affected by rule A.3. The rest of groups are shown in Table B.13 115
5.18 Groups of messages affected by rule A.5 117
5.19 Frequency of each final weight after applying type-A rules 117
5.20 Groups of messages affected by rule B.1 118
5.21 Groups of messages affected by rule B.2 (first step, same sign). 119
5.22 Groups of messages of variable *Competition* affected by rule B.2 (second step, different sign). The rest of groups are detailed in Table B.18 119
5.23 Messages affected by rule B.4 120
5.24 Messages affected by rule B.5 120

5.25 Messages affected by rule C.1 . . . 121
5.26 Messages affected by rule C.2 . . . 121
5.27 Variables needing a special transcription (rule C.3) . . . 122
5.28 Variables with modalities as adjectives (rule C.4) . . . 122
5.29 Number of messages implied in the analysis of each rule . . . 123
5.30 Final planning of the text . . . 124
5.31 Features of messages of group #4 . . . 125
5.32 Best segmentations according to I_B . . . 129
5.33 Best segmentations according to I_C . . . 129
5.34 Best segmentations according to I_D . . . 130
5.35 Best segmentations according to I_A . . . 130

B.1 Best 100 segmentations . . . 141
B.2 Contingency tables of the selected variables (part I) . . . 145
B.3 Contingency tables of the selected variables (part II) . . . 146
B.4 Expected frequencies of the selected variables (part I) . . . 147
B.5 Expected frequencies of the selected variables (part II) . . . 148
B.6 Values of importance associated with the selected variables (part I) . . . 149
B.7 Values of importance associated with the selected variables (part II) . . . 150
B.8 Conditional frequencies of the selected variables (part I) . . . 151
B.9 Conditional frequencies of the selected variables (part II) . . . 152
B.10 Initial messages of the selected variables . . . 153
B.11 Groups of messages affected by rule A.1 . . . 156
B.12 Groups of messages affected by rule A.2 . . . 157
B.13 Groups of messages affected by rule A.3 . . . 158
B.14 Groups of messages affected by rule A.5 . . . 159
B.15 Filtered messages that will be mentioned in the final text . . . 160
B.16 Groups of messages affected by rule B.1 . . . 162
B.17 Groups of messages affected by rule B.2 (first step: same sign). . . . 162
B.18 Groups of messages affected by rule B.2 (second step: different sign) . . . 163
B.19 Messages affected by rule B.4 . . . 164

B.20 Messages affected by rule B.5 164
B.21 Messages affected by rule C.1 164
B.22 Messages affected by rule C.2 164
B.23 Variables needing a special transcription (rule C.3) 165
B.24 Variables with modalities as adjectives (rule C.4) 165

List of Figures

2.1 Some Regular Increasing Monotone (RIM) functions and their corresponding obtained weights 18

3.1 Examples of usefulness degree functions with $K_1 = 4$ and $K_2 = 7$ 29
3.2 Fuzzy concept 'Useful number of classes'with $K_1 = 3$ and $K_2 = 5$ 30
3.3 Function f and its maximum coefficients of variation when $N = 260$ 36
3.4 LAMDA classical function with different values of ρ 39
3.5 Gaussian function with value $s = 0.1$ and different values of ρ 40
3.6 Waissman function with different values of the centre c 41

4.1 Diagram of the designed NLG system. 52
4.2 Diagram of the Signal analysis stage. 53
4.3 The three clusters found in the VoIs of the example. 61

5.1 Summary of the considered variables in the dataset 99
5.2 Histogram of the quantitative descriptors and their chosen density functions 101
5.3 Histogram of the variable "number of classes" for the obtained segmentations 102
5.4 Membership function used of modelling the fuzzy concept 'Useful number of classes'with $K_1 = 3$ and $K_2 = 5$ 103
5.5 Distribution of the modalities within the classes, in addition with their $p-value$ and the decision result of the selection process 108
5.6 Histogram of the conditional frequencies of the dependent variables 111

A.1 LAMDA: hybrid connectives-based classification 139

List of Acronyms

AI Artificial Intelligence . . . 50

B2B business to business . . . 133

BNF Backus-Naur Form . . . 85

BUM Basic Unit-interval Monotone . . . 16

CAIM class-attribute interdependence maximisation . . . 106

EF extreme frequency . . . 106

GAD Global Adequacy Degree . . . 138

IEM Information Entropy Maximisation . . . 54

LAMDA Learning Algorithm for Multivariate Data Analysis . . . 137

MAD Marginal Adequacy Degree . . . 138

MCDM multi-criteria decision making . . . 133

MDLP Minimum Description Length . . . 53

ML machine learning . . . 49

NIC Non-Informative Class . . . 139

NLG natural language generation . . . 133

OWA Ordered Weighted Averaging . . . 134

RBF Radial Base Functions . . . 139

RIM Regular Increasing Monotone . . . 105

RDM Regular Decreasing Monotone . . . 16

SVM Support Vector Machine 104

VoI value of importance 106

Chapter 1

Introduction

This document corresponds to the doctoral thesis carried out by Germán Sánchez Hernández. This thesis corresponds to the doctoral program of Automatic Control, Robotics and Computer Vision (ARV) of the Automatic Control Department (ESAII) and the Institute of Industrial and Control Engineering (IOC), belonging to the Universitat Politècnica de Catalunya – BarcelonaTech (UPC)[1].

This thesis has been developed at ESADE Business School. ESADE is an international academic institution with over fifty years of history. ESADE head quarters are based in Barcelona and its main activities are education, research and social debating in the area of management, economy and business. The research lines of ESADE are mainly focused on entrepreneurship, innovation, leadership and governance, management, skills and knowledge, business social responsibility, etc. ESADE is affiliated with the Ramon Llull University (URL), a private, non profit-making university providing a public service.

The studies conducted in this thesis have been carried out at the Research Group on Knowledge Engineering[2] (GREC). The GREC group was set up in 1994 and is recognised as a consolidated research group (2005 SGR 00943, 2009 SGR 855) by the Government of Catalonia. It is an inter-university group bringing together researchers from the BarcelonaTech and ESADE. Right from the outset, the multidisciplinary facet of the group has allowed it to work on both basic and applied research. The GREC's main activity focuses on research and development of techniques in the area of Artificial Intelligence (AI). The GREC Research Group at ESADE has two broad goals: (1)

[1]ESAII: http://esaii.upc.edu; IOC: http://ioc.upc.edu; UPC: http://www.upc.edu

[2]ESADE: http://www.esade.edu; GREC: http://esade.edu/research-webs/grec; URL: http://www.url.edu

development of AI methodologies in non-structured environments (incomplete information, qualitative and fuzzy); (2) the application of these methodologies to fields bearing on decision-making, finances and marketing.

1.1 Motivation and framework

Intelligent systems for decision support making are especially essential for companies or institutions that adopt strategies based on the use of information and operate in highly complex contexts. These systems must help in understanding and managing the large amount of information available from customers, products, competitors, and in assessing and analysing the alternatives in a explanatory and easily interactive way for the user. The growing interest in the last decades in automatic decision support systems lies in their ability to both synthesising information and obtaining easily interpretable results.

The use of unsupervised learning systems allows the behaviour of certain phenomena to be understood without relying on expert knowledge or information from past situations. Such systems usually offer several ways of segmenting the considered individuals. The selection of the best classification can be faced as a multi-criteria decision making (MCDM) approach. The definition of several criteria to evaluate several alternatives is done in this thesis by aggregating the assessments of each classification in terms of an Ordered Weighted Averaging (OWA) operator. Finally, some tools must be designed in order to interpret the chosen classification, as a qualitative description of the main characteristics of each considered class.

The final scope of this thesis was decided after a three-months visit period to the Centre for Computational Intelligence (CCI) at De Montfort University (DMU), in Leicester, United Kingdom. After having developed projects in which the selection of the best classification and its corresponding description were done in a manual way, the interest of the CCI in the research of operators for aggregating information led to the application of a MCDM approach for ranking classifications. Further research induced the development of a natural language generation (NLG) system to better understand the best classification by qualitatively describing the considered classes.

1.2 Objectives

The main objective of this thesis is to study and develop a complete MCDM system to create, select an understand the best classification on a set of individuals and according a set of criteria. This general objective can be divided into the following specific ones:

1. Generation of several classifications of a set of individuals by the application of unsupervised learning techniques.
2. Evaluation of classifications by analysing and designing a set of fuzzy criteria. Development of an index measuring the degree up to which each criterion is verified by each considered classification.
3. Rank of classifications by aggregating the assessments of each classification in terms of an OWA operator, and selection of the best classification.
4. Interpretation of a classification by designing an developing a NLG system that qualitatively describes the main characteristic of each class.
5. Application of this complete approach into a real marketing problem.

The main contributions of this thesis are related to the second and fourth objectives: the design of a set of fuzzy criteria to assess classifications (Chapter 3) and the development of a system to obtain a natural language-based description of a certain classification (Chapter 4).

1.3 Theoretical background

The use of machine learning (ML) tools within MCDM systems is especially interesting in environment where the available large volume of data has become a negative aspect when analysing alternatives and obtaining useful information. In general, either classification systems (supervised learning) or clustering systems (unsupervised learning) are used. The former are useful in situations in which the patterns have a label indicating their class, obtained from the behaviour, opinion or knowledge of an expert that is tried to be imitated when he is missing. The latter are convenient in scenarios where that label is not available. Frequently, both types of learning are used in combination. Situations can be found where, even being a classification available, existing classes are not separable. In these cases it is necessary to obtain a prior partition of initial classes compat-

ible enough with the original classification in order to group together the more similar examples, increasing the homogeneity of the resulting new classes.

This thesis is situated in the context of unsupervised learning algorithms. The origins of such learning systems date back to the end of the last century. Seminal studies conducted by Ackley et al. (1985) and Barlow (1989) were based on Boltzmann machines. These learning systems imported many of the concepts from multivariate statistics: either those based on the density estimation methods (Grenander and Miller, 1994) or those based on distances between patterns. For this reason, it is considered that there are two groups of unsupervised learning models (Barlow, 1989). On the one hand, models based on the estimation of pattern distribution or density functions and, on the other hand, models based on distances between patterns or between patterns and classes. Models based on connectionist approaches can be found in both directions.

From the application point of view, unsupervised learning systems have been considered in the literature as systems capable to capture knowledge from complex structures (Duda et al., 2001; Figueiredo and Jain, 2002; Jain, 2010). Such methods have been applied in a wide range of domains, among which it is worth mentioning: text categorisation, images recognition, telecommunications fraud detection, stock price forecasting, bioinformatics, fault diagnosis, pollution classification and clinical or socioeconomic systems (Barrón-Adame et al., 2012; Chen et al., 2005; Constantinos and Paris, 2008; Elati and Rouveirol, 2011; Ferraretti et al., 2012; Goldsmith, 2001; Hadavandi et al., 2010; Lee and Yang, 2009; Niebles et al., 2008; Oliver et al., 2011; Yang et al., 2011). In the marketing field, finding new and creative solutions is valuable because these allow for the definition of new strategies and innovation. The use of unsupervised learning algorithms allows us to suggest segmentations that are, in principle, not trivial. In this sense, behavioural patterns of 'interesting' profiles could be established by using this type of algorithms and these may reveal new customer profiles not yet known to experts (Chiu et al., 2009; Hong and Kim, 2012; Mo et al., 2010; Yao et al., 2010; Lu et al., 2012).

Unsupervised learning systems produce in most cases a large number of possible classifications. The development of suitable tools or models for selecting classifications is an important topic of research in such area (Broder et al., 2008; Kukar, 2003; Osei-Bryson, 2010). It is usual to define a set of criteria in order to select the most suitable classification, and to apply them sequentially to the considered classifications (Choi et al., 2005; Osei-Bryson, 2010; Sánchez-Hernández et al., 2007; Sánchez Almeida et al., 2010). This approach discards all those classifications failing to

meet a particular criterion but it is important to note that it could result in classifications being prematurely discarded because they fail to meet one particular criterion but meet the other with a high grade.

An alternative to this sequential approach is that of evaluating the degree up to which each criterion is met by each classification. This can be done by modelling each criterion by means of a membership function, thus associating an index with each criterion. Once each criterion is evaluated on a classification, an overall aggregated value must be obtained for reflecting the degree up to which the set of criteria is satisfied by the classification.

Currently, at least 90 different families of aggregation operators have been studied (Chiclana et al., 2004, 2007; Dubois and Prade, 1985; Fodor and Roubens, 1994; Herrera et al., 2003; Klir and Folger, 1988; Torra, 1997; Torra and Narukawa, 2007; Xu and Da, 2003; Yager, 1988; Zhou et al., 2008). Among them the OWA operator proposed by Yager (1988) is the most widely used. One of the main reasons to support this extensive use is that the OWA operator allows the implementation of the concept of fuzzy majority in the aggregation phase by means of a fuzzy linguistic quantifier (Zadeh, 1983) which indicates the proportion of satisfied criteria 'necessary for a good solution' (Yager, 1996). This is done by using the linguistic quantifier in the computation of the weights associated to the OWA operator. The objective of the aggregation step is to combine a set of criteria in such a way that the final aggregation output takes all the single criterion into account (Dubois and Prade, 1985). The final selection of classifications naturally derives from this set of overall degrees and therefore valuable classifications are not discarded for having failed to meet few criteria.

Interpreting the classes of the chosen classification requires an amount of technical knowledge the end user does not usually possess (Oja, 1983). For this reason, it is desirable to rely on an automated tool for the description of these classes. If this description is done in a qualitative way, it enables the interpretation and understanding of the results, and improves the transmission of useful knowledge to experts.

1.4 Structure of the doctoral thesis

This thesis is structured as follows. In Chapter 2 a literature review of the topics of criteria for assessing and selecting classifications, aggregation functions based on OWA operators and natural

language systems for translating data to qualitative texts is provided. The next two chapters detail the main contributions of this thesis: Chapter 3 analyses and defines a set of five fuzzy criteria for assessing classifications while in Chapter 4 a NLG system to describe qualitatively the most important features of a classification is detailed. All analysed methodologies are applied in Chapter 5, where a case study is presented to generate, select and describe a segmentation from a real business situation. Finally, in Chapter 6 conclusions are drawn and suggestions made for further work. At the end of this document, Appendixes A and B include a brief explanation of LAMDA algorithm, the unsupervised learning method used in this thesis, and result tables of the case study, respectively.

Note that although the improvement or study of new aggregation functions is not within the scope of this thesis, Chapter 2 contains a description and a review of the literature on this topic. These functions have an important role in this thesis when summarising the information provided by the indexes associated with the criteria analysed in Chapter 3. Event though Chapter 5 presents an application of the presented methodology, Chapters 3 and 4 include easy examples with the aim of making this thesis reading more enjoyable and didactic. The end of each chapter contains a summary of conclusions related to the chapter, in addition with further research to be done.

1.5 Publications derived from this thesis

Germán Sánchez-Hernández, Francisco Chiclana, Núria Agell, Juan Carlos Aguado (2013). Ranking and selection of unsupervised learning marketing segmentation. *Knowledge-Based Systems*, 44:20–33.

Francisco J. Ruiz, Albert Samà, Germán Sánchez, José Antonio Sanabria and Núria Agell (2011). An interval technical indicator for financial time series forecasting. *Proceedings of the 25th International Workshop on Qualitative Reasoning (QR).*

Germán Sánchez, Albert Samà, Francisco J. Ruiz and Núria Agell (2010). Moving intervals for nonlinear time series forecasting. *Proceedings of the 13th International Conference of the Catalan Association for Artificial Intelligence (CCIA).*

José Antonio Sanabria, Germán Sánchez, Núria Agell and Josep Sayeras (2010). An application of SVMs to predict financial exchange rate by using sentiment indicators. *Proceedings of the*

V Simposio de Teoría y Aplicaciones de Minería de Datos (TAMIDA).

Germán Sánchez, Juan Carlos Aguado, Núria Agell, Mónica Sánchez (2009). Automatic Comparison and Selection of Classifications in Unsupervised Learning Processes. *XI Jornadas de ARCA Sistemas Cualitativos, Diagnosis, Robótica, Sistemas Domóticos y Computación Ubicua (JARCA).* Almuñécar (Granada), 24-26 June 2009.

Germán Sánchez, Mònica Casabayó, Albert Samà and Núria Agell (2008). Forecasting Customer's Loyalty by Means of an Unsupervised Fuzzy Learning Method. *Electronic proceedings of the 28th International Symposium on Forecasting*, 43. Nice, 22-25 June 2008.

Germán Sánchez, Juan Carlos Aguado and Núria Agell (2007). Forecasting New Customers' Behaviour by Means of a Fuzzy Unsupervised Method. *Artificial Intelligence Research and Development, Frontiers in Artificial Intelligence and Applications. Proceedings of the 10th CCIA.*, 163:368–375. Andorra, 25-26 October 2007. ISBN: 978-1-58603-798.

Germán Sánchez, Núria Agell, Juan Carlos Aguado, Mónica Sánchez and Francesc Prats (2007). Selection Criteria for Fuzzy Unsupervised Learning: Applied to Market Segmentation. In *Foundations of Fuzzy Logic and Soft Computing. Lecture Notes in computer Science*, 4529:307–310.

Cati Olmo, Germán Sánchez, Núria Agell, Mónica Sánchez and Francesc Prats (2007). Using Orders of Magnitude and Nominal Variables to Construct Fuzzy Partitions. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. London, 23-26 July 2007.

Chapter 2

Literature review

The application of unsupervised learning techniques enables the user to obtain new ways of segmenting a data set that were previously unthinkable. The selection of the most suitable classification from the obtained (or considered) ones can be faced as a multi-criteria decision making (MCDM) problem, in which each alternative is assessed according a set of criteria. The application of the considered criteria can be done in a sequential way in which each criterion is applied by discarding those classifications whose evaluation does not reach a predefined threshold. An alternative to this sequential approach is, for each classification, to aggregate each individual assessment and therefore to obtain a global ranking of the classification. Finally, in order to complete the system and fully understand the best classification according to the obtained ranking, a description of the most important characteristics of its classes must be provided.

In this thesis a contribution to the selection of classifications is provided, in terms of the criteria used to assess some aspects of the considered classifications, the functions used to aggregate the assessments and a qualitative description of the chosen classification with the aim of making the result of the MCDM system easily understandable. That is why this chapter includes a review of literature on this three topics. More specifically, Section 2.1 reviews criteria and methods for evaluating classifications that can be obtained from applying any of the available clustering techniques, Section 2.2 reviews aggregation functions with a deeper emphasis in Ordered Weighted Averaging (OWA) operators and finally, in Section 2.3 a review of literature in data-to-text systems is provided.

2.1 Criteria for selecting classifications

Unsupervised learning or clustering is one of the most useful tools in data mining processes for discovering groups that were previously unknown. A clustering technique segments a given data set into groups or clusters such that the individuals in a cluster are more similar to each other than individuals in different clusters (Barlow, 1989; Jain et al., 1999).

The design of suitable systems for selecting classifications is an important topic of research in clustering area (Kukar, 2003; Osei-Bryson, 2010) because such systems produce in most cases different classifications. If the inherent partition of a data set is known, the problem is translated into a search for the optimal clustering scheme that best fits this inherent partition (Halkidi et al., 2002). Other approaches do not select directly the best classification, but define the final partition by grouping the obtained ones in term of voting or averaging (Broder et al., 2008). In general, it is not usual to have *a priori* information of the data set, so the selection of the most suitable classification lies in the application of a single criterion or a set of criteria previously defined.

There are mainly three types of clustering validation criteria: internal, external and relative (Jain et al., 1999; Theodoridis and Koutroumbas, 2008). An *internal* criterion tries to determine if the classification structure is intrinsically appropriate for the considered data (Liu et al., 2010). An *external* criterion of validation compares the considered classification with an *a priori* structure: either a previously known partition of the analysed dataset typically provided by some domain experts or an external variable not participating in the clustering process (Wu et al., 2009). Finally, a *relative* criterion measures the relative similarity between two classifications, usually by comparing them by using the same supervised technique (Jain et al., 1999).

Several works reviewing cluster validation indexes have been published. In Halkidi et al. (2001) a review of clustering algorithms is done. The algorithms are explained by comparing them, including a review of clustering validity measures. Liu et al. (2010) analyse a set of eleven internal criteria that measure compactness and separation of the considered clusters, and give the validation properties of some of these criteria in different scenarios. In Yatskiv and Gusarova (2005) a review of the computation of indexes related to the most used internal and external criteria is done. Osei-Bryson (2010) gives an extensive review of cluster validation and provides a methodology in which the considered classifications are being discarded by applying predefined thresholds on a set of validation criteria. These considered criteria cover the three types of validation criteria. These

works and other using or defining new criteria are shown in Tables 2.1 and 2.2.

Internal criteria for validating classifications can be classified according to the concept they are related to. Criteria associated with the *compactness* concept compute how closely related the individuals in a cluster are; these criteria are usually based on indexes measuring density or variance of the clusters (Bittmann and Gelbard, 2009; Cheng et al., 1999; Halkidi et al., 2001; Liu et al., 2010; Ramze Rezaee et al., 1998; Tibshirani and Walther, 2005; Wang et al., 2009; Xiong et al., 2009). *Separability* criteria determine how distinct or well-separated a cluster is from other clusters (Halkidi et al., 2001; Liu et al., 2010; Ramze Rezaee et al., 1998; Tibshirani and Walther, 2005; Wang et al., 2009; Xiong et al., 2009). Criteria related to the *prediction strength* of the clusters calculate the accuracy rate of a model obtained by applying a supervised learning technique on them (Kukar, 2003; Osei-Bryson, 2010; Tibshirani and Walther, 2005; Xiong et al., 2009). Some criteria are based on the number of important *features* (Osei-Bryson, 2010). Criteria quantifying the achievement of *goals* can be very heterogeneous: from analysing desired structure of the obtained clusters (Cheng et al., 1999), applying economic theories (Choi et al., 2005), being assessed by graphical visualisations (Bittmann and Gelbard, 2009), or checking the existence of outliers clusters or pairs of variables (Osei-Bryson, 2010).

External criteria require the existence of an *a priori* external variable or classification defined for each of the individuals. The computation of an index associated with external criteria can be performed by any of the following indexes: Rand statistic, Jaccard coefficient, Fowlkes and Mallows index, Hubert's statistic and so on (Halkidi et al., 2001; Wu et al., 2009; Yatskiv and Gusarova, 2005).

Finally, the computation of relative criteria implies the pairwise comparison between clusters. This comparison is usually performed by some domain experts (Halkidi et al., 2001; Osei-Bryson, 2010; Yatskiv and Gusarova, 2005). Although there are some methods to guide the search of which comparisons should be made for minimising their number, relative criteria have not been taken into account in this work due to the usual difficulty in getting this feedback from the experts.

As it can be seen in Tables 2.1 and 2.2, all analysed papers give a review of existing criteria or make a definition of new criteria, all of them based on some of the concepts used for clustering evaluation. It is important to note that almost all concepts are covered in the present work.

Paper	Comments	Internal criteria					External criteria	Relative criteria
		Compactness	Separability	Accuracy	Features	Goals		
Ramze Rezaee et al. (1998)	One index for fuzzy c-Means	Yes: compactness	Yes: separation	No	No	No	No	No
Cheng et al. (1999)	Subspace clustering	Yes: high density	No	No	No	Yes: coverage & correlation of dimensions	No	No
Kukar (2003)	Reliability on diagnoses	No	No	Yes: reliability	No	No	No	No
Haldiki et al. (2001)	Review	Yes: several		No	No	No	Yes: several	Yes: several
Choi et al. (2005)	Association rules	No	No	No	No	Yes: Recency, Frequenacy & Monet. Value	No	No
Tibshirani & Walther (2005)	Validation by prediction strength	Yes: variance	Yes: bias	Yes: prediction strength	No	No	No	No
Yatskiv & Gusarova (2005)	Review	No	No	No	No	No	Yes: several	Yes: several
Method presented	Review & application	Yes: I_C (coherence)		Yes: I_A (accuracy)	No	Yes: I_U & I_B (usefulness & balanced)	Yes: I_D (dependency)	No

Table 2.1: Clustering validation criteria (Part I)

Paper	Comments	Internal criteria					External criteria	Relative criteria
		Compactness	**Separability**	**Accuracy**	**Features**	**Goals**		
Bittmann & Gelbard (2009)	Visualisation of hierarchical clustering	Yes: minimal hetereogeneity	No	No	No	Yes: visualisation	No	No
Wang et al. (2009)	Clinical application	Yes: Davies-Bouldin & rel.-free		No	No	No	No	No
Wu et al. (2009)	External criteria for k-Means	No	No	No	No	No	Yes: several	No
Xiong et al. (2009)	k-Means	Yes: Sum of Squared Errors	Yes: entropy and Coef. of Variation	Yes: F-measure	No	No	No	No
Liu et al. (2010)	Internal criteria review	Yes: several	Yes: several	No	No	No	No	No
Osei-Bryson (2010)	Review	No	No	Yes: accuracy	Yes: # of important variables	Yes: outliers, Max/Min	No	Yes: several
Method presented	Review & application	Yes: I_C (coherence)		Yes: I_A (accuracy)	No	Yes: I_U & I_B (usefulness & balanced)	Yes: I_D (dependency)	No

Table 2.2: Clustering validation criteria (Part II)

2.2 Aggregation functions based on OWA operators

This section reviews the literature in aggregation functions, emphasising in the study of OWA operators. Although the improvement or study of new aggregation functions is not one of the contributions of this thesis, a study and review of the literature in this topic is done since these functions have an important role in this thesis. They are responsible for summarising the information provided by the indexes allowing us to select the most suitable classification.

The selection of the most suitable classification among a set of feasible ones and according to a set of predefined criteria can be faced as a MCDM approach. Each classification (alternative) is assessed by each of the considered criteria (evaluations). MCDM problems normally consist of two steps (Fodor and Roubens, 1994): *aggregation* and *exploitation*. The aggregation step consists of combining for each alternative the single evaluations into a collective evaluation in such a way that it summarises the conditions expressed in all the evaluations. The exploitation phase transforms the global evaluation of the alternatives into a ranking of the alternatives. This can be done in different ways, the most common being the use of a ranking method to obtain a score function (Chiclana et al., 1998; Gramajo and Martínez, 2012; Zhang and Guo, 2012; Zhou and Chen, 2012).

Many different families of aggregation operators have been studied (Chiclana et al., 2004, 2007; Dubois and Prade, 1985; Fodor and Roubens, 1994; Herrera et al., 2003; Klir and Folger, 1988; Torra, 1997; Torra and Narukawa, 2007; Xu and Da, 2003; Yager, 1988; Zhou et al., 2008). Among them the OWA operator proposed by Yager (Yager, 1988) is one of the most widely used. Among the reasons to support this extensive use of the OWA operator is that it allows the implementation of the concept of fuzzy majority in the aggregation phase by means of a fuzzy linguistic quantifier (Zadeh, 1983) representing the proportion of satisfied criteria 'necessary for a good solution' (Yager, 1996). This is done by using the linguistic quantifier in the computation of the weights associated with the OWA operator. In addition, Marichal (1998) investigated the aggregation of dependent criteria and the fuzzy integral was found to be the appropriate aggregation operator in these cases. The most representative fuzzy integrals are the Choquet integral and the Sugeno integral. It is well known that the OWA operator is a particular case of Choquet integral, and consequently it is not necessary to assume independence of criteria when using the OWA operator.

Generally speaking, the OWA operator based aggregation process consists of three steps:

(i) the first step is to re-order the input arguments in increasing order. In this way, a particular

element for aggregation is not associated with a particular weight, but rather a weight is associated with a particular ordered position of an aggregated object;

(ii) the second step is to determine the weights for the operator in a proper way;

(iii) finally, the OWA weights are used to aggregate the re-ordered arguments.

Among the three steps, the first step introduces non-linearity into the aggregation process by re-ordering the input arguments, which make Yager's OWA operator significantly different from the classical linear weighted averaging operator.

Definition 2.1. *An OWA operator of dimension n is a mapping $\phi\colon \mathbb{R}^n \to \mathbb{R}$, which has a set of weights $W = (w_1, \cdots, w_n)^T$ associated with it, so that $w_i \in [0,\ 1]$ and $\sum_{i=1}^{n} w_i = 1$,*

$$\phi_W(a_1, \cdots, a_n) = \sum_{i=1}^{n} w_i a_{\sigma(i)} \tag{2.1}$$

where σ is a permutation function such that $a_{\sigma(i)}$ is the i-th highest value in the set $\{a_1, \cdots, a_n\}$.

This OWA operator exhibits the following desirable properties for an aggregation operation:

1. It is commutative:

   $$\phi_W\left(p_{\sigma(1)}, \ldots, p_{\sigma(n)}\right) = \phi_W\left(p_1, \ldots, p_n\right),$$

   being σ any permutation of the set $\{1, \ldots, n\}$.

2. It is an *or-and* operator, i.e., it is located between the minimum and the maximum of the arguments to be aggregated:

   $$min(a_i) \leq \phi_W(a_1, \cdots, a_n) \leq max(a_i).$$

3. It is idempotent:

   $$\phi_W(a, \cdots, a) = a.$$

4. It is monotonic:

   $$\phi_W(a_1, \cdots, a_n) \geq \phi_W(e_1, \cdots, e_n), \text{ if } a_i \geq e_i\ \forall i.$$

An issue in the definition of the *OWA* operator is how to obtain the associated weighting vector (Yager, 1988). In Yager (1988) we can find two ways to do this. The first approach is to use a learning mechanism using some sample data; the second approach is to provide some semantics or meaning to the weights. The latter approach enables applications in the area of quantifier guided aggregations (Yager, 1983; Pei et al., 2012).

In the process of quantifier guided aggregation, given a collection of n criteria represented as fuzzy subsets of the alternatives X, the OWA operator has been used to implement the concept of fuzzy majority in the aggregation phase by means of a *fuzzy linguistic quantifier* (Zadeh, 1983) that indicates the proportion of satisfied criteria 'necessary for a good solution' (Yager, 1996). This implementation is done by using the quantifier to calculate the OWA weights.

Definition 2.2. *A fuzzy subset Q is called a Regular Increasing Monotone (RIM) quantifier if (i) $Q(0) = 0$; (ii) $Q(1) = 1$; (iii) $Q(x) \geq Q(y)$ if $x > y$.*

Definition 2.3. *A fuzzy subset Q is called a Regular Decreasing Monotone (RDM) quantifier if (i) $Q(0) = 1$; (ii) $Q(1) = 0$; (iii) $Q(y) \geq Q(x)$ if $x > y$.*

Definition 2.4. *Given a function $Q : [0,1] \to [0,1]$ such that $Q(0) = 0$, $Q(1) = 1$ and if $x > y$ then $Q(x) \geq Q(y)$, an OWA aggregation operator guided by Q is given as (Yager, 1988):*

$$\phi_Q(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i \cdot a_{\sigma(i)},$$

being $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ a permutation such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$, $\forall i = 1, \ldots, n-1$, i.e., $a_{\sigma(i)}$ is the i-th largest value in the set $\{a_1, \ldots, a_n\}$; and

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), i = 1, \ldots, n. \tag{2.2}$$

These Q functions are called Basic Unit-interval Monotone (BUM) functions in Yager (2003) and 'are particularly useful in situations in which the imperative guiding the OWA aggregation is expressed linguistically by a quantifier'. Note that in Yager (1996) BUM functions are called RIM quantifiers.

Examples of RIM quantifiers are *all*, *most*, *many* and *at least* α, in contrast of RDM quantifiers like *none*, *few* or *at most* α (Yager, 1996). Example 2.1 defines the quantifiers *all* and *there exists*, both of them RIM quantifiers.

Example 2.1. *The quantifier* all *is represented by the fuzzy subset* Q_* *where*

$Q_*(1) = 1$ *and* $Q_*(x) = 0$ *for all* $x \neq 1$.

The quantifier there exist, not one *is represented by the fuzzy subset* Q^* *where*

$Q^*(0) = 0$ *and* $Q^*(x) = 1$ *for all* $x \neq 0$.

Consider the parameterised fuzzy subset defined $[0, \ldots, 1]$ such that

$$Q(r) = r^\alpha, \alpha \geq 0. \tag{2.3}$$

It can be seen that this formulation defines a family of RIM quantifiers. The special cases of this family are worth noting:

- For $\alpha = 1$ we get $Q(r) = r$. This is called the *unitor* quantifier.
- For $\alpha \to \infty$ we get Q_*, the universal quantifier.
- For $\alpha \to 0$ we get $Q*$, the existential quantifier.
- For $0 < \alpha < 1$ we get a concave function.
- For $\alpha > 1$ we get a convex function.

Figure 2.1 depicts some examples of RIM functions of $Q(r) = r^\alpha$ family on the top, and their corresponding vector of weights in the bottom. The first subgraph ($\alpha = 0$) represents the previously named *all* quantifier, in which only the higher value do not have a null weight. Second and third subgraphs correspond to α lower than 1. The concave property of Q provides decreasing weights. Fourth subgraph ($\alpha = 1$) stands for the *unitor* quantifier, obtaining equally-valued weights and therefore representing the *mean* operator. Finally, fifth and sixth subgraphs depict functions with α greater than 1, obtaining convex functions and therefore increasing weights.

Example 2.2 illustrates the use of RIM quantifiers to aggregate a set of values.

Example 2.2. *The aggregation of the set of values* $\{0.5, 0.07, 0.228, 0.057, 0.482\}$ *using an OWA operator guided by the fuzzy linguistic quantifier 'most of' represented via the RIM function* $Q(r) = r^{1/2}$*, whose corresponding weighting vector using (2.2) is*
$(0.447, 0.185, 0.142, 0.120, 0.106)$*, yields*

$$\begin{aligned} \phi_{most\ of}(0.5, 0.07, 0.228, 0.057, 0.482) = &\quad 0.447 \cdot 0.5 + 0.185 \cdot 0.482 + 0.142 \cdot 0.228 + 0.129 \cdot 0.07 + 0.106 \cdot 0.057 \\ = &\quad 0.360. \end{aligned}$$

This collective value is interpreted as the value up to which 'most of' the criteria are verified.

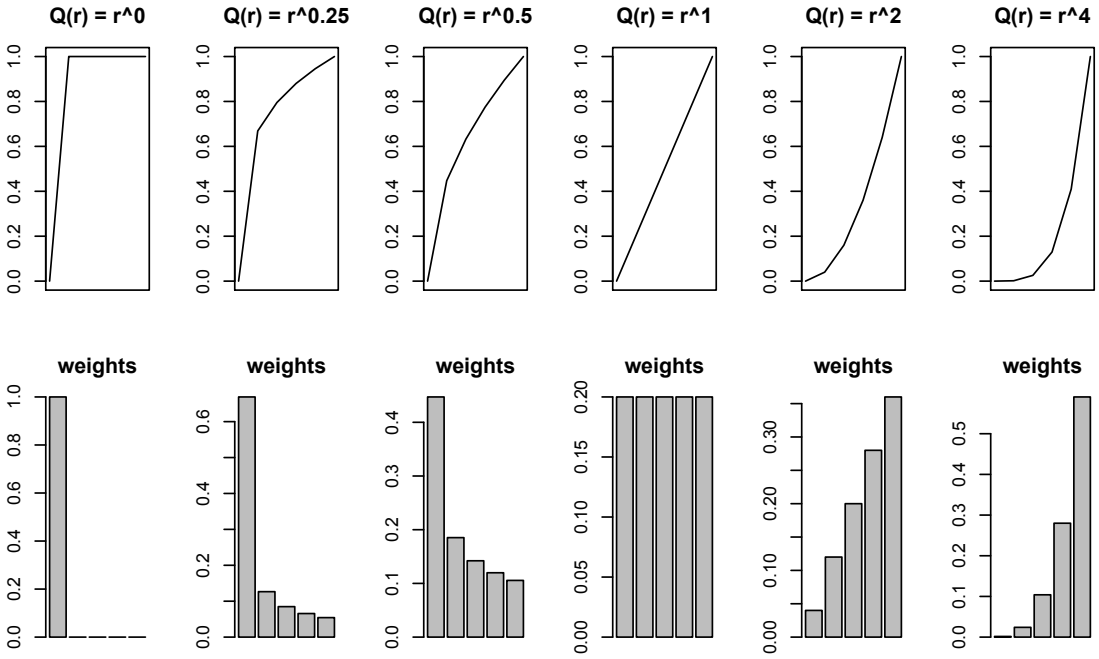

Figure 2.1: Some RIM functions and their corresponding obtained weights

This type of aggregation 'is very strongly dependent upon the weighting vector used' (Yager, 1996), and consequently upon the function expression used to represent the fuzzy linguistic quantifier. The RIM function used in this work (with an α between 0 and 1) guarantees that all the individual valuations contribute to the final aggregated value because it is a strictly increasing function. Moreover, the higher the ranking of a value, the higher the weighting value associated with it. This is a consequence of the concavity property – which was proven in Chiclana et al. (2007) to make a RIM function appropriate for conducting aggregation processes in heterogeneous decision-making problems.

2.3 Data-to-text systems for generating natural language

A growing number of applications require the translation of perceptual, sensory or statistical data into natural language descriptions, therefore increasing the interest in the research of data-to-text systems (Reiter, 2007). These systems summarise qualitative or numeric data into natural language

texts with the motivation of the belief that textual summaries made large amount of numeric data more accessible to human users than traditional ways of presenting data (Hüske-Kraus, 2003b; Reiter et al., 2005). Such systems present not only a practical interest –most of natural language generation (NLG) systems are considered to be data-to-text systems– but also a scientific interest in the study of the relation between language and non-linguistic world.

The most frequent area of application of such systems and the one in which their utility has been proved is the **weather forecasting** domain. Several systems have been designed to produce textual weather forecast from weather data. One of the earliest works in such area and a classical one is *FoG* (Goldberg et al., 1994), which converts weather maps info forecast text by using rules, NLG and linguistic models. It is based on an three-stages architecture: data extraction, conceptual (metereological) processing and linguistic processing, which involves both text planning and text realisation. More recently, *SumTime-Mousam* (Sripada et al., 2003) produces textual marine weather forecasts for offshore oilrig applications. Its input data are mainly time series. It points out the importance of the sensitivity of the text to the end user, and the configurability of the system by the forecasters to easily adjust the output. It uses a three-stages architecture: document planning, micro-planning and realisation, and it shows the big potential of these technologies. This potential has been evaluated in Reiter et al. (2005), where it is concluded that people prefers the computer-generated texts rather than the human-generated ones, mainly due to the best consistency of the automatic texts, as shown in Table 2.3.

There have been also lots of data-to-text systems applied in the **medicine** area mainly because of the large amount of data that human systems have to deal with. In Hüske-Kraus (2003b) a review of some of the applications of these systems in the medicine area is provided, also identifying the main functionalities of NLG applications in health care: producing explanations and giving advice in medical expert systems, generating reports, progress notes and discharge letters, preparing individualised patient information material and generating descriptions of medical concepts. These applications generally deal with raw data, in contrast of systems that are based on discrete events. In this sense, *Suregen-2* (Hüske-Kraus, 2003a) helps medical staff in the elaboration of routine reports by improving the specification and use of the predefined medical ontology, while *Narrative Engine* (Harris, 2008), in its commercial version called *Component-Based Processing*, helps in the creation of summaries during symptoms, tests and prescriptions.

These systems not only help medical personnel in their work but can also offer help to patients.

Table 2.3: Preference results in Reiter et al. (2005) where statisticallly significant results are in **bold**.

Question	Computer/Hybrid	Human	same	p-value
Computer vs. human texts				
More appropriate?	**43% (77)**	**27% (49)**	**30% (53)**	**0.021**
More accurate?	51% (90)	33% (59)	15% (29)	0.011
Easier to read?	41% (74)	36% (65)	23% (41)	>0.1
Hybrid vs. human text				
More appropriate?	38% (68)	28% (50)	34% (62)	>0.1
More accurate?	45% (80)	36% (65)	19% (34)	>0.1
Easier to read?	**51% (91)**	**17% (30)**	**32% (59)**	**<0.0001**

For example, *Stop* (Reiter et al., 2003) tries to help people quit smoking by generating short tailored smoking cessation letters based on responses to a questionnaire. Unfortunately, this system was not effective. In the same way of helping patients, *Piglit* (Cawsey et al., 2000) aids to oncologic patients by elaborating personalised hypertext pages explaining treatments, diseases and measurements related to the patients' condition. As a hybrid system, helping both medical staff and patients, *BT-45* (Portet et al., 2009) builds automatic summaries from data provided by sensors in the Neonatal Intensive Care Unit, both physiological signals and discrete events. This work is based on *Topaz* (Kahn et al., 1991). An example of a system dealing with data as a list of events is described in Hallett and Scott (2005), which generates two main types of report. On the one hand, a longitudinal report provides a quick historical overview of the patient's illness and, on the other hand, a report focussed on a given type of event in a patient's history is supplied.

Data-to-text systems have been applied with success in other domains. *iGraph* (Ferres et al., 2006) improves the **accessibility** of graphical data for the visually-impaired by describing graphs in a simple and repetitive way. This description includes an overview of the axis, maximums and minimums, trends, evolution and so on. In a similar topic, *Atlas.txt* (Thomas and Sripada, 2008) describes geo-referenced information as text also for the visually-impaired by covering the identification of the location of high and low values and trends detected in the data. In **financial** area, *Ana* (Kukich, 1983) generates textual stocks reporting from numeric data of stock market,

basing this construction in three principles: the use of domain-specific semantic and linguistic knowledge, the use of macro-level semantic and linguistic constructs and the production system approach to knowledge representation. One of the standing works in the description of **visual scenes** is *Vitra* (Herzog and Wazinski, 1994), where a knowledge-based system is used to translate visual information into natural language descriptions, focussing in simultaneous scene description and image sequence evaluation. In the same area of application, *Describer* (Roy, 2002), uses machine learning techniques for acquire linguistic structures generalised from training data in the form of domain specific rules of language generation, with the objective of describing objects in computer-generated visual scenes. As a partial data-to-text system in such domain, *Ladder* (Hammond and Davis, 2005) is a sketching language for describing sketch-based user interface. There have been also other areas of application like **Sports**, in which *ScubaText* (Sripada and Gao, 2007) stands out in its helps to scuba divers by making summaries with graphical and textual information oriented to the security of the diving activity.

Most of these systems analyse raw data, typically data gathered from sensors and collected in form of time series. As opposite, *PLANDoc* (McKeown et al., 1994) builds summaries based on the events outputs generated in simulations trying to avoid repetition of similar information and similar phrasing. This work is not framed in a specific domain. Other example of a **generic** system is focussed on labour force surveys (LFS) (Iordanskaja et al., 1992), where bilingual (english and french) summaries of Canadian statistical data are supplied.

The work presented in this thesis in Chapter 4 describes in a qualitative way the most important characteristics of each cluster of a considered data set. It is a based on the analysis and application of a set of rules in order to avoid repetitive information (as seen in Section 4.2), to merge related information into the same sentence (as explained in Subsection 4.3.1) and to obtain a more natural description of the considered classes (as detailed in Subsection 4.3.2). Some of the reviewed papers describe also rule-based systems. A summary of the use of rules in such systems is provided below.

In Iordanskaja et al. (1992) rules are employed in both text planning and realisation stages; Kukich (1983) employs rules in the message generator stage, the discourse organiser and the text generator. In this last stage, for example, rules are used for grouping messages according to a clause-combining grammar; in Goldberg et al. (1994) rules are included in the data extraction stage to adapt the text to the desired type of output, in the linguistic processing stage provided by weather forecasters and for performing the use of the grammar; in Cawsey et al. (2000) rules are

attached to the hierarchy used in the medical knowledge base of the system; the system designed in Roy (2002) differs from the majority because domain-specific rules of language generation are learned from examples produced directly by domain experts. In Portet et al. (2009) expert rules are used to compute the importance of the events collected in the data extraction stage. Also, expert rules helps in finding associations between events while the document planning stage also uses some special-case rules and the microplanning and realisation stages employ rules for matching events against templates. Note that this work is based on the one carried out by Kahn et al. (1991), also based on rules. In Reiter et al. (2003) rules are used for deciding whether to include some content, while Sripada et al. (2003) employ rules in the micro-planning stage between the different parts of the sentence, according to previous or related sentences; in Hallett and Scott (2005) rules are used both in the content selection stage to pick the events to be included in the text, and in the document planning stage to group messages: in Hammond and Davis (2005) the domain shape recognition is performed by a rule-based system while in Ferres et al. (2006) rules are used for describing and querying the input graph.

The NLG system proposed in this thesis tries to make up for two main lacks affecting data-to-text systems. On the one hand, most of the reviewed systems are domain dependent. They have been designed to analyse input data with a known structure and to provide a specific natural language text according to the framework in which they are developed. The proposed NLG system is a generic one. It is able to produce generic natural text by only analysing the provided input clustered data set, without needing the definition of any domain knowledge. But it is important to note that the specification of some optional and short domain information enables the system to produce context-based text and therefore a more attractive description of involved classes. On the other hand, all reviewed works are designed to build summaries or descriptions of an specific data set, without differentiating among any existing subsets. The designed system highlights the most important features of each class by comparing them in terms of conditional and joint distributions of modalities of each considered variable on each class.

Paper	System	Application area	Input data	Users	Rules
Goldberg et al. (1994)	*FoG*	Weather forecasting	Time series	Forecasters	Yes
Reiter et al. (2005)	Forecasting texts		Time series	Forecasters	No
Sripada et al. (2003)	*SumTime-Mousam*		Time series	Forecasters	Yes
Cawsey et al. (2000)	*Piglit*	Medicine	Events	Patients	Yes
Hallett and Scott (2005)	Summaries of events		List of events	Medical staff & patients	Yes
Harris (2008)	*Narrative Engine*: text summaries		Events	Medical staff	No
Hüske-Kraus (2003b)	Review of applications		Raw data	Medical staff	No
Hüske-Kraus (2003a)	*Suregen-2*: Routine reports			Medical staff	No
Kahn et al. (1991)	*Topaz*				Yes
Portet et al. (2009)	*BT-45*: Neonatal summaries		Raw data from sensors	Medical staff & patients	Yes
Reiter et al. (2003)	*Stop*: personalised reports		Manual input	Patients	Yes
Ferres et al. (2006)	*iGraph*	Accessibility	Graphical data	Visually-impaired	Yes
Thomas and Sripada (2008)	*Atlas.txt*		Geo-referenced data	Visually-impaired	No
Kukich (1983)	*Ana*: textual stock market	Financial	Time series	Stock marketers	Yes
Hammond and Davis (2005)	*Ladder*	Image	Sketches		Yes
Herzog and Wazinski (1994)	*Vitru*		Visual scenes		No
Roy (2002)	*Describer*		Visual scenes		Yes
Sripada and Gao (2007)	*ScubaText*	Sports		Scuba divers	No
Iordanskaja et al. (1992)	Summaries	Generic	Statistical data		Yes
McKeown et al. (1994)	*PLANDoc*		List of events		No
Method presented (2013)	Description of groups	Generic	Tabular data	Generic	Yes

Table 2.4: Data-to-text systems

Chapter 3

Fuzzy criteria for selecting classifications

The use of unsupervised learning systems allows the behaviour of certain phenomena to be identified without relying on expert knowledge or information from past situations. Indeed, the main characteristic of this type of learning systems is that they work with patterns without explicitly knowing their output. Because of this, unsupervised learning systems have been considered in the literature as systems capable to capture knowledge from complex structures (Duda et al., 2001; Figueiredo and Jain, 2002; Jain, 2010).

Choosing the most appropriate classification from a given set of classifications of a set of patterns is an important topic on unsupervised systems and, in particular, on clustering analysis. In most cases, the use of these unsupervised techniques leads to several classifications as outputs, i.e. various classifications are compatible with the set of given patterns. For this reason, research in this area aims to develop suitable tools and models for selecting classifications (Broder et al., 2008; Kukar, 2003; Osei-Bryson, 2010).

This chapter presents a set of fuzzy criteria to be integrated into a classification selection methodology. Each fuzzy criterion is modelled via a membership function measuring the degree up to which it is is verified by all considered classifications. With that aim, an index associated with each criterion is designed. This chapter analyses and demonstrates properties and usability of each fuzzy criterion.

The proposed criteria cover most of the concepts historically used to assess classifications, as the ones introduced in Section 2.1. Internal criteria like assessing the achievement of desired goals in terms of obtaining a classification with a useful number of classes or a proper distribution of the considered individuals along the classes are included. Also internal criteria quantifying compactness and separability of the classes, and the potential accuracy of the models associated with the classifications are considered in this chapter. External criteria are also covered in this chapter, by defining an index that evaluates the compatibility or dependency between each classification and an external variable provided by the experts.

In order to select the most suitable classification, all the defined criteria must be taken into account. Previous research use selection criteria as filters: they are applied sequentially to the considered classifications (Choi et al., 2005; Osei-Bryson, 2010; Sánchez-Hernández et al., 2007; Sánchez Almeida et al., 2010). All those failing to meet a particular criterion are discarded and not taken into account in the application of the subsequent criterion. The following drawback can be associated with this type of methodology: because a true-false decision is applied in the application of each of the criterion, this could result in classifications being discarded and not taken into account when they marginally fail to meet one particular criterion but meet other criteria with a high score. Therefore, a classification might be discarded prematurely when its 'overall' score, with respect to the set of criteria, would have been high. In an extreme case, this methodology could produce no result because none of the classifications meet a particular criterion, which could indicate that the criterion in particular might not have been the most adequate or taken into account.

An alternative to the sequential approach described above, which has been successfully applied in multi-criteria decision making (MCDM), is that of evaluating the degree up to which each criterion is met by all classifications and, only after this, obtaining an overall aggregated value for each classification reflecting the degree to which the set of criteria is satisfied by each classification. Note that the objective of the aggregation step is to combine a set of criteria in such a way that the final aggregation output takes all the single criteria into account (Dubois and Prade, 1985). The final selection of classifications naturally derives from this set of overall degrees, and the drawback mentioned above does not apply. Although the improvement or study of further aggregation functions is not within the scope of this thesis, a review of the literature in this area has been included in section 2.1 as it has an important role for defining a complete MCDM approach to select and understand the best classification.

This chapter is structured as follows: Sections 3.1 to 3.5 analyse the usability of a set of five fuzzy criteria for assessing classifications, demonstrating their properties when required, while Section 3.6 details conclusions of this chapter as well as further research to be done in this area.

3.1 First criterion: useful number of classes

The usability of a classification is based on its informativeness and manageability: it is worthwhile examining classifications that have a sufficient number of classes to generate new knowledge, but are small enough to produce an easy and manageable model. For instance, in marketing environments in which these classifications are used to extract behavioural patterns to design market strategies, the number of classes distinguished is usually taken to be between three and five (Casabayó, 2005). This is because marketing campaigns with less than three segments may not be informative; while those with more than five segments may not be manageable.

The assumption of a classification with a number of classes M between K_1 and K_2 to be considered useful for a given problem does not imply that a classification with a number of classes lower than K_1 or higher than K_2 should be automatically discarded. This is specially true in those cases when there is enough evidence to suggest that such classifications perform well with respect to the rest of criteria. A natural approach in these cases would be that of associating a value to each classification to indicate how well they fit with the criterion 'useful number of classes'. By doing this, we move from a crisp to a fuzzy interpretation of the criterion 'useful number of classes', i.e. we move from the use of a characteristic function to the use of a membership function.

Note that a classification with a single class is trivial and therefore not useful, while a classification with a number of classes between K_1 and K_2 is considered totally useful. The minimum number of classes in any classification is 1 (contains all the individuals), while the maximum is N (each class contains just 1 individual). These two classifications are not informative and therefore these classification associated usefulness degree should be 0. A classification usefulness degree therefore should increase as the number of classes increases from 1 to K_1 and should decrease when the number of classes increases from K_2 to N. These restrictions are summarised in Property 3.1.

Property 3.1. *Given a set of N individuals classified into M classes, a classification usefulness degree has the following properties:*

- *It must be 0 for classifications with $M = 1$ or $M = N$ classes.*

- *It must be 1 for classifications with M between K_1 and K_2.*
- *It must increase as the number of classes M increases from 1 to K_1.*
- *It must decrease as M decreases from K_2 to N.*

Therefore the general expression of the membership function associated with the criterion 'useful number of classes' is the following:

Definition 3.1. *Given a classification $\mathcal{C}$, the index of usefulness is characterised by the following membership function:*

$$I_{U,K_1,K_2}(\mathcal{C}) = \begin{cases} f_1(M), & \text{if } 1 \leq M < K_1; \\ 1, & \text{if } K_1 \leq M \leq K_2; \\ f_2(M), & \text{if } K_2 < M \leq N, \end{cases} \tag{3.1}$$

where $M \in \mathbb{N}$ is the number of classes of $\mathcal{C}$; $K_1, K_2 \in \mathbb{N}$ such that $K_1 < K_2$ are two prefixed parameters; and f_1 is a strict increasing function and f_2 is a strict decreasing function verifying $f_1(1) = f_2(N) = 0$ and $f_1(K_1) = f_2(K_2) = 1$.

Figure 3.1 shows some examples of extensions to real numbers of membership functions satisfying the desired properties, with $K_1 = 4$ and $K_2 = 7$. Specifically, from top to bottom and left to right, it is shown a function being linear on its left tail and exponential on the right (3.2), a crisp function with maximum value on the interval $(K_1, \ldots, K_2)$ (3.3) and null on the rest, a function with an exponential left tail and a linear right tile (3.4) and a function with both logarithmic tails (3.5). Functions used in this example are the following:

$$I_{U_1,4,7}(\mathcal{C}) = \begin{cases} \frac{M-1}{4-1}, & \text{if } 1 \leq M < 4; \\ 1, & \text{if } 4 \leq M \leq 7; \\ \frac{e^{(N-M)}-1}{e^{(N-7)}-1}, & \text{if } 7 < M \leq N. \end{cases} \tag{3.2}$$

$$I_{U_2,4,7}(\mathcal{C}) = \begin{cases} 0, & \text{if } 1 \leq M < 4; \\ 1, & \text{if } 4 \leq M \leq 7; \\ 0, & \text{if } 7 < M \leq N. \end{cases} \tag{3.3}$$

$$I_{U_3,4,7}(\mathcal{C}) = \begin{cases} e^{-(4-M)}, & \text{if } 1 \leq M < 4; \\ 1, & \text{if } 4 \leq M \leq 7; \\ \frac{12-M}{12-7}, & \text{if } 7 < M \leq N. \end{cases} \tag{3.4}$$

$$I_{U_4,4,7}(\mathcal{C}) = \begin{cases} \frac{log(M)}{log(4)}, & \text{if } 1 \leq M < 4; \\ 1, & \text{if } 4 \leq M \leq 7; \\ \frac{log(M-7+1)}{log(7)}, & \text{if } 7 < M \leq N. \end{cases} \tag{3.5}$$

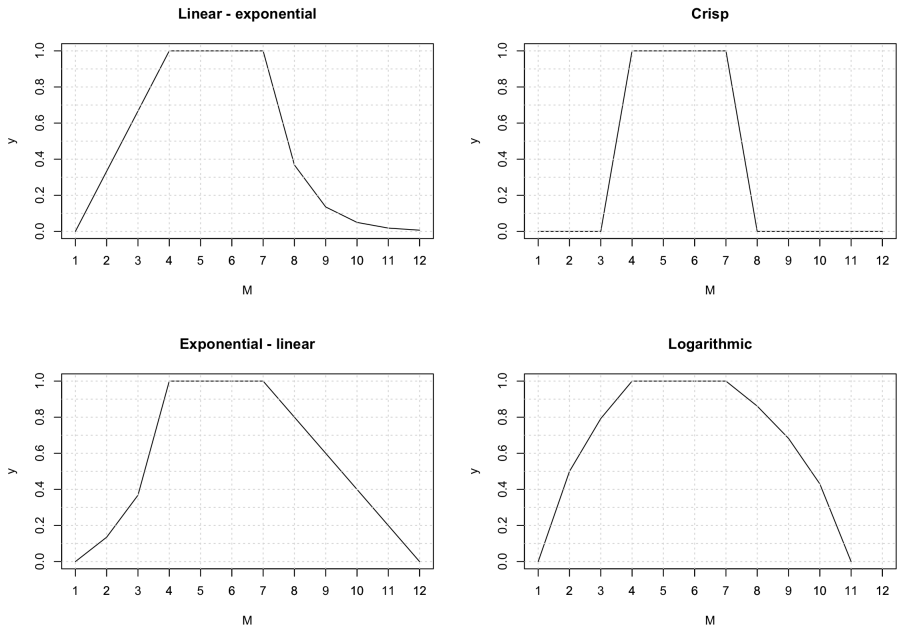

Figure 3.1: Examples of usefulness degree functions with $K_1 = 4$ and $K_2 = 7$

This fuzzy interpretation of the criterion 'useful number of classes' covers a larger number of classifications than the classical approach. In the case study presented in Chapter 5, a left-linear function has been chosen as the simplest example of a strictly increasing function; while a right-exponential function has been selected because the usefulness of a classification should decrease asymptotically when the number of classes increases. Such function is shown in (3.6). Note that the selection of different strict monotonic functions to the ones proposed here would not produce any change in the final ordering, because any two strict monotonic functions are mathematically equivalent in preserving an ordering.

$$I_{U,K_1,K_2}(\mathcal{C}) = \begin{cases} \frac{M-1}{K_1-1}, & \text{if } 1 \leq M < K_1; \\ 1, & \text{if } K_1 \leq M \leq K_2; \\ \frac{e^{(N-M)}-1}{e^{(N-K_2)}-1}, & \text{if } K_2 < M \leq N. \end{cases} \tag{3.6}$$

Figure 3.2 illustrates such a type of membership function with $K_1 = 3$ and $K_2 = 5$. Obviously, different increasing or decreasing functions could be used depending on the specific problem to solve and the preferences of the user: symmetric behaviour on both tails, linear or curve falls, concave or convex functions, etc.

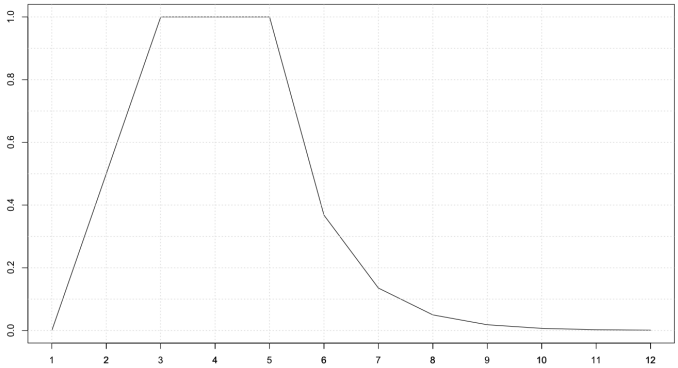

Figure 3.2: Fuzzy concept 'Useful number of classes' with $K_1 = 3$ and $K_2 = 5$

3.2 Second criterion: balanced classes

The second criterion is based on the distribution of individuals within the obtained classes. For this reason, the variable Y = 'number of elements of each class in a given classification' is considered and its associated dispersion will be used to define the fuzzy concept of 'balanced classification'. Note that in some situations, classifications in which one class encompasses most of the individuals (unbalanced) are worth avoiding because they do not contribute to creating new or relevant knowledge. Nevertheless, in other contexts, unbalanced classifications could be desirable.

Let $N \in \mathbb{N}$ be the number of individuals to be classified, and $M \in \{1, \ldots, N\}$ be the number of classes obtained by the classification Y. Given that different classifications can produce a different number of classes, the coefficient of variation, CV_Y, is considered to be a fairer indicator than the

standard deviation, σ_Y, in measuring the concept of 'balanced classification':

$$CV_Y = \frac{\sigma_Y}{\overline{Y}}, \tag{3.7}$$

with

$$\sigma_Y = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(Y_i - \overline{Y})^2} \quad \text{and} \quad \overline{Y} = \frac{1}{M}\sum_{i=1}^{M} Y_i = \frac{N}{M},$$

where Y_i is the number of individuals within the class i. Note that $CV_Y \geq 0$. There are known similarities between coefficient of variation and Gini coefficient (González Abril et al., 2010), usually applied to measure wealth inequalty. Its value is 0 when all individuals have the same wealth, while it is 1 when almost all individuals don't have nothing and one of them concentrates all the wealth.

In order to define a normal membership function (Wu et al., 2009), the *index of balanced classes*, I_B, is proposed as based on coefficient of variation and the minimum and maximum values of CV_Y need to be computed. In the following, the minimum and maximum values of CV_Y are computed by considering all the possible classifications for a given set of elements. Specifically, Lemma 3.1 states some restrictions between the values of variable Y and M and N, Proposition 3.1 uses this lemma to determine the minimum value for CV_Y excluding the trivial classifications, i.e. those with $M = 1$ or $M = N$, Lemma 3.2 gives the values Y_i for which its squared sum is maximum, Proposition 3.2 translates the problem defined in previous lemma to classifications, and Corollary 3.1 computes the maximum value of CV_Y fixed M while Proposition 3.3 establishes the maximum value of CV_Y for any possible classification.

Lemma 3.1. *Let N be a prime number of elements to classify. A classification Y producing M classes such that $1 < M < N$ verifies:*

$$\sum_{i=1}^{M} Y_i^2 \geq \frac{1 + N^2}{M}, \tag{3.8}$$

where $Y_i \in \mathbb{N}$ and $\sum_{i=1}^{M} Y_i = N$.

Proof. Given that $\frac{N}{M} \notin \mathbb{N}$, we can consider

$$Y_i = \frac{N}{M} + k_i,$$

with $\sum k_i = 0$ for $i \in \{1 \dots M\}$ and $|k_i| \geq \frac{1}{M}$. Then, given that $\sum_{i=1}^{M} k_i^2 \geq M \cdot \frac{1}{M^2} = \frac{1}{M}$:

$$\begin{aligned}\sum_{i=1}^{M} Y_i^2 &= \sum_{i=1}^{M}(\frac{N}{M}+k_i)^2 = \sum_{i=1}^{M}(\frac{N^2}{M^2}+2\frac{N}{M}\cdot k_i + k_i^2)\\ &= \frac{N^2}{M} + 2N\sum_{i=1}^{M} k_i + \sum_{i=1}^{M} k_i^2 = \frac{N^2}{M} + \sum_{i=1}^{M} k_i^2\\ &\geq \frac{1+N^2}{M}.\end{aligned}$$

□

Proposition 3.1. *Let N be the number of elements to classify. The minimum value of the coefficient of variation function when applied to the set of all classifications Y producing M classes such that $1 < M < N$ is:*

$$min_Y(CV_Y) = \begin{cases} \frac{1}{N}, & \textit{if } N \textit{ is a prime number};\\ 0, & \textit{otherwise.}\end{cases} \tag{3.9}$$

Proof. If N is not a prime number, we can consider the classification that produces M classes with M being one of the factors of N, such that all the classes consist of $\frac{N}{M}$ individuals. In this case, we have $Y_i - \overline{Y} = 0 \ \ \forall i \in \{1, \ldots, M\}$ and therefore the value of CV_Y is zero.

If N is a prime number, then we have that

$$CV_Y = \frac{\sqrt{\frac{\sum_{i=1}^{M}(Y_i - \frac{N}{M})^2}{M}}}{\frac{N}{M}} = \sqrt{\frac{M\sum_{i=1}^{M} Y_i^2}{N^2} - 1}\ .$$

From Lemma 3.1, we get that:

$$\frac{M\sum_{i=1}^{M} Y_i^2}{N^2} - 1 \geq \frac{1}{N^2},$$

thus,

$$CV_Y \geq \frac{1}{N}.$$

In addition, note that the classification with $M = 2$ classes of cardinal $\lfloor \frac{N}{2} \rfloor$ and $\lceil \frac{N}{2} \rceil$ respectively, has a coefficient of variation of

$$CV_Y = \frac{\sqrt{\frac{(\lfloor \frac{N}{2} \rfloor - \frac{N}{2})^2 + (\lceil \frac{N}{2} \rceil - \frac{N}{2})^2}{2}}}{\frac{N}{2}} = \frac{\sqrt{\frac{(\frac{1}{2})^2 + (\frac{1}{2})^2}{2}}}{\frac{N}{2}} = \frac{1}{N},$$

and therefore we can conclude that

$$min_Y(CV_Y) = \frac{1}{N}.$$

□

As it can be seen, when N (the number of individuals) is a prime number, this minimum depends on the value of N, and it decreases when N increases. Thus, this minimum value of CV_Y can be very small (and therefore negligible) for data sets with many individuals. But in the case of a moderate N, this minimum can be important, as seen in Example 3.1.

Example 3.1. *Let $N = 30$ be the number of individuals within a certain data set. According to Proposition 3.1, the minimum value of the coefficient of variation of any possible distribution of these N individuals is $\frac{1}{N} = \frac{1}{30} = 0.033$. Let's imagine that a certain classification of M individuals has a CV_Y of 0.5. If this value is standardised by knowing the theoretical minimum, it will be obtained a value of 0.52, which is significantly different from the original 0.5 and can affect the decision process.*

Lemma 3.2. *Let $F : \mathbb{R}^M \to \mathbb{R}^+$ be the following function $F(Y_1, \ldots, Y_M) = \sum_{i=1}^{M} Y_i^2$. The solution to the following problem*

$$\begin{aligned} Max: \quad & F(Y_1, \ldots, Y_M) \\ s.t.: \quad & \textstyle\sum_{i=1}^{M} Y_i = N \in \mathbb{N} \\ & Y_i \geq 1 \ \forall i \\ & N > M \end{aligned}$$

is $(Y_{1*}, \ldots, Y_{M*}) = (1, \ldots, 1, N - (M-1))$.

Proof. Let $(Y_1, \ldots, Y_M)$ be such that

$$\sum_{i=1}^{M} Y_i = N > M$$

and

$$1 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_M < N - M + 1.$$

We need to prove:

$$F(Y_{1*}, \ldots, Y_{M*}) > F(Y_1, \ldots, Y_M),$$

or equivalently:

$$\sum_{i=1}^{M} [Y_{i*}^2 - Y_i^2] > 0.$$

Denoting $d_i = Y_{i*} - Y_i$, we have

$$\sum_{i=1}^{M} \left[Y_{i*}^2 - Y_i^2\right] = \sum_{i=1}^{M} (Y_{i*} - Y_i) \cdot (Y_{i*} + Y_i) = \sum_{i=1}^{M} d_i \cdot (Y_{i*} + Y_i).$$

It is clear that $\sum_{i=1}^{M} d_i = 0$, $d_i \leq 0 \ \forall i \in \{1, \ldots, M-1\}$ and $d_M > 0$. Also because $Y_{i*} + Y_i < Y_{M*} + Y_i \leq Y_{M*} + Y_M$, we have that $d_i \cdot (Y_{i*} + Y_i) > d_i \cdot (Y_{M*} + Y_M) \ \forall i \in \{1, \ldots, M-1\}$. Thus, we conclude:

$$\sum_{i=1}^{M} d_i \cdot (Y_{i*} + Y_i) > \sum_{i=1}^{M} d_i \cdot (Y_{M*} + Y_M) = (Y_{M*} + Y_M) \cdot \sum_{i=1}^{M} d_i = 0.$$

□

Proposition 3.2. *Let N be the number of individuals to classify. Considering all the classifications with M classes, those that result in one class with cardinality $N - M + 1$ and the rest of classes with cardinality 1 have the greatest coefficient of variation.*

Proof. Let be Y_* the variable 'number of elements of each class' associated with a classification with one class with cardinality $N - M + 1$ and the rest of classes with cardinality 1. Without lost of generality we can consider:

$$1 = Y_{1*} = Y_{2*} = \cdots = Y_{M-1*} < Y_{M*} = N - M + 1.$$

For any other classification $\mathcal{C}$ with M classes, the range of the variable Y associated with $\mathcal{C}$ can be considered as follows:

$$1 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_M < N - M + 1.$$

Note that $\sum_{i=1}^{M} Y_{i*} = \sum_{i=1}^{M} Y_i = N$, $Y_{i*} \leq Y_i \ \forall i \in \{1, \ldots, M-1\}$, $Y_{M*} \geq Y_M$ and $\overline{Y_*} = \overline{Y} = \frac{N}{M}$.

Therefore, proving $CV_{Y_*} \geq CV_Y$ reduces to prove $\sigma_{Y_*} \geq \sigma_Y$, which in turn reduces to prove $\sum_{i=1}^{M} Y_{i*}^2 \geq \sum_{i=1}^{M} Y_i^2$, which is true according to Lemma 3.2. □

Example 3.2. *Given a data set with $N = 97$ individuals, according to Proposition 3.2 the most unbalanced classification with $M = 5$ classes is that with $Y = \{1, 1, 1, 1, 93\}$. If we compute the value of CV_Y, we get $CV_Y = 1.897$. If one of the individuals moves from the bigger class to another one (for example, $Y_2 = \{1, 1, 1, 2, 92\}$), CV_{Y_2} decreases ($CV_{Y_2} = 1.871$).*

Corollary 3.1. *Let $M \in \mathbb{N}$ be the number of classes used to classify N individuals. The maximum value of CV_Y is:*

$$\frac{1}{N}(N-M)\sqrt{M-1}\ . \tag{3.10}$$

Proof. From Proposition 3.2, we have that maximum value of CV_Y is achieved when the M classes cardinalities are $1 = Y_{1*} = Y_{2*} = \cdots = Y_{M-1*} < Y_{M*} = N-M+1$. A simple algebraic manipulation over Equation (3.2) yields:

$$\sigma_{Y_*} = \sqrt{\frac{\sum_{i=1}^{M} Y_{i*}^2}{M} - \left(\frac{\sum_{i=1}^{M} Y_{i*}}{M}\right)^2} = \sqrt{\frac{N^2 - 2MN + M^2 + 2N - M}{M} - \frac{N^2}{M^2}} =$$

$$= \sqrt{\left(\left(\frac{N}{M}\right)^2 - 2\frac{N}{M} + 1\right)(M-1)} = \sqrt{(\overline{Y}^2 - 2\overline{Y} + 1)(M-1)} =$$

$$= \sqrt{(\overline{Y}-1)^2(M-1)}\ = (\overline{Y}-1)\sqrt{M-1} = \frac{1}{M}(N-M)\sqrt{M-1}$$

and therefore the maximum value for the coefficient of variation is:

$$\frac{\sigma_{Y_*}}{\overline{Y}} = \frac{1}{N}(N-M)\sqrt{M-1}\ .$$

□

Example 3.3. *Given the same data set considered in Example 3.2 and according to Corollary 3.1, the maximum value of CV_Y is $\frac{1}{N}(N-M)\sqrt{M-1} = \frac{1}{97}(97-5)\sqrt{5-1} = 1.897$, whose value agrees with the previously obtained one.*

Proposition 3.3. *Given a number of individuals N, the maximum value of the coefficient of variation for all classifications is:*

$$max_Y(CV_Y) = \begin{cases} \frac{2N-3}{3N}\sqrt{\frac{N}{3}}, & \text{if } N \equiv 0 (mod\ 3); \\ \frac{2N-2}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 1 (mod\ 3); \\ \frac{2N-1}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 2 (mod\ 3). \end{cases} \tag{3.11}$$

Proof. Note that the minimum number of classes in any classification is 1 (contains all the individuals), while the maximum is N (each class contains just 1 individual). These two cases produce a value of zero for CV_Y as they are totally balanced classifications. Therefore, from now on, we assume $M \in \{2, \ldots, N-1\}$.

Let's consider the real function $f : [2, N-1] \to \mathbb{R}$:

$$f(x) = \frac{1}{N}(N-x)\sqrt{x-1},$$

defined as a real extension of (3.10). Figure 3.3 depicts function f with $N = 260$.

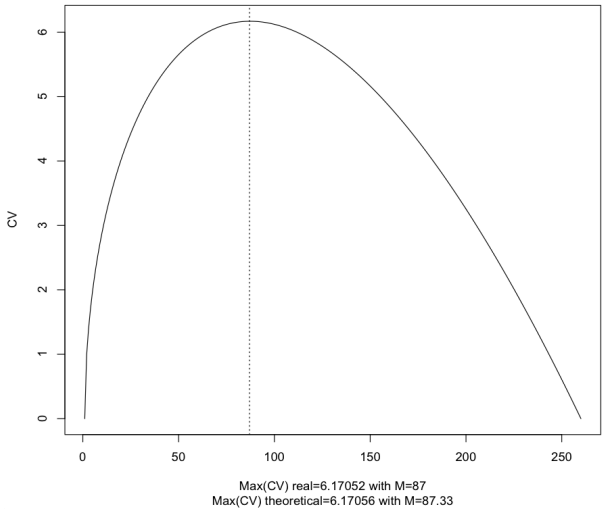

Figure 3.3: Function f and its maximum coefficients of variation when $N = 260$

The derivate of f is:

$$f'(x) = \frac{N+2-3x}{2N\sqrt{x-1}},$$

which is positive in $(2, \frac{N+2}{3})$ and negative in $(\frac{N+2}{3}, N-1)$. Thus, f then reaches its absolute maximum in $\frac{N+2}{3}$.

Nevertheless, the maximum value in the set $\{f(1), f(2), ..., f(N)\}$ depends on whether $\frac{N+2}{3}$ is integer or not. Let a be the integer part of $\frac{N+2}{3}$. Since $a \leq \frac{N+2}{3} < a+1$ and given that f increases in $(1, \frac{N+2}{3})$ and decreases in $(\frac{N+2}{3}, N)$, the maximum is $f(a)$ if $f(a) \geq f(a+1)$ and $f(a+1)$ if $f(a) < f(a+1)$. Simple algebraic manipulations yields:

$$max_Y(CV_Y) = \begin{cases} f(\frac{N+3}{3}) = \frac{2N-3}{3N}\sqrt{\frac{N}{3}}, & \text{if } N \equiv 0(mod\ 3); \\ f(\frac{N+2}{3}) = \frac{2N-2}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 1(mod\ 3); \\ f(\frac{N+1}{3}) = \frac{2N-1}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 2(mod\ 3). \end{cases}$$

Example 3.4. *Given the same data set considered in Example 3.2 and according to Proposition 3.3, the maximum value of CV_Y for any possible classification is* 3.73*, due to $N \equiv 1(mod\ 3)$. This maximum is reached with $M = 33$ classes and therefore with the classification $Y = \{1, 1, \ldots, 1, 65\}$. Indeed, if we compute the coefficient of variation of Y, we get $CV_Y = 3.73$.*

□

The minimum and maximum values of CV_Y are finally used to normalise and define the following *index of balanced classes*:

Definition 3.2. *Given a classification $\mathcal{C}$, the index of balanced classes of $\mathcal{C}$ is defined as:*

$$I_B(\mathcal{C}) = \frac{max_Y(CV_Y) - CV_{\mathcal{C}}}{max_Y(CV_Y) - min_Y(CV_Y)}, \tag{3.12}$$

where $CV_{\mathcal{C}}$ is the coefficient of variation associated with $\mathcal{C}$ and $min_Y(CV_Y)$ and $max_Y(CV_Y)$ are given as per Propositions 3.1 and 3.3, respectively.

The range of index I_B is $[0, 1]$ and it can be interpreted as the membership function associated with the vague concept 'balanced classification'. The higher the value of I_B the more balanced is the classification. Example 3.5 shows the computation of the value of I_B for two different classifications. When unbalanced classifications are preferred in a specific context, the index to use is:

$$I_{\overline{B}} = 1 - I_B\ . \tag{3.13}$$

Example 3.5. *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be the following two different classifications of the same data set consisting of $N = 260$ individuals:*

$$Y_1 = \{90, 80, 90\}; Y_2 = \{110, 30, 20, 90\}$$

According to Propositions 3.1 and 3.3, minimum and maximum values of CV_Y with $N = 260$ individuals are 0 *and* 6.18*, respectively. Computed values for CV_{Y_1} and CV_{Y_2} are* 0.51 *and* 4.43*, respectively. Note than $\mathcal{C}_2$ has a worse value than $\mathcal{C}_1$. The standardised value of CV_{Y_1} and CV_{Y_2}, and therefore the final values of indexes $I_B(\mathcal{C}_1)$ and $I_B(\mathcal{C}_2)$ are* 0.92 *and* 0.28*, respectively. Note that the obtained value for $\mathcal{C}_1$ is almost* 1*, the higher achievable value for I_B.*

3.3 Third criterion: coherent classification

The notion of adequacy of one individual to a class, which is modelled via a membership function, is used in this section to establish the concept of 'coherent classification'. A set of P qualitative or quantitative descriptors $\{D_1, \ldots, D_P\}$ is defined. Each individual to be classified will be represented as $X = (x_1, \ldots, x_P)$, where x_k is the observed value of X for descriptor D_k. Given a classification $\mathcal{C}$ consisting of M classes $\{C_1, \ldots, C_M\}$, the Marginal Adequacy Degree (MAD) of individual X to class C_i according to descriptor D_k, $\text{MAD}_{C_i}(x_k)$, is defined as follows;

$$\text{MAD}_{C_i}(x_k) = \mu_i^k(x_k), \tag{3.14}$$

where μ_i^k is the marginal distribution of descriptor D_k in the class C_i, $i \in \{1, \ldots, M\}$. The MAD is calculated via the density or frequency with which the specific marginal observation appears in the given class (Aguilar-Martin et al., 2002). In the case of a qualitative descriptor, the MAD is computed by taking into account the frequencies of the different modalities that the descriptor exhibits in a certain class. A density function is used if the descriptor is quantitative, in which case the height corresponding to the observed value of the individual inside the density function of the descriptor in the class is measured. The density function to chose has to be estimated for each descriptor, being the three following ones the most frequently used:

LAMDA classical function: This classical distribution function used by LAMDA[1] is based on the distribution function of a binomial variable:

$$\text{MAD}_{C_i}(x_k) = {\rho_{C_i,k}}^{x_k} \cdot (1 - \rho_{C_i,k})^{(1-x_k)}, \tag{3.15}$$

where $\rho_{C_i,k}$ stands for the average value of descriptor D_k in class C_i, and x_k is the normalised observed value of the individual X. Figure 3.4 shows some examples of LAMDA classical function, with different values of ρ: from 0.1 (top-left to bottom-right) to 0.9 (bottom-left to top-right). Note that these functions have one maximum value situated at the extreme, except for $\rho = 0.5$ when the distribution function is constant.

[1]Learning Algorithm for Multivariate Data Analysis (LAMDA) is a technique able to perform both supervised and unsupervised automatic learning. Based on fuzzy hybrid connectives, it's described in Appendix A.

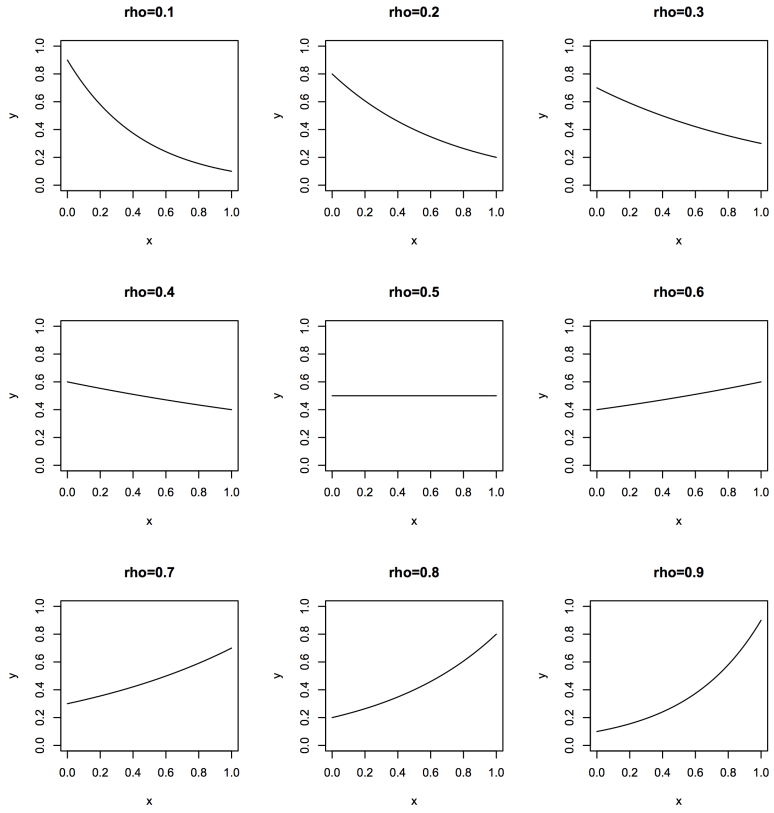

Figure 3.4: LAMDA classical function with different values of ρ

Gaussian function: Given a normal distribution:

$$f(x;\rho,s^2) = \frac{1}{s\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\frac{(x-\rho)^2}{s^2}},$$

with standard deviation value s and mean value ρ, the following normalised function is used:

$$\mathrm{MAD}_{C_i}(x_k) = e^{-\frac{1}{2}\cdot\frac{(x_k-\rho_{C_i,k})^2}{s^2_{C_i,k}}}, \tag{3.16}$$

where $s_{C_i,k}$ and $\rho_{C_i,k}$ are the standard deviation and mean values of descriptor D_k in class C_i, while x_k is the normalised observed value of the individual X. Figure 3.5 shows some examples of the Gaussian function, with a fixed value of $s = 0.1$ and different values of ρ: from 0.1 (left) to 0.9 (right). It is known that the maximum value of the Gaussian functions is located in $x = \rho$.

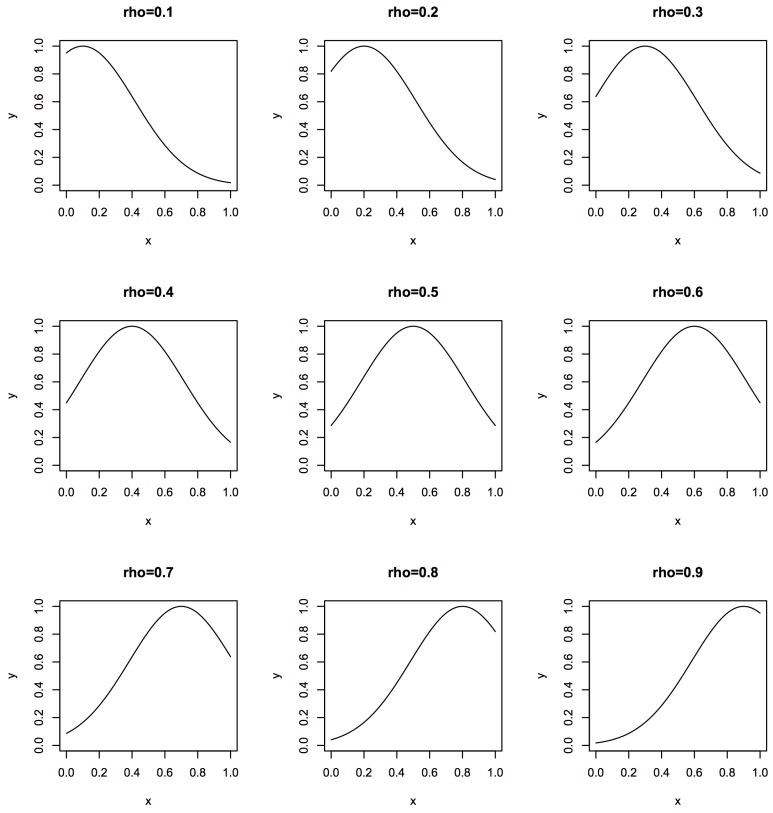

Figure 3.5: Gaussian function with value $s = 0.1$ and different values of ρ

Waissman function: The expression to use is a function defined by Waissman et al. (1998):

$$\mathrm{MAD}_{C_i}(x_k) = \rho_{C_i,k}^{\nu_{c_{i,k}}(x_k)} \cdot (1 - \rho_{C_i,k})^{1-\nu_{c_{i,k}}(x_k)},$$

where x_k is the normalised observed value of the individual X; $c_{i,k}$ is the centre of the distribution of descriptor k in class C_i as the median of the distribution of descriptor k of individuals belonging to class C_i; $\nu_c(x) = 1 - d(x, c)$ is a fuzzy number that measures the amount of presence of each value x around the centre of the distribution c, being $d(x, c)$ the distance between value x and centre c ($d(x, c) = |x - c|$); and $\rho_{C_i,k}$ stands for a measure of dispersion of descriptor k in class C_i computed as the inverse of the mean of distances between each value of descriptor k and the centre $c_{i,k}$, that is to say, $\rho_{C_i,k} = 1 - \frac{1}{N} \cdot \sum d(x_k, c_{i,k})$, being N the total number of individuals. Note that the higher the value of ρ, the more important

is that x_k had a value close to $c_{i,k}$ because x_k is more penalised when it moves away $c_{i,k}$. Also note that the mean of distances between each value and its centre is always lower than or equal to 0.5, therefore $\rho \geq 0.5$.

Figure 3.6 shows some examples of the Waissman function fixed the value of ρ to 0.7 and with different values of c: from 0.1 (top-left) to 0.9 (bottom-right).

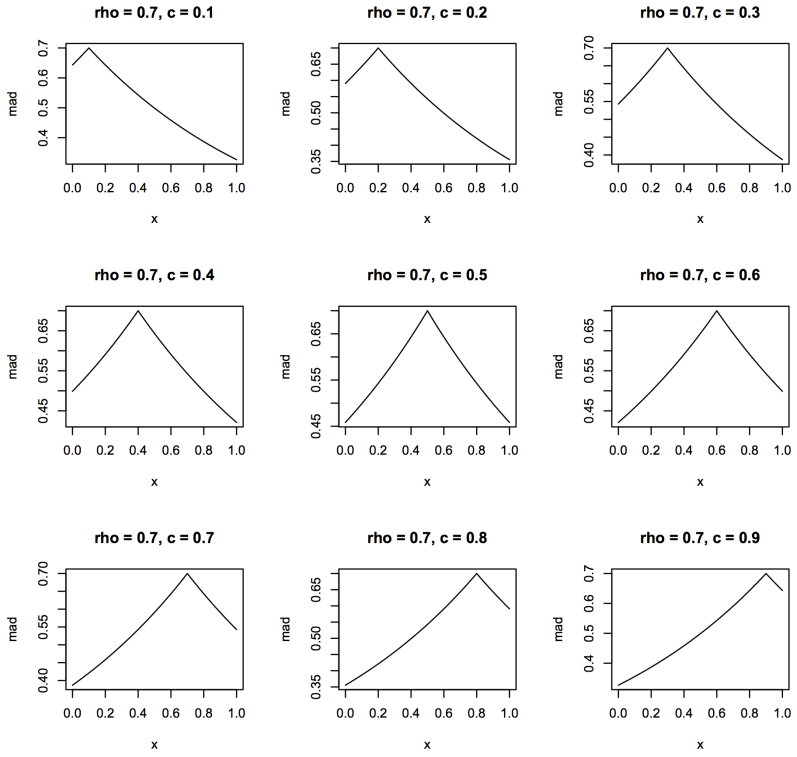

Figure 3.6: Waissman function with different values of the centre c

A classification is considered *coherent* when the differences between MADs are small enough for each class and each individual. The index of coherence will ensure that the Global Adequacy Degrees (GADs) are obtained from similar values of MADs, thus reflecting the fuzzy concept of 'coherent classification'.

Let us consider μ_{ijk} as the MAD of individual j to class i according to descriptor k, the *index of coherence* of classification $\mathcal{C}$ is defined as the following mean of differences ($\mathrm{MD}_{\mathcal{C}}$):

$$\mathrm{MD}_{\mathcal{C}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} max_{k,k'}|\mu_{ijk} - \mu_{ijk'}|}{M} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[max(\mu_{ijk}) - min(\mu_{ijk})]}{M}. \tag{3.17}$$

When all MADs are equal for each individual, then the coherence index will be 0. If each individual has associated a MAD of zero (0) and a MAD of one (1) for each class then the index of coherence will be N, where N is the number of individuals. Thus, the index of coherence range is $[0, N]$. Given that the lower $\mathrm{MD}_{\mathcal{C}}$ the more coherent is the classification $\mathcal{C}$, the following inverse standardisation function is proposed as the membership function of the *index of coherence* (Wu et al., 2009).

Definition 3.3. *The index of coherence of classification is given as follows:*

$$I_C(\mathcal{C}) = 1 - \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}[max(\mu_{ijk}) - min(\mu_{ijk})]}{M \cdot N}, \tag{3.18}$$

where $N \in \mathbb{N}$ is the number of individuals, $M \in \mathbb{N}$ is the number of classes of classification $\mathcal{C}$, and μ_{ijk} is the MAD of individual j to class i according to descriptor k.

Example 3.6 illustrates the computation of I_C in a toy example.

Example 3.6. *Let's consider two classifications $\mathcal{C}_1$ and $\mathcal{C}_2$ consisting of two (A and B) and three classes (C, D and E), respectively. The data set is composed of three individuals (i_1, i_2 and i_3, thus $N = 3$), and three descriptors (d_1, d_2 and d_3). Tables 3.1 and 3.2 show the computed MADs by using any of the functions described above.*

Table 3.1: MADs of individuals in classification $\mathcal{C}_1$

Class A	d_1	d_2	d_3
i_1	0.3	0.4	0.6
i_2	0.2	0.3	0.2
i_3	0.5	0.5	0.3

Class B	d_1	d_2	d_3
i_1	0.7	0.5	0.5
i_2	0.4	0.8	0.5
i_3	0.9	0.6	0.7

It's obvious that, in general, MADs of classification $\mathcal{C}_1$ are much more homogeneous than the ones of classification $\mathcal{C}_2$. Calculation of $I_C(\mathcal{C}_1)$ and $I_C(\mathcal{C}_2)$ implies, for each row of the above tables, the computation of the difference between the maximum and minimum MAD, as shown in Table 3.3. Its last row contains the sum of those differences for each analysed class.

Table 3.2: MADs of individuals in classification $\mathcal{C}_2$

Class C	d_1	d_2	d_3
i_1	0.1	0.6	0.4
i_2	0.7	0.2	0.3
i_3	0.4	0.8	0.3

Class D	d_1	d_2	d_3
i_1	0.1	0.7	0.3
i_2	0.5	0.2	0.7
i_3	0.9	0.3	0.4

Class E	d_1	d_2	d_3
i_1	0.9	0.2	0.7
i_2	0.6	0.9	0.6
i_3	0.7	0.6	0.1

Table 3.3: Differences between maximum and minimum MAD

	$\mathcal{C}_1$		$\mathcal{C}_2$		
	Class A	**Class B**	**Class C**	**Class D**	**Class E**
i_1	0.3	0.2	0.5	0.6	0.7
i_2	0.1	0.4	0.5	0.5	0.3
i_3	0.2	0.3	0.5	0.6	0.6
sum	0.6	0.9	1.5	1.7	1.6

Now $I_C(\mathcal{C}_1)$ and $I_C(\mathcal{C}_2)$ can be computed as the mean of these differences, that is to say, as the sum of the differences of each of their classes, pondering them by the number of considered classes:

$$\begin{aligned} I_C(\mathcal{C}_1) &= 1 - \frac{0.6 + 0.9}{2 \cdot 3} = 1 - \frac{1.5}{6} = 1 - 0.25 = 0.75 \\ I_C(\mathcal{C}_2) &= 1 - \frac{1.5 + 1.7 + 1.6}{3 \cdot 3} = 1 - \frac{4.8}{9} = 1 - 0.53 = 0.47 \end{aligned}$$

As it was expected, the coherence index for $\mathcal{C}_1$ is much higher than for $\mathcal{C}_2$, demonstrating its higher coherence.

3.4 Fourth criterion: dependency on external variables

In many cases, the relevance of the classifications obtained is evaluated by using external variables provided by experts and known as control variables. These control variables can be either a variable not used in the process of generation of the analysed classifications, or another classification with which a high level of compatibility can be required.

The dependency or not of a classification with respect to a control variable can be tested by ap-

plying the χ^2 non-parametric test computed by using the contingency table illustrated in Table 3.4, with $\{C_1 \dots C_i \dots C_M\}$ representing the classes of the considered classification; $\{D_1 \dots D_s \dots D_S\}$ the values of the external variable; and q_{is} the number of observations that take the value D_s in class C_i.

Table 3.4: Contingency table

Class	Descriptors, intervals or linguistic labels				Total classes
	D_1	D_2	...	D_S	
C_1	q_{11}	q_{12}	...	q_{1S}	M_{1+}
...	...	...	...	...	...
C_i	q_{i1}	q_{i2}	...	q_{iS}	M_{i+}
...	...	...	...	...	...
C_M	q_{M1}	q_{M2}	...	q_{MS}	M_{M+}
Total descriptors	M_{+1}	M_{+2}	...	M_{+S}	N

It is important to note that this criterion can be used directly when the control variables are qualitative. In the case of quantitative control variables, they must be previously discretised into intervals (D_s). The discretisation criterion will vary depending on the problem addressed (Dougherty et al., 1995; Kurgan and Cios, 2004; Ruiz et al., 2008).

Under the hypothesis of being the variable independent of the classification, the relative frequency with which members of different classes take different control variable values would not differ significantly. This hypothesis is tested by using

$$\chi^2 = \sum_{i=1}^{N} \sum_{s=1}^{S} \frac{(q_{is} - e_{is})^2}{e_{is}}, \tag{3.19}$$

where e_{is} is the number of expected cases under the hypothesis of independence and is defined as

$$e_{is} = \frac{M_{i+} \cdot M_{+s}}{N}.$$

For each classification, the dependency of each control variable with respect to the classification is studied, and those classifications that have a high dependency on these external variables will be chosen. For this reason, the statistic χ^2 must have a high value.

The range of χ^2 can vary according to the number of classes of the classification. For this reason, Tschuprow's coefficient (Tschuprow and Kantorowitcsch, 1939) is used to evaluate the degree of dependency on the control variable.

Definition 3.4. *Given a classification $\mathcal{C}$, its index of dependency on a control variable is defined as:*

$$I_D(\mathcal{C}) = \frac{\chi^2}{N \cdot \sqrt{M-1} \cdot \sqrt{S-1}}, \tag{3.20}$$

where N is the number of individuals, M is the number of classes of $\mathcal{C}$ and S is the number of unique values of the control variable, if it is qualitative, or the number of considered intervals in the discretisation if it is quantitative.

Note that $0 \leq I_D(\mathcal{C}) \leq 1$, and therefore this degree of dependency of a classification on a control variable could be interpreted directly as the membership function associated with the fuzzy concept 'dependency on a control variable'. Other possible interpretations of the value offered by this criterion rely on the concept of compatibility between the considered classification and the classification defined by the control variable.

3.5 Fifth criterion: accuracy of the predictive model

A high predictability of the model obtained from a classification ensures new individuals to be classified in the proper cluster. To this end, a criterion based on the achieved accuracy when performing supervised learning from a classification is defined.

Following the well-known concepts of precision and recall in machine learning, the fuzzy concept of accuracy of a particular classification is based on the precision and recall of the model. Precision of a class is the proportion of individuals assigned to that class that were correctly classified, while recall is the proportion of individuals of that class that have been classified in that class correctly. Precision and recall of a classification can be defined as the weighted average of precision and recall of its classes, with weights proportional to the cardinality of the classes. The *index of accuracy* of a classification is defined as the harmonic mean of its precision and accuracy values:

Definition 3.5. *Given a classification $\mathcal{C}$, its index of accuracy is defined as:*

$$I_A(\mathcal{C}) = 2 \cdot \frac{precision(\mathcal{C}) \cdot recall(\mathcal{C})}{precision(\mathcal{C}) + recall(\mathcal{C})}, \tag{3.21}$$

where $precision(\mathcal{C})$ and $recall(\mathcal{C})$ are the weighted averages of precision and recall of classes of $\mathcal{C}$, respectively.

The range of precision and recall is $[0, 1]$, so $0 \leq I_A(\mathcal{C}) \leq 1$ and therefore this index of accuracy could be understood as the membership function related to the fuzzy concept 'accuracy of the predictive model of the classification'.

3.6 Conclusions

In this chapter, a set of fuzzy criteria has been proposed and analysed, and modelled using a set of indexes to evaluate the set of considered classifications. Properties and usability of the defined criteria have been explained and proven.

The aims of the analysed criteria have been included in Table 2.1 of Section 2.1 to be compared with criteria used or defined in other works. As it can be seen, criteria developed in this chapter cover almost all the concepts used to assess a classification, including internal and external criteria.

On the one hand, four of the fuzzy criteria taken into account made reference to internal criteria, trying to determine if the classification structure is appropriate for the data. In that sense, first and second criteria are associated with the goals the user aims to achieve with that classification, that is to say, to have a classification with a useful number of classes in terms of information (having sufficient classes to acquire new knowledge) and manageability (not having to deal with too many classes to interpret), and obtaining a balanced or unbalanced classification according to the user preferences and expertise on the analysed data. Coherence index (third criterion) measures both compactness and separability of the classes by rewarding classifications in which individuals show a high adequacy degree in all descriptors of class in which they are included, and a low adequacy in descriptors of the rest of classes. Finally, accuracy index (fifth criterion) quantifies the prediction strength of each model associated with the considered classifications. On the other hand, dependency index (fourth criterion) mades reference to an external criteria in which the compatibility or dependency between each classification and a external variable named control variable is evaluated.

The proposed criteria need the definition a set of parameters in order to be able to adjust their utility to the actual problem analysed, despite default values can be used. For instance, the computation of usefulness index requires on the one hand the definition of the interval of the number

of classes for which this index must take the highest value and, on the other hand, the shape of the fuzzy number in order to define how this index should decrease as the number of obtained classes moves away the interval with maximum usefulness. In the second criterion (balanced classifications) user must decide if prioritising balanced or unbalanced classifications according to his expectations or to the knowledge of a priori distribution of the analysed individuals. The computation of the MADs involved in coherence index (third criterion) needs the selection of the density function to be used on each quantitative descriptor according to their shown distribution. Fourth criterion (dependency index) quantifies the compatibility level between each classification and a control variable provided by the user, while accuracy index (fifth criterion) can be evaluated by using any supervised learning technique preferred by the user. Note that not all indexes are useful in all situations. It is responsibility of the user to decide which indexes to use for assessing the considered classifications.

Once each considered index is computed for each classification, they must be used to choose the best classification according to them. Indexes, for instance, can be analysed sequentially by discarding in each step classifications not reaching a certain minimum threshold defined for each index. This sequential approach presents some deficiencies (Sánchez-Hernández et al., 2013) and therefore an aggregation of the indexes is advised by the author with the aim of selecting the most suitable classification. In next Chapter, a natural language generation (NLG) system to describe in a qualitative way the most important characteristics of classes of this chosen classification is designed and developed.

The proposed criteria are inspired on well-known concepts for clustering validation (Halkidi et al., 2001; Liu et al., 2010; Osei-Bryson, 2010; Yatskiv and Gusarova, 2005). The *useful number of classes* and *balanced classes* criteria have a marketing background, since they were defined to guarantee manageable segmentations (Casabayó, 2005). However, their implementations (I_U and I_B) in a fuzzy environment is a specific contribution of this thesis. The *coherence* criterion measures the compactness and separability of a given segmentation, that are common measures for clustering validation. Its implementation (I_C) is defined in a novel way via a normalised distribution function (Aguilar and López de Mántaras, 1982). Regarding the *dependency* criterion, there are different approaches in the literature to estimate the compatibility between the analysed segmentations and an *a priori* segmentation or an external variable. In the methodology presented, the proposed index I_D relies on the concept of dependency given by a χ^2 distribution. Finally, the *accuracy* criterion

and its associated index I_A have been defined as an aggregation of the widely-known recall and precision indicators (Osei-Bryson, 2010; Tibshirani and Walther, 2005; Xiong et al., 2009).

The most important contributions introduced in this chapter are the definitions of five fuzzy criteria and their corresponding membership functions including, specifically, the definition of a fuzzy number according to the user expectations in the first criterion, the characterisation and mathematical demonstration of the limits of the coefficient of variation used in the definition of the second criterion, the definition of the third criterion by using the concept of adequacy degree, the use of the conception of dependency to assess the relation or compatibility of each classification with an external variable in the fourth criterion and the inclusion of accuracy notion to complete the set of internal criteria for assessing classifications.

As future work, some of the criteria must be improved by generalising them in order to cover all possible needs of the end user. For instance, second criterion (balanced classes) would be generalised to compute the fitness of the distribution of individuals within classes on each distribution. If previous information about the desired distribution of individuals is available, a new index would compare this distribution with each classifications. Dependency index should also be improved by taking into account if the control variable is an ordinal variable. Finally, the parametrisation of the complete process should be formalised with the aim of driving the user to easily reach his expectations.

Chapter 4

Natural language-based automatic qualitative description of clusters

This chapter aims to design a natural language generation (NLG) system to describe qualitatively the most important characteristics of each class or cluster previously defined by means of a classification or clustering process. This system can be used for describing the best classification chosen by using the methodology defined in Chapter 3 in order to define a complete solution to be applied in machine learning (ML) problems. An adaptation of a generic architecture for data-to-text systems consisting in four stages is proposed. It includes the detection of the most relevant patterns of the data and the definition of a grammar that generates the natural language description of the defined clusters.

The use of unsupervised learning systems enables the definition of new classifications that were previously unknown from a set of individuals. Interpretation and description for the obtained classes requires an amount of technical knowledge the end user does not usually possess (Oja, 1983). For this reason, it is desirable to rely on an automated tool for the description of these classes. If this description is also done by using qualitative expressions, it facilitates the interpretation and understanding of the results. This will therefore improve the transmission of useful knowledge to experts who need to understand the profiles of the analysed items of the obtained clusters.

This work aims to provide a solution to this problem by automatically generating a natural language description of the major characteristics of each of the clusters of a segmentation. A tool

has been developed under the R environment (R Core Team, 2012). The base package has been used in addition to several of packages `gdata` (Warnes et al., 2012) and `fpc` (Hennig, 2010).

In order to construct natural sentences according to the specific domain, an ontology must be defined to encode domain knowledge (Gruber, 1993; Gómez-Pérez et al., 2004). This ontology should include the following aspects:

- **Type of the variables**. Knowing the type of the considered variables permits the system to merge properly the modality and variable in a sentence, specially if the specific construction of the sentence is not provided. Modalities of ordinal variables can be treated, in general, as adjectives of the corresponding variable.
- Is the modality an **adjective of the variable?** A more precise alternative this is the last candidate. next esc will revert to uncompleted text. o the previous point is to know if the modalities of a variable can actually be treated as adjectives of the variable. Only placing the modality before the name of the variable gets a proper construction.
- List of **modalities of the variables**. This information can be automatically gathered in the very first step of the system, but the order between modalities in ordinal or numerical variables must be defined. For instance, it allows the system to merge properly sentences of two or more consecutive modalities.
- **Name of individuals**. The individuals are by default referred as "individuals". In order to enrich the text, the proper name of individuals should be defined. With the aim of avoid repetitive mentions, alternative names can be provided. Moreover, the proportions of how many times each alternative appears in the text should be able to be defined in the ontology.
- **Specific construction**. There are always some variables that need the definition of how to construct the proper expression when talking about one of their modalities. For instance, a variable meaning the categorical location of points of sale must be included in the text with the expression "located in".

Reiter (2007) presents an architecture for data-to-text systems and more specifically, for those systems dealing with raw data obtained from sensors. This architecture is based on the one described in Dale and Reiter (2000), an architecture of a NLG system whose inputs are Artificial Intelligence (AI) knowledge bases.

Reiter presents a four-stage architecture with the next stages:

1. **Signal analysis**: detection of basic patterns in the numerical data. This stage is avoidable if input data is not numeric.

2. **Data interpretation**: analysis of the obtained patterns and other discrete data to infer messages about data and relations between patterns, events and messages.

3. **Document planning**: deciding which events will be mentioned in the text and the structure of the text, pointing out how the events will be related to each other.

4. **Microplanning and realisation**: generating the actual text based on the content and structure chosen in the previous stage.

This work adapts this architecture to systems dealing with well-structured data having subsets perfectly defined. Figure 4.1 shows a schematic diagram of the adapted architecture used in this work. Signal analysis stage receives data in a matrix form where each row corresponds to an individual that is described by variables stored by columns. The set of individuals is partitioned in classes. This stage select the variables that best explain the differences between the analysed classes, and computes the extreme frequencies (EF) and values of importance (VoI) described in Section 4.1. Data interpretation stage filters these extreme frequencies and values of importance by analysing type-A rules, as shown in Section 4.2. Document planning stage finds relations between the filtered extreme frequencies and values of importance by examining type-B rules according to the domain-specific ontology, and points out lexical modifications to be carried out on the messages associated to the values by employing type-C rules. Finally, activated rules detailed in Section 4.3 in conjunction with the ontology are used in Microplanning and realisation stage to sort and group the messages associated with each EF and VoI and to obtain the final text, processes detailed in Section 4.4. Next, the four stages are explained among the tasks involved in them.

4.1 Signal analysis

At this stage a data preprocessing step to select the variables that best explain the differences between the defined classes is conducted.

In figure 4.2 a diagram of the tasks included in this stage is shown. Quantitative variables are initially discretised in order to carry out the same analysis for all variables, to filter the significant

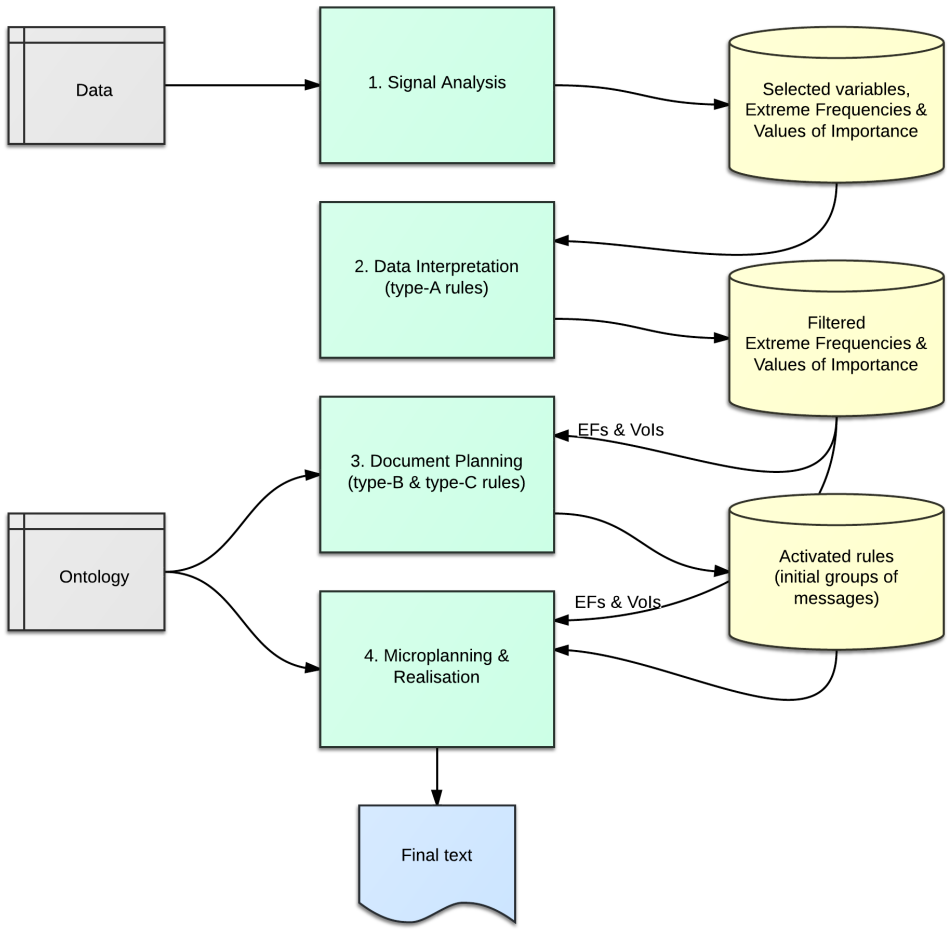

Figure 4.1: Diagram of the designed NLG system.

ones. After that, relevant values of importance and extreme frequencies are computed. Below the concrete tasks to be performed are detailed.

4.1.1 Discretisation

Discretisation is the process by which a continuous variable is transformed into a finite number of intervals associated with a discrete variable. The use of discrete variables, besides decreasing the computational cost of most data analysis processes, also facilitates the interpretation of the obtained results (Dougherty et al., 1995; Liu et al., 2002). Moreover, a **discretisation step** unifies

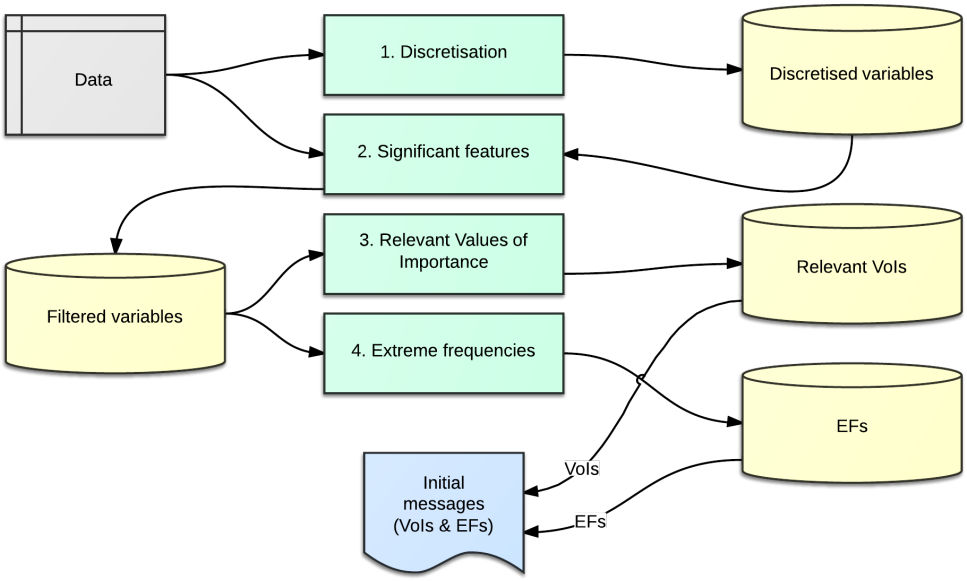

Figure 4.2: Diagram of the Signal analysis stage.

the methods to be used over any type of variables.

The studied techniques that have been applied in this approach are located in the area of qualitative reasoning. The use of qualitative modelling to represent the type of reasoning that people use to understand continuous aspects of the world is one of the main aspects of qualitative reasoning (Forbus, 2011). Discretisation becomes a mandatory step on qualitative methods when data presents excess of precision.

For instance, if the range of a numeric variable is unknown, it is not possible to decide if a concrete value is high or low. If this variable is discretised in advance, it would be easier to understand the actual extent of any value. Although its range is unknown, the variable can be discretised.

The existing methods of discretisation can be classified mainly into two categories: unsupervised and supervised Dougherty et al. (1995). Unsupervised methods do not consider the class to which the patterns belong. The used discretisation method has to take advantage that a class variable is provided within the data set, that is to say, it has to be supervised to ensure the maximum dependence between the discretised variable and the defined class.

One of the simplest supervised algorithms for discretisation is *1R* (Holte, 1993); *Minimum De-*

scription Length (MDLP) criterion (Rissanenm, 1978), *Information Entropy Maximisation (IEM)* (Fayyad and Irani, 1993), *maximum-entropy* (Wong and Chiu, 1987) and *D2* (Catlett, 1991) are entropy-based algorithms; there are also clustering-based algorithms like *K-means* (Tou and González, 1974); *IDD* (Ruiz et al., 2008) takes into account the order of the classes (if available) and is able to dealing with a large number of classes; statistics-based algorithms like *ChiMerge* (Kerber, 1992), *Chi2* (Liu and Setiono, 1997) and *StatDist* (Richeldi and Rossotto, 1995) provide justification of using the χ^2 statistic to measure the class-attribute interdependence. From among them, *CAIM* algorithm (Kurgan and Cios, 2004) presents excellent results in improving the performance of supervised ML, is able to automatically select the number of discrete intervals and maximises mutual class-attribute interdependence by using the same concepts introduced in the selection of significant features presented in the next subsection. *CAIM* is the general recommendation of the author of this work when there is no any other specific technique that better deals with peculiarities of the data.

4.1.2 Significant features

The qualitative description to be generated should emphasise the most available relevant features, i.e. the variables that best differentiate one class from others. That is why a test of independence between each variable and defined classes is conducted, discarding those variables that do not have a sufficient level of dependence with the class variable.

In order to measure the class-attribute interdependence, a Chi-square test of independence is performed on each of the considered variables. These non-parametric tests use a cross table of the analysed variable and the class variable called contingency table, like the one shown in Example 4.1. They are based on the χ^2 statistics and test if the observed distribution of the modalities of the variable along the class is due to chance. Each χ^2 distribution has a degree of freedom associated with it which depends on the number of the modalities of the variable and the number of classes. Under the hypothesis of being the variable independent to the classification the expected frequencies are computed and differences between them and the observed frequencies are aggregated into the χ^2 statistic. The null hypothesis of the test claims that the observed frequencies are not statistically different to the expected values, that is to say, variable and class are independent. If the χ^2 statistic is large enough, there are sufficient evidences to reject the null hypothesis, so the variable is considered dependent on the class.

There are two main requirements to apply there Chi-square tests of independence: the sample size is reasonably large and there should be at least 5 expected cases per modality and class. Each Chi-square test has its own peculiarities according to the nature of the analysed data. For instance, Yates' chi-squared test (Yates, 1934) performs a correction for continuity in which the analysed continuous variable has been discretised by rounding it. Cochran-Mantel-Haenszel test (Mantel, 1963) allows the comparison of two groups on a dichotomous response, the same as occurs with McNemar's test (McNemar, 1947): they use a $2x2$ contingency table. Portmanteau tests (Hosking, 1980) test for the presence of autocorrelation in time-series analysis. Finally, the best-known and most-used of these chi-squared tests is Pearson's chi-squared test (Pearson, 1900). It tests the dependence of categorical variables with any number of modalities and is the chosen test to be performed in this work.

Example 4.1. *Let H be the height of an international sampling of people, and N the nationality of each one of the people included in the study. Table 4.1 shows the frequency of the couple country-height in the data set used in this study, considering H as a discretised variable, with modalities "High", "Average" and "Low", that is to say, its absolute joint frequencies.*

Table 4.1: Contingency table of the example

	High	Average	Low	Total
Norway	30	9	1	40
Spain	30	50	20	100
Japan	20	70	80	170
Total	80	129	101	310

If H and N where independent, the absolute frequencies would be very similar to those shown in Table 4.2, which are the expected frequencies according to the total individuals of each modality and country.

The chi-squared test of independence evaluates the differences between the observed values (Table 4.1) and expected ones (Table 4.2) by applying Equation (4.1)

$$\chi^2 = \sum_{i=1}^{3}\sum_{h=1}^{3} \frac{(q_{ih} - e_{ih})^2}{e_{ih}}, \tag{4.1}$$

Table 4.2: Expected values in the example

	High	Average	Low	Total
Norway	10.3	16.6	13.0	40
Spain	25.8	41.6	32.6	100
Japan	43.9	70.7	55.4	170
Total	80	129	101	310

where q_{ih} is the number of cases observed with modalities $i \in N$ and $h \in H$, and e_{ih} is the number of cases expected under the null hypothesis of independence. The p-value obtained by comparing the value of the statistic ($\chi^2 = 58.97$) to a chi-squared distribution is almost zero, so null hypothesis can be rejected and the dependence between variable N and H is proved.

Despite discarded variables may not have a dependence on defined classes, combinations of them could present a level of dependence on the class variable. Example 4.2 shows a pair of variables that individually don't have any kind of dependence with defined class but the combination of their modalities shows a dependence on classes.

Example 4.2. *Let's consider a data set consisting on 200 individuals described by two variables and classified in two classes. Both variables have modalities "Low" and "High". If we compute the contingency table of Variable 1, independence of each separately variable and class is evident, as shown in Table 4.3. The same applies to Variable 2.*

Table 4.3: Example: variables independent on class but dependent when combining them

Var. 1	**Var. 2**	**Class**	**# of indiv.**
Low	Low	1	50
High	High	1	50
Low	High	2	50
High	Low	2	50

Class	**Variable 1**		Total
	Low	High	
1	50	50	100
2	50	50	100
Total	100	100	200

Let Variable 3 be a new variable consisting of combining modalities of Variables 1 and 2. Table 4.4 shows the new obtained data set and the contingency table corresponding with this new variable, demonstrating the dependence of this new variable and the class variable. Obtained value of χ^2

statistic with 3 degrees of freedom is 200*, whose corresponding p-value is almost* 0*.*

Table 4.4: Example: combination of independent variables becoming dependent on class

Variable 3	**Class**	**# of individuals**
Low–Low	1	50
High–High	1	50
Low–High	2	50
High–Low	2	50

Class	**Variable 3**				Total
	Low–Low	Low–High	High–Low	High–High	
1	50	0	0	50	100
2	0	50	50	0	100
Total	50	50	50	50	200

Example 4.3 demonstrates that variables not having any dependence on the class, even combining them, will be discarded.

Example 4.3. *Let's consider a similar data set to the previous one, consisting of 200 individuals described by two variables with modalities "Low" and "High", and classified in two classes. The contingency table included in Table 4.5 shows that Variable 1 (and even Variable 2) are independent on the class variable.*

Table 4.5: Example: variables independent on class and also independent when combining them

Var. 1	**Var. 2**	**Class**	**# of indiv.**
Low	Low	1	50
High	High	1	50
Low	Low	2	50
High	High	2	50

Class	**Variable 1**		Total
	Low	High	
1	50	50	100
2	50	50	100
Total	100	100	200

Again, Variable 3 is added by combining modalities of Variables 1 and 2. Table 4.6 shows the new obtained data set in addition with contingency table corresponding with this new variable, that

still remains independent on the class variable. Obtained value of χ^2 *statistic with 1 degree of freedom is* 0*, whose corresponding p-value is* 1*.*

Table 4.6: Example: combination of independent variables remaining independent on class

Variable 3	**Class**	**# of individuals**
Low–Low	1	50
High–High	1	50
Low–Low	2	50
High–High	2	50

Class	**Variable 3**				Total
	Low–Low	Low–High	High–Low	High–High	
1	50	0	0	50	100
2	50	0	0	50	100
Total	100	0	0	100	200

Therefore, instead of directly discarding non-dependent variables, a new process testing the dependence of new variables obtaining by combining these non-dependent variables must be carried out. In a first step, pairs of non-dependent variables are tested. If one of the new variables proves to be dependent on the class variable, it is introduced in the system as a new variable, discarding variables with which this new variable has been created. A second step tries to reproduce the process by testing the dependence of trios of non-dependent variables.

4.1.3 Relevant values of importance

Once selected those features showing more dependence with the considered classes, the goal is to develop a measure that assesses the differences in the distribution of the modalities of each variable along each class and the population. These measures point out, within each variable, the importance of each modality in the classes. To do this, the contingency table obtained in the stage of selecting significant features is used to compute a **value of importance (VoI)** for each of the absolute joint frequencies.

Definition 4.1. *Given a contingency table with the absolute joint frequencies, a VoI for each of the frequencies is computed as follows:*

$$VoI_{is} = sign(q_{is} - e_{is}) \cdot \frac{(q_{is} - e_{is})^2}{e_{is}}, \tag{4.2}$$

where q_{is} is the number of cases observed with modality s in class i and e_{is} is its expected value. Thus, a VoI is computed for each class, variable and modality.

Each VoI is the addend of the chi-squared statistic taking into account if the observed value is greater than the expected one or vice versa. In other words, the VoIs are related to the importance of their contribution in the χ^2 test of independence between the distribution of the modalities of a variable along the defined classes. The magnitude of each value of importance inside a variable defines which ones must be included in the subsequent stage. Each variable is analysed separately because the magnitudes of the VoIs are not comparable between variables.

A need for analysing some thresholds defining the relevance of the values of importance is shown in Example 4.4. Moreover, the range of values of importance varies when computed for different variables so these thresholds would have to be computed independently for each variable. Two different levels of relevance of the VoIs are defined by using the Definitions 4.2 and 4.3.

Example 4.4. *Let's consider the study conducted in Example 4.1. Table 4.7 shows the values of importance associated with the absolute frequencies of table 4.1, computed by applying Equation (4.2).*

Table 4.7: Values of importance in the example

	High	Average	Low
Norway	37.5	-3.5	-11.1
Spain	0.7	1.7	-4.9
Japan	-13.0	0.0	10.9

As it can be seen, the most important difference lies in norwegian tall people ($VoI_{Norway,High} = 37.5$): they are much taller than people in the sample. Its relevance in the differentiation between classes and population is greater than other high values of importance, like the related to "Japan-High" ($VoI = -13.0$) and "Norway-Low" ($VoI = -11.1$) in a negative way, i.e. their proportion

in their classes are much lower than in the population, and in "Japan-Low" (VoI = 10.9). It is convenient to note the different relevance of these two groups of values of importance. The latter ones are relevant but not as much as the first one.

Definition 4.2. *Given a variable V, which has proved to be dependent on the classification $\mathcal{C}$, the set Z_V is defined as the combination of the VoIs in absolute values of the variable V according to the classification $\mathcal{C}$.*

Note that the number of VoIs computed for variable V is $M * S$, where M is the number of classes of classification $\mathcal{C}$ and S is the number of modalities of the variable V.

Definition 4.3. *Given the set Z_V containing the VoIs of the variable Z according to the classification $\mathcal{C}$, and a discretisation of its values in K subsets ($Z_{V,1}$ to $Z_{V,k}$ such that $\forall z_{V,i} \in Z_{V,i}$ and $\forall z_{V,j} \in Z_{V,j}$, it satisfies $z_{V,i} < z_{V,j}$ $\forall i, j | i < j$, the VoI very relevant are those that belong to $Z_{V,K}$ while the VoI relevant are those that belong to $Z_{V,K-1}$.*

Any clustering technique can be applied to detect the relevance of the VoIs associated with a variable. In this application K clusters are obtained and therefore $K-1$ thresholds. In this work a univariate K-means is applied to obtain $K = 3$ clusters and 2 thresholds. This technique has been chosen due to be one of the most commonly known clustering techniques, to the simplicity of its implementation, to its linear computational cost, and to the fact that it works fine when the clusters to detect are similar in size (Xiong et al., 2009). VoIs assigned to the cluster with higher values are treated as very relevant VoIs (high relevance) while the ones assigned to the next cluster are considered simply relevant VoIs (medium relevance). Finally, VoIs allocated to the rest of clusters are considered non-relevant (low relevance) and therefore discarded of next stages. Example 4.5 illustrates the application of this method.

Example 4.5. *Let's consider the VoIs shown in Table 4.7. The variable Z_H associated with the variable 'Height' is*

$$Z_H = \{37.5, 3.5, 11.1, 0.7, 1.7, 4.9, 13.0, 0.0, 10.9\}.$$

Note the positive signs of all values. This has be done because the importance of a VoI doesn't depend on its sign. It is as important an extremely low frequency than an extremely high one.

The application of a K-means technique on Z_H, with $K = 3$, produces the thresholds $h_1 = 25.25$ and $h_2 = 7.9$. These values are shown in Figure 4.3 in the form of an horizontal line, alongside the analysed VoI.

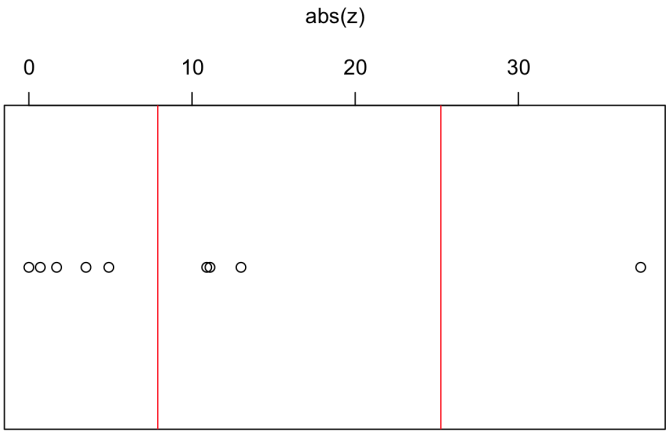

Figure 4.3: The three clusters found in the VoIs of the example.

With this analysis the VoIs greater than $h_1 = 25.25$ *are defined as very relevant, as in the case of the pair "Norway-High", with an associated VoI of* 37.5*. It also recognises as simply relevant VoIs the ones greater than* $h_2 = 7.9$*, as the pairs "Norway-Low", "Japan-High" and "Japan-Low", with the related VoIs* -11.1*,* -13.0 *and* 10.9*, respectively. Finally, the messages associated a VoIs with an absolute value less than* $H_2 = 7.9$ *are definitely filtered and discarded.*

4.1.4 Extreme frequencies

In the previous subsection the values that more differentiate the classes between them have been emphasised, considering the joint frequency distribution, i.e. taking into account all classes at once. In this subsection each class is managed separately, i.e. the conditional frequencies on each of the classes are examined.

The qualitative description must include a reference to those variables presenting some modalities with **extreme frequency (EF)** in order to highlight those modalities that encompass most of individuals in the class, or in which it have been observed very few individuals with that particular modality.

With that goal a matrix of conditional distributions of the variable in reference of each class is computed. It is necessary to study some thresholds in order to get a formal definition of what does mean that a frequency conditioned to a class is an extreme value, as shown in Example 4.6.

Example 4.6. *Let's consider again the study presented in Example 4.1. Table 4.8 shows the*

conditional relative frequencies, that is to say, the distribution of the modalities (High, Average and Low) along the three classes (Norway, Spain and Japan). Also a matrix with the joint relative distribution (and the marginal distributions) is included.

Table 4.8: Conditional and joint distributions in the example

	High	Average	Low	*sum*
Norway	0.75	0.22	0.02	1
Spain	0.3	0.5	0.2	1
Japan	0.12	0.41	0.47	1

	High	Average	Low	*sum*
Norway	0.097	0.029	0.003	0.13
Spain	0.097	0.161	0.065	0.32
Japan	0.065	0.226	0.258	0.55
sum	0.26	0.42	0.33	1

The most extreme frequency seems to be located in the pair "Norway-Low": almost none of norwegian people is short (2%). Some other frequencies could be highlighted so a threshold to decide which ones to remark is required.

Definition 4.4. *Given a classification $\mathcal{C}$, the set W is defined as the combination of the relative frequencies conditioned to the class of all the variables.*

Definition 4.5. *Given the set W containing the conditional frequencies of all the variables according to classification $\mathcal{C}$, the extreme frequencies are defined as those which are less or greater than the percentiles 1% and 99% of the set W.*

Note that the fact of working with percentiles leads the number of detected extreme frequencies to be directly proportional to the number of analysed variables, thus increasing the length of the descriptive text in those cases where the number of variables is high.

Example 4.7. *Let's consider the conditional frequencies shown in Table 4.8. To detect the extreme frequencies all the analysed dependent variables must be considered. In this case, the use of this unique variable is sufficient to illustrate the detection of the extreme frequencies. Thus, the set W associated with this table is*

$$W = \{0.75, 0.22, 0.02, 0.3, 0.5, 0.2, 0.12, 0.41, 0.47\}.$$

The percentiles 1% and 99% of W are $p_1 = 0.028$ and $p_{99} = 0.73$, respectively. Therefore, the conditional frequencies considered as extreme frequencies are 0.02 (less than p_1) and 0.75 (greater than

p_{99}), associated respectively to the pairs "Norway-Low" and "Norway-High". Then, the messages associated with relative frequencies inside the interval $]p_1, p_{99}[$ are discarded.

4.1.5 Initial messages

Each VoI and EF has an associated initial message. Each message has the following attributes:

- **Type of message**: either VoI or EF.
- **Class**: the class to which the message refers.
- **Variable**: the variable to which the message refers.
- **Modality**: the modality of the variable to which the message refers.
- **Sign**: low or high (negative or positive), it indicates the semantics of the message. For example, a negative message suggests that the number of individuals showing a certain modality of a variable in a certain class is very low either in relative terms by contrasting the class with the rest of the population (VoI) or in absolute terms by comparing the frequency of the modality with the other modalities of the variable (EF).
- **Value**: associated value of the message, either its own VoI or the relative frequency.

VoIs will have an extra attribute, which will be referred to the significance of the VoI: high, medium or low.

Also note that in the examples the associated messages have been abbreviated for showing their purpose clearer. The phrases in the final text must be complete. These complete sentences are shown below, beside the notation used in the examples included in the sections below.

- **(VoI+)**, the message is related to a VoI of positive sign: *the proportion of individuals with the modality M in the variable V is (much) greater than in the population.*
- **(VoI–)**, the message is related to a VoI of negative sign: *the proportion of individuals with the modality M in the variable V is (much) lower than in the population.*
- **(EF+)**, the message is related to an EF of positive sign: *(almost) all individuals have the modality M in the variable V.*
- **(EF–)**, the message is related to an EF of negative sign: *(almost) none of the individuals has the modality M in the variable V.*

4.2 Data interpretation

Data interpretation is the second stage of the four-stage architecture defined by Reiter and introduced previously. In the current work, this stage is responsible for detecting basic relations between the messages obtained in the previous stage.

The main aim of this is stage is to look for messages offering similar information in order to avoid repetitions. Once two related messages have been detected, if one of them is implied by or is redundant with the other one, it is discarded and the first one is prioritised. For instance, if a message says that all individuals of a class present a certain modality, while other one says that the proportion of individuals showing this same modality is higher than in the population, the second message is discarded because it is implied by the first one and therefore it is redundant.

Rules for discovering these relations and for discarding redundant messages look for groups of messages sharing the attributes are summarised in Table 4.9. This table informs, for each rule, which attributes of the messages are used to detect the relations. Remember that type of a message is either EF or VoI, the type of value which the message comes from, and sign stands for the sign of the value (positive or negative). For example, rule A.1 will analyse messages that at least share the same class, variable and type.

Table 4.9: Summary of type-A rules, showing which attributes are checked by each rule

Rule	Class	Variable	Modality	Type	Sign
A.1	Yes	Yes		Yes	
A.2	Yes	Yes	Yes		Yes
A.3	Yes	Yes			
A.4		Yes	Yes	Yes	Yes
A.5		Yes	Yes	Yes	

At the end of this stage, a weight is assigned to each message, initially set to 1. After applying each type-A rule, the weight of each discarded message is decreased by one unit, and the weights of messages that have caused this discard (prioritised messages) are increased. Once explored all rules, messages with a null or negative weight are definitely discarded: only messages with more prioritisations than discards are passed to the next stage.

In the following subsections, an exhaustive description of the different rules for discarding mes-

sages is introduced. In each subsection a different rule is explained and analysed. In addition, examples are given to better understand its meaning and aim.

4.2.1 Rule A.1: discarding negative messages

This rule avoids negative messages of a variable in reference to a class when there are other positive messages of the same variable and class but about a different modality. In general, if most of the individuals of a class present a certain modality (positive message), the other modalities are also highlighted in the form of a negative message. That is to say, the existence of these positive messages about a modality of a variable implies the existence of negative messages of the other modalities of the variable, so these latter messages must be avoided.

Given that EFs are more descriptive, messages of this type have to be treated in a different way than the messages of typeVoIs. Below the different way in what messages derived from EFs or VoIs is detailed.

Discarding messages related to EF: if there are positive and negative messages of type EF, the negative ones must be discarded and the positive ones, prioritised. A positive EF means that (almost) all the individuals in a class presents a certain modality, so it's very usual getting negative EF for the rest of modalities of the variable. These negative messages can be avoided. Examples 4.8 and 4.9 detail this statement in the cases of variables with two or three modalities, respectively.

Example 4.8. *Let V be a binary variable with the two modalities "No" and "Yes", and the following messages obtained from EFs related to the same class:*

- *All individuals are "No" (EF+);*
- *None of the individuals is "Yes" (EF–).*
  Conduct to:
- *All individuals are "No" (discard the negative EF message).*

Example 4.9. *Let V be a variable with the modalities "Red", "Green" and "Blue". The following EFs are related to the same class:*

- *Almost all individuals are "Red" (EF+);*

- *Almost none of the individuals is "Green" (EF–);*
- *Almost none of the individuals is "Blue" (EF–).*

  Conduct to:
- *Almost all individuals are "Red" (discard the negative EFs messages).*

Discarding messages related to VoI: if there are positive and negative messages of type VoI for the same binary variable (two modalities), the negative ones must be discarded and the positive ones, prioritised. A positive VoI of a binary variable means that the proportion of individuals presenting a certain modality is (very) high. On the one hand, this implies that positive VoI of the other modality can't be obtained. On the other hand, it is very probable to get a negative VoI for that other modality. In this case, this negative message must be discarded. Note that the variable must be binary because otherwise it is possible to count with two positive VoIs so discarding the negative VoIs can lead to a loss of information. Examples 4.10 and 4.11 detail this statement in the cases of variables with two or three modalities, respectively.

Example 4.10. *Let V be a binary variable with the two modalities "No" and "Yes". The following VoIs are related to the same class:*

- *The proportion of "No" is high (VoI+);*
- *The proportion of "Yes" is low (VoI–).*

  Conduct to:
- *The proportion of "No" is high (discard the negative VoI message).*

Example 4.11. *Let V be a variable with the modalities "Red", "Green" and "Blue". The following VoIs are related to the same class:*

- *The proportion of "Red" is high (VoI+);*
- *The proportion of "Green" is low (VoI–);*
- *The proportion of "Blue" is low (VoI–).*

  Conduct to:
- *Negative VoIs can't be discarded due to a loss of information.*

To sum up, different messages can be affected by this rule if they are related to the same class, variable and type.

4.2.2 Rule A.2: discarding messages obtained from VoIs

This rule is made to avoid VoI messages about a modality of a variable when there are other EF messages of the same modality and variable with the same sign. As it was said before, EF messages are more descriptive than VoI so the latter ones can be discarded in favour of the first ones.

Proposition 4.1. *Given a positive (negative) message coming from type EF, it implies the same positive (negative) message of type VoI.*

Proof. Let's consider a positive EF message corresponding to the class $\mathcal{C}$, the variable V and the modality M. It means that most of individuals in class $\mathcal{C}$ have the modality M for the variable V. If there not exists the same positive VoI (same class, variable and modality), i.e. the proportion of M in $\mathcal{C}$ is not important, then the proportion of M in the other classes is similar to the proportion in $\mathcal{C}$. This implies that variable V has the same distribution in all classes, so variable V is independent on the class variable. This leads us to a contradiction because in the previous stage only variables dependent on the class variable were selected, so the existence of the positive VoI is proved. The same proof can be applied for negative messages. □

Examples 4.12 and 4.13 instances the application of rule A.2 in the case of positive and negative messages, respectively.

Example 4.12. *Let V be a variable with the modalities "Red", "Green" and "Blue". The following messages are related to the same class:*

- *Almost all individuals are "Red" (EF+);*
- *The proportion of "Red" is high (VoI+).*
  Conduct to:
- *Almost all individuals are "Red" (discard the VoI message).*

Example 4.13. *Let V be a variable with the modalities "Red", "Green" and "Blue". The following messages are related to the same class:*

- *Almost none of the individuals is "Green" (EF–);*
- *The proportion of "Green" is low (VoI–).*
  Conduct to:
- *Almost none of the individuals is "Green" (discard the VoI message).*

4.2.3 Rule A.3: discarding modalities

Rule A.3 avoids information overload that occurs when all modalities of a variable are highlighted within a certain class. If this occurs, positive messages are prioritised giving priority to EFs over VoIs, discarding the rest of messages.

Examples 4.14 to 4.17 illustrate the application of this rule. Let V be a variable with the modalities "small", "medium" and "big". The following messages are related to the same class.

Example 4.14. *Mixed-type messages with almost one positive EF:*

- *Almost all individuals are "medium" (EF+);*
- *Almost none of the individuals is "big" (EF–);*
- *The proportion of "small" is very low (VoI–).*
  Conduct to:
- *Almost all individuals are "medium" (discard the negative messages).*

Example 4.15. *Mixed-type messages without any positive EF:*

- *Almost none of the individuals is "medium" (EF–);*
- *Almost none of the individuals is "big" (EF–);*
- *The proportion of "small" is very high (VoI+).*
  Conduct to:
- *The proportion of "small" is very high (discard the negative messages).*

Example 4.16. *All messages are related to EFs:*

- *Almost all individuals are "medium" (EF+);*
- *Almost none of the individuals is "big" (EF–);*
- *Almost none of the individuals is "small" (EF–).*
  Conduct to:
- *Almost all individuals are "medium" (discard the negative messages).*

Note that this example is equivalent to Ex. 4.9. As the rules are applied sequentially, rule A.1 will be the one to be used.

Example 4.17. *All messages are related to VoIs:*

- *The proportion of "medium" is high (VoI+);*
- *The proportion of "big" is low (VoI–);*
- *The proportion of "small" is low (VoI–).*

  Conduct to:
- *The proportion of "medium" is high (discard the negative messages).*

Note that this example is equivalent to Ex. 4.11. As the rules are applied sequentially, rule A.1 will be the one to be used.

4.2.4 Rule A.4: discarding variables (messages of the same sign)

Rule A.4 has been designed to avoid situations in which the same information is highlighted in all the analysed classes, i.e. there exist messages about all classes concerning the same variable, modality, type and sign. In this case, rule A.4 will discard all the involved messages. These messages share variable and modality, but in order to ensure that there is no loss of information by discarding too many messages, they also must be the same type and even the same sign. Furthermore, they cannot all be positive at a time, because if so, the variable would have been discarded in previous steps for its independence with the classes. Examples 4.18 and 4.19 illustrate the behaviour of this rule when dealing with messages coming from EFs and VoIs, respectively.

Example 4.18. *Discarding extreme frequencies:*

- *Almost none of the individuals in class $\mathcal{C}_1$ is "medium" (EF–);*
- *None of the individuals in class $\mathcal{C}_2$ is "medium" (EF–);*
- *Almost none of the individuals in class $\mathcal{C}_3$ is "medium" (EF–).*

  Conduct to:
- *Discard all messages.*

Example 4.19. *Discarding values of importance:*

- *The proportion of "medium" in class $\mathcal{C}_1$ is low (VoI–);*
- *The proportion of "medium" in class $\mathcal{C}_2$ is very low (VoI–);*
- *The proportion of "medium" in class $\mathcal{C}_3$ is low (VoI–).*

  Conduct to:
- *Discard all messages.*

4.2.5 Rule A.5: discarding variables (messages of different sign)

Analysing in more detail rule A.4, it is detected that messages sharing type, variable and modality –when they cover all the classes– may be redundant even when their sign is different.

In cases where only one of the classes (let's call it $\mathcal{C}$) has a positive message (therefore, the other classes have it negative), it means that $\mathcal{C}$ not only has a high percentage of individuals with the shared modality, but most of elements that present that modality are classified in that class. This new information is attached to the positive messages affected by rule A.5, after discarding the negative ones.

Examples 4.20 and 4.21 show the behaviour of this rule when dealing with messages coming from EFs and VoIs, respectively. Note the addition made in the sentences conducted.

Example 4.20. *Discarding extreme frequencies concerning the same variable:*

- *Almost none of the individuals in class $\mathcal{C}_1$ is "medium"(EF–);*
- *Almost all individuals in class $\mathcal{C}_2$ are "medium" (EF+);*
- *None of the individuals in class $\mathcal{C}_3$ is "medium" (EF–).*
  Conduct to:
- *Almost all individuals in class $\mathcal{C}_2$ are "medium" (moreover, almost all "medium" individuals are in this class) (discard the negative messages and amplify the positive sentence).*

Example 4.21. *Discarding values of importance concerning the same variable:*

- *The proportion of "medium" in class $\mathcal{C}_1$ is low (VoI–);*
- *The proportion of "medium" in class $\mathcal{C}_2$ is high (VoI+);*
- *The proportion of "medium" in class $\mathcal{C}_3$ is very low (VoI–).*
  Conduct to:
- *The proportion of "medium" in class $\mathcal{C}_2$ is high (moreover, most of "medium" individuals are in this class) (discard the negative messages and amplify the positive sentence).*

4.3 Document planning

Document planning is the third stage of the analysed four-stage architecture. This stage is responsible for two basic tasks. On the one hand, to discover relations among messages in order to combine

them in the final text. On the other hand, to identify specific modifications to be carried out on the sentences according to the semantics of the modalities or variables of these messages. It's important to note that these tasks are only addressed to detect and list relations and possible modifications. The effective execution of these modifications is carried out in the next stage corresponding to the realisation.

To identify the modifications to be carried out on the messages, an ontology must be defined to encode domain knowledge. This ontology will be used again in the fourth stage of the NLG system.

In the following subsections these two tasks are analysed and detailed, providing examples for each step.

4.3.1 Discovering relations among messages

In order to achieve a natural and compact final text, messages that are related to each other must appear together in one sentence. The discovery of the relations among messages is performed through the application of the rules defined in the following subsections. All these rules merge messages concerning the same class, because the final text is organised by classes. These rules are summarised in Table 4.10. This table shows, for each rule, which attributes of the messages have to match to be considered in the same sentence. For example, rule B.1 will merge messages sharing class, variable, type and sign.

Table 4.10: Summary of the activation of type-B rules, showing which attributes are checked by each rule

Rule	**Class**	**Variable**	**Modality**	**Type**	**Sign**
B.1	Yes	Yes		Yes	Yes
B.2	Yes	Yes			Yes
B.3	Yes		Yes	Yes	
B.4	Yes	Yes			
B.5	Yes			Yes	Yes

At the end of this phase, each message will be associated with a vector of activated rules to be applied in the final stage of realisation. Following subsections describe exhaustively the different rules designed for discovering relations among messages. Each subsection explains and analyses a

different rule, including examples to better understand its meaning and aim.

Rule B.1: merging modalities of an ordinal variable

This rule groups messages sharing the same modality of an ordinal variable. If some modalities of an ordinal variable are highlighted, they are merged in the proper way by applying this rule. In addition, if the modalities highlighted are two or more consecutive ones, the constructed sentence will take into account the interval containing these modalities.

As it is shown in Example 4.22, EFs and VoIs cannot be merged because if so, the obtained sentence would be overcomplicated and there would be no gain by merging those messages of different type.

Example 4.22. *Analysing messages of different type:*

- *Almost all individuals are "weak" (EF+);*
- *The proportion of "strong" is high (VoI+).*
  Would conduct to:
- *Almost all individuals are "weak" and the proportion of "strong" is low (it's better still not to merge these different type messages).*

The same occurs with messages of different sign, as seen in Example 4.23, so this rule only takes into account messages sharing not only class and variable, but also sign and type.

Example 4.23. *Analysing messages of different sign:*

- *The proportion of "weak" is high (VoI+);*
- *The proportion of "strong" is low (VoI–).*
  Would conduct to:
- *The proportion of "weak" is high while the proportion of "strong" is low (it's better still not to merge these different sign messages).*

Examples 4.24 to 4.26 illustrate the behaviour of rule when dealing with messages of the same sign. In all the cases the subject has had to be changed to plural in order to construct a truth sentence. In Example 4.24 messages to be merged have two non-consecutive modalities Note that the grammatical number of the subject must be changed to plural.

Example 4.24. *Merging messages of the same sign with non-consecutive modalites. Let's assume that the modalities of the analysed variable are "inexistent", "weak", "regular", "strong" and "very strong":*

- *The proportion of "weak" is high (VoI+);*
- *The proportion of "strong" is high (VoI+).*
  Conduct to:
- *The proportions of "weak" and "strong" are high (merge messages and change the grammatical number of the subject to plural).*

In example 4.25 the messages considered have consecutive modalities and some of them have extreme modalities. Note that only the non-extreme modality is mentioned.

Example 4.25. *Merging messages of the same sign with consecutive and extreme modalites. Let's assume that the modalities of the analysed variable are "inexistent", "weak", "regular", "strong" and "very strong":*

- *The proportion of individuals with the modality "strong" in variable competition is low (VoI–);*
- *The proportion of individuals with the modality "very strong" in variable competition is low (VoI–).*
  Conduct to:
- *The proportions of individuals with a competition greater than or equal to "strong" are high (merge messages and change the grammatical number of the subject to plural).*

In example 4.26 messages with more than two consecutive and non-extreme modalities are merged. In this case, the interval formed with the consecutive modalities is mentioned.

Example 4.26. *Merging messages with more than two consecutive modalities. Let's assume that the modalities of the analysed variable are "very low", "low", "regular", "high" and "very high":*

- *The proportion of individuals with a "low" competition is high (VoI+);*
- *The proportion of individuals with a "regular" competition is high (VoI+);*
- *The proportion of individuals with a "high" competition is high (VoI+).*
  Conduct to:
- *The proportions of individuals with a competition between "low" and "high" are high (merge messages).*

Rule B.2: merging modalities

Rule B.2 groups messages in a similar way than in rule B.1, but in this case not only ordinal variables are considered. Moreover, this rule also merges messages sharing class and variable, without the need of sharing type of message.

This rule deals only with messages not affected by previous rule and consists of two steps. In a first step messages sharing sign are merged while a second step is needed for merging the rest of messages (those with different sign). Note that this rule is a very generic rule that will affect most of the analysed messages.

Examples 4.27 and 4.28 show the behaviour of this rule when dealing with messages of the same sign and type. When sharing sign, the obtained sentence is much more compact than when dealing with messages of different sign.

Example 4.27. *VoIs of the same sign:*

- *The proportion of individuals with modality "green" is low (VoI–);*
- *The proportion of individuals with modality "red" is low (VoI–).*
  Conduct to:
- *The proportions of individuals with modalities "green" and "red" are low (merge messages).*

Example 4.28. *EFs of the same sign:*

- *None of the individuals is "green" (EF–);*
- *None of the individuals is "red" (EF–).*
  Conduct to:
- *None of the individuals is "green" nor "red" (merge messages).*

Examples 4.29 and 4.30 illustrate the behaviour with messages of the same type but with different sign.

Example 4.29. *VoIs of different sign:*

- *The proportion of individuals with modality "green" is low (VoI–);*
- *The proportion of individuals with modality "red" is high (VoI+).*
  Conduct to:

- *The proportion of individuals with modality "green" is low while the proportion of them with modality "red" is high (merge messages).*

Example 4.30. *EFs of different sign:*

- *None of the individuals is "green" (EF–);*
- *Almost all individuals are "red" (EF+).*
  Conduct to:
- *None of the individuals is "green" while almost all of them are "red" (merge messages).*

Note that the negative EF would have been discarded by rule A.1 if the current variable had only these two modalities. In this case, rule B.2 would not apply.

Finally, Examples 4.31 and 4.32 instances the application of rule B.2 with messages of different type.

Example 4.31. *Messages with the same sign but different type:*

- *Almost all individuals are "weak" (EF+);*
- *The proportion of "strong" is high (VoI+).*
  Conduct to:
- *Almost all individuals are "weak" and the proportion of "strong" is high (merge messages).*

Example 4.32. *Messages of different type and sign:*

- *Almost all individuals are "weak" (EF+);*
- *The proportion of "strong" is low (VoI–).*
  Conduct to:
- *Almost all individuals are "weak" while the proportion of "strong" is low (merge messages).*

Rule B.3: merging variables with the same modalities

It's usual to deal with different variables having the same modalities. Often quantitative variables are discretised by using the same labels ("low", "normal", "high", etc.) or ordinal variables with a similar meaning use the same modalities ("deficient", "regular", "good", "excellent" and so on). Rule B.3 merge messages of the same type which share the modality. It firstly filters messages

affected by previous type-B rules. In a first step, messages of the same sign are analysed and in a second step, the rest of messages are taken into account.

When dealing with categorical variables, two or more messages can be merged by rule B.3 if they refer to the same modality, regardless of the rest of modalities. In the case of handling with messages relative to ordinal variables, messages can be merged only if their corresponding variables share all the modalities, because the meaning of a certain modality can differ between variables according for example to the number of modalities of the variable.

Examples 4.33 to 4.37 show the behaviour of this rule when dealing with messages of the same sign and type. Among them, Examples 4.33 to 4.36 deal with messages coming from EFs. When trying to merge EF messages associated with absolute values (they are related to the 100% or 0% of the individuals of the class), the compaction is carried out in a simple way, as shown in Examples 4.33 and 4.34.

Example 4.33. *Absolute positive EFs of the same sign about two ordinal variables with the same exact modalities.*

- *All individuals have an "excellent" communication (EF+);*
- *All individuals have an "excellent" quality (EF+).*
  Conduct to:
- *All individuals have an "excellent" communication and quality.*

Example 4.34. *Absolute negative EFs of the same sign about two categorical variables sharing the analysed modality.*

- *None of the individuals has a "green" image (EF–);*
- *None of the individuals has a "green" building (EF–).*
  Conduct to:
- *None of the individuals has a "green" image nor building.*

In the case of dealing with non-absolute EF messages (those referring to less than 100% or more than 0% of the individuals of the class), the compaction of the sentence cannot be done in a simple way due to a possible lack of information or to an untrue statement, as shown in Examples 4.35 and 4.36.

Example 4.35. *Non-absolute positive EFs of the same sign about two ordinal variables with the same exact modalities.*

- *Almost all individuals have an "excellent" communication (EF+);*
- *Almost all individuals have an "excellent" quality (EF+).*
  May conduct to:
- *Almost all individuals have an "excellent" communication and quality (false statement: the conjunction of both conditions may not be related to "almost all individuals"),*
  or to:
- *Almost all individuals have an "excellent" communication or quality (lack of information),*
  but the correct way is:
- *Regarding communication and quality, almost all individuals are excellent.*

Example 4.36. *Non-absolute negative EFs of the same sign about two ordinal variables with the same exact modalities.*

- *Almost none of the individuals has an "excellent" communication (EF+);*
- *Almost none of the individuals has an "excellent" quality (EF+).*
  Conduct to:
- *In reference to communication and quality, almost none of the individuals is excellent.*

Example 4.37 illustrates the merging of messages coming from VoIs. In this case, the grammatical number of the composed sentence has to be changed to plural.

Example 4.37. *VoIs of the same sign about two ordinal variables with the same exact modalities:*

- *The proportion of individuals with a "small" size is low (VoI–);*
- *The proportion of individuals with a "small" showcase is low (VoI–).*
  Conduct to:
- *The proportions of individuals with a "small" size and showcase are low (merge messages and change the grammatical number of the sentence).*

Examples 4.38 and 4.39 illustrate the behaviour of this rule when dealing with messages of the different sign. In this case, the sentence cannot be compacted.

Example 4.38. *EFs of different sign:*

- *Almost all individuals have an "excellent" communication (EF+);*
- *None of the individuals has an "excellent" quality (EF–).*
  Conduct to:
- *Almost all individuals have an "excellent" communication but none of them has an "excellent" quality (merge messages).*

Example 4.39. *VoIs of different sign:*

- *The proportion of individuals with a "small" size is high (VoI+);*
- *The proportion of individuals with a "small" showcase is low (VoI–).*
  Conduct to:
- *The proportion of individuals with a "small" size is high but the proportion of them with a "small" showcase is low (merge messages).*

Rule B.4: adding single modalities

Once reached this step, most messages have been associated with other messages. It is not natural, when talking about the most relevant features of a certain class, to include a reference to a variable that has already been previously referred. Rule B.4 avoids those situations by associating single messages (those without relations) to sentences (groups of messages) that share the variable which are related to, with the one of the single message. To this aim, this rule looks for single messages and tries to associate them with any of the existing sentences. That is why this rule only analyses the class and variable to which messages are related to.

Examples 4.40 and 4.41 illustrate the addition of single messages to sentences detected by rules B.1 and B.2, respectively.

Example 4.40. *B.1 sentence and EF:*

- *The proportions of "weak" and "strong" are high (positive sentence detected by rule B.1);*
- *None of the individuals is "inexistent" (EF–).*
  Conduct to:
- *The proportions of "weak" and "strong" are high and none of the individuals is "inexistent" (merge messages).*

Example 4.41. *B.2 sentence and VoI:*

- *None of the individuals is "green" nor "red" (negative sentence detected by rule B.2);*
- *The proportion of "blue" is high (VoI+).*
  Conduct to:
- *None of the individuals is "green" nor "red" and the proportion of "blue" is high (merge messages).*

Rule B.5: merging single messages

As the last step for merging messages, rule B.5 merges single messages (those without relations) of the same type and sign. These messages don't have any relation with other messages, but their sign (and type) can define a slight relation with other single messages. Therefore, this rule looks for single messages and tries to find relations among them, linking messages of the same type and sign.

When merging messages coming from EFs, the absoluteness of them, as said in section 4.3.1, must be taken into account. EF messages only can be compacted in a simple way if they are related to absolute values (they affect to the 100% or 0% of the individuals of the class). Examples 4.42 and 4.43 show the behaviour of this rule when dealing with messages coming from absolute EFs and non-absolute EFs, respectively.

Example 4.42. *Merging single absolute EF messages:*

- *None of the individuals has a "strong" competition (EF–);*
- *None of the individuals has a "good" evaluation (EF–).*
  Conduct to:
- *None of the individuals has a "strong" competition nor a "good" evaluation (merge and compact messages).*

In order to avoid a possible lack of information or an untrue statement, non-absolute messages must be compacted as follows:

Example 4.43. *Merging single non-absolute EF messages:*

- *Almost all individuals have a "weak" competition (EF+);*

- *Almost all individuals have a "bad" evaluation (EF+).*
  Conduct to:
- *Regarding communication and evaluation, almost all individuals are "weak" and "bad", respectively (merge messages).*

Example 4.44 illustrates the way of behaving of rule B.5 when merging messages coming from VoIs. The grammatical number of the composed sentence must be changed to plural in order to compact it.

Example 4.44. *Merging single VoI messages:*

- *The proportion of individuals with a "strong" competition is high (VoI+);*
- *The proportion of individuals with a "good" evaluation is high (VoI+).*
  Conduct to:
- *The proportions of individuals with a "strong" competition and a "good" evaluation are high (merge messages).*

4.3.2 Using the semantics of variables and modalities

In order to obtain a natural description of the main features of the analysed classes, semantics of variables and modalities must be used. To reach this objective, an ontology encoding the knowledge of the domain must be defined. The following rules have been designed to detect which messages or sentences must be modified with the aim of producing natural language phrases. Each rule includes examples to better understand its meaning and aim.

Rule C.1: use of the semantics of modality "no"

Messages obtained from both the EFs and VoIs must change their verbal form when dealing with "no" modalities. Examples 4.45 and 4.46 detail the application of this rule when treating positive and negative EF messages, respectively. The application of rule C.1 with EF messages is direct, only having to modify the sign of the sentence.

Example 4.45. *Modifying the message of a positive EF:*

- *All individuals have the modality "no" in variable internet (EF+).*
  Conduct to:

- *None of the individuals has internet (invert the sign of the message).*

Example 4.46. *Modifying the message of a negative EF:*

- *None of the individuals has the modality "no" in variable internet (EF–).*
  Conduct to:
- *All individuals have internet (invert the sign of the message).*

Instead, the application of rule C.1 with messages coming from VoIs is not as direct. Examples 4.47 and 4.48 illustrate its application to positive and negative VoIs, respectively.

Example 4.47. *Modifying the message of a positive VoI:*

- *The proportion of individuals with the modality "no" in variable internet is high (VoI+).*
  Conduct to:
- *The proportion of individuals without internet is high (invert the sign of the preposition).*

Example 4.48. *Modifying the message of a negative VoI:*

- *The proportion of individuals with the modality "no" in variable internet is very low (VoI–).*
  Conduct to:
- *The proportion of individuals without internet is very low (invert the sign of the preposition).*

Rule C.2: use of the semantics of modality "yes"

By using a similar reasoning as in rule C.1, rule C.2 modifies sentences of both type EF and VoI messages that are associated with the modality "yes". In general, this rule will only avoid the inclusion of the modality in the sentences, as shown in examples 4.49 and 4.50 in the case of type EF messages and 4.51 and 4.52 in the case of type VoI messages.

Example 4.49. *Modifying the message of a positive EF:*

- *All individuals have the modality "yes" in variable internet (EF+).*
  Conduct to:
- *All individuals have internet (delete the mention of the modality).*

Example 4.50. *Modifying the message of a negative EF:*

- *None of the individuals has the modality "yes" in variable internet (EF–).*
  Conduct to:
- *None of the individuals has internet (delete the mention of the modality).*

Example 4.51. *Modifying the message of a positive VoI:*

- *The proportion of individuals with the modality "yes" in variable internet is high (VoI+).*
  Conduct to:
- *The proportion of individuals with internet is high (delete the mention of the modality).*

Example 4.52. *Modifying the message of a negative VoI:*

- *The proportion of individuals with the modality "yes" in variable internet is very low (VoI–).*
  Conduct to:
- *The proportion of individuals with internet is very low (invert the sign of the preposition and delete the mention of the modality).*

Rule C.3: use of the semantics of variables

Tule C.3 is one of the most important rules for obtaining a natural text. Generic rules can be designed for discarding messages or finding relations among them, but a specific use of the semantics of each variable can increase considerably the naturalness of the generated text.

The application of this rule implies the need of designing an ontology where the generation of the messages associated with EFs or VoIs is defined for each variable. Examples 4.53 to 4.55 show several cases in which the semantics of variables must be used.

Example 4.53. *Variable "location":*

- *Almost all individuals have the modality "city" in the variable location (EF+).*
  Conduct to:
- *Almost all individuals are located in "cities".*

Example 4.54. *Variable "size":*

- *None of the individuals has the modality "medium" in the variable size (EF–).*
  Conduct to:

- *None of the individuals is medium-sized.*

Example 4.55. *Variable "number of assistants":*

- *The proportion of individuals with the modality "few" in the variable number of assistants is very high (VoI+).*
  Conduct to:
- *The proportion of individuals with "few" assistants is very high.*

Rule C.4: modalities as adjectives

One way to easily "naturalise" the generated text is taking into account if the modality acts as an adjective of the variable. In this case, just placing the modality before the variable generates a more natural sentence. This must be defined in the ontology of the current domain, however the application of this rule is very easy. Marking which of the variables have modalities corresponding to adjectives of their associated variable is the only requirement to apply this rule. These variables don't need the definition mentioned in the above rule C.3.

Examples 4.56 to 4.59 illustrate some variables of this kind, and how the message is modified to take the meaning of the modality into account.

Example 4.56. *Variable "communication":*

- *All individuals have the modality "excellent" in the variable communication (EF+).*
  Conduct to:
- *All individuals have an excellent communication (placing modality before the variable).*

Example 4.57. *Variable "competition":*

- *Almost none of the individuals has the modality "strong" in the variable competition (EF–).*
  Conduct to:
- *Almost none of the individuals has a strong competition (placing modality before the variable).*

Example 4.58. *Variable "weight":*

- *The proportion of individuals with the modality "minimal" in the variable weight is very high (VoI+).*
  Conduct to:

- *The proportion of individuals with a minimal weight is very high (placing modality before the variable).*

Example 4.59. *Variable "sensitivity to promotions":*

- *The proportion of individuals with the modality "high" in the variable sensitivity to promotions is low (VoI–).*
  Conduct to:
- *The proportion of individuals with a high sensitivity to promotions is low (placing modality before the variable).*

Rule C.5: use of linguistic quantifiers

Most of the examples used to illustrate how the rules work use linguistic quantifiers like "all", "none", "very", "low", etc. to magnify the importance or relevance of the messages. The following show how to use the quantifiers depending on type, sign and relevance of messages.

The quantifiers "high" and "low" are used to transform the sign of messages obtained from VoIs into a natural sentence, as shown in Example 4.60.

Example 4.60. *Messages using the quantifiers "high" and "low":*

- *The proportion of "strong" is high (VoI+);*
- *The proportion of "weak" is low (VoI–).*

The quantifier "very" is used to exhibit the high relevance of some VoIs, as shown in Example 4.61.

Example 4.61. *Messages using the quantifier "very":*

- *The proportion of "red" is very high (highly relevant VoI+);*
- *The proportion of "green" is very low (highly relevant VoI–).*

The quantifiers "all" and "none" are used in messages coming from EFs, when their associated value is 1 and 0 (100% and 0% of the individuals of the class), respectively, as shown in Example 4.62.

Example 4.62. *Messages using the quantifiers "all" and "none":*

- *All individuals are "excellent" (EF+, with a value of* 1*);*
- *None of the individuals is "poor" (EF–, with a value of* 0*).*

The quantifier "almost" is used in EF messages when they are not highly relevant, i.e. when their associated value is below 1 for positive messages and above 0 for negative ones. Example 4.63 illustrate these cases.

Example 4.63. *Messages using the quantifier "almost":*

- *Almost all individuals are "excellent" (EF+, with a value lower than* 1*);*
- *Almost none of the individuals is "poor" (EF–, with a value lower than* 0*).*

4.4 Microplanning and realisation

This stage is the final stage of the architecture of the presented NLG system. It is responsible for stating the definitive structure of the text and transcribing the analysed messages into the final text, by enriching them in order to obtain a natural text. This stage is divided into two substages: microplanning and realisation. Microplanning stage is responsible for deciding the structure of the final text. This structure must take into account the attributes of messages that come from previous stages and the relations detected by type-B rules. Realisation stage receives this planning and produces the final text by transforming messages and sentences obtained in previous stages according to activated rules among types B and C.

Finally, in order to generate the phrases of the final text, a grammar should be formalised (Dale and Reiter, 2000; Turner et al., 2006) and a tool should be developed to apply the designed grammar. A formal grammar is a set of production rules that define the sentences accepted by a certain language. A context-free grammar is a grammar in which the left-hand side of each production rule consists of only a single nonterminal symbol. Despite the study of the appropriate grammar in not within the scope of this thesis and will be presented as future work, a shorten version of a Backus-Naur Form (BNF) grammar is introduced in Example 4.64.

Example 4.64. *Let's consider the following grammar:*

$$
\begin{aligned}
S &\rightarrow Dc \mid DcS \\
Dc &\rightarrow E \mid D \mid EDc \mid DDc \\
E &\rightarrow Qe\ X\ Vb\ M\ V\ . \\
D &\rightarrow N\ of\ X\ with\ a\ M\ V\ Vb\ Q\ Vc\ . \\
Qe &\rightarrow All \mid Almost\ all \mid Almost\ none\ of\ the \mid None\ of\ the \\
X &\rightarrow individuals \\
Vb &\rightarrow has \mid have \mid is \mid are \\
N &\rightarrow The\ proportion \\
M &\rightarrow list\ of\ modalities \\
V &\rightarrow list\ of\ variables \\
Q &\rightarrow very \mid \lambda \\
Vc &\rightarrow high \mid low
\end{aligned}
$$

where S is the initial symbol, Dc is the description of a certain class, E and D are sentences obtained from a EF or a VoI, respectively, Qe is a quantifier for EF sentences, X is the name of the analysed individuals, Vb is a verbal form, N is a nominal syntagm, M is a modality of a variable, V is a variable, Q is a quantifier adverb, Vc is an adjective and "very", 'ìndividals", "None of the" and so on are the terminal symbols of the grammar. The sentence "The proportion of individuals with a good communication is high." is generated as follows:

$$
\begin{aligned}
S &\rightarrow Dc \\
&\rightarrow D \\
&\rightarrow N\ of\ X\ with\ a\ M\ V\ Vb\ Q\ Vc\ . \\
&\rightarrow The\ proportion\ of\ X\ with\ a\ M\ V\ Vb\ Q\ Vc\ . \\
&\rightarrow The\ proportion\ of\ individuals\ with\ a\ M\ V\ Vb\ Q\ Vc\ . \\
&\rightarrow The\ proportion\ of\ individuals\ with\ a\ good\ V\ Vb\ Q\ Vc\ . \\
&\rightarrow The\ proportion\ of\ individuals\ with\ a\ good\ communication\ Vb\ Q\ Vc\ .
\end{aligned}
$$

$\rightarrow$ *The proportion of individuals with a good communication is* $Q\,Vc$.

$\rightarrow$ *The proportion of individuals with a good communication is* $\lambda\,Vc$.

$\rightarrow$ *The proportion of individuals with a good communication is high.*

Following subsections describe in detail the different steps carried out in substages of microplanning and realisation.

4.4.1 Microplanning

The natural text to be generated by the NLG system presented in this work has been designed to describe the most important features of the classes or clusters analysed. Having this in mind, the text has been organised considering each class separately. Therefore, messages concerning different classes are not mixed in the same paragraph.

Within each class, messages are sorted according to the following criteria:

1. **Weight**: type-A rules discard messages that are redundant by prioritising the more descriptive ones. Each time a message is prioritised, its weight is increased. Those prioritised messages must be mentioned earlier in the text.

2. **Type**: as EF messages are more descriptive than VoI ones, messages coming from EFs are included earlier.

3. **Sign**: positive messages are usually more important than negative ones. For example, as it has been explained in section 4.2.1, the existence of a positive message in one class usually implies the existence of as many negative messages as the rest of classes. So, positive messages must be incorporated earlier in the text.

4. **Relevance**: finally, if messages are still tied after applying the above criteria, their relevance is checked. Messages with a higher relevance must be mentioned earlier in the text.

Once messages are sorted, a group number is assigned sequentially to them and their related messages according to type-B rules because all of them will belong to the same phrase, as shown in the Example 4.65.

Example 4.65. *Table 4.11 shows an example of several messages sorted by taking into account the above criteria and the result of grouping them.*

Table 4.11: Example of messages ordered according to the defined criteria and the resulting groups

Message ID	Class	Related to	Rules
1	1	4	B.1
2	1	-	-
3	1	5	B.2
4	1	1	B.1
5	1	3	B.2
6	1	-	-
7	2	-	-
8	2	10	B.1
9	2	-	-
10	2	8	B.1
11	2	8, 10	B.4
12	2	-	-
13	3	16	B.2
14	3	15	B.3
15	3	14	B.3
16	3	13	B.2
17	3	-	-

Group ID	Class	Messages
1	1	1, 4
2	1	2
3	1	3, 5
4	1	6
5	2	7
6	2	8, 10, 11
7	2	9
8	2	12
9	3	13, 16
10	3	14, 15
11	3	17

The first subtable shows a sorting of 17 messages according to the criteria. Messages can only be grouped with messages of the same class. Messages marked with a hyphen in the column "Related to" are not going to be grouped with any message. For instance, message #1 must be grouped with message #4, while message #2 will appear alone in the text. The second subtable illustrates the obtained groups (phrases). For instance, message #4 will be included in the first phrase thanks to its related message #1 and despite having been sorted as the fourth message.

In the next substage a phrase is generated for each group of related messages.

4.4.2 Realisation

The realisation of the final text is carried out by a sequential process that performs a transcription in the form of natural language phrases of the different groups of messages obtained in the previous microplanning substage. This substage explores sequentially the defined groups and generates a phrase with the messages of the group according to the rules to be applied on them.

As type-A rules have been used for discarding messages in previous stages, the first type of rules to be analysed are the type-B ones. Table 4.12 shows the compatibility among rules, i.e. if two rules can co-exist in the same group of messages.

Table 4.12: Compatibility between type-B rules

	B.1	**B.2**	**B.3**	**B.4**	**B.5**
B.1	-	No	No	Yes	No
B.2	No	-	No	Yes	No
B.3	No	No	-	No	No
B.4	Yes	Yes	No	-	No
B.5	No	No	No	No	-

Rules B.3 and B.5 are the most independent of type-B rules because if they are activated, it means that the other rules aren't. This is why these two rules are the first ones to be checked and applied if needed.

As shown in Table 4.12, rules B.1 and B.2 are independent between them, but they both can co-exist with rule B.4. Then, in those cases in which rules B.1 or B.2 are activated, they are applied for the messages of the group that are affected by them and, if rule B.4 is activated in that group, the obtained sentence is merged to the rest of the messages of the group.

The following sketch summarises the way of applying type-B rules for each of the groups of messages. Example 4.66 illustrates the way of applying type-B rules for messages presented in Table 4.11.

- If rules B.3 or B.5 are active → apply them on the whole group (end).
- If not →
  - If rules B.1 or B.2 are active → apply them on the messages affected by them.

    - If rule B.4 is active → apply it on the sentence generated by rule B.1 or B.2 and the rest of messages of the group (end).
    - If not → (end).
  - If not → generate a generic sentence with the message of the group (end).

Example 4.66. *Let's analyse how to apply type-B rules for composing phrases #3 and #5 presented in Table 4.11.*

- *Messages of phrase #3 (message IDs #3 and #5) are related by rule B.2. Looking at the above sketch, rules B.3 and B.5 are not activated, so the next step is analysed. Rule B.2 is activated, so it's applied to build a sentence with both messages. Given that rule B.4 is not activated, the process has finished.*
- *Messages of phrase #5 (message IDs #8, #10 and #11) are related by rules B.1 and B.4. More specifically, messages #8 and #10 are related by rule B.1, and the sentence created by merging them is related to message #11 by rule B.4. Let's apply the sketch. Given that rules B.3 and B.5 are not activated, the next step is considered. Given that rule B.1 is activated in messages #8 and #10, a sentence merging both messages is created. Moreover, rule B.4 is activated in message #11, so it is applied on the previous sentence to merge it with message #11, finishing then the process.*

Once decided which kind of sentence must be generated according to type-B rules, the type-C rules are checked and applied in order to generate the proper components of the sentences that will form the final phrases. The components of a sentence used in this work are described in detail below:

- **Subject**: subject of messages differs depending on whether dealing with EF messages or VoI messages. Moreover, there is a special case presented in section 4.3.1 (rule B.3), by Examples 4.35 and 4.36.
  - In case of messages coming from EFs, the subject is very simple and it refers to how many individuals present a certain modality: "*(all | almost all | none of | almost none of) individuals*". Just in case that the message is not the first message of the sentence, the word for referring to individuals is replaced by the pronoum "them": "*[...] and all of them [...]*".

  - In case of dealing with messages coming from VoIs, the subject is the most elaborated part of the sentence, because it has to mention the variable or the modality and therefore, the ontology must be used to compose the sentence. Sentences of this type of messages begin with the nominal phrase "*The proportion of individuals*". This nominal phrase is completed with the variable of modality which the message refers to, usually beginning with prepositions "with" or "without": "*with a strong competition | with few assistants | without competition*".

    When merging messages coming from VoIs (all type-B rules except B.4) with the same sign, the grammatical number of the composed sentence must be changed to plural in order to properly compact it: "*The proportions of individuals*".

  - When treating with non-absolute negative messages of type EF affected by rule B.3, the subject must be preceded by the expression "*(Regarding | In reference to) variables,*", in order to avoid untrue sentences or lacks of information.

- **Verb**: the verb used in sentences depends not only on the type of message, but on the semantics of the variable or modality associated with the message.

  - When dealing from EF messages, it's usual to use verb forms like "have" or "are", depending on the semantics of the variable or modality: "*All individuals are | have*". Just in case of dealing with negative EF messages, the verb number is changed to singular: "*None of the individuals is | has*".

  - In case of translating VoI messages, the verb "to be" is always used in third person singular, as it refers to "the proportion": "*The proportion of individuals with a certain modality is*". As in the elaboration of the subject, when merging VoI messages of the same sign, the grammatical number of the verb is changed to plural: "*The proportion of individuals with modalities | variables are*".

- **Predicate**: the predicate of messages coming from EFs is the part of sentence responsible for referring the variable or modality of the message, in a similar way that the subject of VoI messages. Its construction must use the defined ontology and therefor this part of the sentence depends on the semantic of the variable or modality: "*(All individuals) have a strong competition — are located in little towns ...*". In case of dealing with messages coming from

VoIs, the predicate is very simple as it only must reflect the sign and relevance of the message, thus using linguistic quantifiers: "*(The proportion of individuals with a certain modality is) very high | high | low | very low*".

Table 4.13 summarises the peculiarities of the application of each rule, as explained in section 4.3.1. Column "Features of messages" indicates the characteristics that have the messages to be merged, apart from the requirements defined for each rule shown in Table 4.10. For instance, B.1 rule merges messages related to the same ordinal variable and sharing type and sign. Then, the composition of the sentence takes into account if the modalities are consecutive or not, and if there is any extreme modality. The number of Examples is included, and the used connectives to merge the different messages are detailed in the next columns. Finally, the last column states if the obtained sentence has been compacted or not.

4.5 Conclusions

This chapter presents the design and development of a natural language generation (NLG) system to describe in a qualitative way the most important characteristics of the classes of a clustered data set. The system highlights the most relevant patterns of each class, on the one hand by analysing separately the joint frequency distribution of each class and, on the other hand, by comparing their conditional frequencies.

The designed system is based on an adaptation of a well known architecture for data-to-text systems consisting in four stages: signal analysis, data interpretation, document planning and microplanning and realisation. Each stage is widely described in this chapter, including exhaustive examples to better understand the aim of each involved task. Theses tasks include the process of the initial data in order to standardise the analysed variables, the selection of the significant features that best differentiate each class, the identification of most relevant values (either extreme frequency (EF) or value of importance (VoI)) and the transcription of these values into a natural text.

The process involved in the translation of EFs and VoIs is based on the analysis and application of rules. The values are firstly translated into messages, and rules are applied sequentially on them in order to avoid redundant messages, to merge related messages to be combined in the final text and to identify specific modifications on the messages to obtain a final natural text.

An ontology to encode domain knowledge is introduced in addition to a grammar used to implement the proper transcription of the messages. The system has been developed under the R environment into a tool that automatically obtains the final text.

The most important contributions introduced in this chapter are the adaptation of a generic four-stage architecture for data-to-text systems in order to design and develop a rule-based system for dealing with well-structured data in which examples are segmented into several classes, the use of the concept of dependence to detect the most relevant characteristics of each class, the analysis and design of a set of rules to discard messages offering similar information in order to obtain a cohesive description, the analysis and design of a set of rules to merge related messages with the aim of generating a compact description and the analysis and design of a set of rules to naturalise the generated description.

As future research, there are two main areas to further investigate. On the one hand, to improve the detection of important values. This implies, for instance, to take into account the different levels of importance of analysed variables with the aim of emphasise them in the final text, and to identify other ways of highlighting other important values. On the other hand, to complete the definition of the grammar to be used in the realisation stage and to develop the needed mechanisms to incorporate it to the process.

Table 4.13: Summary of type-B rules

Rule	Features of messages	Examples	Used connectives	Compacted
B.1	Non-consecutive modalities	4.24	and	Yes
	Consecutive modalities			
	+ extreme modality	4.25	(greater than | lower than) or equal to	Yes
	+ non-extreme modality	4.26	between ... and	Yes
B.2	Same type			
	+ same sign + VoIs	4.27	and | nor	Yes
	+ same sign + EFs	4.28	and | nor	Yes
	+ diff. sign + VoIs	4.29	while	No
	+ diff. sign + EFs	4.30	while	No
	Different type			
	+ same sign	4.31	and	No
	+ different sign	4.32	while	No
B.3	Same sign			
	+ absolute EFs	4.33, 4.34	and | nor	Yes
	+ non-absolute EFs	4.35, 4.36	(Regarding | In reference to) ... and	Yes
	+ VoIs	4.37	and	Yes
	Different sign			
	+ EFs	4.38	but	No
	+ VoIs	4.39	but	No
B.4	B.1 + message	4.40	and | .	No
	B.2 + message	4.41	and | .	No
B.5	Absolute EFs	4.42	and | nor	Yes
	Non-absolute EFs	4.43	(Regarding | In reference to) ... and	Yes
	VoIs	4.44	and	Yes

Chapter 5

Application to market segmentation

This chapter describes a case study addressing a challenge in a marketing environment, and solves it by using the methodologies introduced in Chapters 3 and 4. The case presented shows the relation between the theoretical study done in previous chapters and its connection to real applications. The study takes place in a business to business (B2B) environment over nine months, where information about the retailers of a commercial firm was provided by the firms' sales representatives. B2B environments appear when a firm distributes its products via other firms (retailers), and they are characterised by marketing activities of organisations exchanging commerce transactions with other organisations (Turnbull and Leek, 2003).

The challenge to solve in this case study is motivated to comprehend fluctuations in orders made by points of sale that distribute the firm's products. The impossibility of extensively analysing each store due to the scarcity of resources makes necessary an automatic study that optimises the needed resources for performing marketing campaigns intelligently oriented to the set of shops expected to maximise customers' (retailers) satisfaction and loyalty. The main objective of this study is to identify and then segment a set of retailers (points of sale) of an industrial company, considering behavioural, relational and descriptive variables.

There are three main actions to perform in order to solve the detected challenge. Firstly, an automatic unsupervised learning process is carried out to get different ways of segmenting the

analysed stores. Secondly, the methodology presented in Chapter 3 is used to select the best market segmentation according to marketing experts and firm expectations. The main requirement of the study is to obtain a segmentation as consistent as possible with the sensitivity to promotions of the customers of each shop. Finally, in order to get a fully understandable segmentation, the system introduced in Chapter 4 is employed to obtain a qualitative description in natural language of the main relevant characteristics of the obtained segments. The obtained new segmentation and the final text interpreting the considered classes will give an opportunity to marketing executives and managers to understand their customers' behaviour. In addition, it will enable them to design or define appropriate and common marketing strategies for each segment.

5.1 Dataset

The study conducted is based on data collected using the observations, knowledge, and experience of the sales representatives working for Textil Seu SA, an outdoor sporting equipment firm (Grifone, http://www.grifone.com) established in La Seu d'Urgell (in Catalonia, Spain) for more than 25 years. Grifone works in a B2B environment and distributes clothes through points of sale and not directly to customers.

Data was collected as a result of a visiting period of one week to the main offices of Textil Seu, located in La Seu d'Urgell. Some of the data was obtained by extracting information of the database in conjunction with the sales' department. Subjective information was provided by the representatives of the firm, that know in general the main characteristics of each shop. Obtained data had to be preprocessed by discarding some non-informative variables and by recoding some ordinal variables to achieve a sufficient diversity in order to be able to get segments sufficiently heterogeneous between them.

This section presents the results obtained by considering a database of information from 260 shops that distribute Grifone products (Sánchez-Hernández et al., 2013). According to marketing experts, 16 variables were selected to describe these points of sale (3 quantitative, 5 qualitative, and 8 qualitative ordinal). The selected variables are summarised in Table 5.1, and include the following:

- *Aesthetics*: aesthetics quality of the display window.
  Values: "deficient", "regular", "good", "excellent".

- *Antiquity*: how many months the shop is distributing our products.
  Values: natural numbers.
- *Assistants*: number of full-time sellers in the shop.
  Values: natural numbers.
- *Communication*: communicative quality of the display window.
  Values: "deficient", "regular", "good", "excellent".
- *Competition*: level of competition within the area in which the store is located.
  Values: "no", "weak", "strong".
- *DisplayGrifone*: states if the display window of the store includes Grifone products.
  Values: "no", "yes".
- *DisplaySize*: qualitative size of the display window.
  Values: "small", "medium", "big".
- *Evaluation*: subjective quantitative assessment provided by our Grifone representatives, in terms of an overall impression of the point of sale.
  Values: $0, \ldots, 10$.
- *GrifoneWeight*: qualitative weight of Grifone products in the store.
  Values: "minimal", "secondary", "main".
- *Internet*: revels if the shop commercialises its products through the Internet.
  Values: "no", "yes".
- *Location*: type of the town in which the store is located, specially if this location is related to the mountain and outdoor sports.
  Values: "inner cities", "seaside cities", "no mountain towns", "mountain towns", "ski towns".
- *Maintenance*: maintenance quality to which is subjected the store.
  Values: "deficient", "regular", "good", "excellent".
- *PromosSensit*: level of sensitivity to promotions exhibited by the clients' store.
  Values: "low", "medium", "high".
- *Size*: size of the store in term of square meters dedicated to selling products.
  Values: "small", "medium", "big".
- *Specialists*: indicates if the shop is whether specialised in the sector of outdoor sporting equipment.

Values: "no", "yes".

- *ThermalExhibitor*: specifies if the store has a thermal exhibitor, that is to say, an exhibitor displaying thermal clothes.
  Values: "no", "yes".

Consequently, each of the points of sale is described by a vector of dimension 16. Figure 5.1 plots the distribution of each variable. At this point, it should be noted that *PromosSensit* variable is not going to be used in the unsupervised learning process. The aim of this approach is to obtain classifications not related to promotional activities and then select the most compatible one with *PromosSensit* variable. Therefore, this variable will be considered as external variable to be used in criteria number 4 for classification selection.

Table 5.1: Description of variables

Type	**Number**	**Name**	**Description**
Quantitative	3	Antiquity	*Duration of commercial relationship*
		Assistants	*Number of full-time sales assistants*
		Evaluation	*Assessment by Grifone representatives*
Qualitative	5	Specialists	*Specialist store*
		Location	*Geographic location*
		DisplayGrifone	*Grifone products in the display window*
		ThermalExhibitor	*Thermal product display*
		Internet	*Use of the Internet for e-Commerce*
Ordinal	8	Competition	*Level of competition*
		Size	*Store size*
		Maintenance	*Store maintenance*
		DisplaySize	*Display window size*
		Communication	*Communicative quality*
		Aesthetics	*Aesthetics quality*
		GrifoneWeight	*Grifone products' importance*
		PromosSensit	*Promotions sensitivity*

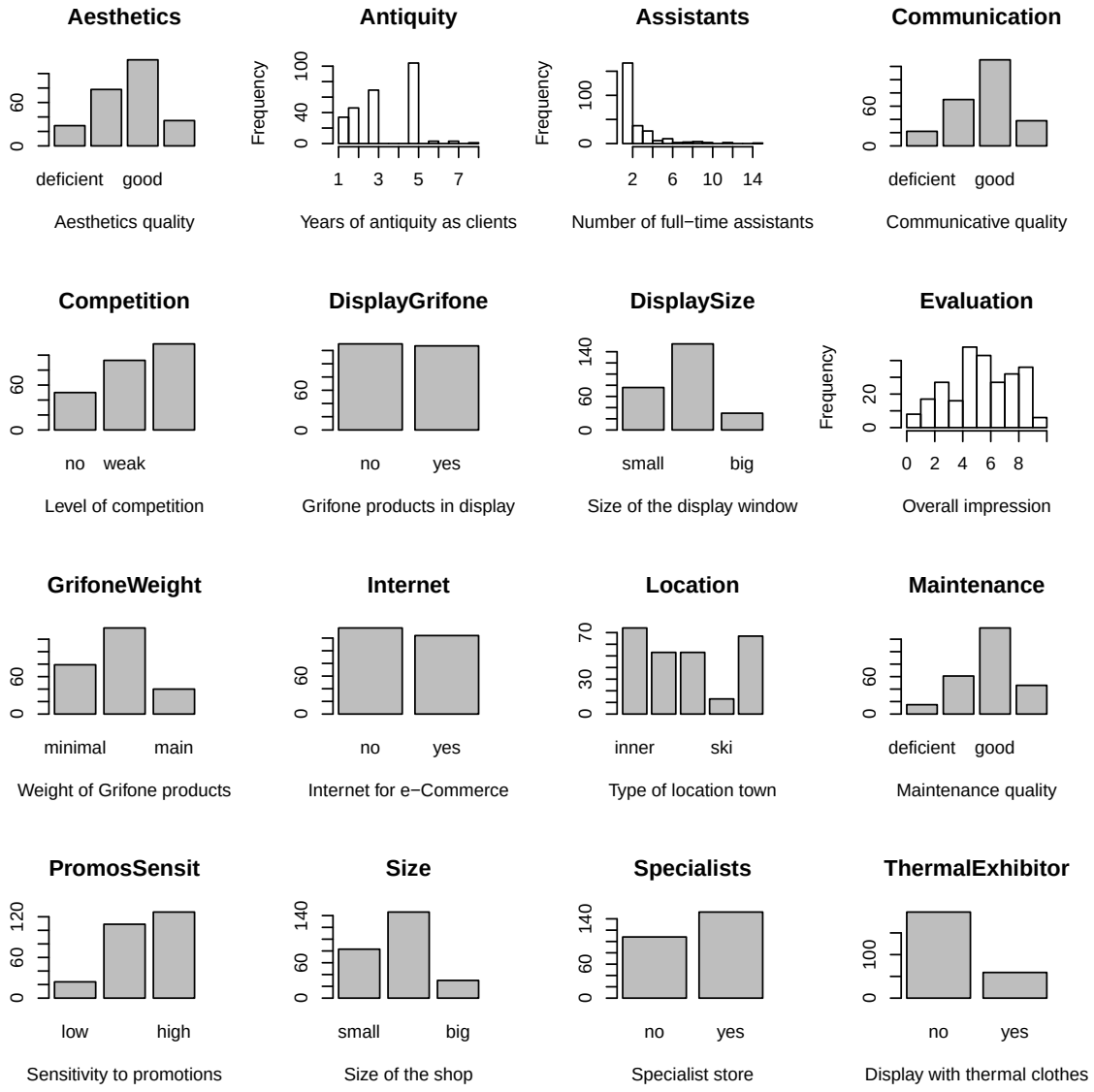

Figure 5.1: Summary of the considered variables in the dataset

5.2 Obtaining segmentations

In order to obtain a set of different segmentations without taking into account any a priori information, an unsupervised learning process must be performed. The unsupervised learning technique

used in this work is based on the algorithm Learning Algorithm for Multivariate Data Analysis (LAMDA) (Aguado, 1998; Aguado et al., 1999; Aguilar and López de Mántaras, 1982). LAMDA is a classification method based on fuzzy hybrid connectives that combines some of the most interesting capabilities of both purely numeric and purely symbolic algorithms. Its basic operation and components are detailed in Appendix A.

All connectives mentioned in Appendix A have been considered, by automatically exploring different degrees of tolerance. Fixed a connective, tolerance levels were obtained by varying it between 0 and 1, ensuring each tested tolerance to produce a different segmentation (Aguado, 1998). LAMDA is an iterative process, that is to say, it processes each individual several times, modifying its class until achieving a stability in which the classifications remains unalterable or until reaching a maximum number of iterations prefixed to 10.

In order to compute the Marginal Adequacy Degree (MAD) of each individual to each class according to each variable, density functions must be chosen for each quantitative variables according to their actual distributions shown in Figure 5.1 (see page 99). *Antiquity* distribution has a single maximum value not located in the extremes, so Waissman function is the chosen one. *Assistantes* variable presents a left-skewed distribution, as the Classical function. Finally, variable *Evaluation* shows a gaussian shape, so Gaussian function is the associated density function with it. These distributions and the corresponding functions with parameters obtained from the actual distribution are detailed in Figure 5.2.

In this case, the unsupervised version of LAMDA is used to generate a set of segmentations. Note that, as said before, variable *Promotions sensitivity* has not been included in this process because the objective of this learning phase is to obtain segmentations naturally related to this concept. The unsupervised learning process produced 566 segmentations, as shown in Table 5.2.

Note that despite the process of analysing different degrees of tolerance was carried out on each considered connective, some of the obtained segmentations were discarded due to their instability, as it can be seen in the case of Lukasiewicz connective. In this case, Lukasiewicz connective has generated a single stable segmentation, that corresponds to tolerance 0. As seen in connectives detailed in page 138, Frank n-norms have an extra parameter s whose value has been modified between -5 and 5, avoiding the cases $s = 0$ and $s = 1$.

The 566 obtained segmentations are composed of a number of classes between 1 and 244. Considering all segmentations obtained in the unsupervised process, Figure 5.3 shows the distribution of

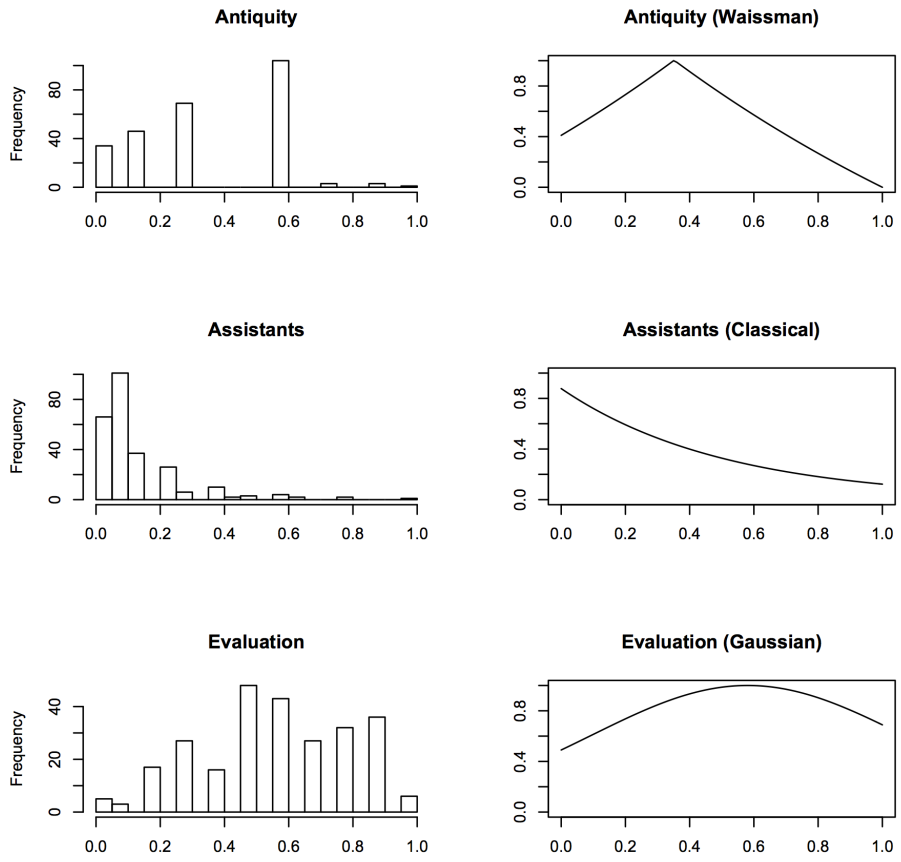

Figure 5.2: Histogram of the quantitative descriptors and their chosen density functions

variable "number of classes" assigned to each segmentation. As it can be seen, this is a left-skewed distribution because most of the obtained classifications have a low number of classes.

5.3 Ranking and selecting segmentations

Criteria defined in Chapter 3 are applied to choose the most appropriate points of sale segmentation according to them. The corresponding index of each criterion is computed for each segmentation, and only after this, the assessments obtained by each segmentation are aggregated by employing an Ordered Weighted Averaging (OWA) operator. This section reviews parameters used in the application of each criterion.

Table 5.2: Number of segmentations obtained by hybrid connective

Connective	**Tolerance**	**# of segmentations**
MinMax	Between 0 and 1	244
Frank $s = -5$	Between 0 and 1	98
Frank $s = -4$		58
Frank $s = -3$		51
Frank $s = -2$		27
Frank $s = -1$		22
Frank $s = -0.5$		31
Frank $s = 0.5$		10
Frank $s = 1.5$		3
Frank $s = 3$		3
Prob. prod.	Between 0 and 1	18
Lukasiewicz	0	1

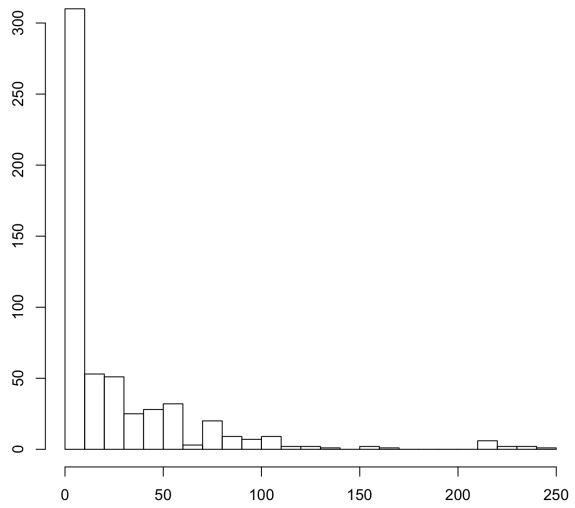

Figure 5.3: Histogram of the variable "number of classes" for the obtained segmentations

First criterion (usefulness index) needs, on the one hand, setting the values of K_1 and K_2 that define the interval considered to have the highest value in its associated index. In this marketing environment, the most desirable number of classes is set between three and five (Casabayó, 2005) and therefore $K_1 = 3$ and $K_2 = 5$. On the other hand, the shape of the fuzzy number to use must be defined: a left-linear function has been chosen as the simplest example of a strictly increasing function while a right-exponential function has been selected because the usefulness of a classification should decrease asymptotically when the number of classes increases, as shown in (5.1).

$$I_{U,3,5}(\mathcal{C}) = \begin{cases} \frac{M-1}{3-1}, & \text{if } 1 \leq M < 3; \\ 1, & \text{if } 3 \leq M \leq 5; \\ \frac{e^{(260-M)}-1}{e^{(260-5)}-1}, & \text{if } 5 < M \leq 260. \end{cases} \tag{5.1}$$

Figure 5.4 shows a graphical representation of an extension to real numbers of the membership function used in this study. Segmentations with a number of classes between the preferred $K_1 = 3$ and $K_2 = 5$ get the maximum degree of usefulness. Segmentations with only 1 class are not useful so their degree is 0, while it has been considered that segmentations with 2 classes are not too useful, having a usefulness of 0.5. On the right side, it has been decided to decrease exponentially the degree of usefulness as the number of obtained classes increases.

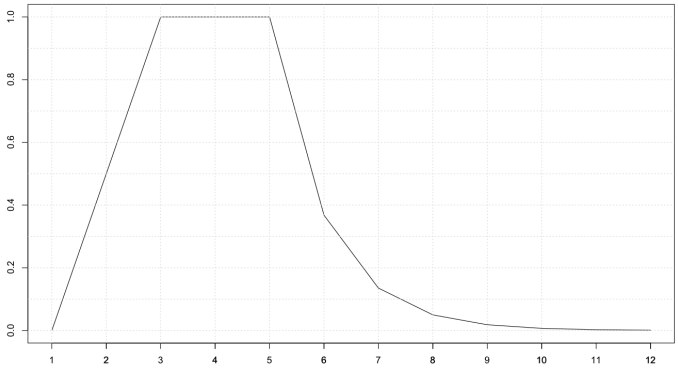

Figure 5.4: Membership function used of modelling the fuzzy concept 'Useful number of classes' with $K_1 = 3$ and $K_2 = 5$

In order to avoid segmentations in which one of the classes encompasses most of the shops and with the aim of getting manageable classes with a non too large number of shops per class, balanced classes are required. Therefore, criterion (3.12) is used to compute the balanced index with the

second part of (3.11) because $N = 260 \equiv 2(mod\ 3)$.

Third criterion (coherence index) needs to choose the density function to be used on each quantitative variable. This selection is carried out by examining the histogram of each of the three quantitative variables. These histograms are shown in Figure 5.2 (see page 101). Distribution of variable *Antiquity* presents a single maximum value close to the centre of the distribution, as in Waissman function. Most of values of variable *Assistants* are located on the left side, so Classical function has been chosen to represent it. Finally, variable *Evaluation* shows a gaussian shape, so Gaussian function is the associated density function with it. These functions have been used to compute the index defined in the third criterion of coherence.

According to the fourth criterion of dependency of external variables, variable *PromosSensit* is used as the control variable to contribute in the computation of the fourth index. Table 5.3 details the distribution of the modalities within this variable.

Table 5.3: Distribution of variable *PromosSensit*

Low	**Medium**	**High**
24	109	127

Segmentations obtained by analysing the variables included in the learning phase, with a high degree of dependency or compatibility with this variable are desired. Therefore, the dependency index between each segmentation and variable *PromosSensit* is computed.

Following the fifth criterion description, a supervised learning process is performed on the obtained segmentations in order to compute the accuracy of the predictive model associated with them. This step involves partitioning the dataset by means of a stratified cross-validation process with 10 folds. This means that each randomly obtained fold has the same proportion of the classes than in the whole dataset. Support Vector Machines (SVMs) are considered for supervised learning due to their good performance on high dimensional spaces (Chapelle et al., 1999). In order to obtain a statistically significant value for accuracy, the cross-validation process has been performed 30 times, and the accuracy index is the mean of the 30 accuracy indexes computed in each iteration.

Once the five indexes have been computed for the 566 different segmentations, each one of them is represented by a vector of five components corresponding to this five indexes. Then, they are aggregated for each segmentation by using an OWA operator. This operator is guided by the fuzzy

linguistic quantifier 'most of', represented using the Regular Increasing Monotone (RIM) function $Q(r) = r^{1/2}$ as justified in Section 2.2, which has associated the following weighting vector in the case of aggregating five values:

$$(0.447, 0.185, 0.142, 0.120, 0.106)\,.$$

Table 5.4 shows an extract of the best segmentations obtained by using this methodology, being sorted by their achieved OWA value. The 100 best segmentations are shown in Table B.1 (see Appendix B, page 141).

Table 5.4: Extract of the best segmentations using fuzzy selection criteria OWA methodology

Rank	**ID**	**Conn.**	**Toler.**	**Classes**	**Criterion assessment**					**OWA**
				M	I_U	I_B	I_C	I_D	I_A	
1	#259	Minmax	0.439	3	1	0.928	0.251	0.528	0.936	0.8423
2	#260	Minmax	0.469	3	1	0.929	0.254	0.363	0.922	0.8207
3	#258	Minmax	0.422	4	1	0.885	0.226	0.425	0.875	0.8103
4	#243	Minmax	0.290	3	1	0.909	0.256	0.008	0.965	0.7868
5	#244	Minmax	0.304	3	1	0.920	0.255	0.012	0.948	0.7855
6	#257	Minmax	0.411	3	1	0.933	0.279	0.032	0.856	0.7786
7	#256	Minmax	0.400	4	1	0.872	0.246	0.064	0.906	0.7751
8	#253	Minmax	0.359	4	1	0.883	0.231	0.021	0.915	0.7722
9	#252	Minmax	0.352	4	1	0.884	0.237	0.025	0.904	0.7712
10	#255	Minmax	0.393	4	1	0.939	0.263	0.063	0.768	0.7686

Note that despite all segmentations shown in Table 5.4 present a value of 1 in I_U, there are some segmentations ranked in the top 100 having a lower value in this index, as segmentations #248, #78 and #51, shown in Table B.1. Such segmentations would have been discarded if the sequential approach would have been taken into account.

Therefore, segmentation #259 has been selected, formed by three classes with 35, 98 and 127 elements, respectively. Before assigning a qualitative name to each class, it is necessary to understand which are the most important characteristics of each one, both by analysing them separately and by comparing them with each other. This analysis is precisely what is carried out when obtaining

the qualitative description of the classes.

5.4 Qualitative class description

The natural language generation (NLG) system described in Chapter 4 has been used to obtain a qualitative description of the most important characteristics of the segments obtained in the previous stage. Therefore, the application of this methodology will result in a final text explaining each segment individually. Below the peculiarities of each stage of the adapted architecture are detailed.

5.4.1 Signal analysis

The objective of this first stage, as seen in Figure 4.2 (page 53) is to obtain the list of the most relevant characteristics of the classes of classification #259 in form of initial messages with their associated values. This stage consists of the following steps: discretisation, features selection and detection of relevant values of importance (VoIs) and extreme frequencies (EFs).

Three of the variables considered are quantitative (Table 5.1). With the aim of unifying the methods to be used over all variables, a **discretisation** step is carried out over variables *Antiquity*, *Assistants* and *Evaluation*. As it was justified in Subsection 4.1.1, the chosen supervised discretisation technique is the class-attribute interdependence maximisation (CAIM) method (Kurgan and Cios, 2004). The three variables have been discretised into 3 intervals, whose distributions are detailed in Table 5.5.

Table 5.5: Frequencies of each obtained interval in the discretisation process

Variable	**Intervals**		
	P1	**P2**	**P3**
Antiquity	80	69	111
Assistants	204	32	24
Evaluation	52	107	101

In order to enrich the final text, their modalities have been properly chosen, as shown in Table

5.6.

Table 5.6: Chosen modalities for the discretised variables

Variable	**Intervals**		
	P1	**P2**	**P3**
Antiquity (years)	less than two	three	more than three
Assistants	few	many	a lot of
Evaluation	bad	good	excellent

The selection of the **significant variables** is done by computing the dependence level of the variables in respect to the class. A Pearson's Chi-square test of independence is carried out for each one of the 16 variables. Variables for which the dependence with the class variable is not significant are discarded (the significance of the computed statistic is lower than 0.05). Figure 5.5 shows, for each variable, the distribution of their modalities within each class. In addition, the statistical significance ($p-value$) of the computed statistic is given. Variables *Antiquity*, *Specialists* and *Internet* are discarded because their modalities are distributed in a similar way along the three analysed classes.

The detection of the **relevant VoIs** and the **EFs** implies the computation of several tables for each considered variable. Both VoIs and EFs are based on the contingency table that include the joint distribution of each variable and the analysed classes. Table 5.7 shows the contingency table of variable *Competition*. The rest of the contingency tables are included in Tables B.2 and B.3 (see Appendix B, pages 145 and 146). In general, all the computed tables are included in Appendix B with the aim of facilitating the reading of this text.

Table 5.7: Contingency table of variable *Competition*

Competition	*No*	*Weak*	*Strong*
1	0	13	21
2	26	38	34
3	24	42	60

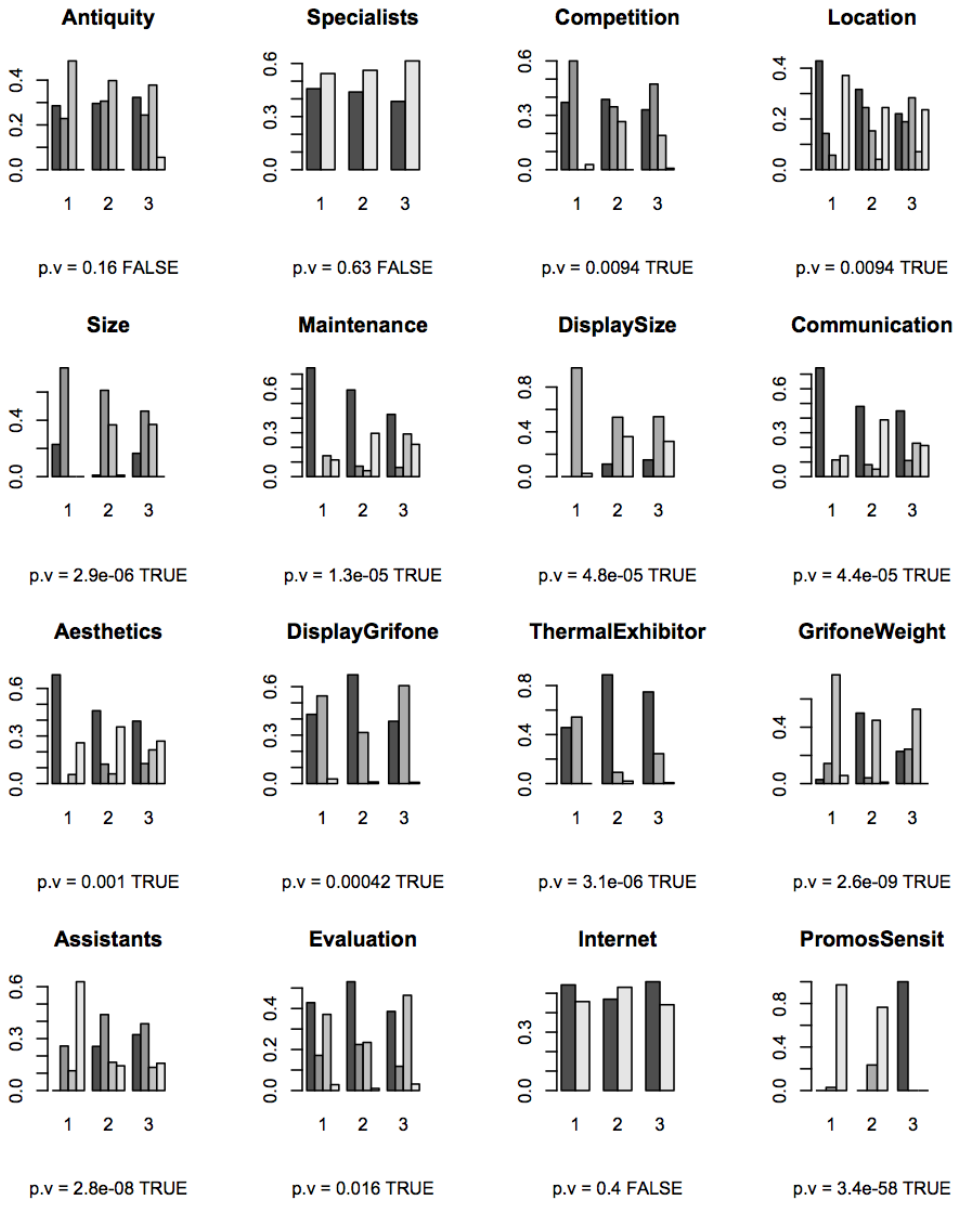

Figure 5.5: Distribution of the modalities within the classes, in addition with their $p - value$ and the decision result of the selection process

Each frequency is associated with a VoI, computed as the addend of the chi-squared statistic, that is to say, by using the difference between this observed frequency and the expected one. In

Table 5.8 the expected frequencies of variable *Competition* are included. The expected frequencies of the rest of variables are detailed in Tables B.4 and B.5 (see Appendix B, pages 147 and 148). It is important to note that, despite some variables present, in some of their modalities, an expected frequency below 5, these lower frequencies are very close to this required minimum, so no further action is done.

Table 5.8: Expected frequencies of variable *Competition*

Competition	*No*	*Weak*	*Strong*
1	6.6	12.3	15.2
2	19.0	35.3	43.7
3	24.4	45.4	56.2

Table 5.9 displays the VoIs of variable *Competition*. Note that, despite the VoIs are always positive numbers by definition (Equation 4.2), Table 5.9 includes their associated signs. The result of the computation of the VoIs is included in Tables B.6 and B.7 (see Appendix B, pages 149 and 150).

Table 5.9: VoIs associated with variable *Competition*

Competition	*No*	*Weak*	*Strong*
1	-6.66	0.05	2.25
2	2.59	0.20	-2.15
3	-0.01	-0.26	0.26

Looking at the absolute values of Table 5.9, it is clear that variable *Competition* will have three type of VoIs. The greater one is the pair "Class 1 – No", which will be considered as highly relevant. The other group will contain three VoIs greater than 2, and will be considered simply relevant. The rest of VoIs, close to 0, will be considered non relevant and therefore discarded.

A process for defining these different levels of relevance of VoI is carried out on each variable as introduced in Chapter 4. VoI of each variable are grouped and a K-means method (Xiong et al., 2009) with $K = 3$ is applied on each set associated with each variable. Among the three obtained

clusters for each variable, the cluster of lower VoIs is discarded while the VoIs of the medium and higher clusters are considered medium relevant and highly relevant, respectively. Table 5.10 shows, for each variable, the computed cut points and the number of relevant and highly relevant detected VoIs. This process has identified in total 67 VoIs, 38 of them relevant and 29 highly relevant.

Table 5.10: Cut points obtained for each variable and number of relevant VoIs detected

Variable	**Cut point 1**	**Cut point 2**	**# of relevant VoIs**	**# of highly relevant VoIs**
Competition	1.2	4.6	3	1
Location	1.1	3.8	4	2
Size	1.6	6.6	4	2
Maintenance	1.1	6.2	5	2
DisplaySize	2.7	6.2	1	2
Communication	1.2	3.5	3	4
Aesthetics	0.8	2.7	3	4
DisplayGrifone	1.8	4.7	2	2
ThermalExhibitor	3.4	11.3	2	1
GrifoneWeight	3.7	10.3	4	1
Assistants	1.7	10.7	3	2
Evaluation	1.4	3.3	1	3
PromosSensit	19.3	37.9	3	3

The obtained cut points for variable *Competition* are 1.2 and 4.6. That means that VoIs with an associated absolute value higher than 4.6 are defined to be highly relevant, while VoIs with a value between 1.2 and 4.6 are defined to be simply relevant. Thus, application of these cut points defines one highly relevant VoI and three relevant VoIs, as it can be seen when looking at Table 5.9.

As explained in chapter 4, the analysis of conditional frequencies of each variable will result in the detection of extreme frequencies. Table 5.11 show the conditional frequencies of variable *Competition*, while the conditional frequencies of the rest of variables are presented in Tables B.8 and B.9 (see Appendix B, pages 151 and 152).

The absence of individuals with modality "No" belonging to class 1 must be mentioned in the

Table 5.11: Conditional frequencies of variable *Competition*

Competition	*No*	*Weak*	*Strong*
1	0.00	0.38	0.62
2	0.27	0.39	0.35
3	0.19	0.33	0.48

final text and it seems that the rest of frequencies are not too extreme to be highlighted. In contrast to the process of the detection of relevant VoIs, the identification of EFs is done by performing a unique process over all variables. Conditional frequencies of each selected variable are combined into set W. A histogram of the values of this set is displayed in Figure 5.6.

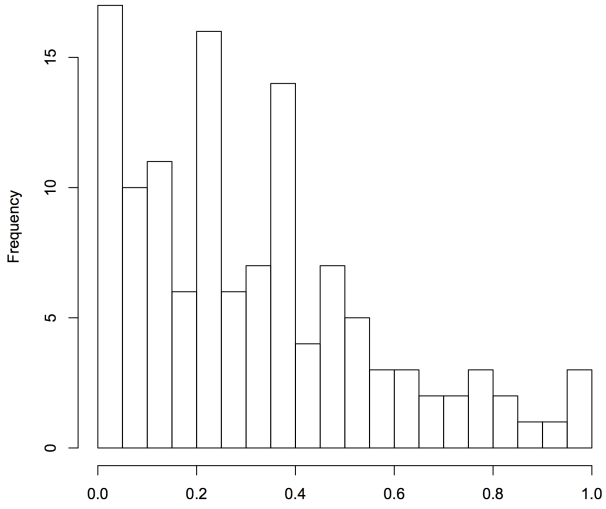

Figure 5.6: Histogram of the conditional frequencies of the dependent variables

The EFs are defined as those having a value lower than or equal to the percentile 1%, or those being greater than or equal to the percentile 99% of set W, that is to say, those values lower or equal to 0 or greater or equal to 0.971. Table 5.12 presents the values associated with some percentiles.

This process results in the identification of 13 EFs, three of them positive (almost all individuals of the class present a certain modality) and the other 10 negative (none of the individuals has a certain modality). Table 5.13 exhibits, for each variable and class, the number of obtained EFs.

Table 5.12: Percentiles of conditional frequencies to detect EFs

1%	5%	10%	25%	50%	75%	90%	95%	99%
0.000	0.000	0.029	0.114	0.265	0.455	0.685	0.814	0.971

Table 5.13: Number of obtained extreme frequencies

Variable	**Class 1**	**Class 2**	**Class 3**
Competition	1	0	0
Location	0	0	0
Size	1	0	0
Maintenance	1	0	0
DisplaySize	2	0	0
Communication	1	0	0
Aesthetics	1	0	0
DisplayGrifone	0	0	0
ThermalExhibitor	0	0	0
GrifoneWeight	0	0	0
Assistants	0	0	0
Evaluation	0	0	0
PromosSensit	2	1	3

In conclusion, in this first stage three numerical variables have been discretised, three non-dependent variables out of 16 have been discarded, and 67 relevant VoIs and 13 EFs have been identified. All values are combined into the same data frame. From now on, the values (both VoIs and EFs) together with their associated information (type of value, class, variable, modality, sign and relevance) are called initial messages. Table 5.14 includes the five initial messages of variable *Competition* (1 highly relevant VoI, 3 relevant VoIs and 1 EFs), while the rest of messages are shown in Table B.10 (see Appendix B, page 153).

Table 5.14: Initial messages detected for variable *Competition*. The rest of messages are included in Table B.10

ID	Class	Variable	Modality	Type	Sign	Relev.	Value
#1	1	*Competition*	*no*	VoI	neg.	high	6.6
#30	1	*Competition*	*strong*	VoI	pos.	normal	2.3
#47	2	*Competition*	*no*	VoI	pos.	normal	2.6
#48	2	*Competition*	*strong*	VoI	neg.	normal	2.1
#68	1	*Competition*	*no*	EF	neg.	-	0.0
...							

5.4.2 Data interpretation

The aim of this second stage is to apply a series of rules on the detected VoIs and EFs in order to identify messages offering similar information for discarding the less descriptive ones. Section 4.2 defines a collection of five rules to be applied on the set of initial messages. These rules are applied in this section on the identified VoIs and EFs by previous stage.

Before applying type-A rules, a initial weight of 1 is assigned to each of the 80 messages. Each time a message is discarded, its weight is decreased one unit with the aim of decreasing the importance of the message while the prioritisation of a message implies to increase its weight by one unit in order to increase its importance.

Rule A.1 discards negative messages when there are other positive messages related to the same class and variable containing similar information. In the case study analysed, this rule is activated 6 times, as shown in Table 5.15. This Table shows, for each of the 6 groups of messages analysed, which messages are discarded (the negative ones) and which of them are prioritised (the positive ones).

The following paragraph analyses two of the identified groups of messages. Text associated with message #71 is "*In class 1, all shops have a medium display size*". Note that if all individuals of a class present a certain modality, the other modalities cannot be present in the class, so the message #72 "*None of the shops has a big display size*" will be discarded. Message #64 says "*In class 3, the proportion of individuals displaying Grifone products is high*". Again, if a modality of a binary

Table 5.15: Groups of messages affected by rule A.1

Class	Variable	Type	ID	Modality	Sign	Action	Weigth
1	*DisplaySize*	EF	#71	*medium*	pos.	Prioritise	2
			#72	*big*	neg.	Discard	0
	ThermalExhibitor	VoI	#9	*yes*	pos.	Prioritise	2
			#42	*no*	neg.	Discard	0
	PromosSensit	EF	#75	*medium*	pos.	Prioritise	2
			#76	*high*	neg.	Discard	0
2	*DisplayGrifone*	VoI	#17	*no*	pos.	Prioritise	2
			#18	*yes*	neg.	Discard	0
3	*DisplayGrifone*	VoI	#64	*yes*	pos.	Prioritise	2
			#63	*no*	neg.	Discard	0
	PromosSensit	EF	#80	*high*	pos.	Prioritise	2
			#78	*low*	neg.	Discard	0
			#79	*medium*	neg.	Discard	0

variable involves a great number of the individuals in a class, then the other modality presents a low proportion of individuals, as said by message #62: "*The proportion of individuals not displaying Grifone products is low*". As a result, rule A.1 has discarded 7 negative messages and prioritised 6 positive ones.

Rule A.2 discards messages of type VoI when there are other EF messages of the same variable, modality and sign, because the latter ones are more descriptive than the first ones. This rule is activated for 11 pairs EF-VoI, as shown in Table B.12 (see Appendix B, page 157). In Table 5.16 only messages related to variable *Competition* are highlighted.

Both messages #68 and #1 are related to the same class, variable and modality and have the same sign. Text associated with message #68 stands "*None of the individual has a 'no' competition*[1]" while message #1 speaks about the same individuals but in a different way: "*The proportion*

[1]The proper construction of this sentence is "*All individuals have competition*", but its standard way is shown for better understanding the aim of the rule.

Table 5.16: Groups of messages of variable *Competition* affected by rule A.2. The rest of groups are shown in Table B.12

Class	Variable	Modality	Sign	ID	Type	Action	Weight
1	*Competition*	*no*	neg.	#68	EF	Prioritise	2
				#1	VoI	Discard	0

of individuals with a 'no' competition is very low[2]". If none of the individuals (or all of them) presents a certain modality, it is trivial that the proportion of individuals with that modality will be very low, so the redundant information will be discarded. Summarising, rule A.2 has discarded 11 messages related to VoIs, prioritising their related 11 EFs.

Rule A.3 discards messages when all modalities of a variable are highlighted within a certain class, prioritising positive messages and messages related to VoIs. This rules affects 8 groups of messages, shown in Table B.13 (see Appendix B, page 158). An extract of those groups are exemplified in Table 5.17, showing a group related to variable *PromosSensit*.

Table 5.17: Messages of variable *PromosSensit* affected by rule A.3. The rest of groups are shown in Table B.13

Class	Variable	ID	Modality	Type	Sign	Action	Weight
2	*PromosSensit*	#55	*low*	VoI	pos.	Prioritise	2
		#56	*medium*	VoI	pos.	Prioritise	2
		#22	*high*	VoI	neg.	Discard	-1
		#77	*high*	EF	neg.	Discard	1

Sentences associated with the first group of messages (class 2) are the following:

- "*The proportion of shops with a low sensitivity to promotions is high*"
- "*The proportion of shops with a medium sensitivity to promotions is high*"
- "*The proportion of shops with a high sensitivity to promotions is very low*"

[2]Again, the proper construction of the sentence is: "*The proportion of individuals without competition is very low*".

- "*None of the shops has a high sensitivity to promotions*".

Variable *PromosSensit* has 3 modalities: "low", "medium" and "high". This rule stands that having positive messages on two of the modalities, any mention to the third one can be avoided. In this case, the two first positive messages (#55 and #56) say that two out of three modalities ("low" and "medium") present a high proportion of shops. It is reasonable to infer that the third modality (in this case, "high") will present a very low proportion (as it is looking at message #22) and even that very few shops will have that modality (message #77). In conclusion, rule A.3 has discarded 14 negative messages, prioritising their related 13 positive messages.

Rule A.4 discards messages when the same information is mentioned for all classes, that is to say, when there are messages sharing variable, modality, type and sign related to all classes. It is difficult, in general, that this situation occurs very often, but if it happens, it must be avoided. In the case study analysed, this rule has not been activated by any group of messages.

Rule A.5 is related to previous rule A.4 because its objective is the same: to avoid mentioning information when this information is related to all classes. This rule is less restrictive than A.4 because it doesn't take into account sign of messages. Having this relaxation in mind, Table 5.18 shows the 4 groups of messages affected by rule A.5.

Text associated with positive message #14 of variable *Communication* is "*The proportion of shops with a regular communication is very high in class 2*", while the related negative messages #40 and #61 say "*The proportion of shops with a regular communication is low in class 1*" and "*The proportion of shops with a regular communication is low in class 3*", respectively. Information about the negative messages can be avoided by remarking in sentence of message #14 that "*moreover, most of regular shops are in this class*". To sum up, rule A.5 has discarded 8 negative messages, prioritising their corresponding 4 positive ones.

After applying all type-A rules, each message has an associated weight regarding its discards and prioritisations. Table 5.19 shows the frequency of each final weight, that is to say, how many messages have finally each weight. Table B.15 (see Appendix B, page 160) shows the list of the filtered 52 messages that will be mentioned in the final text.

The 28 messages having a negative or null weight are directly discarded. The list of these directly discarded messages is the following: #1, #2, #3, #4, #7, #16, #18, #22, #28, #35, #38, #39, #40, #41, #42, #45, #46, #49, #61, #63, #72, #76, #77, #78, #79. It has been decided, among messages with weight = 1, to discard those having at least one discard: these messages have

Table 5.18: Groups of messages affected by rule A.5

Variable	Modality	Type	ID	Class	Sign	Action	Weight
Aesthetics	*excellent*	VoI	#26	3	pos.	Prioritise	2
			#16	2	neg.	Discard	0
			#41	1	neg.	Discard	0
Communication	*regular*	VoI	#14	2	pos.	Prioritise	2
			#40	1	neg.	Discard	0
			#61	3	neg.	Discard	0
Location	*mountain/sky run towns*	VoI	#23	3	pos.	Prioritise	2
			#2	1	neg.	Discard	0
			#49	2	neg.	Discard	0
PromosSentit	*high*	EF	#80	3	pos.	Prioritise	5
			#76	1	neg.	Discard	-1
			#77	2	neg.	Discard	0

Table 5.19: Frequency of each final weight after applying type-A rules

Weight	-1	0	1	2	3	4	5
Number of messages	9	16	36	13	4	1	1

the same number of discards and prioritisations. The list of these secondly discarded messages is the following: #5, #29, #69.

5.4.3 Document planning

The third stage of the NLG system described in this work consists of two tasks. On the one hand, to discover relations between the 52 filtered messages to combine them into the same sentence (type-B rules). On the other hand, to identify specific modifications to be made on the messages according to the semantics of their variable or modality (type-C rules). These two tasks are widely explained in Section 4.3. Rules are sequentially analysed and the identified relations are annotated to be used

in the final stage. Note that although the merging process is performed in the final stage, this stage only takes note of which messages are affected by which type-B or type-C rules.

Rule B.1 merges messages of the same ordinal variable with consecutive modalities. Messages must also share the type and sign in order to be merged into the same sentence. This rule is activated in three pairs of messages, as shown in Table 5.20. The order of each modality within its variable (in addition to the number of modalities of the variable) is included to better understand the application of this rule.

Table 5.20: Groups of messages affected by rule B.1

Class	**Variable**	**Type**	**Sign**	**ID**	**Modality**	**Order**
1	*Assistants*	VoI	pos.	#10	*many*	2/3
				#11	*a lot of*	3/3
	Size	VoI	pos.	#33	*medium*	2/3
				#34	*big*	3/3
2	*PromosSensit*	VoI	pos.	#55	*low*	1/3
				#56	*medium*	2/3

For instance, texts associated with both messages #55 and #56 are very similar: "*The proportion of shops with a low sensitivity to promotions is high*" and "*The proportion of shops with a medium sensitivity to promotions is high*". They can easily be merged by taking into account the order of their modalities: "*The proportions of shops with a sensitivity to promotions lower than or equal to medium are high*". As said before, this merging process will be detailed in the last stage.

Rule B.2 is a less restrictive version of rule B.1 in the sense that it groups messages of the same variable where order is not required. Moreover, it associates messages of different sign and even not sharing type. These rule is quite generic, and 16 pairs of messages have activated it. In a first step, messages sharing sign are taken into account, as shown in Table 5.21.

Sentences obtained from merging messages sharing sign can be more compacted, but they must have the same sign. The sentence associated with messages #31 and #32, both of them VoIs, is: "*The proportions of shops located in inner cities and no mountain towns are high*", while messages #36 and #70 cannot be compacted: "*The proportion of shops with a regular maintenance is low and none of the shops has a deficient maintenance*".

Table 5.21: Groups of messages affected by rule B.2 (first step, same sign).

Class	Variable	Sign	ID	Modality	Type	Order
1	*Location*	pos.	#31	*inner cities*	VoI	-
			#32	*no mountain*	VoI	-
	Maintenance	neg.	#36	*regular*	VoI	2/4
			#70	*deficient*	EF	1/4

The result of a second step, in which messages with different sign are analysed, is detailed in Table B.18 (see Appendix B, page 163). An extract of messages concerning variable *Competition* is displayed in Table 5.22.

Table 5.22: Groups of messages of variable *Competition* affected by rule B.2 (second step, different sign). The rest of groups are detailed in Table B.18

Class	Variable	ID	Modality	Type	Sign	Order
1	*Competition*	#30	*strong*	VoI	pos.	3/3
		#68	*no*	EF	neg.	1/3
2	*Competition*	#47	*no*	VoI	pos.	1/3
		#48	*strong*	VoI	neg.	3/3

Rule B.3 merges messages having the same modality, with variables with the same exact list of modalities. It only takes into account messages not being affected by previous type-B rules. Grouped messages must share their type, and messages are processed in two steps: it firstly analyses messages of the same sign and it secondly takes into account all messages regardless their sign. This rule has not been activated by any group of messages.

Rule B.4 merges single messages (those not having any relation with other messages) with messages or groups of messages sharing the related variable. Therefore, the final text will mention in the same sentence all messages related to the same variable. Only message #37 has been merged into one of the existing groups. That group was created by the application of rule B.2 on messages #36 and #70. Table 5.23 show the features of these three messages.

Table 5.23: Messages affected by rule B.4

Class	Variable	ID	Modality	Type	Sign	Order
1	*Maintentance*	#37	*good*	VoI	pos.	3/4
		#36	*regular*	VoI	neg.	2/4
		#70	*deficient*	EF	neg.	1/4

The result of the merging process of these three messages is detailed in the next stage.

Rule B.5, as the last rule for merging messages, merges the last single messages (those not having any relation with other messages) with other single messages that share the same type and sign. This rule is activated in three groups of messages: EF–positive, VoI–positive and VoI–negative, as shown in Table 5.24.

Table 5.24: Messages affected by rule B.5

Class	Type	Sign	ID	Variable	Modality
1	EF	pos.	#71	*DisplaySize*	*medium*
			#75	*PromosSensit*	*medium*
2	VoI	neg.	#52	*ThermalExhibitor*	*yes*
			#54	*Assistants*	*a lot of*
3	VoI	pos.	#64	*DisplayGrifone*	*yes*
			#65	*GrifoneWeight*	*main*

For instance, the first group of messages can be merged in the same sentence as follows: "*Almost all shops have a medium-sized display and a medium sensitivity to promotions*".

In conclusion, type-B rules have been activated in 45 out of the 52 filtered messages, thus leaving 7 single messages.

Type-C rules detect which sentences will have to be modified according to the semantics of variables or modalities of the associated messages. For this reason, the ontology previous defined must be used. In concrete, rules C.3 and C.4 must know which variables need a special construction and if modalities of variables can be treated as adjectives of the variables.

Rule C.1 looks for messages having the modality "no". This kind of messages must adjust their transcription according to the meaning of modality "no" within the variable. There are three messages affected by this rule (all three messages with modality "no"), detailed in Table 5.25.

Table 5.25: Messages affected by rule C.1

Modality	ID	Class	Variable	Type	Sign
	#17	2	*DisplayGrifone*	VoI	pos.
no	#47	2	*Competition*	VoI	pos.
	#68	1	*Competition*	EF	neg.

The composition of the texts, as it will be explained in the final stage, depends on the sign of the message. For instance, message #17 can be transcribed as "*The proportion of shops not showing Grifone products is high*", but negative messages must change the sign of the sentence, as in message #68: from "*None of the shops has a "no" competition*" to "*All shops have competition*".

Analogously to rule C.1, **rule C.2** takes into account the semantics of modality "yes". Three of the messages present this modality, as shown in Table 5.26.

Table 5.26: Messages affected by rule C.2

Modality	ID	Class	Variable	Type	Sign
	#9	1	*ThermalExhibitor*	VoI	pos.
yes	#52	2	*ThermalExhibitor*	VoI	neg.
	#64	3	*DisplayGrifone*	VoI	pos.

For instance, text associated with message #9 can be "*The proportion of shops with thermal exhibitor is high*", without mentioning the modality.

Rule C.3 uses the defined ontology in order to identify which messages need a special treatment to be transcribed in a natural way. Among the 13 relevant variables, seven of them need this special treatment, as shown in Table 5.27, in which a comment about the way of transcribing these messages when related to both VoIs and EFs is included.

Thus, 22 messages related to variables included in Table 5.27 have activated this rule.

Table 5.27: Variables needing a special transcription (rule C.3)

Variable	**VoIs transcription**	**EFs transcription**
Competition	with/without competition	have/have not competition
Location	located in	are located in
DisplaySize	with a *medium*-sized display	have a *medium*-sized display
DisplayGrifone	displaying Grifone products	display Grifone products
ThermalExhibitor	with thermal exhibitor	have thermal exhibitor
Assistants	with *few* assistants	have *few* assistants

Rule C.4 uses the ontology to detect which of the messages that do not need any special management when transcribed can be easily treated by just placing their modality before the variable. This is correct because modalities are adjectives of their corresponding variables. Table 5.28 lists and exemplifies variables whose messages are affected by this rule. Note that the construction of all these variables is very similar.

Table 5.28: Variables with modalities as adjectives (rule C.4)

Variable	**VoIs transcription**	**EFs transcription**
Size	with a *small* size	have a *small* size
Maintenance	with a *good* maintenance	have a *good* maintenance
Communication	with an *excellent* communication	have an *excellent* communication
Aesthetics	with a *regular* aesthetics	have a *regular* aesthetics
GrifoneWeight	with a *main* Grifone weight	have a *main* Grifone weight
Evaluation	with a *bad* evaluation	have a *bad* evaluation
PromosSensit	with a *high* sensitivity to promotions	have a *high* sensitivity to promotions

In general, rules C.3 and C.4 do not have to be complementary, that is to say, there can exist variables not affected by them. But in this case, all variables not affected by previous rule C.3 are affected by rule C.4, thus having 30 messages affected by it.

Rule C.5 identifies messages that need the use of linguistic quantifiers. There are two main

type of quantifiers used by this system. On the one hand, highly relevant VoIs use the quantifier "very" to emphasise their relevance: "*The proportion of shops with a good maintenance is very high*". On the other hand, all EFs use the quantifiers "all", "almost all', "almost none" and "none" to specify how many shops the sentence refers to: *"All individuals have a medium sensitivity to promotions"*. In this study, 25 of the messages have activated this rule and therefore will use these quantifiers in their transcription in the final text.

Table 5.29 summarises the number of messages analysed by each of the rules of this stage, and how many of them have activated each rule. Sequential analysis of type-B rules, where each rule only takes into account messages not affected by previous rules, is shown in their decreasing number of analysed messages.

Table 5.29: Number of messages implied in the analysis of each rule

Rule	**# analysed messages**	**# affected messages**
B.1	52	6
B.2	46	32
B.3	14	2
B.4	12	3
B.5	11	4
C.1	52	3
C.2	52	3
C.3	52	22
C.4	52	30
C.5	52	25

5.4.4 Microplanning and realisation

The last stage of the system is responsible for deciding the final structure of the text and transcribing groups of messages into proper text. In the first substage, messages are sorted and grouped according to the criteria defined in Section 4.4. Table 5.30 details the 29 obtained groups of mes-

sages.

Table 5.30: Final planning of the text

Class	Group	Messages	B Rules	Group	Messages	B Rules
1	1	#71, #75	B.3	6	#74, #8	B.2
	2	#9	-	7	#10, #11	B.1
	3	#68, #30	B.2	8	#33, #34	B.1
	4	#70, #36, #37	B.2, B.4	9	#31, #32	B.2
	5	#73, #6	B.2	10	#44, #43	B.2
2	11	#17	-	16	#47, #48	B.2
	12	#14, #15	B.2	17	#50, #13	B.2
	13	#55, #56	B.1	18	#51	-
	14	#19, #53	B.2	19	#12	-
	15	#20, #21	B.2	20	#52, #54	B.5
3	21	#80	-	26	#25	-
	22	#64, #65	B.5	27	#27, #67	B.2
	23	#23, #57	B.2	28	#59, #58	B.2
	24	#26, #62	B.2	29	#66	-
	25	#24, #60	B.2			

Realisation of the text is a sequential process that explores each group of messages transcribing them into a phrase. This transcription is carried out according to the way of merging messages inside a group defined in Table 4.13, and taking into account the specific way of generating each sentence identified by type-C rules.

The construction of the phrase corresponding to group #4 is exemplified below. Table 5.31 shows the features of its components, messages #70, #36 and #37.

Standard sentences corresponding to the three messages are the following:

- (#37) *The proportion of shops with modality "good" in variable "maintenance" is high.*
- (#70) *None of shops has modality "deficient" in variable "maintenance".*
- (#36) *The proportion of shops with modality "regular" in variable "maintenance" is low.*

Table 5.31: Features of messages of group #4

ID	Class	Variable	Modality	Type	Sign	Relev.	Value
#37			*good*	VoI	pos.	normal	3.0
#70	1	*Maintenance*	*deficient*	EF	neg.	-	0.0
#36			*regular*	VoI	neg.	normal	2.2

All messages are affected by rule C.4, because modalities of variable *Maintenance* can be treated as their adjectives. Therefore, sentences are transformed into the following:

- (#37) *The proportion of shops with a good maintenance is high.*
- (#70) None of shops has a deficient maintenance.
- (#36) The proportion of shops with a regular maintenance is low.

Messages #70 and #36 are affected by rule B.2, so a sentence between them must be firstly created. Looking at Table 4.13 both messages correspond to different types but their sign is the same. Their sentence cannot be totally compacted and the conjunction to be used between both messages must be "and", as in Example 4.31:

- (#70 $\wedge$ #36) *None of shops has a deficient maintenance and the proportion of shops with a regular maintenance is low.*

The unique modification that can be made is the use of the pronoun "them" in the second sentence:

- (#70 $\wedge$ #36) *None of shops have a deficient maintenance and the proportion of them with a regular maintenance is low.*

The result of this merging step has to be joined with message #37 by using conjunction "and" of a full stop, depending on if previous sentence has been compacted or not. As previous sentence could not be compacted, a full stop is used between them:

- (#37 $\wedge$ (#70 $\wedge$ #36)) *The proportion of shops with a good maintenance is high. None of PoSs has a deficient maintenance and the proportion of them with a regular maintenance is low.*

Finally, in order to avoid repetitive mentions to the name of individuals several alternatives have been provided: "PoSs", with a probability of 0.6, "shops" (0.25) and "stores" (0.15). The final qualitative description obtained from classification #259 is the following:

```
Class 1
=======
Almost all shops have a medium-sized display and a medium sensitivity to promotions.
The proportion of PoSs with thermal product display is very high.
All PoSs have competition and the proportion of them with a strong competition is high.
The proportion of shops with a good maintenance is high. None of PoSs has a
deficient maintenance and the proportion of them with a regular maintenance is low.
None of PoSs has a deficient communication and the proportion of them with a good
communication is very high.
None of PoSs has a deficient aesthetics and the proportion of them with a good
aesthetics is very high.
The proportions of shops with a number of assistants greater than or equal to many
are very high.
The proportions of stores with a size greater than or equal to medium are high.
The proportions of shops located in inner cities and no mountain towns are high.
The proportion of PoSs with a secondary Grifone weight is high while the proportion
of them with a minimal Grifone weight is low.

Class 2
=======
The proportion of PoSs not displaying Grifone products is very high.
The proportion of stores with a regular communication is very high (moreover, most
of PoSs with a regular communication are in class 2) while the proportion of them
with an excellent communication is very low.
The proportions of PoSs with a sensitivity to promotions lower than or equal to
medium are high.
The proportion of shops with a minimal Grifone weight is very high while the
proportion of them with a main Grifone weight is low.
```

The proportion of PoSs with a good evaluation is very high while the proportion of them with an excellent evaluation is very low.
The proportion of PoSs without competition is high while the proportion of them with a strong competition is low.
The proportion of shops with a regular maintenance is high while the proportion of them with an excellent maintenance is very low.
The proportion of shops with a regular aesthetics is high.
The proportion of PoSs with a big size is very low.
The proportion of shops with thermal product display or with a lot of assistants is low.

Class 3
=======
All PoSs have a high sensitivity to promotions (moreover, all PoSs with a high sensitivity to promotions are in class 3).
The proportion of shops displaying Grifone products or with a main Grifone weight is high.
The proportion of shops located in mountain or ski towns is very high (moreover, most of PoSs located in mountain or ski towns are in class 3) while the proportion of them located in inner cities is low.
The proportion of stores with an excellent aesthetics is very high (moreover, most of stores with an excellent aesthetics are in class 3) while the proportion of them with a good aesthetics is low.
The proportion of PoSs with an excellent maintenance is very high while the proportion of them with a good maintenance is low.
The proportion of PoSs with an excellent communication is very high.
The proportion of PoSs with an excellent evaluation is very high while the proportion of them with a good evaluation is low.
The proportion of stores with a big size is high while the proportion of them with a medium size is low.
The proportion of stores with many assistants is low.

5.5 Discussion and managerial implications

This section firstly discusses some aspects of the ranking and selection method while in a second part managerial implications derived by analysing in detail the qualitative description of the three considered classes are described.

5.5.1 Discussion

The real case study presented in Subsection 5.3 illustrates one of the main advantages of the proposed methodology for ranking and selecting classifications that is its capability to deal with the ambiguity that appears when managing multiple criteria associated with fuzzy concepts. However, one of the drawbacks of the study carried out in this chapter is that it considers subjective variables obtained from different sales representatives. This can lead to incongruences if, for instance, some of the representatives tends to give worse (or lower) values than the others, causing his/her shops to have an overall worse evaluation. Nevertheless, this problem has been minimised by clearly defining the meaning of linguistic labels used to evaluate each feature of the stores prior to data collection process. One could think that a possible solution could be to have a unique person assessing all the shops but it is almost impossible to count on one representative encompassing knowledge about all considered individuals (260 points of sale).

Despite this approach does not take into account relative criteria for assessing classifications, it covers almost all the validation concepts considered in the literature for evaluation classifications (see Table 2.1 in page 12). In contrast, most of the reviewed approaches in clustering validation only take into account some of those aspects.

The fuzzy indicators considered in this work can cause an increase of complexity in assessing each of the classifications, but at the same time they provide to the process the inherent flexibility of human reasoning. In addition, the consideration of OWA operators introduces the concept of majority "most of", widely employed in assessment processes and, in particular, in marketing evaluations. The RIM function used in this work guarantees that all the individual valuations contribute to the final aggregated value. It is important to note that the higher the ranking of a value, the higher the weighting value associated with it, which is appropriate for conducting aggregation processes in heterogeneous decision-making problems (Chiclana et al., 2007).

Below a discussion about the hypothetical rank defined by each criterion is discussed. Table

5.32 shows the best segmentations looking exclusively at the values obtained from the *balanced classes* criterion. Although in this study it has been desirable to obtain a balanced segmentation, it is not indispensable, as indicated by the fact that best overall segmentations (#259, #260 and #258) perform below segmentations #272, #189 and #172, the best ones according to I_B.

Table 5.32: Best segmentations according to I_B

ID	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**	**rank**
#272	Minmax	0.958	2	0.5	0.995	0.283	0.010	0.802	0.6996	75
#189	Minmax	0.782	3	1	0.992	0.277	0.018	0.168	0.6923	124
#172	Minmax	0.772	3	1	0.990	0.283	0.008	0.400	0.7222	44

Best segmentations according to the *coherence* criterion are included in Table 5.33. Values obtained by this criterion are fairly homogeneous and therefore exhibiting a low influence in the process of selecting the best segmentation in the methodology presented here. Indeed, the classifications discarded with the present methodology are caused by their low performance in other criteria, as illustrated by segmentation #246, which performs very well in all criteria but I_D.

Table 5.33: Best segmentations according to I_C

ID	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**	**rank**
#214	Minmax	0.904	2	0.5	0.914	0.295	0.006	0.795	0.6632	233
#241	Minmax	0.843	3	1.0	0.965	0.290	0.012	0.108	0.6815	204
#246	Minmax	0.911	3	1	0.921	0.289	0.013	0.770	0.7633	22

The best three segmentations with the present methodology are kept using the *dependency* criterion and shown in Table 5.34, which indicated a high importance of this criterion in this methodology. However, it is important to note that this criterion is no decisive, as segmentation #258 is ranked lower than segmentation #260 in the overall ranking due to its lower performance in the other criteria despite having a higher value in I_D.

Table 5.35 includes the best segmentations according to their *accuracy*. The fact that these segmentations do not show a high overall rank, and the best overall segmentations in Table 5.4 exhibit

Table 5.34: Best segmentations according to I_D

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#259	Minmax	0.439	3	1	0.928	0.251	0.528	0.936	0.8423	1
#258	Minmax	0.422	4	1	0.885	0.226	0.425	0.875	0.8103	3
#260	Minmax	0.469	3	1	0.929	0.255	0.363	0.922	0.8208	2

a very high accuracy degree illustrates that in this case study a great number of segmentations that are quite accurate also perform well in the rest of criteria.

Table 5.35: Best segmentations according to I_A

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#003	Frank (0.5)	0.055	2	0.5	0.841	0.193	0.001	0.988	0.6921	125
#002	Frank (1.5)	1	2	0.5	0.841	0.193	0.001	0.988	0.6921	126
#001	Frank (3)	0.5	2	0.5	0.842	0.149	0.001	0.988	0.6871	164

The present methodology avoids, on the one hand, the predefinition of arbitrary thresholds associated with each considered criterion in order to decide which segmentations are taken into account in the application of subsequent criteria. On the other hand, it is clear that the sequential application of the above set of criteria would have prevented any of the ten best overall classifications shown in Table 5.4 and obtained with the methodology to have been ranked in those positions. For instance, classification #243 is the fourth best one according to its aggregated value, but it would have been discarded due to its low assessment according to the dependency criterion I_D. Therefore, the case study clearly illustrates that the methodology presented avoids discarding segmentations that could be potentially useful for marketing experts when observed globally.

5.5.2 Managerial implications

The purpose of the generation of an automatic qualitative description of the chosen segmentation is to allow marketing managers to understand the considered segments in order to define marketing actions to improve their customers' satisfaction. In general, analysing the obtained description,

marketing experts will build their understanding of each segment and even they can give them a label to make clearer the meaning of each class.

Taking into account the qualitative description of segmentation #259 automatically obtained by applying the presented methodology, in the following paragraphs a label is chosen for each class and interpretation of the main features and implications are exposed.

Class 1: "Multi-sports shops".

Consists of 35 points of sale with a strong competition, good qualities (in terms of maintenance, communication and aesthetics), big stores (in size and number of assistants), non located in mountain towns and a secondary Grifone weight.

Class 1 might correspond to multi-sports shops having large stores, selling many different products and not being located in mountain cities. As they are not mountain-sports specialists, so the marketing campaigns evolving this type of cities should be made to enhance the attraction of mountain sports, taking into account the medium sensitivity to promotions of the customers of these shops.

Class 2: "They don't like us".

This class has 98 shops with a minimal presence of Grifone products. Qualities of stores are regular, although our commercials rated them pretty good. They do not use to have a high level of competition and the points of sale are not too big (in size and number of assistants).

From our brand's point of view, shops of Class 2 seem to be the less interesting shops. Grifone is not well situated and customers of these shops demonstrate a low sensitivity to promotions.

Class 3: "The top shops".

This is the largest class with 127 shops generally sector specialists, being situated in mountain or ski towns. The points of sale are usually large without too assistants, their qualities are the best of the analysed stores and Grifone brand is well situated.

The main feature os these shops is the high sensitivity to promotions of their customers. This means that these points of sale are good candidates to give the best revenues to our marketing campaigns. Being the shops of Class 3 the Grifone's favourite clients, an analysis of their sales should be carried out in order to detect the ones with the lower purchases level as a possible target of a marketing campaign.

5.6 Conclusions

This chapter details the application of the methodologies described in Chapters 3 and 4 on a B2B real case study. The study analyses a dataset of 260 shops that distribute outdoor sporting equipment of Grifone, the brand considered. Three actions have been carried out to solve the challenge identified in this case study: to automatically obtain segmentations of the set of points of sale, to select the best market segmentation according to the marketing department and to obtain a qualitative description of this chosen segmentation in order to understand the defined segments and therefore design specific and optimal marketing campaigns for each segment.

The application of the unsupervised version of LAMDA algorithm permits us to automatically obtain more than 500 stable ways of segmenting the considered shops. Through the application of criteria explained in Chapter 3, segmentations have been assessed, globally ranked and the best one according to those criteria has been chosen. Finally, after explicit short domain information about involved variables, an automatic qualitative description of the shops made in natural language has been provided in order to better understand the considered segmentation.

In summary, a complete marketing challenge that includes segmentation of customers, selection and description of the best segmentation has been faced in this chapter by applying methodologies developed in this thesis, demonstrating their power and usability in a marketing case study. The main contribution of this chapter is to show the global capacity of the introduced methodologies in a real case study. In addition, it is demonstrated the goodness of the use of OWA operators to aggregate a set of criteria instead of analysing each criterion sequentially, showing This approach avoids the need of defining arbitrary thresholds for each considered criterion in order to discard classifications with an assessment below those thresholds.

As future work, scalability of the system must be analysed with the aim of avoiding situations in which for instance a number of obtained segmentations excessively high causes the problem to be unmanageable. Moreover, from a general point of view, improvements detailed in the conclusions section of each chapter must be analysed and developed.

Chapter 6

Limitations and future research

This chapter collects and expands the conclusions of each of the chapters contained in this doctoral dissertation by considering a more general point of view. Further research is also included at the end of the chapter.

A complete multi-criteria decision making (MCDM) system has been presented in this thesis with the objective of creating, selecting and making understandable the best classification among a set of individuals. In order to justify the relevance of the research done, a synthetic overview of the state of the art is provided in Chapter 2 on the topics of criteria for selecting classifications, aggregation functions for summarising individual assessments according to the defined criteria, and data-to-text systems for generating natural language texts. A theoretical framework has been proposed in two chapters that attempted to analyse a set of fuzzy criteria for assessing classifications (Chapter 3) and to design a natural language generation (NLG) system to qualitatively describe the most important characteristics of the considered clusters of a segmentation (Chapter 4). The proposed methodologies have been applied in Chapter 5 by developing a real marketing case study framed in a business to business (B2B) environment.

More specifically, in Chapter 3 a set of five fuzzy criteria to assess classifications has been proposed, covering almost all the concepts employed when evaluating classifications. Each criterion is modelled by means of an index that measures the degree up to which the criterion is met by each classification. The properties and usability of the defined criteria have been explained and proven. The most important contributions of this chapter are related to the mathematical characterisation

and formal definition of the indexes associated with the proposed criteria. The indexes are proposed to be aggregated by using an Ordered Weighted Averaging (OWA) operator guided by a fuzzy linguistic quantifier that is used to implement the concept of fuzzy majority in the process. All the theoretical background on which this process is based is detailed in Chapter 2, more specifically in Section 2.2.

The aggregation of the defined indexes permits us to overall assess each alternative classification and therefore the selection of the most suitable one. The design of a NLG system to describe qualitatively the most important characteristics of this best classification is detailed in Chapter 4. This rule-based automatic system is able to produce context-based text by specifying some short domain-specific information but the lack of this information does not prevent the generation of a generic but informative text. The adaptation of a four-stage architecture for data-to-text systems, the detection of the most relevant characteristics of each class by means of the concept of dependence and the design of a set of rules to filter and merge messages of the final text are the main contributions of this chapter.

These methodologies have been applied to a real marketing case based on a B2B environment and presented in Chapter 5. This study encompasses all processes described in this dissertation: generating segmentations of points of sale by employing the unsupervised learning capabilities of LAMDA algorithm, assessing each segmentation by applying the defined criteria and by computing their corresponding indexes, aggregating the indexes in order to obtain an overall evaluation of each segmentation, ranking those segmentations and therefore selecting the best one, and generating a natural text describing the considered classes. The main contribution of this application chapter is the assessment of segmentations by means of an overall degree up to which most of the defined criteria are met by each one, instead of employing the sequential approach.

Further research con be summarised in three lines: to improve the defined criteria and their corresponding indexes, to further study other functions for aggregating information and finally, to improve the qualitative description quality and naturalness. Regarding the fuzzy criteria for assessing classifications, the following specific objectives can be taken into account for further research:

- Generalisation of the defined criteria in order to consider all possible needs of the end user.
- Design of a user-friendly method for choosing parameters needed in the computation of the indexes.

- Definition and analysis of alternative selection criteria.

Regarding the aggregation of the indexes in order to rank classifications, these are the objectives considered for future research:

- Study of the different types of OWA operators with the aim of improving the aggregation of indexes. For instance, the importance of the considered criteria can be taken into account, or even the computed indexes can take the form of fuzzy numbers.
- Study of other linguistic quantifiers to obtain the weights vector used by OWA operators, by extending the study from quantifier "most" to others like "at least α" or "almost all".

With regard to the natural text generation for describing a set of clusters, the following objectives are being considered to further investigate:

- Improvement of the detection of important values in each cluster.
- Formalisation of the ontology to be used to generate a more natural description.
- Completing the definition of the grammar to be used to realise the text, and design and implementation of the system to use the defined grammar.

Finally, from the application point of view, there are three specific objectives to continue the research:

- Use of the defined criteria to assess the performance of unsupervised learning techniques in different problems in order to understand their limitations in theoretical and real-world problems.
- Use of the defined methodology to ensemble several machine learning techniques in order to aggregate their outputs and to design an ensemble-based system.
- Application of the considered methodologies to other real problems. These methodologies will be used to segment a set of users according to their profile. This profile will be constructed using social networks (e.g. Facebook, LinkedIn) to capture their interest. This segmentation can be used to boost innovation by recommending bridges between users from different segments.

Appendix A

Learning Algorithm for Multivariate Data Analysis (LAMDA)

The machine learning technique used in this thesis is Learning Algorithm for Multivariate Data Analysis (LAMDA). Despite its study is not within the scope of this thesis, it has been decided to introduce a description of LAMDA in order to help the reader to understand the basic operation of such an important tool for this work.

LAMDA was originally developed by Joseph Aguilar (Aguilar and López de Mántaras, 1982), and implemented by Juan Carlos Aguado (Aguado et al., 1999). LAMDA is based on fuzzy hybrid connectives, and employs the interpolation capabilities of logic operators over fuzzy environments (Klir and Yuan, 1995). A linearly compensated hybrid connective, considered as an interpolation between a t-norm T and its dual t-conorm T^* ($T^*(x,y) = 1 - T(1-x, 1-y)$) is used:

$$H = (1-\beta)T + \beta T^*,$$

where $\beta \in [0,1]$ is known as the level of tolerance of the classification. It can be noted that for $\beta = 0$, the t-norm is obtained, and for $\beta = 1$, the t-conorm is the result. Taking into account that the t-norms are fuzzy operators associated to an intersection or conjunction and the t-conorms are associated to a union or disjunction, the parameter β determines the exigency level of the

classification. Obviously, we can define $\lambda = 1-\beta$ as the tolerance level of a given classification. The exploration of different tolerance levels is done automatically by taking into account all tolerances that cause the generation of different classifications (Aguado et al., 1999).

The employed hybrid connectives have been obtained from t-norms Min, probabilistic product, Lukasiewicz, and Frank n-norms together with their dual t-conorms. Below the proper expression for each t-norm and t-conorm is detailed.

- MinMax:
  - Min: $M(x_1,\ldots,x_n) = min\{x_1,\ldots,x_n\}$
  - Max: $M^*(x_1,\ldots,x_n) = max\{x_1,\ldots,x_n\}$
- Probabilistic Product:
  - $\prod(x_1,\ldots,x_n) = x_1 \cdot \ldots \cdot x_n$
  - $\prod^*(x_1,\ldots,x_n) = 1 - \prod_{i=1}^{n}(1-x_i)$
- Lukasiewicz:
  - $W(x_1,\ldots,x_n) = max\{1-n+\sum_{i=1}^{n} x_i, 0\}$
  - $W^*(x_1,\ldots,x_n) = min\{\sum_{i=1}^{n} x_i, 1\}$
- Frank n-norms:
  - $F_s(x_1,\ldots,x_n) = log_s(\frac{1-\prod_{i=1}^{n}(s^{x_i}-1)}{(s-1)^n-1}+1)$
  - $F_s^*(x_1,\ldots,x_n) = 1 - log_s(\frac{1-\prod_{i=1}^{n}(s^{1-x_i}-1)}{(s-1)^n-1}+1)$,
    for $s \in (0,+\infty)$ and $s \neq 1$.

Note that the t-norms M, $\sum$ and $\prod$ are obtained from F_s taking limits when $s \to 0$, $s \to 1$ and $s \to \infty$, respectively.

Each segmentation depends, on the one hand, on the hybrid connective employed and, on the other hand, on the selected exigency or tolerance level. Minimum tolerance means that an individual is assigned to a class only if every descriptor points to that, whereas a maximum tolerance implies that the individual is assigned to a class if this is the indicated by at least one of the descriptors. LAMDA is able to automatically explore any in-between tolerance degree. The basic LAMDA operation is depicted in Figure A.1.

Marginal Adequacy Degree (MAD) is computed for each descriptor, class and individual. These partial results are aggregated by means of the hybrid connectives H to supply the Global Ade-

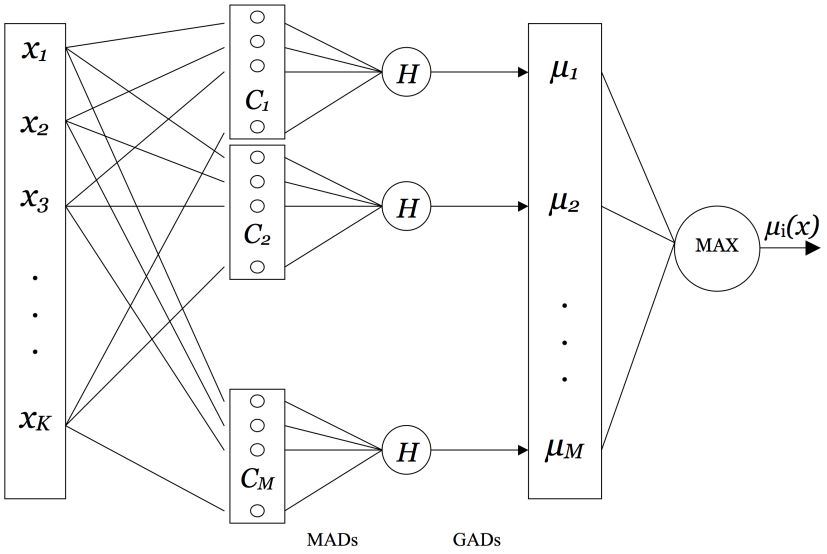

Figure A.1: LAMDA: hybrid connectives-based classification

quacy Degree (GAD) of an individual to a class. The visible structural resemblance between the LAMDA algorithm and the Artificial Neural Networks is worth noting, especially for the Radial Base Functions (RBF) type. LAMDA exhibits greater flexibility than neural networks, for example, in its ability to perform either a supervised or an unsupervised learning process indistinctly and its capability to combine pattern recognition with a simple, non-iterative class upgrading.

Whenever the unsupervised learning LAMDA capabilities are employed, the first individual is always placed in a class by itself. Subsequent individuals can then be be assigned to the already existing class(es) or a new one is created. In order to determine when this occurs, the algorithm generates, first of all, a special class called Non-Informative Class (NIC) which represents maximum entropy, with the characteristic of returning the same (low) GAD for every possible individual. As such, the decision-making process consists of comparing the GAD of the individual to the NIC class with the GADs to every other existing class. If one of the real classes returns the maximum GAD, the new individual will be assigned to it and the class will be modified accordingly. But if the NIC is the one with the highest GAD, this means that none of the existing classes are close enough to the individual and so a new class has to be created by the LAMDA algorithm.

Appendix B

Tables of the case study

Table B.1: Best 100 segmentations

Rank	**ID**	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
1	#259	MinMax	0.439	3	1.0	0.928	0.251	0.528	0.936	0.8423
2	#260	MinMax	0.469	3	1.0	0.929	0.255	0.363	0.922	0.8208
3	#258	MinMax	0.422	4	1.0	0.885	0.226	0.425	0.875	0.8103
4	#243	MinMax	0.290	3	1.0	0.909	0.257	0.008	0.965	0.7868
5	#244	MinMax	0.304	3	1.0	0.920	0.256	0.012	0.948	0.7856
6	#257	MinMax	0.411	3	1.0	0.933	0.280	0.032	0.856	0.7787
7	#256	MinMax	0.400	4	1.0	0.872	0.246	0.064	0.906	0.7752
8	#253	MinMax	0.359	4	1.0	0.883	0.232	0.021	0.915	0.7722
9	#252	MinMax	0.352	4	1.0	0.884	0.238	0.025	0.904	0.7714
10	#255	MinMax	0.393	4	1.0	0.939	0.264	0.063	0.768	0.7687
11	#254	MinMax	0.381	4	1.0	0.942	0.262	0.047	0.772	0.7677
12	#263	MinMax	0.511	3	1.0	0.894	0.279	0.013	0.842	0.7672
13	#262	MinMax	0.500	3	1.0	0.894	0.279	0.012	0.835	0.7662
14	#250	MinMax	0.350	5	1.0	0.872	0.242	0.112	0.820	0.7660
15	#276	MinMax	0.963	3	1.0	0.946	0.264	0.027	0.763	0.7654

Continued on next page

Table B.1 – continued from previous page

Rank	**ID**	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
16	#266	MinMax	0.944	3	1.0	0.955	0.264	0.019	0.754	0.7648
17	#261	MinMax	0.491	3	1.0	0.883	0.274	0.009	0.845	0.7646
18	#70	Frank	0.949	3	1.0	0.957	0.250	0.024	0.753	0.7640
19	#249	MinMax	0.343	5	1.0	0.887	0.239	0.120	0.782	0.7639
20	#246	MinMax	0.911	3	1.0	0.921	0.290	0.013	0.770	0.7635
21	#258	MinMax	0.940	3	1.0	0.923	0.278	0.045	0.742	0.7617
22	#248	MinMax	0.340	5	1.0	0.887	0.235	0.119	0.767	0.7612
23	#224	MinMax	0.827	3	1.0	0.923	0.266	0.009	0.743	0.7565
24	#278	MinMax	0.965	3	1.0	0.928	0.281	0.034	0.687	0.7541
25	#269	MinMax	0.947	3	1.0	0.928	0.278	0.026	0.689	0.7533
26	#207	MinMax	0.802	3	1.0	0.903	0.264	0.005	0.746	0.7527
27	#164	MinMax	0.731	4	1.0	0.955	0.255	0.015	0.663	0.7506
28	#135	MinMax	0.663	4	1.0	0.946	0.203	0.038	0.678	0.7473
29	#133	MinMax	0.662	4	1.0	0.965	0.217	0.024	0.652	0.7472
30	#140	MinMax	0.665	4	1.0	0.933	0.239	0.013	0.682	0.7471
31	#2	Frank	0.001	3	1.0	0.862	0.147	0.041	0.757	0.7365
32	#239	MinMax	0.842	3	1.0	0.931	0.269	0.020	0.538	0.7305
33	#199	MinMax	0.794	3	1.0	0.950	0.281	0.034	0.488	0.7298
34	#3	Prob.	0.001	3	1.0	0.843	0.141	0.032	0.741	0.7289
35	#191	MinMax	0.789	3	1.0	0.952	0.259	0.011	0.508	0.7279
36	#192	MinMax	0.788	3	1.0	0.949	0.255	0.012	0.511	0.7275
37	#249	MinMax	0.934	3	1.0	0.949	0.275	0.006	0.492	0.7264
38	#43	Frank	0.980	3	1.0	0.965	0.249	0.026	0.475	0.7262
39	#69	Frank	0.956	3	1.0	0.963	0.249	0.020	0.483	0.7261
40	#280	MinMax	0.964	3	1.0	0.939	0.287	0.022	0.473	0.7251
41	#75	Frank	0.969	3	1.0	0.964	0.252	0.028	0.460	0.7244
42	#96	MinMax	0.624	3	1.0	0.968	0.270	0.010	0.452	0.7243
43	#227	MinMax	0.832	3	1.0	0.954	0.279	0.015	0.458	0.7240

Continued on next page

Table B.1 – continued from previous page

Rank	**ID**	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
44	#188	MinMax	0.783	3	1.0	0.979	0.267	0.020	0.431	0.7239
45	#172	MinMax	0.772	3	1.0	0.990	0.284	0.008	0.400	0.7224
46	#201	MinMax	0.801	3	1.0	0.969	0.281	0.033	0.379	0.7177
47	#267	MinMax	0.946	3	1.0	0.979	0.274	0.013	0.374	0.7158
48	#243	MinMax	0.845	3	1.0	0.967	0.274	0.013	0.383	0.7150
49	#271	MinMax	0.974	3	1.0	0.905	0.271	0.011	0.459	0.7137
50	#165	MinMax	0.727	4	1.0	0.914	0.262	0.014	0.450	0.7133
51	#193	MinMax	0.790	3	1.0	0.966	0.259	0.014	0.374	0.7119
52	#186	MinMax	0.781	4	1.0	0.919	0.275	0.009	0.425	0.7118
53	#122	MinMax	0.659	4	1.0	0.945	0.261	0.013	0.398	0.7115
54	#262	MinMax	0.935	3	1.0	0.939	0.283	0.003	0.383	0.7099
55	#260	MinMax	0.934	3	1.0	0.941	0.278	0.061	0.338	0.7094
56	#253	MinMax	0.922	3	1.0	0.959	0.274	0.031	0.324	0.7069
57	#98	MinMax	0.653	4	1.0	0.956	0.258	0.011	0.354	0.7067
58	#290	MinMax	0.996	3	1.0	0.933	0.284	0.012	0.359	0.7065
59	#248	MinMax	0.918	2	0.5	0.989	0.290	0.011	0.843	0.7052
60	#282	MinMax	0.969	3	1.0	0.933	0.272	0.017	0.355	0.7049
61	#118	MinMax	0.656	4	1.0	0.956	0.259	0.011	0.338	0.7044
62	#109	MinMax	0.634	4	1.0	0.971	0.241	0.023	0.320	0.7038
63	#288	MinMax	0.995	3	1.0	0.943	0.266	0.031	0.329	0.7038
64	#102	MinMax	0.626	4	1.0	0.934	0.243	0.020	0.366	0.7034
65	#277	MinMax	0.967	3	1.0	0.956	0.281	0.033	0.291	0.7028
66	#174	MinMax	0.775	3	1.0	0.949	0.281	0.001	0.312	0.7012
67	#106	MinMax	0.631	4	1.0	0.960	0.241	0.022	0.316	0.7011
68	#208	MinMax	0.805	3	1.0	0.959	0.290	0.009	0.283	0.7009
69	#179	MinMax	0.776	3	1.0	0.950	0.278	0.003	0.308	0.7006
70	#78	Frank	0.988	2	0.5	0.989	0.281	0.008	0.823	0.7003
71	#51	Frank	0.996	2	0.5	0.989	0.277	0.008	0.823	0.6999

Continued on next page

Table B.1 – continued from previous page

Rank	**ID**	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
72	#274	MinMax	0.962	3	1.0	0.955	0.265	0.018	0.296	0.6998
73	#196	MinMax	0.791	3	1.0	0.959	0.280	0.008	0.279	0.6990
74	#198	MinMax	0.794	3	1.0	0.953	0.281	0.012	0.283	0.6989
75	#182	MinMax	0.779	3	1.0	0.960	0.284	0.009	0.271	0.6989
76	#143	MinMax	0.667	4	1.0	0.966	0.256	0.016	0.284	0.6988
77	#209	MinMax	0.805	3	1.0	0.963	0.282	0.014	0.263	0.6987
78	#289	MinMax	0.995	3	1.0	0.950	0.282	0.007	0.287	0.6986
79	#202	MinMax	0.795	3	1.0	0.972	0.276	0.027	0.244	0.6985
80	#190	MinMax	0.783	3	1.0	0.969	0.290	0.008	0.248	0.6985
81	#149	MinMax	0.679	3	1.0	0.959	0.286	0.013	0.259	0.6980
82	#142	MinMax	0.668	4	1.0	0.970	0.251	0.012	0.279	0.6978
83	#124	MinMax	0.660	4	1.0	0.953	0.260	0.014	0.291	0.6978
84	#108	MinMax	0.632	4	1.0	0.965	0.248	0.020	0.279	0.6976
85	#183	MinMax	0.780	3	1.0	0.963	0.283	0.011	0.255	0.6975
86	#147	MinMax	0.675	3	1.0	0.972	0.286	0.016	0.232	0.6974
87	#195	MinMax	0.792	3	1.0	0.963	0.280	0.010	0.255	0.6971
88	#113	MinMax	0.641	3	1.0	0.976	0.275	0.020	0.232	0.6970
89	#111	MinMax	0.639	3	1.0	0.975	0.278	0.015	0.232	0.6967
90	#137	MinMax	0.672	3	1.0	0.975	0.285	0.018	0.221	0.6967
91	#146	MinMax	0.669	4	1.0	0.983	0.253	0.012	0.250	0.6965
92	#148	MinMax	0.677	3	1.0	0.964	0.282	0.017	0.240	0.6964
93	#163	MinMax	0.767	3	1.0	0.986	0.288	0.011	0.203	0.6963
94	#116	MinMax	0.642	3	1.0	0.973	0.282	0.010	0.232	0.6963
95	#293	MinMax	0.997	3	1.0	0.968	0.281	0.026	0.225	0.6961
96	#139	MinMax	0.664	4	1.0	0.932	0.240	0.013	0.324	0.6961
97	#171	MinMax	0.771	3	1.0	0.987	0.284	0.015	0.199	0.6958
98	#211	MinMax	0.807	3	1.0	0.988	0.279	0.019	0.199	0.6957
99	#170	MinMax	0.735	3	1.0	0.977	0.285	0.008	0.217	0.6956

Continued on next page

Table B.1 – continued from previous page

Rank	**ID**	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
100	#187	MinMax	0.788	3	1.0	0.945	0.252	0.010	0.296	0.6955

Table B.2: Contingency tables of the selected variables (part I)

Competition	*No*	*Weak*	*Strong*
1	0	13	21
2	26	38	34
3	24	42	60

Location	*IC*	*SC*	*M*	*NMC*
1	15	5	2	13
2	31	24	19	24
3	28	24	45	30

Size	*Small*	*Medium*	*Big*
1	0	27	8
2	36	60	1
3	47	59	21

DisplaySize	*Small*	*Medium*	*Big*
1	1	34	0
2	35	52	11
3	40	68	19

Maint.	*Def.*	*Reg.*	*Good*	*Exc.*
1	0	4	26	5
2	7	29	58	4
3	8	28	54	37

Comm.	*Def.*	*Reg.*	*Good*	*Exc.*
1	0	5	26	4
2	8	38	47	5
3	14	27	57	29

Table B.3: Contingency tables of the selected variables (part II)

Aesthetics	*Def.*	*Reg.*	*Good*	*Exc.*
1	0	9	24	2
2	12	35	45	6
3	16	34	50	27

DisplayGrifone	*No*	*Yes*
1	15	19
2	66	31
3	49	77

ThermalExhibitor	*No*	*Yes*
1	16	19
2	87	9
3	95	31

GrifoneWeight	*Minimal*	*Secondary*	*Main*
1	1	27	5
2	49	44	4
3	29	67	31

Assistants	*Few*	*Many*	*A lot of*
1	13	12	10
2	84	11	3
3	107	9	11

Evaluation	*Bad*	*Good*	*Excellent*
1	8	13	14
2	20	54	24
3	24	40	63

PromosSensit	*Low*	*Medium*	*High*
1	1	34	0
2	23	75	0
3	0	0	127

Table B.4: Expected frequencies of the selected variables (part I)

Competition	*No*	*Weak*	*Strong*
1	6.6	12.3	15.2
2	19.0	35.3	43.7
3	24.4	45.4	56.2

Location	*IC*	*SC*	*M*	*NMC*
1	9.8	7.0	8.4	8.8
2	28.1	20.1	24.3	25.4
3	36.1	25.9	31.3	32.7

Size	*Small*	*Medium*	*Big*
1	10.8	19.3	3.8
2	30.9	55.1	10.9
3	40.2	71.6	14.2

DisplaySize	*Small*	*Medium*	*Big*
1	9.9	20.1	4.0
2	28.3	57.4	11.3
3	36.8	74.5	14.7

Maint.	*Def.*	*Reg.*	*Good*	*Exc.*
1	2.0	7.9	18.0	6.1
2	5.7	22.6	51.3	17.4
3	7.4	29.4	66.7	22.6

Comm.	*Def.*	*Reg.*	*Good*	*Exc.*
1	2.8	9.3	16.9	5.0
2	7.9	26.4	48.3	14.3
3	10.3	34.3	62.8	18.6

Table B.5: Expected frequencies of the selected variables (part II)

Aesthetics	*Def.*	*Reg.*	*Good*	*Exc.*
1	3.7	10.2	15.5	4.6
2	10.6	29.1	44.2	13.2
3	13.7	37.8	57.4	17.2

DisplayGrifone	*No*	*Yes*
1	16.6	16.4
2	48.4	47.6
3	63.0	62.0

ThermalExhibitor	*No*	*Yes*
1	25.6	7.4
2	73.0	21.0
3	96.3	27.7

GrifoneWeight	*Minimal*	*Secondary*	*Main*
1	9.6	16.4	5.0
2	28.9	49.1	15.0
3	38.5	65.5	20

Assistants	*Few*	*Many*	*A lot of*
1	24.7	3.5	2.7
2	74.2	10.5	8.2
3	99.0	14.0	11.0

Evaluation	*Bad*	*Good*	*Excellent*
1	6.2	12.4	12.4
2	18.7	37.1	37.1
3	25.0	49.5	49.5

PromosSensit	*Low*	*Medium*	*High*
1	2.9	12.6	15.5
2	8.6	37.9	46.5
3	11.5	50.5	62.0

Table B.6: Values of importance associated with the selected variables (part I)

Competition	*No*	*Weak*	*Strong*
1	-6.66	0.05	2.25
2	2.59	0.20	-2.15
3	-0.01	-0.26	0.26

Location	*IC*	*SC*	*M*	*NMC*
1	2.55	-0.64	-5.33	1.76
2	0.35	0.81	-1.39	-0.06
3	-1.84	-0.14	5.05	-0.23

Size	*Small*	*Medium*	*Big*
1	-11.22	2.68	3.84
2	0.78	0.52	-9.32
3	0.98	-2.21	2.69

DisplaySize	*Small*	*Medium*	*Big*
1	-8.33	8.49	-4.04
2	1.41	-0.63	-0.01
3	0.22	-0.69	1.29

Maint.	*Def.*	*Reg.*	*Good*	*Exc.*
1	-2.02	-2.16	2.97	-0.23
2	0.32	1.57	0.69	-10.26
3	0.06	-0.11	-2.67	9.40

Comm.	*Def.*	*Reg.*	*Good*	*Exc.*
1	-2.96	-2.08	4.13	-0.24
2	-0.01	5.11	-0.08	-6.07
3	0.99	-1.51	-0.67	5.87

Table B.7: Values of importance associated with the selected variables (part II)

Aesthetics	*Def.*	*Reg.*	*Good*	*Exc.*
1	-3.77	-0.21	3.98	-1.56
2	0.20	1.07	0.00	-3.92
3	0.39	-0.44	-1.14	5.74

DisplayGrifone	*No*	*Yes*
1	-0.28	0.29
2	5.84	-5.98
3	-3.41	3.49

ThermalExhibitor	*No*	*Yes*
1	-4.46	14.96
2	2.30	-7.71
3	-0.04	0.15

GrifoneWeight	*Minimal*	*Secondary*	*Main*
1	-8.24	4.86	-0.00
2	12.34	-1.26	-8.16
3	-2.58	-0.02	6.38

Assistants	*Few*	*Many*	*A lot of*
1	-7.62	13.74	14.18
2	0.66	-0.09	-4.04
3	0.54	-2.81	-0.04

Evaluation	*Bad*	*Good*	*Excellent*
1	0.14	-0.14	0.01
2	0.01	4.63	-5.20
3	-0.08	-2.88	3.79

PromosSensit	*Low*	*Medium*	*High*
1	-1.54	25.46	-17.10
2	21.52	28.00	-47.87
3	-11.72	-53.24	68.03

Table B.8: Conditional frequencies of the selected variables (part I)

Competition	*No*	*Weak*	*Strong*
1	0.00	0.38	0.62
2	0.27	0.39	0.35
3	0.19	0.33	0.48

Location	*IC*	*SC*	*M*	*NMC*
1	0.43	0.14	0.06	0.37
2	0.32	0.24	0.19	0.24
3	0.22	0.19	0.35	0.24

Size	*Small*	*Medium*	*Big*
1	0.00	0.77	0.23
2	0.37	0.62	0.01
3	0.37	0.46	0.17

DisplaySize	*Small*	*Medium*	*Big*
1	0.03	0.97	0.00
2	0.36	0.53	0.11
3	0.32	0.54	0.15

Maint.	*Def.*	*Reg.*	*Good*	*Exc.*
1	0.00	0.11	0.74	0.14
2	0.07	0.30	0.59	0.04
3	0.06	0.22	0.43	0.29

Comm.	*Def.*	*Reg.*	*Good*	*Exc.*
1	0.00	0.14	0.74	0.11
2	0.08	0.39	0.48	0.05
3	0.11	0.21	0.45	0.23

Table B.9: Conditional frequencies of the selected variables (part II)

Aesthetics	*Def.*	*Reg.*	*Good*	*Exc.*
1	0.00	0.26	0.69	0.06
2	0.12	0.36	0.46	0.06
3	0.13	0.27	0.39	0.21

DisplayGrifone	*No*	*Yes*
1	0.44	0.56
2	0.68	0.32
3	0.39	0.61

ThermalExhibitor	*No*	*Yes*
1	0.46	0.54
2	0.91	0.09
3	0.75	0.25

GrifoneWeight	*Minimal*	*Secondary*	*Main*
1	0.03	0.82	0.15
2	0.51	0.45	0.04
3	0.23	0.53	0.24

Assistants	*Few*	*Many*	*A lot of*
1	0.37	0.34	0.29
2	0.86	0.11	0.03
3	0.84	0.07	0.09

Evaluation	*Bad*	*Good*	*Excellent*
1	0.23	0.37	0.40
2	0.20	0.55	0.24
3	0.19	0.31	0.50

PromosSensit	*Low*	*Medium*	*High*
1	0.03	0.97	0.00
2	0.23	0.77	0.00
3	0.00	0.00	1.00

Table B.10: Initial messages of the selected variables

ID	Class	Variable	Modality	Type	Sign	Relev.	Value
#1	1	*Competition*	*no*	VoI	neg.	high	6.6
#2	1	*Location*	*mount./ski run towns*	VoI	neg.	high	5.3
#3	1	*Size*	*small*	VoI	neg.	high	11.2
#4	1	*DisplaySize*	*small*	VoI	neg.	high	8.3
#5	1	*DisplaySize*	*medium*	VoI	pos.	high	8.5
#6	1	*Communication*	*good*	VoI	pos.	high	4.1
#7	1	*Aesthetics*	*deficient*	VoI	neg.	high	3.8
#8	1	*Aesthetics*	*good*	VoI	pos.	high	4.0
#9	1	*ThermalExhibitor*	*yes*	VoI	pos.	high	15.0
#10	1	*Assistants*	*many*	VoI	pos.	high	13.7
#11	1	*Assistants*	*a lot of*	VoI	pos.	high	14.2
#12	2	*Size*	*big*	VoI	neg.	high	9.3
#13	2	*Maintenance*	*excellent*	VoI	neg.	high	10.3
#14	2	*Communication*	*regular*	VoI	pos.	high	5.1
#15	2	*Communication*	*excellent*	VoI	neg.	high	6.1
#16	2	*Aesthetics*	*excellent*	VoI	neg.	high	3.9
#17	2	*DisplayGrifone*	*no*	VoI	pos.	high	5.8
#18	2	*DisplayGrifone*	*yes*	VoI	neg.	high	6.0
#19	2	*GrifoneWeight*	*minimal*	VoI	pos.	high	12.3
#20	2	*Evaluation*	*good*	VoI	pos.	high	4.6
#21	2	*Evaluation*	*excellent*	VoI	neg.	high	5.2
#22	2	*PromosSensit*	*high*	VoI	neg.	high	47.9
#23	3	*Location*	*mount./ski run towns*	VoI	pos.	high	5.1
#24	3	*Maintenance*	*excellent*	VoI	pos.	high	9.4
#25	3	*Communication*	*excellent*	VoI	pos.	high	5.9
#26	3	*Aesthetics*	*excellent*	VoI	pos.	high	5.7
#27	3	*Evaluation*	*excellent*	VoI	pos.	high	3.8

Continued on next page

Table B.10 – continued from previous page

ID	Class	Variable	Modality	Type	Sign	Relev.	Value
#28	3	*PromosSensit*	*medium*	VoI	neg.	high	53.2
#29	3	*PromosSensit*	*high*	VoI	pos.	high	68.0
#30	1	*Competition*	*strong*	VoI	pos.	normal	2.3
#31	1	*Location*	*inner cities*	VoI	pos.	normal	2.5
#32	1	*Location*	*no mountain towns*	VoI	pos.	normal	1.8
#33	1	*Size*	*medium*	VoI	pos.	normal	2.7
#34	1	*Size*	*big*	VoI	pos.	normal	3.8
#35	1	*Maintenance*	*deficient*	VoI	neg.	normal	2.0
#36	1	*Maintenance*	*regular*	VoI	neg.	normal	2.2
#37	1	*Maintenance*	*good*	VoI	pos.	normal	3.0
#38	1	*DisplaySize*	*big*	VoI	neg.	normal	4.0
#39	1	*Communication*	*deficient*	VoI	neg.	normal	3.0
#40	1	*Communication*	*regular*	VoI	neg.	normal	2.1
#41	1	*Aesthetics*	*excellent*	VoI	neg.	normal	1.6
#42	1	*ThermalExhibitor*	*no*	VoI	neg.	normal	4.5
#43	1	*GrifoneWeight*	*minimal*	VoI	neg.	normal	8.2
#44	1	*GrifoneWeight*	*secondary*	VoI	pos.	normal	4.9
#45	1	*Assistants*	*few*	VoI	neg.	normal	7.6
#46	1	*PromosSensit*	*medium*	VoI	pos.	normal	25.5
#47	2	*Competition*	*no*	VoI	pos.	normal	2.6
#48	2	*Competition*	*strong*	VoI	neg.	normal	2.1
#49	2	*Location*	*mount./ski run towns*	VoI	neg.	normal	1.4
#50	2	*Maintenance*	*regular*	VoI	pos.	normal	1.6
#51	2	*Aesthetics*	*regular*	VoI	pos.	normal	1.1
#52	2	*ThermalExhibitor*	*yes*	VoI	neg.	normal	7.7
#53	2	*GrifoneWeight*	*main*	VoI	neg.	normal	8.2
#54	2	*Assistants*	*a lot of*	VoI	neg.	normal	4.0
#55	2	*PromosSensit*	*low*	VoI	pos.	normal	21.5

Continued on next page

Table B.10 – continued from previous page

ID	Class	Variable	Modality	Type	Sign	Relev.	Value
#56	2	*PromosSensit*	*medium*	VoI	pos.	normal	28.0
#57	3	*Location*	*inner cities*	VoI	neg.	normal	1.8
#58	3	*Size*	*medium*	VoI	neg.	normal	2.2
#59	3	*Size*	*big*	VoI	pos.	normal	2.7
#60	3	*Maintenance*	*good*	VoI	neg.	normal	2.7
#61	3	*Communication*	*regular*	VoI	neg.	normal	1.5
#62	3	*Aesthetics*	*good*	VoI	neg.	normal	1.1
#63	3	*DisplayGrifone*	*no*	VoI	neg.	normal	3.4
#64	3	*DisplayGrifone*	*yes*	VoI	pos.	normal	3.5
#65	3	*GrifoneWeight*	*main*	VoI	pos.	normal	6.4
#66	3	*Assistants*	*many*	VoI	neg.	normal	2.8
#67	3	*Evaluation*	*good*	VoI	neg.	normal	2.9
#68	1	*Competition*	*no*	EF	neg.	-	0.0
#69	1	*Size*	*small*	EF	neg.	-	0.0
#70	1	*Maintenance*	*deficient*	EF	neg.	-	0.0
#71	1	*DisplaySize*	*medium*	EF	pos.	-	1.0
#72	1	*DisplaySize*	*big*	EF	neg.	-	0.0
#73	1	*Communication*	*deficient*	EF	neg.	-	0.0
#74	1	*Aesthetics*	*deficient*	EF	neg.	-	0.0
#75	1	*PromosSensit*	*medium*	EF	pos.	-	1.0
#76	1	*PromosSensit*	*high*	EF	neg.	-	0.0
#77	2	*PromosSensit*	*high*	EF	neg.	-	0.0
#78	3	*PromosSensit*	*low*	EF	neg.	-	0.0
#79	3	*PromosSensit*	*medium*	EF	neg.	-	0.0
#80	3	*PromosSensit*	*high*	EF	pos.	-	1.0

Table B.11: Groups of messages affected by rule A.1

Class	**Variable**	**Type**	**ID**	**Modality**	**Sign**	**Action**	**Weigth**
1	*DisplaySize*	EF	#71	*medium*	pos.	Prioritise	2
			#72	*big*	neg.	Discard	0
	ThermalExhibitor	VoI	#9	*yes*	pos.	Prioritise	2
			#42	*no*	neg.	Discard	0
	PromosSensit	EF	#75	*medium*	pos.	Prioritise	2
			#76	*high*	neg.	Discard	0
2	*DisplayGrifone*	VoI	#17	*no*	pos.	Prioritise	2
			#18	*yes*	neg.	Discard	0
3	*DisplayGrifone*	VoI	#64	*yes*	pos.	Prioritise	2
			#63	*no*	neg.	Discard	0
	PromosSensit	EF	#80	*high*	pos.	Prioritise	2
			#78	*low*	neg.	Discard	0
			#79	*medium*	neg.	Discard	0

Table B.12: Groups of messages affected by rule A.2

Class	Variable	Modality	Sign	ID	Type	Action	Weight
1	*Aesthetics*	*deficient*	pos.	#74	EF	Prioritise	2
				#7	VoI	Discard	0
	Communication	*deficient*	neg.	#73	EF	Prioritise	2
				#39	VoI	Discard	0
	Competition	*no*	neg.	#68	EF	Prioritise	2
				#1	VoI	Discard	0
	DisplaySize	*big*	neg.	#72	EF	Prioritise	1
				#38	VoI	Discard	0
	DisplaySize	*medium*	pos.	#71	EF	Prioritise	3
				#5	VoI	Discard	0
	Maintenance	*deficient*	neg.	#70	EF	Prioritise	2
				#35	VoI	Discard	0
	PromosSensit	*medium*	pos.	#75	EF	Prioritise	3
				#46	VoI	Discard	0
	Size	*small*	neg.	#69	EF	Prioritise	2
				#3	VoI	Discard	0
2	*PromosSensit*	*high*	neg.	#77	EF	Prioritise	2
				#22	VoI	Discard	0
3	*PromosSensit*	*high*	pos.	#80	EF	Prioritise	3
				#29	VoI	Discard	0
	PromosSensit	*medium*	neg.	#79	EF	Prioritise	1
				#28	VoI	Discard	0

Table B.13: Groups of messages affected by rule A.3

Class	Variable	ID	Modality	Type	Sign	Action	Weight
1	*Assistants*	#10	*many*	VoI	pos.	Prioritise	2
		#11	*a lot of*	VoI	pos.	Prioritise	2
		#45	*few*	VoI	neg.	Discard	0
	DisplaySize	#5	*medium*	VoI	pos.	Prioritise	1
		#71	*medium*	EF	pos.	Prioritise	4
		#4	*small*	VoI	neg.	Discard	0
		#38	*big*	VoI	neg.	Discard	-1
		#72	*big*	EF	neg.	Discard	0
	ThermalExhibitor	# 9	*yes*	VoI	pos.	Prioritise	3
		#42	*no*	VoI	neg.	Discard	-1
	Size	#33	*medium*	VoI	pos.	Prioritise	2
		#34	*big*	VoI	pos.	Prioritise	2
		#3	*small*	VoI	neg.	Discard	-1
		#69	*small*	EF	neg.	Discard	1
2	*DisplayGrifone*	#17	*no*	VoI	pos.	Prioritise	3
		#18	*yes*	VoI	neg.	Discard	-1
	PromosSensit	#55	*low*	VoI	pos.	Prioritise	2
		#56	*medium*	VoI	pos.	Prioritise	2
		#22	*high*	VoI	neg.	Discard	-1
		#77	*high*	EF	neg.	Discard	1
3	*DisplayGrifone*	#64	*yes*	VoI	pos.	Prioritise	3
		#63	*no*	VoI	neg.	Discard	-1
	PromosSensit	#29	*high*	VoI	pos.	Prioritise	1
		#80	*high*	EF	pos.	Prioritise	4
		#28	*medium*	VoI	neg.	Discard	-1
		#78	*low*	EF	neg.	Discard	-1
		#79	*medium*	EF	neg.	Discard	0

Table B.14: Groups of messages affected by rule A.5

Variable	**Modality**	**Type**	**ID**	**Class**	**Sign**	**Action**	**Weight**
Aesthetics	*excellent*	VoI	#26	3	pos.	Prioritise	2
			#16	2	neg.	Discard	0
			#41	1	neg.	Discard	0
Communication	*regular*	VoI	#14	2	pos.	Prioritise	2
			#40	1	neg.	Discard	0
			#61	3	neg.	Discard	0
Location	*mountain/sky run towns*	VoI	#23	3	pos.	Prioritise	2
			#2	1	neg.	Discard	0
			#49	2	neg.	Discard	0
PromosSentit	*high*	EF	#80	3	pos.	Prioritise	5
			#76	1	neg.	Discard	-1
			#77	2	neg.	Discard	0

Table B.15: Filtered messages that will be mentioned in the final text

ID	Class	Variable	Modality	Type	Sign	Relev.	Value	Weight
#6	1	*Communication*	*good*	VoI	pos.	high	4.1	1
#8	1	*Aesthetics*	*good*	VoI	pos.	high	4.0	1
#9	1	*ThermalExhibitor*	*yes*	VoI	pos.	high	15.0	3
#10	1	*Assistants*	*many*	VoI	pos.	high	13.7	2
#11	1	*Assistants*	*a lot of*	VoI	pos.	high	14.2	2
#12	2	*Size*	*big*	VoI	neg.	high	9.3	1
#13	2	*Maintenance*	*excellent*	VoI	neg.	high	10.3	1
#14	2	*Communication*	*regular*	VoI	pos.	high	5.1	2
#15	2	*Communication*	*excellent*	VoI	neg.	high	6.1	1
#17	2	*DisplayGrifone*	*no*	VoI	pos.	high	5.8	3
#19	2	*GrifoneWeight*	*minimal*	VoI	pos.	high	12.3	1
#20	2	*Evaluation*	*good*	VoI	pos.	high	4.6	1
#21	2	*Evaluation*	*excellent*	VoI	neg.	high	5.2	1
#23	3	*Location*	*mount./ski*	VoI	pos.	high	5.1	2
#24	3	*Maintenance*	*excellent*	VoI	pos.	high	9.4	1
#25	3	*Communication*	*excellent*	VoI	pos.	high	5.9	1
#26	3	*Aesthetics*	*excellent*	VoI	pos.	high	5.7	2
#27	3	*Evaluation*	*excellent*	VoI	pos.	high	3.8	1
#30	1	*Competition*	*strong*	VoI	pos.	normal	2.3	1
#31	1	*Location*	*inner cities*	VoI	pos.	normal	2.5	1
#32	1	*Location*	*no mountain*	VoI	pos.	normal	1.8	1
#33	1	*Size*	*medium*	VoI	pos.	normal	2.7	2
#34	1	*Size*	*big*	VoI	pos.	normal	3.8	2
#36	1	*Maintenance*	*regular*	VoI	neg.	normal	2.2	1
#37	1	*Maintenance*	*good*	VoI	pos.	normal	3.0	1
#43	1	*GrifoneWeight*	*minimal*	VoI	neg.	normal	8.2	1
#44	1	*GrifoneWeight*	*secondary*	VoI	pos.	normal	4.9	1

Continued on next page

Table B.15 – continued from previous page

ID	Class	Variable	Modality	Type	Sign	Relev.	Value	Weight
#47	2	*Competition*	*no*	VoI	pos.	normal	2.6	1
#48	2	*Competition*	*strong*	VoI	neg.	normal	2.1	1
#50	2	*Maintenance*	*regular*	VoI	pos.	normal	1.6	1
#51	2	*Aesthetics*	*regular*	VoI	pos.	normal	1.1	1
#52	2	*ThermalExhibitor*	*yes*	VoI	neg.	normal	7.7	1
#53	2	*GrifoneWeight*	*main*	VoI	neg.	normal	8.2	1
#54	2	*Assistants*	*a lot of*	VoI	neg.	normal	4.0	1
#55	2	*PromosSensit*	*low*	VoI	pos.	normal	21.5	2
#56	2	*PromosSensit*	*medium*	VoI	pos.	normal	28.0	2
#57	3	*Location*	*inner cities*	VoI	neg.	normal	1.8	1
#58	3	*Size*	*medium*	VoI	neg.	normal	2.2	1
#59	3	*Size*	*big*	VoI	pos.	normal	2.7	1
#60	3	*Maintenance*	*good*	VoI	neg.	normal	2.7	1
#62	3	*Aesthetics*	*good*	VoI	neg.	normal	1.1	1
#64	3	*DisplayGrifone*	*yes*	VoI	pos.	normal	3.5	3
#65	3	*GrifoneWeight*	*main*	VoI	pos.	normal	6.4	1
#66	3	*Assistants*	*many*	VoI	neg.	normal	2.8	1
#67	3	*Evaluation*	*good*	VoI	neg.	normal	2.9	1
#68	1	*Competition*	*no*	EF	neg.	-	0.0	2
#70	1	*Maintenance*	*deficient*	EF	neg.	-	0.0	2
#71	1	*DisplaySize*	*medium*	EF	pos.	-	1.0	4
#73	1	*Communication*	*deficient*	EF	neg.	-	0.0	2
#74	1	*Aesthetics*	*deficient*	EF	neg.	-	0.0	2
#75	1	*PromosSensit*	*medium*	EF	pos.	-	1.0	3
#80	3	*PromosSensit*	*high*	EF	pos.	-	1.0	5

Table B.16: Groups of messages affected by rule B.1

Class	Variable	Type	Sign	ID	Modality	Order
1	*Assistants*	VoI	pos.	#10	*many*	2/3
				#11	*a lot of*	3/3
	Size	VoI	pos.	#33	*medium*	2/3
				#34	*big*	3/3
2	*PromosSensit*	VoI	pos.	#55	*low*	1/3
				#56	*medium*	2/3

Table B.17: Groups of messages affected by rule B.2 (first step: same sign).

Class	Variable	Sign	ID	Modality	Type	Order
1	*Location*	pos.	#31	*inner cities*	VoI	-
			#32	*no mountain*	VoI	-
	Maintenance	neg.	#36	*regular*	VoI	2/4
			#70	*deficient*	EF	1/4

Table B.18: Groups of messages affected by rule B.2 (second step: different sign)

Class	Variable	ID	Modality	Type	Sign	Order
1	*Aesthetics*	#8	*good*	VoI	pos.	3/4
		#74	*deficient*	EF	neg.	1/4
	Communication	#6	*good*	VoI	pos.	3/4
		#73	*deficient*	EF	neg.	1/4
	Competition	#30	*strong*	VoI	pos.	3/3
		#68	*no*	EF	neg.	1/3
	GrifoneWeight	#43	*minimal*	VoI	neg.	1/3
		#44	*secondary*	VoI	pos.	2/3
2	*Communication*	#14	*regular*	VoI	pos.	2/4
		#15	*excellent*	VoI	neg.	4/4
	Competition	#47	*no*	VoI	pos.	1/3
		#48	*strong*	VoI	neg.	3/3
	Evaluation	#20	*good*	VoI	pos.	2/3
		#21	*excellent*	VoI	neg.	3/3
	GrifoneWeight	#19	*minimal*	VoI	pos.	1/3
		#53	*main*	VoI	neg.	3/3
	Maintenance	#13	*excellent*	VoI	neg.	4/4
		#50	*regular*	VoI	pos.	2/4
3	*Aesthetics*	#26	*excellent*	VoI	pos.	4/4
		#62	*good*	VoI	neg.	3/4
	Evaluation	#27	*excellent*	VoI	pos.	3/3
		#67	*good*	VoI	neg.	2/3
	Location	#23	*mount./ski*	VoI	pos.	-
		#57	*inner cities*	VoI	neg.	-
	Maintenance	#24	*excellent*	VoI	pos.	4/4
		#60	*good*	VoI	neg.	3/4
	Size	#58	*medium*	VoI	neg.	2/3
		#59	*big*	VoI	pos.	3/3

Table B.19: Messages affected by rule B.4

Class	Variable	ID	Modality	Type	Sign	Order
1	*Maintentance*	#37	*good*	VoI	pos.	3/4
		#36	*regular*	VoI	neg.	2/4
		#70	*deficient*	EF	neg.	1/4

Table B.20: Messages affected by rule B.5

Class	Type	Sign	ID	Variable	Modality
1	EF	pos.	#71	*DisplaySize*	*medium*
			#75	*PromosSensit*	*medium*
2	VoI	neg.	#52	*ThermalExhibitor*	*yes*
			#54	*Assistants*	*a lot of*
3	VoI	pos.	#64	*DisplayGrifone*	*yes*
			#65	*GrifoneWeight*	*main*

Table B.21: Messages affected by rule C.1

Modality	ID	Class	Variable	Type	Sign
no	#17	2	*DisplayGrifone*	VoI	pos.
	#47	2	*Competition*	VoI	pos.
	#68	1	*Competition*	EF	neg.

Table B.22: Messages affected by rule C.2

Modality	ID	Class	Variable	Type	Sign
yes	#9	1	*ThermalExhibitor*	VoI	pos.
	#52	2	*ThermalExhibitor*	VoI	neg.
	#64	3	*DisplayGrifone*	VoI	pos.

Table B.23: Variables needing a special transcription (rule C.3)

Variable	**VoIs transcription**	**EFs transcription**
Competition	with/without competition	have/have not competition
Location	located in	are located in
Size	*medium*-sized	are *medium*-sized
DisplaySize	with a *medium*-sized display	have a *medium*-sized display
DisplayGrifone	displaying Grifone products	display Grifone products
ThermalExhibitor	with thermal exhibitor	have thermal exhibitor
Assistants	with *few* assistants	have *few* assistants

Table B.24: Variables with modalities as adjectives (rule C.4)

Variable	**VoIs transcription**	**EFs transcription**
Maintenance	with a *good* maintenance	have a *good* maintenance
Communication	with an *excellent* communication	have an *excellent* communication
Aesthetics	with a *regular* aesthetics	have a *regular* aesthetics
GrifoneWeight	with a *main* Grifone weight	have a *main* Grifone Weight
Evaluation	with a *bad* evaluation	have a *bad* evaluation
PromosSensit	with a *high* sensitivity to promotions	have a *high* sensitivity to promotions

Bibliography

Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, 9:147–169.

Aguado, J. C. (1998). *A Mixed Qualitative-Quantitative Self-Learning Classification Technique Applied to Situation assessment in Industrial Process Control*. PhD thesis, Universitat Politècnica de Catalunya.

Aguado, J. C., Catalá, A., and Parra, X. (1999). Comparison of structure and capabilities between a non-standard classificarion technique and the radial basis bunction neural networks. In *Proceedings of the 13th European Simulation Multiconference (ICQFN 99)*, volume II, pages 442–448, Warsaw, Poland.

Aguilar, J. and López de Mántaras, R. (1982). The process of classification and learning the meaning of linguistic descriptors of concepts. *Approximate Reasoning in Decision Analysis*, pages 165–175.

Aguilar-Martin, J., Agell, N., Sánchez, M., and Prats, F. (2002). Analysis of tensions in a population based on the adequacy concept. In *Topics in Artificial Intelligence, 5th Catalonian Conference on AI, CCIA 2002, Castellón, Spain, October 24-25, 2002, Proceedings*, volume 2504 of *Lecture Notes in Computer Science*, pages 17–28. Springer.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Barrón-Adame, J., Cortina-Januchs, M., Vega-Corona, A., and Andi, D. (2012). Unsupervised system to classify SO_2 pollutant concentrations in Salamanca. *Expert Systems with Applications*, 39(1):107–116.

Bittmann, R. M. and Gelbard, R. (2009). Visualization of multi-algorithm clustering for better economic decisions - the case of car pricing. *Decision Support Systems*, 47(1):42–50.

Broder, A. Z., Ciccolo, P., Fontoura, M., Gabrilovich, E., Josifovski, V., and Riedel, L. (2008). Search advertising using web relevance feedback. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1013–1022, New York, NY.

Casabayó, M. (2005). *Shopping behaviour forecasts: Experiments based on a fuzzy learning technique in the Spanish food retailing industry.* PhD thesis, University of Edinburgh.

Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In *Proceedings of the European Working Session Learning (EWSL'91)*, pages 164–178.

Cawsey, A., Jones, R., and Pearson, J. (2000). The evaluation of a personalised health information system for patients with cancer. *User Modelling and User-Adapted Interaction*, 10:47–72.

Chapelle, O., Haffner, P., and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10:1055–1064.

Chen, Y., Wang, J. Z., and Krovetz, R. (2005). CLUE: cluster-based retrieval of images by unsupervised learning. *IEEE Transactions on Image Processing*, 14:1187–1201.

Cheng, C.-H., Wai-chee Fu, A., and Zhang, Y. (1999). Entropy-based subspace clustering for mining numerical data. In *KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 84–93.

Chiclana, F., Herrera, F., and Herrera-Viedma, E. (1998). Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. *Fuzzy Sets and Systems*, 97(1):33–48.

Chiclana, F., Herrera-Viedma, E., Herrera, F., and Alonso, A. (2004). Induced ordered weighted geometric operators and their use in the aggregation of multiplicative preference relations. *International Journal of Intelligent Systems*, 19:233–255.

Chiclana, F., Herrera-Viedma, E., Herrera, F., and Alonso, A. (2007). Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations. *European Journal of Operational Research*, 182:383–399.

Chiu, C.-Y., Chen, Y.-F., Kuo, I.-T., and Ku, H. C. (2009). An intelligent market segmentation system using k-means and particle swarm optimization. *Expert Systems with Applications*, 36(3, Part 1):4558–4565.

Choi, D. H., Ahn, B. S., and Kim, S. H. (2005). Prioritization of association rules in data mining: Multiple criteria decision approach. *Expert Systems with Applications*, 29(4):867–878.

Constantinos, S. and Paris, A. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowledge-Based Systems*, 21(7):721–726.

Dale, R. and Reiter, E. (2000). *Building natural language generation systems*. Cambridge University Press, Cambridge, UK.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the 12th International Conference on Machine Learning*, pages 194–202, San Francisco, CA.

Dubois, D. and Prade, H. (1985). A review of fuzzy set aggregation connectives. *Information Sciences*, 36:85–121.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification (2nd ed.)*. John Wiley & Sons, New York.

Elati, M. and Rouveirol, C. (2011). Unsupervised learning for gene regulation network inference from expression data: A review. In Elloumi, M. and Zomaya, A. Y., editors, *Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications*, pages 955–978. John Wiley & Sons, Inc.

Fayyad, U. and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI´93)*, pages 1022–1027.

Ferraretti, D., Gamberoni, G., and Lamma, E. (2012). Unsupervised and supervised learning in cascade for petroleum geology. *Expert Systems with Applications*, 39(10):9504–9514.

Ferres, L., Parush, A., Roberts, S., and G., L. (2006). Helping people with visual impairments gain access to graphical information through natural language: The igraph system. In *Proceedings of*

the 10th International Conference on Computers Helping People with Special Needs (ICCHP'06), pages 1122–1130. Springer-Verlag.

Figueiredo, M. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396.

Fodor, J. and Roubens, M. (1994). *Fuzzy Preference Modelling and Multicriteria Decision Support*. Kluwer.

Forbus, K. D. (2011). Qualitative modeling. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(4):374–391.

Goldberg, E., Driedger, N., and Kittredge, R. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, 9(2):45–53.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2004). *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web (1 ed.)*. Springer.

González Abril, L., Velasco Morente, F., Gavilán Ruiz, J. M., and Sánchez-Reyes Fernández, L. M. (2010). The similarity between the square of the coefficient of variation and the gini index of a general random variable. *Revista de métodos cuantitativos para la economía y la empresa*, 10:5–18.

Gramajo, S. and Martínez, L. (2012). A linguistic decision support model for QoS priorities in networking. *Knowledge-Based Systems*, 32:65–75.

Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56:549–603.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5:199–220.

Hadavandi, E., Shavandi, H., and Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8):800–808.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 2:107–145.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002). Cluster validity methods: Part i. *ACM SIGMOD Record*, 31(2):40–45.

Hallett, C. and Scott, D. (2005). Structural variation in generated health reports. In *Proceedings of the 3rd International Workshop on Paraphrasing (IWP2005)*, pages 33–40.

Hammond, T. and Davis, R. (2005). LADDER, a sketching language for user interface developers. *Computers and Graphics*, 29:518–532.

Harris, M. (2008). Building a large-scale commercial NLG system for an EMR. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 157–160.

Hennig, C. (2010). *fpc: Flexible procedures for clustering*. R package version 2.0-3.

Herrera, F., Herrera-Viedma, E., and Chiclana, F. (2003). A study of the origin and uses of the ordered weighted geometric operator in multicriteria decision making. *International Journal of Intelligent Systems*, 18:689–707.

Herzog, G. and Wazinski, P. (1994). VIsual TRANslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, pages 175–187.

Holte, R. (1993). Very simple classification rules perform well on most commonly used dataset. *Machine Learning*, 11:63–91.

Hong, T. and Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer's intention to purchase. *Expert Systems with Applications*, 39:2127–2131.

Hosking, J. R. (1980). The multivariate portmanteau statistic. *Journal of the American Statistical Association*, 75(371):602–608.

Hüske-Kraus, D. (2003a). Suregen-2: A shell system for the generation of clinical documents. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 215–218, Budapest, Hungary.

Hüske-Kraus, D. (2003b). Text generation in clinical medicine: A review. *Methods of Information in Medicine*, 42:51–60.

Iordanskaja, L., Kim, M., Kittredge, R., Lavoie, B., and Polguere, A. (1992). Generation of extended bilingual statistical reports. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING-92)*, pages 1019–1023, Nantes, France.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31:651–666.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31:264–323.

Kahn, M., Fagan, L., and Sheiner, L. (1991). Combining physiologic models and symbolic methods to interpret time-varying patient data. *Methods of Information in Medicine*, 30:167–187.

Kerber, R. (1992). Chimerge: Discretization of numeric attributes. In *Proceedings of the Ninth International Conference on Artificial Intelligence (AAAI´92)*, pages 123–128.

Klir, G. J. and Folger, T. A. (1988). *Fuzzy Sets, Uncertainty and Information.* Prentice-Hall.

Klir, J. and Yuan, B. (1995). *Fuzzy sets and fuzzy logic, Theory and Applications.* Prentice Hall.

Kukar, M. (2003). Transductive reliability estimation for medical diagnosis. *Artificial Intelligence in Medicine*, 29(1-2):81–106.

Kukich, K. (1983). Design of a knowledge-based report generator. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics (ACL-83)*, pages 145–150, Cambridge, MA.

Kurgan, L. A. and Cios, K. J. (2004). CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16:145–153.

Lee, C. H. and Yang, H. C. (2009). Construction of supervised and unsupervised learning systems for multilingual text categorization. *Expert Systems with Applications*, 36:2400–2410.

Liu, H., Hussain, F., Lim, C., and Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423.

Liu, H. and Setiono, R. (1997). Feature selection via discretization. *IEEE Transactions on Knowledge and Data Engineering*, 9(4):642–645.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *IEEE 10th International Conference on Data Mining (ICDM), 2010*, pages 911–916.

Lu, Z., Wang, S., Li, X., Yang, L., Yang, D., and Wu, D. (2012). Online shop location optimization using a fuzzy multi-criteria decision model – case study on taobao.com. *Knowledge-Based Systems*, 32:76–83.

Mantel, N. (1963). Chi-square tests with one degree of freedom; extensions of the mantel-haenszel procedure. *Journal of the American Statistical Association*, 58(303):690–700.

Marichal, J.-L. (1998). *Aggregation Operators for Multicriteria Decision Aid*. PhD thesis, Institute of Mathematics, University of Liège, Liège, Belgium.

McKeown, K., Kukich, K., and Shaw, J. (1994). Practical issues in automatic document generation. In *Proceedings of ANLP-1994*, pages 7–14.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Mo, J., Kiang, M. Y., Zou, P., and Li, Y. (2010). A two-stage clustering approach for multi-region segmentation. *Expert Systems with Applications*, 37(10):7120–7131.

Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal. *International Journal of Computer Vision*, 79(3):299–318.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1:327–332.

Oliver, J. L., Tortosa, L., and Vicent, J. F. (2011). A neural network model to develop actions in urban complex systems represented by 2d meshes. *International Journal of Computer Mathematics*, 88:3361–3379.

Osei-Bryson, K.-M. (2010). Towards supporting expert evaluation of clustering results using a data mining process model. *Information Sciences*, 180:414–431.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pei, Z., Ruan, D., Liu, J., and Xu, Y. (2012). A linguistic aggregation operator with three kinds of weights for nuclear safeguards evaluation. *Knowledge-Based Systems*, 28:19–26.

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., and Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173:789–816.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramze Rezaee, M., Lelieveldt, B., and Reiber, J. (1998). A new cluster validity index for the fuzzy c-mean. *Pattern Recognition Letters*, 19(3-4):237–246.

Reiter, E. (2007). An architecture for data-to-text systems. In *Proceedings of the 11th European Workshop on Natural Language Generation (ENLG-07)*, pages 97–104, Schloss-Dagstuhl, Germany.

Reiter, E., Robertson, R., and Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144:41–58.

Reiter, E., Sripada, S., Hunter, J., Yu, J., and Davy, I. (2005). Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Richeldi, M. and Rossotto, M. (1995). Class-driven statistical discretization of continuous attributes. In *Proceedings of the Eighth European Conference on Machine Learning (ECML´95)*, pages 335–338.

Rissanenm, J. (1978). Modeling by shortest data description. *Automatica*, 14:445–471.

Roy, D. (2002). Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16:353–385.

Ruiz, F., Angulo, C., and Agell, N. (2008). IDD: A supervised interval distance-based method for discretization. *IEEE Transactions on Knowledge and Data Engineering*, 40:1230–1238.

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz Tuñón, C., and de Vicente, A. (2010). Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra. *The Astrophysical Journal*, 714(1):487–504.

Sánchez-Hernández, G., Agell, N., Aguado, J. C., Sánchez, M., and Prats, F. (2007). Selection criteria for fuzzy unsupervised learning: Applied to market segmentation. In *Lecture Notes in Computer Science*, volume 4529, pages 307–317, Berlin. Springer-Verlag.

Sánchez-Hernández, G., Chiclana, F., Agell, N., and Aguado, J. C. (2013). Ranking and selection of unsupervised learning marketing segmentation. *Knowledge-Based Systems*, 44:20–33.

Sripada, S. and Gao, F. (2007). Summarizing dive computer data: A case study in integrating textual and graphical presentations of numerical data. In *Proceedings of the Workshop on Multimodal Output Generation (MOG-2007)*, pages 149–157.

Sripada, S., Reiter, E., and Davy, I. (2003). SumTime-Mousam: Configurable marine weather forecast generator. *Expert Update*, 6:4–10.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press, Cambridge, United Kingdom, 4th edition edition.

Thomas, K. and Sripada, S. (2008). What's in a message? Interpreting georeferenced data for the visually impaired. In *Proceedings of the Fifth International Natural Language Generation Conference (INLG´08)*, pages 113–120. Association for Computational Linguistics.

Tibshirani, R. and Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 15(3):511–528.

Torra, V. (1997). The weighted OWA operator. *International Journal of Intelligent Systems*, 12:153–166.

Torra, V. and Narukawa, Y. (2007). A view of averaging aggregation operators. *IEEE Transactions on Fuzzy Systems*, 16:1063–1067.

Tou, J. and González, R. (1974). *Pattern Recognition Principles*. Addison-Wesley.

Tschuprow, A. A. and Kantorowitcsch, M. (1939). Principles of the mathematical theory of correlation. *Journal of the American Statistical Association*, 34:755.

Turnbull, P. W. and Leek, S. (2003). Business-to-business marketing: Organizational buying behavior, relationships and networks. In Market, M., editor, *The marketing book*, pages 142–169. Butterworth-Heinemann, Oxford, UK.

Turner, R., Sripada, S., Reiter, E., and Davy, I. (2006). Generating spatio-temporal descriptions in pollen forecasts. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL´06)*, pages 163–166.

Waissman, J., Aguilar, J., Dahhou, B., and Roux, G. (1998). Généralisation du degré d'adequation marginale dans la méthode de classification LAMDA. In *6èmes Rencontres de la Société Francophone de Classification*, Montpellier, France.

Wang, H., Ding, M., Li, X., and Shen, B. (2009). Clinical information driven ensemble clustering for inferring robust tumor subtypes. In *2nd International Conference on Biomedical Engineering and Informatics, 2009*, pages 1–4. IEEE.

Warnes, G. R., with contributions from Ben Bolker, Gorjanc, G., Grothendieck, G., Korosec, A., Lumley, T., MacQueen, D., Magnusson, A., Rogers, J., and others (2012). *gdata: Various R programming tools for data manipulation*. R package version 2.11.0.

Wong, A. and Chiu, D. (1987). Synthesizing statistical knowledge from incomplete mixed-mode data. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9:796–805.

Wu, J., Xiong, H., and Chen, J. (2009). Adapting the right measures for k-means clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, pages 877–886. ACM Press.

Xiong, H., Wu, J., and Chen, J. (2009). K-means clustering versus validation measures: a data-distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39(2):318–331.

Xu, Z. S. and Da, Q. L. (2003). An overview of operators for aggregating information. *International Journal of Intelligent Systems*, 18:953–969.

Yager, R. R. (1983). Quantifiers in the formulation of multiple objective decision functions. *Information Sciences*, 31:107–139.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 18:183–190.

Yager, R. R. (1996). Quantifier guided aggregation using OWA operators. *International Journal of Intelligent Systems*, 11:49–73.

Yager, R. R. (2003). Induced aggregation operators. *Fuzzy Sets and Systems*, 137:59–69.

Yang, Y., Liao, Y., Meng, G., and Lee, J. (2011). A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis. *Expert Systems with Applications*, 38(9):11311–11320.

Yao, Z., Holmbom, A. H., Eklund, T., and Back, B. (2010). Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis. In *Advances in Data Mining: Applications and Theoretical Aspects*, volume 6171 of *Lecture Notes in Artificial Intelligence*, pages 292–307.

Yates, F. (1934). Contingency tables involving small numbers and the $\chi 2$ test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235.

Yatskiv, I. and Gusarova, L. (2005). The methods of cluster analysis results validation. In *Transport and Telecommunication*, pages 75–80.

Zadeh, L. A. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computational & Applied Mathematics*, 9:149–184.

Zhang, Z. and Guo, C. (2012). A method for multi-granularity uncertain linguistic group decision making with incomplete weight information. *Knowledge-Based Systems*, 26:111–119.

Zhou, L. and Chen, H. (2012). A generalization of the power aggregation operators for linguistic environment and its application in group decision making. *Knowledge-Based Systems*, 26:216–224.

Zhou, S., Chiclana, F., John, R., and Garibaldi, J. (2008). Type-1 OWA operators for aggregating uncertain information with uncertain weights induced by type-2 linguistic quantifiers. *Fuzzy Sets and Systems*, 159:3281–3296.

List of Keywords

Aggregation Functions, 5
    OWA, 2, 5, 14, 134

Chi-square test of independence, 54
Coefficient of variation, 30
Contingency table, 44

Data-to-text systems
    Applications, 19
    Architecture
        Data interpretation, 64
        Document planning, 70
        Microplanning and realisation, 85
        Signal analysis, 51
    Use of rules, 21, 65
Discretisation, 44, 52

Extreme frequencies, 51, 61

Gini coefficient, 31

LAMDA, 38, 134, 137
    Marginal Adequacy Degree, 47, 138
    Marginal Adequacy Degree (MAD), 38
Linguistic quantifier, 16
    RDM, 16
    RIM, 16, 17

Machine learning, 3, 49
    Supervised learning, 3
    Unsupervised learning, 3, 10, 25, 49
        Applications, 4
MCDM, 2, 3, 14, 26, 133
Membership function, 25, 28

NLG, 2, 19, 47, 49, 133
    Architecture, 50
    Grammar, 85

Ontology, 50, 71, 91, 120

Selecting classifications, 4, 10, 25
    Validation criteria, 10, 133
        Accuracy, 45
        Balanced classes, 30
        Coherent classification, 38
        Dependency on external variables, 43
        External criteria, 11, 26
        Internal criteria, 11, 26
        Relative criteria, 11
        Useful number of classes, 27

Tschuprow's coefficient, 45

Values of importance, 51, 58