University of Nottingham

PhD

2015

Potential of Psychological Information to Support Knowledge Discovery in Consumer Debt Analysis

Supervisors:

Author:

Alexandros Ladas

Uwe Aickelin

Jon Garibaldi

Eamonn Ferguson

A thesis submitted in fulfilment of the requirements

for the degree of PhD

Department of Computer Science

March 2016

Declaration of Authorship

I, Alexandros Ladas, declare that this thesis titled, 'Potential of Psychological Information to Support Knowledge Discovery in Consumer Debt Analysis' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: ______________________________

Date: ______________________________

"As we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns the ones we don't know we don't know. It is the latter category that tend to be the difficult ones."

Donald Rumsfeld

UNIVERSITY OF NOTTINGHAM

Abstract

Faculty Of Science

Department of Computer Science

PhD

Potential of Psychological Information to Support Knowledge Discovery in Consumer Debt Analysis

by Alexandros Ladas

In this work, we develop a Data Mining framework to explore the multifaceted nature of consumer indebtedness. Data Mining with its numerous techniques and methods poses as a powerful toolbox to handle the sensitivity of these data and explore the psychological aspects of this social phenomenon. Thus, we begin with a series of transformations that deal with any inconsistencies the data may contain but more importantly they capture the essential psychological information hidden in the data and represent it in a new feature space as behavioural data. Then, we propose a novel consensus clustering framework to uncover patterns of consumer behaviour which draws upon the ability of cluster ensembles to reveal robust clusters from difficult datasets. Our Homals Consensus, models successfully the relationships between different clusterings in the cluster ensemble and manages to uncover representative clusters that are more suitable for explaining the complex patterns of a socio-economic dataset. Finally under a supervised learning approach the behavioural aspects of consumer indebtedness are assessed. In more detail, we take advantage of the flexibility Neural Networks provide in determining their architecture in order to propose a novel Neural Network solution, named TopDNN, that can handle non-linearities in the data and takes into account the extracted behavioural knowledge by incorporating it in the model. All the above sketch an elaborate framework that can reveal the potential of the behavioural data to support Knowledge Discovery in Consumer Debt Analysis on one hand and the ability of Data Mining to supplement existing models and theories of complex and sensitive nature on the other.

Acknowledgements

As this chapter of my life comes to an end, I want to thank Experian for funding these doctoral studies and my supervisors, Uwe, Jon and Eamonn. Special thanks to John Gathergood for providing us the data and to Rodrigo Scarpel for all the help and guidance he gave me the in second year of my studies. Also I would like to thank Stefanos and James who were the first to welcome me in this city I now call home and of course all the beautiful people I met here these 4 years. You know who you are. Cheers!

Contents

List of Figures

List of Tables

To my mother,
Eleni…

Chapter 1

Introduction

1.1 The Landscape of Consumer Indebtedness

As consumer debt takes the form of a social phenomenon in developed countries, whose materialistic environment encourages consumers to go into debt in rational grounds (Gardharsdottir and Dittmar, 2012), it becomes imminent to establish the proper balance between spending and saving that will allow us to thrive economically and maintain the high standards of living we currently enjoy (Watson, 2003). Relatively recent developments in data technology that allow the usage of advanced computational methods to store and analyse vast amounts of socio-economic data can be utilised in the scope of getting a better insight of this global economic force that keeps gaining a momentum in modern societies (Kamleitner et al., 2012). The acquired knowledge will supplement newer interdisciplinary approaches that study the complex nature of this multifaceted phenomenon (Kamleitner et al., 2012; Stone and Maury, 2006) and can assist into identifying potential risks of household over-indebtedness that will protect households and economies against the severe consequences these may cause (Betti et al., 2001).

1.1.1 A Complex Social Phenomenon on the Rise

As defined in Wikipedia, economics consider consumer debt as the outstanding debt of consumers, in contrast with that of businesses or governments. In macroeconomic terms, it is debt which is used to fund consumption rather than investment and it includes debts incurred on purchase of goods that are consumable (Wikipedia, 2011). Consumer debt is on the rise, in the developed countries (Webley and Nyhus, 2001) and particularly in the United States and the United Kingdom (Wikipedia, 2011), where consumer indebtedness is becoming a common theme in most of the households and at the same time it gains a momentum in emerging markets (Kamleitner et al., 2012).

The reason behind this upcoming trend in developed consumer cultures is the fact that they are repleted with idealised images of affluent lifestyles, where the consumption of the right material goods is associated with positive outcomes such as success, happiness and other indicators of psychological health (Gardharsdottir and Dittmar, 2012). This materialistic environment allowed banks to promote unsecured lending for the acquisition of luxury goods like cars and holidays and nurtured attitudes that made consumer debt socially acceptable. In (Betti et al., 2001) it is made clear that being indebted is now considered a normal behaviour and that a certain level of debt is inevitable for the majority of the households. The resulting present oriented lifestyle is longed by consumers of emerging markets like the new member states of European Union (Webley and Nyhus, 2001) strengthening the popularity of consumer debt around the world in a way that can now be considered as a global economic force (Kamleitner et al., 2012). On the other hand consumer debt is almost unknown in countries like Japan and China, because of the long-standing cultural taboos against personal debt (Gardharsdottir and Dittmar, 2012).

What makes consumer debt so popular in consumer cultures is the benefits it provides to consumers and households. The latter group of people can now utilise debt to smooth consumption throughout their lives, to borrow in order to finance expenditures (particularly housing and schooling) earlier in their lives and to pay down debt during higher-earning periods as the life-cycle theory dictates (Wikipedia, 2011). In (Betti et al., 2001) the consumer debt is characterised as a common instrument used to maintain a stable level of consumption, compatible with its life-time resources over different stages of the life-cycle. That means that consumer debt allows consumers and families to increase consumption in order to secure necessary expenditure but also acquire luxury goods earlier in life in a more present oriented manner. But in deciding to borrow, households have to make assumptions about their ability to repay the loan over its lifetime. Such assumptions are usually linked to their employment and income prospects, interest rates and future house prices (Hunt et al., 2015).

In addition to this, the purchase of material goods and the accumulation of consumer debt has become easier now with tools like credit cards. As described in (Kim and DeVaney, 2001) a credit card is now considered both a payment tool and a convenient source of credit and at the same time it is linked with the exponential growth of consumer debt in both developed and developing countries (Wang et al., 2011). Credit cards, originally created in developed countries, have spread rapidly in developing countries and have become a vital payment tool for consumers around the world. They are easy and quick to use and provide comforts to consumers that can now buy material goods without having to worry if they can afford a potential purchase or to carry enough cash. The comforts they provide appeal to the impulsive nature of individuals who can acquire goods they desire immediately and without any effort (Dittmar and Drury, 2000).

Despite all these benefits, the dangers for household over-indebtedness continue to lurk. As reasonable the assumptions of permanent income hypothesis and life-cycle decisions might be, they do not always hold as circumstances might develop in such a way as to undermine the ability of households to meet their financial obligations. The actual occurrence of these circumstances, such as unanticipated long spells of unemployment, a pregnancy or a divorce, would automatically outweigh any non-insurance precautions that consumer might have taken (Betti et al., 2001). Moreover the impulsive nature of individuals who get used to live in a lifestyle beyond their means (Gardharsdottir and Dittmar, 2012) might result in a loss of control in expenditure and in an accumulation of debt that is not repayable in a further extent. Household over-indebtedness has severe consequences for the affected households and for the financial industry, as well as broader economic and social implications (Consulting, 2013). In (Hunt et al., 2015) they state that high and rapidly rising levels of household debt can be risky as households suffer easier from financial shocks. This causes a decrease in spending which might cause financial institutions to suffer direct losses. From a social point of view, the excessive accumulation of debts accompanied by household liquidity constraints causes a deterioration in households social and economic well-being, thus leading in the short and medium term to social exclusion and poverty (Consulting, 2013). On a macro-economic level this might cause economic collapses of countries like Iceland (Gardharsdottir and Dittmar, 2012) and Italy (D'Alessio and Iezzi, 2013) whereas in (Webley and Nyhus, 2001) an increasing number of bankruptcies in private households in countries like USA, UK, France and Germany is being mentioned.

From all the above it becomes apparent that is of great social value to develop a strong financial indebtedness model to accurately identify individuals who are at risk to develop personal financial management problems (Stone and Maury, 2006). This way we will be able to gain a better understanding of the causes that drive consumer indebtedness and to prevent economic crises from happening. Accepting that the rational decisions behind the life-cycle theory are insufficient to explain alone the over-indebtedness of consumers, the model needs also to take account of the psychological aspects of consumer indebtedness and adopt a more holistic approach in the effort to understand and explain the complex nature of this emerging social phenomenon. Some of these aspects have been already revealed by the extensive credit use while the more complicated ones remain to be studied under a multidisciplinary approach.

1.2 Psychological Aspects of Consumer Indebtedness

1.2.1 A Multifaceted Phenomenon

The traditional economic and consumer models assume a “rational”, discerning and thoughtful consumer who gathers information strategically and buys goods according to functional cost-considerations (Dittmar and Drury, 2000). It is the life-cycle theory that assumes that consumers try to maximise utility from lifetime consumption by weighing the cost of living goods in the present against the expected cost in the future and choose the cost that minimises the consumption (Kim and DeVaney, 2001). However, this rational view of consumer indebtedness has been challenged by recent studies in the literature as it is limiting the research only on socio-economic variables (Ottaviani and Vandone, 2011).

Recent evidence supports that the behaviour of consumers deviates from the “rational” model (Ottaviani and Vandone, 2011) and that the analysis of consumer indebtedness should adopt a broader approach. Some studies have shown that the level of debt prediction is not a function of economic factors exclusively (Webley and Nyhus, 2001; Lea et al., 1995; Wang et al., 2011) while others emphasise the importance of psychological factors in improving the ability to predict the level of debt (Watson, 2003; Wang et al., 2011).

Now Consumer Indebtedness is considered a phenomenon with distinct facets (Kamleitner et al., 2012; Stone and Maury, 2006), which is influenced by several psychological processes and entails a magnitude of privetal, societal and economic implications. The incorporation of disciplines from psychology in consumer debt analysis revealed the significance of psychological factors in modelling consumer Indebtedness, by associating a series of personality traits, attitudes, beliefs and behaviours to consumer debt. This places the personality of individuals in the center of consumer debt analysis together with the traditional economic “rational” model.

1.2.2 Personality Psychology and Economics

Recent studies from areas of economics and personality psychology highlight the significant role personality psychology can play in the modern economics. Personality psychology offers a rich theoretical background that can possibly enhance the existing models in economics in a similar way cognitive psychology did by helping research to reach influential breakthroughs. This is clear in (Ferguson et al., 2011) where the authors emphasise the importance of accounting individual differences in economics, claiming that information that stems from the careful study of psychological characteristics of individuals, when applied in theoretically meaningful way, can provide further insight in interpreting complexities and patterns of economic behaviour. The collaboration of these two fields can lead to new areas of research, like the study of anti-social traits within an economic framework, an area of personality psychology that was long undermined by traditional economic analyses. Moving one step further, they claim that both disciplines can benefit each other, encouraging economists and psychologists to invest on this collaboration. As an additional argument for this collaboration, an important research work that associates economic parameters with personality traits, is presented in (Almlund et al., 2011).

1.3 Computational Economics

In a similar way that the incorporation of disciplines from the rich theory of psychology proved to be useful for the purposes of economics, so can be the application of Data Mining techniques in economic data. From the careful data pre-processing to improve the quality of the data by tackling inconsistencies they may contain and the diverse methods to explore relationships and patterns in the data to the strong predictive models and reliable evaluation techniques, Data Mining offers a variety of techniques and algorithms that comprise a complete and sophisticated toolbox to analyse complex real world data, like socio-economic data, and can guarantee representative and meaningful Knowledge Discovery. Such Knowledge can be used to produce economic benefits (Chen, 2006).

Data mining is defined as the process of discovering interesting patterns and knowledge from large amounts of data (Han et al., 2006). The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically. Usually it is treated as merely an essential step in the process of knowledge discovery which can be seen in Fig 1.1. In practice however the other steps are influenced by the decisions made in the Data Mining step, which is considered the core of the process. So it is common now to adopt a broader definition of Data Mining which is a synonym of Knowledge Discovery containing the necessary pre-processing steps, transformation, Data Mining models and data visualisation as seen in Fig 1.1.

1.3.1 Data Mining in Economics

Due to the lack of data, the lack of computational power and the lack of computationally tested institutional designs in the past, Data Mining approaches in social and economic sciences have been notoriously difficult (Helbing and Balietti, 2011). However, recent

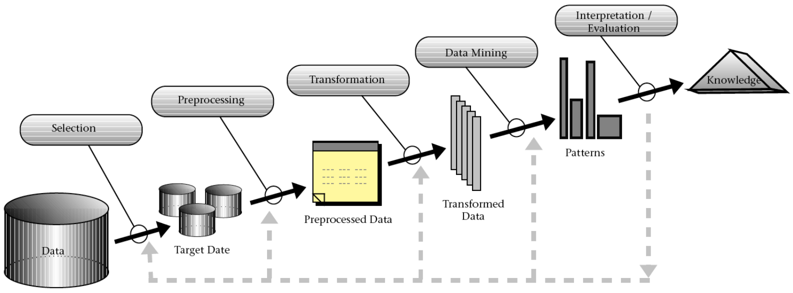


Figure 1.1: Process of Knowledge Discovery (Han et al., 2006)

advancements in Data Technology can change the scenery of Knowledge Discovery in Economics. Now we have data generated massively as more and more decentralised data sources are added every day capturing almost every notion of human activity, whether that is searching the World Wide Web for information or using digital sources of credit to buy products. In a lot of cases people leave behind a series of "digital traces" that form a rich dataset suitable for the purposes of complex social and economic analyses.

The enormous size and the broad nature of the data captured and stored nowadays require the sophisticated forms of analysis Data Mining has to offer in order to reduce serious gaps in our knowledge and understanding of techno-social-economic systems and prevent economic crises from happening by identifying systemic weaknesses (Helbing and Balietti, 2011). Since the grand challenges mankind is facing in the 21st century are to a large extent of socio-economic nature, such contributions can be of utmost significance. On emphasising upon this importance the authors in (Helbing and Balietti, 2011) state that the use of Data Mining techniques has to spread out in economic sciences if we want to find better answers. Moving one step further, they state that with the employing of Data Mining, economic science can be pushed to a new methodological paradigm which transcends the boundaries between theories and experiments, supplementing this way the current theoretical and statistical practices.

1.3.2 Towards Behavioural Mining

Despite all the advantages Data Mining offers, a "blind" application of the techniques and the methods can be dangerous (Helbing and Balietti, 2011) and a deeper understanding of the underlying assumptions related to the field of economics is required. An understanding that can be gained by incorporating theoretical knowledge of the domain of application into the Data Mining process. In doing so, Data Mining has to respect semantics of the data and try to include them in the modelling procedure by taking into account their underlying structure and the potential relationships between data points, in order to successfully mine behaviours within an economic context (Chen, 2006).

In such way Behaviour Mining, a special case of Data Mining, can be seen as the process that explains the phenomena that give rise to the data rather than just uncovering relations among the data. Behaviour is defined as the action or reaction of an entity or human to situations or stimuli in its environment (Cao, 2010). Within a socio-economic context and especially in Consumer Debt Analysis where consumer indebtedness is governed by rational decisions or psychological behaviours and attitudes, such level of analysis is essential for understanding the underlying forces that drive this social phenomenon. As a result, in order to understand the causes of such phenomena like consumer indebtedness, facts about the relations of the data must be supplied (Chen, 2006). The theoretical background provided by the disciplines of Economics and Psychology can be used to supply the required facts by squeezing out the behavioural elements hidden in the data and thus transforming them to a behavioural feature space. (Cao, 2010). Moving towards behavioural data is the necessary step to support Behaviour Mining, since on this representation Data Mining methods and techniques can explicitly and more effectively analyse patterns of behaviour.

1.4 A Data Mining framework to analyse Consumer Indebtedness

1.4.1 Purpose of this PhD

In this work we explore the multifaceted aspect of consumer indebtedness within a complete Data Mining framework and we try to verify the importance of psychological and behavioural characteristics of consumers in modelling consumer indebtedness. In other words, we utilise Data Mining techniques to research the potential of psychological information to support Knowledge Discovery in Consumer Debt Analysis.

Data Mining offers the necessary mechanisms and methods to handle the sensitive nature of this exploration. In more detail, by utilising data pre-processing techniques and algorithms we can identify and extract psychological behavioural aspects of consumer indetedness respecting its multifaceted nature, deal with the inconsistencies complex real world data such as socio-economic data possess and prepare them to contribute in the analysis. Then by using strong predictive models we can obtain reliable answers to significant questions of Consumer Debt Analysis.

Therefore we begin to build our framework by performing a series of data transformations in order to improve the quality of the data on one hand but also to explore the patterns of behaviour in the data in an exploratory manner and extract behavioural elements from socio-economic data. Then unsupervised learning approaches like clustering are used in order to build behavioural profiles of consumers based on economic/financial behaviours and social characteristics which could also include psychological characteristics like personality traits and attitudes if those are available on the data. These two techniques consist the first step of our two-step process that can be seen in Fig 1.2 named Behavioural Extraction, and it results in the production of behavioural data and behavioural profiles, which are capable of representing different behaviours of consumers. Moving to the second of step of the process, the Behavioural Modelling, the potential of Data Mining models to replace the traditional statistical modelling widely used in econometrics is assessed. The new strong and accurate models, capable of handling complex relations in the data are used in order to test the contribution of behavioural data and profiles in improving the modelling of consumer indebtedness and they provide insights to the problem of consumer indebtedness. Ultimately our proposed framework contains all the necessary steps to test whether the behavioural data and profiles can improve the existing traditional econometric models of consumer indebtedness which rely exclusively on socio-economic variables and provide a deeper insight on the “nature” of this complex problem. With that being said our research hypothesis can be stated more formally:

- **Hypothesis:** Psychological Information has the potential to achieve Knowledge Discovery in Consumer Debt Analysis.

This can be alternatively pictured as the process of including an an additional step of behavioural transformations to the original data in contrast with the traditional approach that is indicated by the green arrow in the Fig 1.2, that is also implemented within a Data Mining framework that will reliably assess its impact on modelling consumer indebtedness.

The proposed Data Mining framework can provide answers to the following research questions:

1. How do we process the data in order to extract psychological features of consumers from socio-economic data?
2. How do we build behavioural profiles that can produce individual behavioural differences of consumers?
3. Can Data Mining models successfully replace statistical modelling which is commonly used in economic and social Sciences?

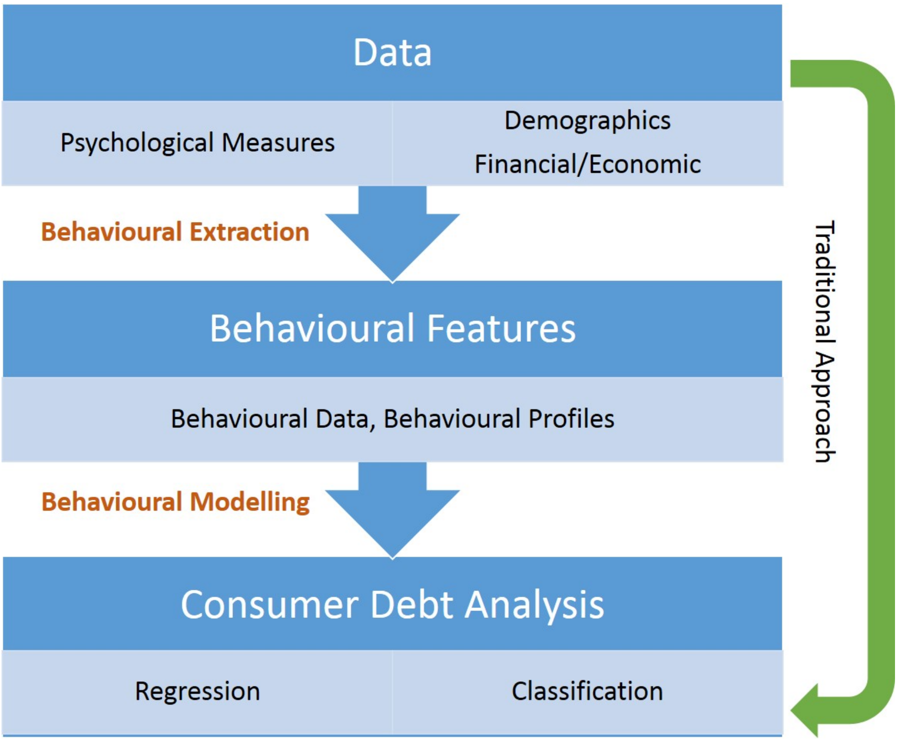


Figure 1.2: Diagram of the Data Mining framework for Consumer Debt Analysis

4. Can behavioural data enhance the process of Knowledge Discovery in Consumer Debt Analysis?

5. And finally can the incorporation of behavioural data and behavioural profiles improve the modelling of consumer indebtedness?

1.4.2 Methods and Techniques

1.4.2.1 Behavioural Feature Extraction through clustering

On implementing the first step of our framework we perform a series of data transformations on a socio-economic dataset (CCCS) in order to represent them in a behavioural feature space. The resulting behavioural data are free of several inconsistencies that characterise a socio-economic dataset and are able to express economic/financial and spending behaviours of consumers. Under a consensus clustering framework that utilises the agreement of different clustering algorithms to produce robust and representative clusters, the behavioural data are shown to improve the quality of the clustering and produce seven behavioural profiles with different patterns of behaviours. The final groups of debtors emerge under a careful examination of several evaluation techniques that point out the importance of modelling the agreement of the cluster ensemble and are shown to contribute in the Consumer Debt Analysis by grouping consumers based on nuanced observations of well researched factors (Kamleitner et al., 2012). The agreement of the cluster ensemble poses as an alternative validation measure and seems to uncover patterns of consumer behaviour that transcend the assumptions of Compactness and Separation.

After the the contribution of behavioural data on our unsupervised learning approach has been established, we continue to improve the framework of consensus clustering we used by introducing a new consensus function. Our proposed consensus solution models successfully the agreements and disagreements of the cluster ensemble in order to produce well defined clusters that describe patterns that try to maximise the Compactness and Separation of clusters on one hand and the agreement of the cluster ensemble on the other. Our method is evaluated in different datasets and is applied on the second socio-economic dataset (DebtTrack) improving the process of behavioural profiling to a greater extent and producing meaningful behavioural profiles that provide a better understanding of the complex "nature" of consumer indebtedness.

1.4.2.2 Behavioural Modelling through Regression and Classification

As the need to develop fairly accurate quantitative prediction models that can provide a deeper insight in the nature of consumer indebtedness becomes apparent (Atiya, 2001) popular Data Mining models are utilised for the purposes of the level of debt prediction. Random forests and especially Neural Networks outperform the traditional statistical modelling of Linear Regression which is widely used in economic and social sciences and suffers from its inability to handle incosistencies of real world applications. These models are also used in order to verify the significance of behavioural data and behavioural profiles, previously extracted from the CCCS dataset, into modelling consumer indebtedness. Furthermore we take advantage of the ability to design the topology of Neural Networks and we introduce a novel way to incorporate into their architecture meaningful knowledge that derives from these explanatory techniques applied on data, like clustering and factor analysis, improving more the predictive ability of our models.

Finally utilising the comprehensive nature of DebtTrack dataset, a survey that contains psychological items, we explore the multifaceted nature of consumer indebtedness. A factor analysis method extracts personality factors from the data and their impact on modelling consumer indebtedness is verified by Data Mining models whose superior performance has been acknowledged from previous steps of our research. More accurately, extracted psychological factors are proven to be important for models that separate debtors from non-debors and for models that predict the level of debt, two important research questions of Consumer Debt Analysis (Wang et al., 2011).

1.4.3 Contribution

Our proposed framework introduce novel and interesting methods and techniques of Knowledge Discovery. Regarding the clustering approaches, a new method to evaluate the clustering results is presented that challenges the ideas of *Compactness* and *Separation* existing internal validation criteria adopt. The proposed metric improves the quality of clustering by revealing robust and representative patterns when applied in real-world data where usually the assumptions of *Compactness* and *Separation* do not hold. In addition to this, a new consensus solution is devised that can improve the existing framework of consensus clustering and enhance the process of behavioural profiling. The proposed consensus solution uses Homogeneity Analysis to model the disagreements and agreements of the cluster ensemble in a new representation that can empower the clustering process to identify well defined consensus clusters in a two tier clustering structure. Homals Consensus, as it is named, in contrast with traditional clustering approaches proves to be more suitable for extracting behavioural clusters from socio-economic data that are characterised by large number of inconsistencies.

Concerning the supervised modelling strong and accurate models are required to deal with the non-linearities in the data and examine the multifaceted nature of consumer indebtedness. Therefore, we take advantage of the flexibility of Neural Networks in the design of their topology as it offers a way to incorporate important steps of the Data Mining process into a regression model. Such steps are evident in the exploratory techniques to uncover associations between variables and between data points. The resulting Neural Network can maintain the superior performance of Neural Networks but it can also be more interpretable as the neuronal architecture reflects the knowledge extracted from unsupervised learning approaches. This latter fact is much desired in social sciences. All the above, sketch a complete framework for the Consumer Debt Analysis including necessary transformations of data, exploratory models and reliable and transparent modelling that may be extended to any real world application problem that contains a dataset with similar inconsistencies and sensitive behavioural characteristics as the socio-economic datasets of CCCS and DebtTrack.

Our results confirm the beneficial role Data Mining can play in Consumer Debt Analysis in two socio-economic datasets. The Behavioural Extraction methods and techniques of data pre-processing are able to handle a number of inconsistencies that complex real world data possess and in combination with exploratory unsupervised learning approaches they produce meaningful behavioural data and profiles that are able to capture psychological and behavioural information hidden in the data. At the same time powerful Data Mining models introduced during the Behavioural Modelling stage pose as competitors to traditional statistical modelling as not only they exhibit a better predictive ability, but they also possess explanatory power to provide deeper understanding of the “nature” of this social phenomenon.

The analysis of the models either from unsupervised learning or supervised learning reveals the importance of psychological information to support Knowledge Discovery in the field of Consumer Debt Analysis. The extracted behavioural data and personality factors help clustering identify a novel classification of consumers on one hand and improve the performance of the predictive modeling on the other, verifying the multifaceted nature of Consumer Indebtedness.

1.5 Outline of this thesis

In the 2nd Chapter we discuss the developments of the research regarding the Consumer Debt Analysis and the Data Mining modelling in an extensive literature review that starts with the work conducted for the purposes of Consumer Debt Analysis and ends with the technical presentation of Data Mining models.

In the 3rd Chapter we present in every detail all the methods and techniques included in our research to form our Consumer Debt Analysis framework covering data preprocessing, clustering approaches and classification and regression modelling.

In Chapters 4 and 5 we describe our work for the Behavioural Extraction stage. Starting with the necessary behavioural transformations and the alternative method to evaluate clustering, we produce seven behavioural profiles under a consensus clustering framework as it is shown in Chapter 4. In Chapter 5 we develop a new consensus function and verify its performance in multiple datasets. Then it is used to extract knowledge from the DebtTrack dataset enhancing the process of behavioural profiling.

Our Behavioural Modelling stage is analysed in Chapters 6 and 7. The superior performance of Data Mining models over statistical modelling is explained in Chapter 6 together with our novel method of Neural Networks whereas in chapter 7 we confirm the multifaceted nature of consumer indebtedness and the potential of psychological information to support Knowledge discovery in Consumer Debt Analysis.

Finally, in Chapter 8 we conclude our work and we discuss of potential extensions of this framework to other fields of economic research.

Chapter 2

Literature Review

2.1 Introduction

In this chapter we present our research in bibliography regarding the problem of modelling consumer indebtedness within a Data Mining framework. Findings of great significance for the Consumer Debt Analysis are being shown, which associate consumer indebtedness with diverse aspects of consumer behaviour revealing its complex and multifaceted nature. We then analyse ways we can utilise more efficiently Data Mining techniques that can handle the complexity of the problem and support the process of Knowledge Discovery on this field. For this reason we deal with technical issues of clustering, like clustering evaluation and consensus clustering that can uncover meaningful patterns of behaviour of consumers and we study the potential of Data Mining regression/classification algorithms to replace statistical models and improve the modelling of consumer indebtedness in a further extent. Finally, we explore ways in order to preprocess socio-economic data so that they can be prepared for the Data Mining processes free of any inconsistency they may contain and also to be able to capture behavioural and psychological information hidden in the data and make it available further use in our Data Mining framework.

2.2 Analysis of Consumer Debt

Consumer debt has risen to be a significant problem of our modern society. Especially in developed countries this emerging social problem has raised concern among the academic community, which tries to provide reasonable explanations for the over-indebtedness of consumers. As traditional economic models failed to predict the rapid growth of this problem, research has focused on alternative ways to find answers, like

Psychology. Now consumer debt is considered a complex problem that is associated with many distinct facets, is influenced by psychological processes and entails a multitude of personal, societal and economic factors (Kamleitner et al., 2012; Stone and Maury, 2006). The inclusion of psychological factors in the analysis of consumer debt has been greeted with great optimism by the research community as it has been shown throughout the literature that it can improve the level of debt prediction, which is not a function of exclusively economic factors any more (Webley and Nyhus, 2001; Lea et al., 1995; Wang et al., 2011) and that it can help in explaining sufficiently the behaviour of consumers that deviates from the rational model that it was assumed so far by traditional economics (Ottaviani and Vandone, 2011). Part of the economic psychology field, the research for Consumer Debt Analysis has been mainly focused on answering three fundamental questions (Wang et al., 2011):

1. Which factors discriminate debtors from non-debtors?
2. Which factors affect how deep consumers go into debt?
3. Which factors influence the repayment of debt?

Answering these questions has led to the discovery of a series of diverse factors that are associated with consumer indebtedness. A variety of socio-economic, psychological, attitudinal factors are studied throughout the literature. Besides this, research also revealed other associations of consumer indebtedness with expressions of consumer behaviour like consumption and financial literacy, supporting further the multifaceted view of this social phenomenon.

2.2.1 Factors Associated with Consumer Indebtedness

While an abundance of factors has been suggested to be related to consumer debt, articles like (Wang et al., 2011; Stone and Maury, 2006; Kamleitner et al., 2012; Kamleitner and Kirchler, 2007) provide a nice review of the factors proposed and a nice overview of the ongoing research in the field. Here we present the factors in a three-category structure:

1. Socio-economic.
2. Environmental
3. Psychological

A complete list of the factors associated with consumer indebtedness can be viewed in Table 2.1.

Socio-economic
Income
Age
Number of children in the household
Marital Status
Employment Status
Education
Social Class
Sex
Ethnicity
Number of credit cards
Real Assets
Environmental
Job Changes
Interest rates
Life events
Financial events
Psychological
Conscientiousness
Locus of Control
Sensation Seeking
Social Comparison
Present Orientations
Attitudes towards credit/debt
Impulsivity

Table 2.1: A list of factors associated with consumer indebtedness

2.2.1.1 Socio-economic Factors

Among the socio-economic factors that influence consumer indebtedness is the *income*, household or personal. The amount of money a consumer earns increases the likelihood of that person to be in debt (Livingstone and Lunt, 1992) whereas other findings (Stone and Maury, 2006; Ottaviani and Vandone, 2011; Kim and DeVaney, 2001; Lea et al., 1993) associate *income* with the amount of accumulated debt. However not all findings in literature accept the relationship between *income* and consumer indebtedness (Norvilitis et al., 2003). Authors in (Kamleitner et al., 2012) claim that this relationship is not clear and that *income* is indeed important only when combined with other factors.

Age is another well studied factor in literature and is shown to be associated with consumer indebtedness (Stone and Maury, 2006; Livingstone and Lunt, 1992). It is supported by the Life-cycle theory that dictates that the younger a consumer is the more likely to owe (Webley and Nyhus, 2001). In (Kamleitner et al., 2012) the authors emphasise on the complex relationship between *age* and consumer debt showing that it can be approximated by curvilinear function.

Two factors that indicate the family status of a consumer, *marital status* and *number of children in the household*, seem also to be associated with debt. A consumer that is married is more prone to debt (Kamleitner et al., 2012) while the formation of a family is usually followed by an increase in consumption since the household needs to provide the necessary goods to the children. Therefore the *number of children in the household* is also shown to predict the level of household indebtedness (Kim and DeVaney, 2001; Livingstone and Lunt, 1992; Lea et al., 1993; Ottaviani and Vandone, 2011).

Other socio-economic factors include the *employment status* of the person (Wang et al., 2011) especially in the case of self-employed (Ottaviani and Vandone, 2011), *education* (Kim and DeVaney, 2001) but only when is considered in terms of specific financial education (Kamleitner et al., 2012), *social class* (Wang et al., 2011; Lea et al., 1993), *sex* and *ethnicity* (Stone and Maury, 2006) from demographic variables and *real assets* and *number of credit cards* (Kim and DeVaney, 2001) from economic variables .

2.2.1.2 Environmental Factors

A special case of factors is presented in (Kamleitner et al., 2012), which refer to circumstances that occur in the life-stage of a consumer and result in an unexpected increase in spending and in a greater extent to the accumulation of debt. This category of factors describes the effect of the environment on the indebtedness of a consumer and includes factors like *job changes*, *interest rates*, *financial events* (Stone and Maury, 2006) and *life events*. *Interest rates* are also mentioned in (Kim and DeVaney, 2001) where it is explained that high interest rates cause an increase in outstanding debt.

2.2.1.3 Psychological Factors

After presenting a series of factors, socio-economic and environmental, that are closely related with the “rational” approach to consumer indebtedtedness according to life-cycle theory, we continue with the discussion of psychological factors. Throughout the literature a series of personality traits, behaviours, attitudes and other psychological factors provide an alternative approach to analyse consumer debt. They place the personality of the individual in the center of the attention and reveal the psychological aspect of consumer indebtedness that causes a person to go into debt not based on a “rational” cognitive process but based on individual preferences.

The personality of an individual, according to (Larsen and Buss, 2008), is the set of psychological traits and mechanisms within the individual that are organised and relatively enduring and that influence his or her interaction with the intrapsychic, physical and social environment. Personality is based on stable and consistent traits that describe the average tendencies of a person and define how unique and how similar a person is in comparison with others. Defined in the dispositional domain of personality psychology, traits play a vital role in describing different individuals.

Traits are usually adjectives used to describe characteristics of people. Numerous traits have been studied in literature covering all possible aspects of human nature but the more interesting fact is how these traits are organised into a single quantifiable taxonomy. A very popular taxonomy of traits is the Big Five or the Five-Factor Model (Costa and MacCrae, 1992) that was defined by several independent sets of researchers. These researchers began by studying known personality traits and then factor-analysing hundreds of measures of these traits (in self-report and questionnaire data, peer ratings, and objective measures from experimental settings) in order to find the underlying factors of personality. The factor analysis resulted in five dimensions of personality orthogonal to each other that organise a big amount of personality traits in such an objective way that makes the model robust even when it is applied in different cultures. The quantified aspect of this model makes it ideal for the purposes of quantitative analyses like the ones that take place in economic sciences. The five dimensions of personality are these:

- **Openness to experience**: (inventive/curious vs. consistent/cautious). Appreciation for art, emotion, adventure, unusual ideas, curiosity, and variety of experience. Openness reflects the degree of intellectual curiosity, creativity and a preference for novelty and variety a person has. It is also described as the extent to which a person is imaginative or independent, and depicts a personal preference for a variety of activities over a strict routine. Some disagreement remains about how to interpret the openness factor, which is sometimes called ”intellect” rather than openness to experience.

- **Conscientiousness**: (efficient/organized vs. easy-going/careless). A tendency to be organized and dependable, show self-discipline, act dutifully, aim for achievement, and prefer planned rather than spontaneous behavior.

- **Extraversion**:(outgoing/energetic vs. solitary/reserved). Energy, positive emotions, surgency, assertiveness, sociability and the tendency to seek stimulation in the company of others, and talkativeness.

- **Agreeableness**: (friendly/compassionate vs. analytical/detached). A tendency to be compassionate and cooperative rather than suspicious and antagonistic towards others. It is also a measure of one’s trusting and helpful nature, and whether a person is generally well tempered or not.

- **Neuroticism**: (sensitive/nervous vs. secure/confident). The tendency to experience unpleasant emotions easily, such as anger, anxiety, depression, and vulnerability. Neuroticism also refers to the degree of emotional stability and impulse control and is sometimes referred to by its low pole, “emotional stability”.

Among the personality traits that are researched for the purposes of Consumer Debt Analysis one can find *conscientiousness*, *locus of control* and *sensation seeking*. However their contribution in explaining the consumer indebtedness is not clear in most of the cases. The relationship between *conscientiousness* and consumer indebtedness is not strong enough to indicate a clear association (Webley and Nyhus, 2001). *Locus of control* is a trait that refers to the extent to which individuals believe they can control events affecting them and it has shown to affect consumer indebtedness (Wang et al., 2011), in contrast with the findings in (Lea et al., 1995; Kamleitner et al., 2012) that suggest otherwise. Finally *sensation seeking* is a trait defined by the search for experiences and feelings, that are varied, novel, complex and intense, and by the readiness to take physical, social, legal, and financial risks for the sake of such experiences. According to (Kamleitner et al., 2012) *sensation seeking* encourages credit use and high sensation seekers are expected to risk more and engage in problematic financial behaviours than low sensation seekers. Findings in (Wang et al., 2011) however do not support this claim.

In contrast with personality traits whose contribution to Consumer Debt Analysis remains controversial, psychological factors that describe behavioural aspects of individuals seem to have a bigger impact on analysing consumer indebtedness, like *social comparison* and *present orientation*. *Social comparison* is perceived as the need to express one’s identity through social relations and consumption. It is showed to increase the propensity of credit use (Kamleitner et al., 2012) and to be predictive of the level of debt of a consumer (Lea et al., 1995). *Present orientation*, a term that describes individuals who tend to systematically overvalue immediate costs and benefits and undervalue future ones, is also associated with the accumulation of consumer debt (Webley and Nyhus, 2001).

Finally very significant is the contribution of attitudinal factors in the analysis of consumer debt. As shown in (Robb and Sharpe, 2009) positive attitudes towards credit increase the credit use and the likelihood of a consumer to go in debt and high affective credit attitude scores are found to be related with outstanding balance. At the same time positive attitudes towards debt are related to consumer indebtedness (Webley and Nyhus, 2001) and are among the important predictors of the level of indebtedness (Livingstone and Lunt, 1992).

2.2.1.4 Impulsivity

Unlike the personality traits whose contribution in the understanding of consumer indebtedness demands further research, *impulsivity* or *self control* has a more clear association with consumer debt. Impulsivity is a facet of personality that involves a tendency to act on a whim, displaying behavior characterised by little or no forethought, reflection, or consideration of the consequences. Within economic context impulsive subjects overestimate the duration of the time intervals and as consequence discount the value of delayed rewards more than self-controlled individuals do (Ottaviani and Vandone, 2011).

Considering the involvement of *impulsivity* or *self control* in Consumer Debt Analysis, the work in (Gathergood, 2012) revealed a significant association with consumer debt as it is shown to be a significant predictor of the level of indebtedness. It is also associated with the greater use of specific types of credit like credit cards, mail order catalogues, home credit and payday loans. Moreover impulsive spending, in the same work has been linked to impulsive personality that can lead to income shocks.

In (Ottaviani and Vandone, 2011) *impulsivity* was found a strong predictor of unsecured debt but not secured debt, like mortgages and car loans. The rationale behind this, stems from the fact that secured debt affects decisions that last for a long time and therefore it complies with life-cycle theory (Webley and Nyhus, 2001), according to which a consumer enters into debt on rational grounds in order to maximise utility in a time consisted manner. Thus it is not associated with the impatient sort sighted behaviour impulsive individuals adopt which favours short run benefits like the ones unsecured debt usually provides.

Other findings in literature also support the impact of *impulsivity* on consumer indebtedness (Noone et al., 2012; Wang et al., 2011), while in (Kamleitner et al., 2012) is showed to increase the credit use and finally the work of (Dittmar and Drury, 2000) provides further evidence between impulsive individuals and the use of credit cards.

2.2.2 Consumption

A significant discovery being presented in the findings of this research, is the importance of consumption data for the analysis of consumer debt. In (Vissing-Jorgensen, 2011) it is shown that what a consumer buys provides substantial information about potential default losses on a loan. The reason is simple and is based on the fact that states that what individuals choose to buy can reveal significant characteristics of their personality. Among the reasons people choose to consume and decide where to spend their money on are to satisfy their survival needs, to establish social significance with the power and prestige that money and possessions represent and to compensate for individuals shortcomings such as faulty self-esteem (Watson, 2003). Different types of spending reveal different aspects of personality and direct or indirect links to consumer indebtedness.

From the above types of spending, the most clear division of expenditure is the one that distinguishes spending money to cover survival needs from spending money to establish social significance. This difference between necessary and luxury purchases is analysed thoroughly in literature of consumer indebtedness. (Kamleitner et al., 2012) compares consumption patterns of borrowers and non-borrowers with the first group to choose to spend less money on necessities and more to luxuries. In (Lea et al., 1995) they authors claim that people with serious debt problems regard certain kind of expenditures on children as necessities even though these are commonly considered as luxuries, while the authors in (Watson, 2003) try to confirm the link between luxury spending and consumer indebtedness (Ottaviani and Vandone, 2011) but their results are not able to support this hypothesis. In addition to this, luxury spending reveals materialistic aspects of individuals (Watson, 2003) who are likely to exhibit favorable attitudes towards borrowing money to make luxury purchases like furnishings, second holiday houses and jewelry as well as aspects of impulsivity (Dittmar and Drury, 2000) when it is expressed on purchases like clothes, jewelry and ornaments.

Materialism has been defined as the importance ascribed to ownership and acquisition of material goods (Gardharsdottir and Dittmar, 2012), with materialistic individuals seeing the acquisition of material goods as a central life goal and as the primary way of attaining access, happiness and self-identity (Gardharsdottir and Dittmar, 2012; Pham et al., 2012). Consumers with high materialistic values are also more likely to have more credit and positive attitudes towards debt (Watson, 2003), associating materialism with consumer indebtedness. This argumentative relationship in literature, though unclear in the findings of Kamleitner et al. (2012), is further strengthened by its relationship with compulsive buying.

Compulsive buying is a psychiatric disorder in which individuals lose control over their buying behaviour, their impulse to buy is experienced as irresistible and they continue with excessive buying despite adverse consequences (Gardharsdottir and Dittmar, 2012). Not to be confused with impulsive buying which is influenced by environment and is infrequent, compulsive buying occurs frequently, is triggered by internal stimuli like anxiety, stress and self-esteem and leads to financial problems and indebtedness. (Pham et al., 2012). Since the relationship between compulsive buying and materialism is well established (Pham et al., 2012), the indirect relationship between materialism and consumer indebtedness is explored and confirmed in (Gardharsdottir and Dittmar, 2012), where high levels of materialism are also associated with a greater spending tendency, higher compulsive scores, greater financial worries and less self reported skills in money management.

From this, it is made clear that spending patterns are related with self esteem, survival need and social significance but they can provide deeper insight of the behaviours of debtors who usually regard certain types of expenditure as necessities while they are commonly considered as luxuries (Lea et al., 1995; Ottaviani and Vandone, 2011). Furthermore compulsive spending which is considered a psychiatric disorder of individuals who lose control over their spending behaviour (Gardharsdottir and Dittmar, 2012) has been shown to be caused by anxiety and stress and is associated with materialism in a complex relationship (Pham et al., 2012). All these point out the potential of consumption data to reveal aspects of the personality of the consumer.

2.2.3 Financial Illiteracy

Finally, consumer indebtedness is closely related to the concepts of *financial literacy* or *financial education*. *Financial literacy* refers to the consumer's understanding of financial concepts and to the consumer's ability to correctly interpret financial data. Low levels of *financial literacy* are associated with limited participation in the stock market, limited financial preparation for retirement and high risk in developing outstanding debt (Gathergood, 2012). Further findings in (Kamleitner et al., 2012) show evidence that the association between *financial literacy* and consumer indebtedness is becoming clearer.

The reason for this is that *financial literacy* influences the ability of a consumer to create and follow *money management* or *financial management* practices. *Financial management* practices as defined in (Pham et al., 2012) refer to behaviours like budgeting, making payments on time, saving money, credit debt management and knowing net's worth. As a consequence they help individuals to keep track of their expenditure in order to improve their financial well being, a theoretical construct to describe both objective indicators such as *income* or debt and subjective indicators such as *financial worry* (Gardharsdottir and Dittmar, 2012). These practices have been shown to moderate the relationship between materialism and compulsive buying and therefore can lead indirectly to less financial problems. On the other hand, the absense of these practices can result in the over-indebtedness of the consumer (Lea et al., 1995).

2.2.4 A broader view on Consumer Indebtedness

All the above findings introduce a broader view on the social phenomenon of consumer indebtedness, which include traditional rational approaches, psychological characteristics of individuals, financial education and spending behaviours. Its multifaceted nature is evident in numerous notions of consumer's activity leaving traces in different organisations. Since the recent developments of data technology provide the means to capture a big proportion of human activity in a digital form, the challenge of our time is to capture all the possible digital traces consumers produce as part of their economic activity, merge them into a complete dataset and find the appropriate techniques to analyse these complex data in order to enhance the process of Knowledge Discovery in Consumer Debt Analysis. On the significance of this task, in the next section we focus on the potential of Data Mining techniques to explore and identify patterns of complex nature for the purposes of Consumer Debt Analysis.

2.3 Unsupervised Learning for explaining Consumer Indebtedness

2.3.1 Classifications in Consumer Debt Analysis

In literature several diverse groupings of consumers are identified that serve as alternative debtor profiles and provide a deeper insight into the problem of consumer indebtedness by describing different types of consumers based on different criteria.

Based on the level of debt, the authors in (Lea et al., 1995) describe three groups of consumers, non-debtors, mild debtors and serious debtors. Based on the repayment behaviour of consumers the work in (Webley and Nyhus, 2001) identifies the model consumer, the bad consumer, the temporarily indebted, the chronically indebted and the defaulting consumers. More importantly, chronic debtors that exhibit a tendency to remain in debt are a small group of consumers that are charaterised by having more limited resources, being more present oriented and finding more difficult to control their expenditure. In (Kim and DeVaney, 2001) consumers are differentiated based on their credit usage, whereas in (Pham et al., 2012) consumers are split into big spenders and value seekers based on their spending behaviour. Big spenders are characterised by high levels of *materialism* and low *financial management practices* exhibiting signs of compulsive behaviour as opposed to value seekers who also demonstrate high levels of *materialism* but better *financial management practices*.

Another important classification of consumers is the well known *Socio-Economic Status* (SES) or *social class*, which formulates a construct that characterises consumers based on their differences across several socio-economic variables. SES has been shown to possess the ability to explain credit card usage (Watson, 2003), to be an important predictor of *financial planning* and to influence levels of retirement investments and the likelihood of seeing professional advice (Noone et al., 2012).

All the above point out that the potential classifications of consumers may have to explain aspects of economic activity that are related to the complex nature of consumer indebtedness. Towards this direction, clustering, a traditional unsupervised learning algorithm for grouping together objects that are similar, can further assist in the analysis of consumer debt by either verifying the same groups that are described in the literature of economic psychology in a massive dataset or by uncovering new groups of debtors with more nuanced observations of well researched factors. This is highlighted in (Kamleitner et al., 2012), where the inability of single factors to sufficiently explain indebtedness on their own is explicitly stated and it is proposed that researchers should focus on identifying particular consumer groups characterised by more nuanced observation of well researched variables. It is possible that a combination of factors may have more explanatory power beyond the sum of their individual effects and it is of great importance to establish the impact of combinations of factors on analysing consumer indebtedness.

In a similar way, clustering has been proven useful in other fields of economics, like in marketing (Otto et al., 2009), where it was utilised to provide a customer segmentation on spending patterns that could be used to improve the marketing strategies of companies. Also in (Hennig and Liao, 2013) clustering was used to partition the population in different social classes.

2.3.2 Clustering

Clustering refers to the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those of other groups. The resulting groups are often used to develop an understanding of the underlying structure in the data (Eshghi et al., 2011), an important task of Data Mining as described in (Dimitriadou et al., 2002). It usually requires the definition of a similarity measure between objects in order to uncover the natural structure in the data. For that reason, various algorithms have been proposed in the literature exhibiting different notions of similarity, capabilities and strengths and clustering gained a lot or popularity as it has been widely used in many application areas. However, as powerful as this mrthod can be, the application of clustering in order to discover similar groups of observations requires a number of important decisions to be made, including choosing the right clustering method, determining the correct numbers of clusters and finding ways to evaluate the quality of the clustering (Hennig and Liao, 2013).

The last is a very difficult task due to the ill defined nature of clustering, because it is difficult to measure the quality and the importance of the resulting clusters. Unlike supervised learning where labels lead to crisp performance criteria like accuracy, the external information is absent in clustering and thus we cannot assess the quality of the clusters. This also makes difficult to determine what is the optimal number of clusters that best describe the dataset and which are these as this decision must be based upon the purpose of clustering (Caruana et al., 2006). In a similar way, choosing the appropriate clustering method requires deep understanding of the domain of application. That stems from the fact that each clustering technique poses its own strengths and limitations, and is unable to identify all cluster shapes and structures that are encountered in practice in real world data (Fred and Jain, 2005). Therefore, selecting the best algorithm for each application requires the understanding of the purpose of clustering and the definition of the interpretative meaning of the clustering solution that can be translated into data analytic characteristics of the clustering method (Hennig and Liao, 2013).

2.3.3 Clustering in practice

2.3.3.1 Choosing the appropriate clustering algorithm

In literature a variety of clustering algorithms has been proposed based on different methodologies and assumptions. A basic categorisation of the clustering methods can be found in (Han et al., 2006) where they are split into *partitioning*, *hierarchical* and *density based* methods. More advanced clustering methodologies introduces new types of clustering like *soft* clustering and *model based* methods. All these clustering methods impose different views on the data based on different grouping criteria and assumptions. In (Eshghi et al., 2011) it is shown that different clustering methods produce different results both in terms of the number of groups and in terms of group membership. This suggests that each of the clustering methods can lead to a potentially different interpretation of the underlying structure of the data based on the assumptions associated with the methodology of the clustering method. Therefore, it requires a careful approach to formulate the interpretative meaning of the clusters and choose the appropriate clustering method that will produce the desired view on the data.

One of the most popular clustering algorithms is a *partitioning* algorithm named K-means (Lloyd, 1982). More accurately K-means is a *centroid-based* algorithm that partitions the data based on their distance from k representative centroids. Due to its simplicity K-means became very popular in numerous applications but it has also received a lot of criticism at the same time. Its major limitation is its inability to identify clusters with arbitrary shapes ultimately imposing hyper-spherical shaped clusters on the data (Fred and Jain, 2005). In addition to this, K-means performs miserably in several situations where the data cannot be accurately characterised by a mixture of K-gaussians with identical covariance matrices (Strehl and Ghosh, 2003).

However, it remains a very efficient method for clustering large datasets (Huang, 1998), while experimental comparisons have shown K-means to outperform modern clustering approaches like *model based* methods and Self Organising Maps (Eshghi et al., 2011). That means that K-means remains a very useful tool for unsupervised Knowledge Discovery in the data if the desired interpretative meaning of a clustering application agrees with the hyperspherical assumpions a *centroid-based* algorithm makes. Then some of its limitations and technical weaknesses can be dealt with careful data pre-processing techniques.

An example of such weakness is that its performance relies on the random initialisation of the centroids in its early stage. Different centroid selection in the beginning of the algorithm might result in different clustering results. This is usually dealt with a repeated application of the algorithm on the data in order to obtain the best of the clustering results which could be closer to the optimal solution. Another approach is to involve hierarchical clustering to obtain a suitable number of clusters and initial seeds that can be used to define the initial centroids of K-means. This technique has been shown to produce very good results (Eshghi et al., 2011).

An alternative *centroid-based* algorithm to K-means that is more robust to the presence of outliers is K-medoids. K-medoids uses medoids as the centers of gravity that partition the data and theseare based on the median and not the average, and therefore they are not susceptible to the effects of outliers. K-medoids has also been shown to outperform *model based* methods in a comparison presented in (Hennig and Liao, 2013).

Both algorithms require the number of clusters to be given as an input and that remains one of their biggest drawbacks. It is needed then to define validation criteria to assess the quality of clustering in order to determine the number of clusters that describe best the data if that is not known beforehand.

2.3.3.2 Evaluating the quality of the clustering

Recognised as one of the vital issues essential to the success of clustering applications (Liu et al., 2010), identifying what makes a good clustering has been a controversial subject that has been lengthy discussed throughout the literature. Assumptions regarding the perception of similarity and definition of good clusters led to the creation of certain validation measures known as validation indices which fall into two main categories, external validation and internal validation (Liu et al., 2010; Dimitriadou et al., 2002). External validation demands the presence of external class information which serves as the ground truth to compare with the clustering. Internal validation measures on the other hand, rely only on the information in the data. Assuming that the goal of clustering is to partition the objects in such way that the ones that fall into the same cluster are as similar as possible and the ones that belong to different clusters as distinct as possible, they try to assess the goodness of fit of the clustering structure by measuring the degrees of *Compactness* and *Separation* of the clusters. *Compactness* measures how closely related the objects in the clusters are and separation how distinct and well separated a cluster is from the other clusters of the grouping.

Since class information is usually not available in a traditional clustering problem, the discussion is reduced to detecting the best validation index to measure the quality of the clustering. From the large number of internal validation indices presented, each one offers certain advantages and certain drawbacks. A detailed review of these indices, together with a comparative analysis of their performances and the suggested method of using them in order to define the optimal number of clusters in a dataset can be found in (Vendramin et al., 2010; Liu et al., 2010). Through a series of experiments and tests two of the indices, *Calinski* (Caliński and Harabasz, 1974) and *Silhouette* (Rousseeuw, 1987), are shown to achieve very good results. Different from the theoretical concept of goodness of fit, *Calinski* seems to favour homogeneity of clusters focusing on *Compactness* while *Silhouette* emphasises the gaps between neighbouring clusters focusing on *separation* (Hennig and Liao, 2013). Despite their good performance both indices suffer from their own limitations. *Silhouette* cannot handle the case of subclusters while *Calinski* is susceptible to noise and skewed data (Liu et al., 2010).

This enhances the view that there is no single validation index that is superior compared to all the rest and that the experimental comparison of the validation indices cannot single out such an index. For this reason the authors in (Eshghi et al., 2011) devised their own evaluation metrics to compare three different clustering methodologies that follow the notions of Compactness and *separation*.

But, the assumptions of *Compactness* and *Separation* as an indication of a good clustering are argumentative. Assumptions of traditional indices are unsatisfactory in most cases (Hennig and Liao, 2013) while in (Caruana et al., 2006), it is shown that clusterings that would be of most interest to users are not very compact usually. That stems from the fact that any reasonable partitioning of the data can be potentially useful for some purpose regardless of whether it is optimal according to some criterion or not. Thus, searching for the best clustering according to some criteria does not make sense since the criteria require understanding of the data but clustering is the principal tool to understand the data resembling the famous chicken-or-the-egg problem. Instead we should focus on devising clustering criteria that capture the information need of the users and that the best clustering of a dataset should be dependant on how the clustering will be used. The latter is being further supported by (Eshghi et al., 2011; Hennig and Liao, 2013), articles that serve as general guides to conduct clustering on data that derive from real world applications. The work of (Hennig and Liao, 2013) in particular concentrates on formalising the *aim of clustering* that dictates that the purpose of clustering should be considered when a clustering gets evaluated, enhancing the view that the produced classification should reveal meaningful patterns for the domain of application.

2.3.3.3 Determining the number of clusters

Since the number of clusters of a dataset is not usually known beforehand, the estimation of clusters contained in the data remains a critical issue in the agenda (Vendramin et al., 2010). It is considered to be an unsolved problem of great significance and the success of a research is depending on this crucial decision (Dimitriadou et al., 2002; Hennig and Liao, 2013). Such decision involves the production of several clustering solutions with different number of clusters each and the selection of the clustering solution that best describes the data. From that is is made clear that this decision is closely related to the task of evaluating the clustering result we discussed above. Therefore, it is very common to utilise internal validation indices to serve as a mechanism to find the optimal number of clusters when this is not known beforehand (Dimitriadou et al., 2002). Usually when validation indices are used for this purpose they have to follow certain rules, like the elbow method (Thorndike, 1953).

However not everyone in literature agrees with these rules and there are examples of using the validation indices differently. In (Vendramin et al., 2010) the number of clusters that maximises *Calinski* is chosen instead of applying the elbow method (Dimitriadou et al., 2002) raising questions regarding the ability of these indices to determine the number of clusters in a clear and objective manner. This comes along with the controversial ability of validation indices to evaluate the clustering solution since the concepts of *Compactness* and *Separation* do not always hold in real world applications.

2.3.3.4 Difficulties of Traditional Clustering

All the above summarise a series of important decisions that have to be made when clustering is applied. The insufficient evidence of a superior clustering algorithm or a clear evaluation technique make it very difficult for someone to design the best clustering application for every given data specific problem. It is clear that the concept of optimal clustering remains related to the aim of clustering and to the domain it will be applied upon. Thus, the meta-clustering technique described in (Caruana et al., 2006) does not try to find the optimal clustering because such target is considered unrealistic but rather a collection of good clusterings organised in an hierarchical meta-clustering approach that will help the user detect the clustering that suits his needs.

This idea leads to a new concept of clustering. Since each algorithm and validation index suffers from certain assumptions and limitations while their performance is affected by the domain of application, a new method has been proposed that tries to identify clusters by finding the consensus of different clustering algorithms. This method, named consensus clustering or cluster ensembles share similar ideas with the meta-clustering approach.

2.3.4 Consensus Clustering

Consensus clustering formalizes the idea that combining different clusterings into a single representative would emphasize the common organization in the data (Goder and Filkov, 2008) and extract as much information as possible from the data (Soria and Garibaldi, 2010). Its idea is based on the assumption that patterns belonging to a natural cluster are likely to be in the same cluster in different data partitions. Since a large number of clustering algorithms exists but no single algorithm has been yet able to identify all of cluster shapes and structures in practice (Fred and Jain, 2005) and capture all the desired properties of the data (Caruana et al., 2006), a combined clustering using different clustering algorithms can neutralise the limitations and assumptions each one of them possesses (Fred and Jain, 2005) and therefore, improve the clustering robustness and quality (Duarte et al., 2010).

Consensus clustering has been shown to outperform single clustering algorithms if the *diversity* is ensured and comparable clusterings are acquired (Strehl and Ghosh, 2003) but also to demonstrate improved accuracy on a number of artificial and real world datasets (Topchy et al., 2004). Its most interesting characteristic though is its ability to detect novel cluster structures (Topchy et al., 2004) of various shapes, sizes and densities (Fred and Jain, 2005).

Apart from improving the quality of an ensemble of clusterings, consensus clustering has been used for the purposes of knowledge reuse and fusion (Strehl and Ghosh, 2003). Knowledge reuse refers to the combination of different classifications based on different features on the same data into a combined clustering while fusion refers to the combination of different clusterings on the same features but different subsets of the data in a distributed computing sense. Consensus clustering has been successfully applied in real world applications like gene expression (Monti et al., 2003) and (Filkov and Skiena, 2004) microarray data in order to find robust and representative clusters of similar proteins and genes and to group breast cancer patients based on the differences in the expression of histone markers (Soria and Garibaldi, 2010).

A typical structure of a consensus clustering solution consists of two steps (Fern and Brodley, 2004):

- Building a clustering ensemble
- Consensus function to create a consensus partition

The first step involves the utilisation of clustering algorithms to create an ensemble of clusterings on the same dataset. The obtained clusterings from the cluster ensemble should be large in number and as different as possible in order to provide the desired *diversity*. *Diversity* is needed in order for the consensus clustering solution to be able to explore a larger proportion of the solution space and avoid being stuck in a local optimum (Caruana et al., 2006). This idea is supported from several works in literature (Strehl and Ghosh, 2003; Hennig and Liao, 2013; Fern and Brodley, 2004) and some even favour *diversity* over *quality*, like (Topchy et al., 2004) where authors claim that the requirements for individual clustering algorithms can be relaxed in favour of weaker and inexpensive partition generation, since we can obtain a good combined clustering solution from a number of partitions generated by simple and weak clusterers which are efficient and computationally cheap, like K-means. In more detail, they prove that the consensus solution converges to a true underlying clustering solution as the number of partitions in the ensemble increases. However, in (Fern and Lin, 2008) authors claim that both *quality* and *diversity* are needed for a consensus clustering solution to achieve its purpose and that we might be able to achieve a better performance if we select a smaller ensemble of greater *quality*. An extensive comparison of consensus clustering solutions in (Kuncheva et al., 2006) revealed that both qualities are essential for the performance of consensus clustering further supporting this argument.

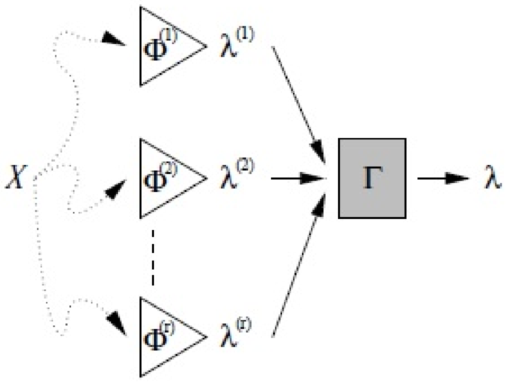


FIGURE 2.1: A consensus function (Γ) that combines clusterings (λ^i) from a variety of clusterers (Φ^i) performed on the same data (X) (Strehl and Ghosh, 2003)

After generating the cluster ensemble the clusters of different clusterings need to be aligned in order to assess their similarity with each other or with the consensus partition later on. This task is described as solving the correspondence problem and is identified as a demanding and difficult task for every consensus function (Topchy et al., 2004) whose accuracy depend on the correct relabeling of the matching problem.

Finally a consensus function has to be devised in order to combine all the generated clusterings into one consensus as it can be seen in Fig 2.1. As the goal of consensus clustering is to improve the quality and robustness of the clustering results, a consensus function is needed to map a set of given partitions on the data into a consensus partition that would be representative of the given clusterings and would share as much information as possible with the cluster ensemble (Strehl and Ghosh, 2003; Goder and Filkov, 2008). More formally in (Fred and Jain, 2005) the problem is reduced to finding an "optimal" data partition that should satisfy the following properties:

1. Consistency with the clustering ensemble
2. Robustness to small variations in clustering ensemble
3. Goodness of fit with ground truth information, if available

2.3.4.1 Consensus Clustering Solutions

The authors in (Strehl and Ghosh, 2003) propose three algorithms for obtaining the consensus partition from an ensemble of clusterings. These three heuristic algorithms try to maximise the similarity of the consensus partition with the cluster ensemble. The first one, named CSPA, creates a similarity matrix between the objects based on the co-occurence in the same cluster that resembles a bipartite graph. They use a graph clustering algorithm to cluster this table and obtain the consensus clusters. The second algorithm, HPGA tries to find the min-cut of the graph and the third proposed algorithm, MCLA constructs a meta-graph that represents the relationships between clusters and uses a graph clustering algorithm to obtain the consensus classes. The three presented approaches, construct a hypergraph depicting the relations between clusterings and then perform a clustering on this abstraction in order to find a consensus.

A similar solution is presented in (Fern and Brodley, 2004) where they construct a bipartite graph from a given cluster ensemble that models both the relationships between observations and the relationships between clusters. In contrast with the methods proposed in (Strehl and Ghosh, 2003) where each kind of relationship is treated separately from the three proposed algorithms, this method takes into account both relationships and reduces the problem to a graph partitioning problem that is solved efficiently. They argue that both types of relationships have to contribute in the production of consensus clusters as it is shown that focusing only on one type of relationship results in information loss.

In (Fred and Jain, 2005) they also adopt the same approach of taking into account the co-occurrence of pair of observations and they propose to use hierarchical clustering to cluster a co-association matrix. Their proposed method achieves better results than a graph partitioning algorithm that is adopted by the previous method.

In contrast with these methods, the consensus clustering framework proposed in (Soria and Garibaldi, 2010) focuses on creating a smaller cluster ensemble of good quality as this is guaranteed by the use of a series of validation indices. Then they check the agreement of the partitions in order to obtain the core classes. Their approach results in clear and good clusters but it focuses only on a subset of observations, that is the observations that co-occur in the clusters.

Finally, in (He et al., 2005) the authors emphasise the resemblance between the consensus clustering problem and clustering categorical data. That is because a cluster partition can be viewed as a categorical variable and therefore we can use a clustering for categorical data approach to obtain consensus clusters from a matrix of memberships of observations in cluster partitions.

All of the proposed methods work exclusively on the information given in the clusterings and they do not take into account the characteristics of the objects in the dimensional space focusing only on the co-associations between objects or clusters. In some cases (Strehl and Ghosh, 2003) the number of consensus clusters is required to be given as a parameter apriori. Moreover all these methods demand to solve the correspondence problem first and align the clusters of different partitions before they can be applied and with the only exception being the work in (Soria and Garibaldi, 2010) all the other consensus clustering solutions focus on building a big and diverse cluster ensemble disregarding the property of clusterings of high quality.

2.3.4.2 Consensus Clustering Evaluation

A consensus clustering solution can be evaluated on three levels according to (Duarte et al., 2010):

1. Original data representation
2. Clustering ensemble space
3. Learned pairwise similarity

The first level evaluates a consensus clustering solution in a similar way with any other clustering. It tries to assess the quality of the final grouping of observations provided by the consensus partition by the same validation criteria that define the goodness of a clustering result in traditional clustering approaches. The second level focuses on assessing the consistency of the final consensus partition with the initial cluster ensemble. The final partition must be as similar as possible with the initial partitions in order to be characterised as consensus. Typical proposed measurements, try to measure the level of agreement between the resulting consensus partition and initial partitions and include the Average Normalised Mutual Information (ANMI)(Strehl and Ghosh, 2003) between the consensus partition and the cluster ensemble, the Average Cluster Consistency (ACC) and the Rand Index (Rand, 1971) or Jaccard Coefficient. However ANMI is not reliable when the different partitions produce different number of clusters (Fred and Jain, 2005). Finally the third level measures the coherence between clustering solutions and the co-association matrix induced by the clustering ensemble and it evaluates the ability of the consensus solution to obtain the pairs of objects that tend to co-occur. It is a specialised case of evaluation assuming that the consensus clustering solution focuses on the similarities of pairs of observations.

Apart from these three levels of evaluation and assuming the desired properties of a consensus partitions as these are specified in (Fred and Jain, 2005), the Consensus Clustering solution can also evaluated upon its robustness and its accuracy when compared to the ground truth, if that is available.

2.3.5 Clustering socio-economic data

When it comes to real world applications and especially socio-economic data, clustering, either traditional or consensus, faces several problems. As any other real world dataset, it is contaminated with noise, missing values and outliers, suffers from high dimensionality, it contains mixed data, categorical together with numerical attributes, and there is a large imbalance in the expression of the numerical variables. The presence of categorical attributes is prohibitive for most of the clustering algorithms since they do not have inherent geometric properties (Zhang et al., 2000) and therefore euclidean distance fails to capture similarities between objects (Ahmad and Dey, 2007). Since euclidean distance is the cornerstone of these algorithms, numerical attributes with relatively larger values shape the clustering procedure, as they are able to produce significant differences in the calculation of distances, forcing clustering to rely exclusively on them disregarding attributes with smaller values. All these degrade the quality of the clustering and require a set of pre-processing steps as vital transformations of data before the process of clustering takes place.

Dealing with mixed data in literature we can find modifications of clustering algorithms that successfully incorporate mixed data. In (Huang, 1998) K-modes is introduced that extends K-means to cluster categorical data by using a simple matching dissimilarity measure for categorical objects that is modes instead of means and a frequency-based method to update modes in similar fashion K-means updates its centroids. Then K-prototypes is presented, which combines K-modes and K-means for dealing with mixed data while it preserves the efficiency and scalability of K-means. In (Ahmad and Dey, 2007) an extension of K-prototypes is proposed, suggesting a more reliable distance function and more representative cluster centers that are not defined only by the mode. Finally in (Hennig and Liao, 2013) a dissimilarity measure that is appropriate for both numerical and categorical is proposed. Gower distance was integrated into PAM, a k-medoids approach, to create the dissimilarity matrix it works with. However the problem with these proposals is that conventional validation criteria cannot be used to assess the quality of their results. Finally in (Zhang et al., 2000) another approach is presented that is similar to the correspondence analysis. The categorical dataset is viewed like a sum of tuples and the purpose of the approach focuses on assigning weights to each possible categorical value in an iterative process until it converges. The weight assignment that is calculated based on the co-occurrence of a tuple with other tuples, consists of a configuration whose orthogonality ensures the presence of negative and positive values. Then the objects are split into a negative and positive group. From this it is clear that this approach lacks the ability to produce more than two groups of objects.

2.4 Data Transformations

2.4.1 Importance of data transformations for clustering

Data transformations are essential prior to clustering in order to prepare the data for the clustering process by handling the inconsistencies the data may contain. For that reason a lot of methods and techniques have been developed for the purposes of preprocessing. For the socio-economic context the necessity of unit standardisation and log transformations is emphasised in (Hennig and Liao, 2013) as it deals successfully with noise, outliers and the asymmetry of expression of numerical attributes. The importance of standardisation prior to clustering is further supported by (Caruana et al., 2006; Duarte et al., 2010). In addition to this the reduction of dimensionality is considered an essential transformation (Vyas and Kumaranayake, 2006) where PCA (Hotelling, 1933) is utilised for this reason and in (Cortinovis et al., 1993) where attribute elimination strategies are adopted.

In addition to this, data transformations can assist in handling mixed data, a very problematic issue for clustering. Two strategies that can be developed to deal with mixed datasets according to (Ahmad and Dey, 2007) are to either transform the categorical data into numerical and apply conventional clustering or discretise the numeric attributes and apply categorical data clustering. The second strategy results in loss of information and therefore degrades the quality of the clustering results, whereas the first strategy demands a meaningful transformation of the categorical attributes into numerical where the distance between objects can reflect the similarity between data points.

2.4.2 Homogeneity Analysis

This transformation can be carried away by Homogeneity Analysis (Homals) (De Leeuw and Mair, 2009, 2007), a non-linear multivariate analysis with low computational costs that has the ability to handle vast amounts of data. In its strict sense, Homogeneity Analysis can be viewed as a descriptive tool to analyse categorical data, very similar correspondence analysis (Lebart and Salem, 1988). But instead of using SVD (Singular Value Decomposition), it relies on optimising a loss function, it is faster and it capitalises sparseness in data, successfully dealing with missing data.

Given a dataset with categorical variables expressing the information across objects, Homals tries to find a low-dimensional space in which objects and categories are positioned in such a way that as much information as possible is retained from the original data. As explained in (Michailidis and de Leeuw, 1998)the goal becomes to construct a low-dimensional joint map of objects and categories in Euclidean space. The choice of low dimensionality is because the map can be plotted and thus can be interpreted and understood whereas the choice of Euclidean space stems from its nice properties (projections, triangle inequality) and our familiarity with Euclidean geometry.

The desired properties in this low dimensional joint space dictate that the category points serve as centers of gravity of the object points that share the same category. The larger the spread between category points the better a variable discriminates and thus, it indicates how much a variable contributes to relative loss. The distance between two object scores is related to the"similarity" between their response patterns. A"perfectly homogenous" solution would imply that all object points coincide with their category points (De Leeuw and Mair, 2009).

Homogeneity Analysis tries to achieve the maximum *homogeneity* by quantifying or rescaling the objects and the variables in a joint p-dimensional space where similar objects and categories with similar content will be placed close together. It does so by truing to minimise the departure from homogeneity. This departure is measured by the sum of squares of the distances between the object scores and their corresponding categories. Thus the problem is reduced to minimize these distances in the p-dimensional space. Certain constraints guarantee that the representation will be centered and that the object scores will be orthogonal.

The geometrical properties of this joint representation created by Homogeneity Analysis are ideal for clustering purposes. As the distance between object points reflects the similarity of these two object points, a clustering algorithm can uncover similar groups of objects when it performed upon the transformed data.

In its broad sense, Homogeneity Analysis can incorporate different scale levels keeping the order and preserving the distances. This way it can fit more parsimonious models as it can handle ordinal and numerical variables increasing its functionality. With the ability to handle all types of data Homogeneity Analysis can serve as *optimal scaling* and *dimensionality reduction* technique. Optimal scaling refers to the procedure which transforms the observed response categories according to some specified criterion (loss function) as part of an optimisation process that can find the optimal summarisation of the data. The fact that this functionality can extend to categorical data gives the opportunity to Homogeneity Analysis to serve as a non-linear PCA or non-linear canonical correlation analysis.

2.4.3 Exploratory Factor Analysis

Another interesting pre-processing technique that tries to represent the underlying structure of the data into linear constructs that are called factors is Exploratory Factor Analysis (EFA). As defined more formally in (Fabrigar and Wegener, 2011) EFA refers to a set of statistical procedures designed to determine the number of distinct constructs, usually refered as factors, that are needed to account for the pattern of correlations among a set of measures. Factors are considered as unobservable constructs that exert linear influences on one or more measured variables on the dataset (Ferguson and Cox, 1993).

In order to achieve that, Exploratory Factor Analysis aims to understand and represent the structure of correlations among observed scores on a set of variables. Therefore the goal of Factor Analysis is to arrive at a relatively parsimonious representation of the structure that summarises the original data and reveals the associations among the variables. It can be seen as a process of clustering the correlations matrix of the variables into factors. Therefore Exploratory Factor Analysis can serve as *dimenionality reduction* technique that reveals potential associations in the data.

(Fabrigar et al., 1999) summarises the five basic steps that are need in order to apply EFA on a dataset:

1. Choose the variables of the dataset.
2. Decide whether EFA is fitting procedure.
3. Determing the number of factors.
4. Decide which rotation is suitable for representing the factors.

Usually the first and the second steps involve techniques to decide whether the data are suitable for EFA. For that reason *Stability Coefficient*, which indicates how stable a factor structure is relative to the population drawn, can determine if there are enough observations and the *KMO test*, which indicates whether the associations between the variables in the correlation matrix can be accounted for by a smaller set of factors, checks the appropriateness of the correlation matrix. The third step involves a decision upon the algorithm used for EFA and the fourth step involves a number of techniques to determine the best number of factors that can represent the data. The criteria for this decision is the statistical utility, interpretability and stability or robustness. Statistical utility formalises a logic of identifying the suitable number of factors based on the ideas that one more factor would not explain more while one factor less cannot explain enough.

Scree plot and parallel analysis are the most common techniques used for this purpose. Interpretability makes it crucial to avoid situations where none of the resulting factors cannot express substantially more than one variables of the data, which is referred as *underfactoring* or situations where factors explain too many variables of the data, which is referred as *overfactoring.* Finally the representation of the factors in the final decision of the EFA is important to find out if the resulting factors would be interdependent to each other or not. That will affect the interpretation of the factors in the end.

2.4.4 Towards Behavioural Data

The most interesting fact of both pre-processing techniques, discussed in this section, is their ability to represent the original dataset in a new dimensional space where associations and patterns in the data are more clear. Observing the underlying structure in this represented space can lead into squeezing out behavioural elements hidden in the original data and forming a new behavioural dataset that will express well defined patterns of behaviour and will enable Data Mining methods to successfully mine behaviours from complex data.

2.5 Understanding Consumer Indebtedness

Interpreting our findings from our search in bibliography, a way for developing *Behavioural Extraction* is highlighted. On one hand data transformations can deal with a number of inconsistencies real world data may contain, but more importantly they possess the ability to reveal behavioural elements hidden in the data moving towards behavioural data. On the other hand clustering methods and especially consensus clustering methods that possess certain advantages when applied to real world applications, they can discover behavioural profiles of consumers and thus provide a deeper understanding of the “nature” of consumer indebtedness by validating the theoretical classifications of consumers discussed throughout the literature or by discovering novel groups of consumers based on patterns in the data that were not explored before.

Therefore in our work we apply a series of pre-processing techniques on socio-economic data in order to find out a method to extract psychological features of consumers, answering the first research question of our research, and we employ clustering approaches in order to devise behavioural profiles of consumers in an effort to provide answers to the second research question. The resulting behavioural data and profiles need to reveal insightful patterns of consumer behaviour in order to assess the impact of these techniques on Consumer Debt Analysis answering the fourth research question as it is defined in Chapter 1 in subsection 1.4.1.

While understanding the data that describe consumer indebtedness is considered an important task, it is of greater significance to try to provide answers to the three important research questions stated in section 2.2 earlier in this chapter (Wang et al., 2011). These questions can easily be translated into data mining tasks and therefore we can utilise supervised learning approaches to build strong and reliable models in order to find meaningful answers. Furthermore we can take into account all the extracted behavioural information from unsupervised learning approaches and include it in the supervised learning process in order to improve the performance of our modelling. For this reason in the next section we discuss ways we can implement strong and accurate models for Consumer Debt Analysis.

2.6 Supervised Learning for modelling Consumer Indebtedness

2.6.1 Limitations of Statistical Modelling

So far in Consumer Debt Analysis answering the three fundamental questions was mainly carried out by traditional statistical models like linear regression, which has the ability to reveal linear associations between variables. However, as common as the utilisation of these models in the field of Economics might be, so is their limited ability to deal with characteristics that data from real world applications possess. Their difficulty to handle non-linearity in the data makes them unable to solve non-linear classification problems (Refenes et al., 1994), while the colinearity between the independent variables can lead to incorrect identifications of most predictors (Sousa et al., 2007). In (Gromping, 2009), it is argued that linear regression requires explicit modelling of non-linearities and that it is robust only when the number of instances is much larger than the number of predictors. These limitations make them inappropriate to model successfully consumer indebtedness since socio-economic datasets exhibit strong non-linearity among several other inconsistencies.

Apart from its limited predictive ability, the way that statistical modelling is applied in practice as shown in literature raises questions regarding the validity of the relationships uncovered by the proposed models. A significant amount of the work is summarised in (Stone and Maury, 2006) where they also provide a model for separating debtors from non-debtors. However, their suggested logit model suffers from a low R^2 (33%). In a similar way, a linear regression model built for estimating the outstanding credit card balance in (Kim and DeVaney, 2001) exhibits 30% R^2. The works of (Gathergood, 2012; Ottaviani and Vandone, 2011) present models that take into account psychological factors as predictors and exhibit even lower R^2 in their probit models (around 10% to 20%) together with the hierarchical regression model of financial planning presented in (Noone et al., 2012), the structural equation modelling in (Gardharsdottir and Dittmar, 2012) and the linear regression model in (Wang et al., 2011). Surprisingly enough the linear regression model presented in (Livingstone and Lunt, 1992) achieves a remarkable 66% R^2 but as it is explained in (Stone and Maury, 2006), this big proportion of explained variance, is due to the small number of respondents.

But the limited number of observations is considered to be the rule and not the exception in these case studies. In fact most of the models are built based on a very small number of observations that is usually below 500. The only exception is the logistic regression separating debtors from non-debtors in (Webley and Nyhus, 2001) built upon 1873 subjects. Based on this fact, we are unsure whether to regard these findings as reliable since the small number of instances cannot be considered representative enough and since the study of of complex socio-economic sustems must take into account their sometimes counter-intuitive behaviours which imply that linear, conventional and straight-forward modelling attempts will usually be considered inappropriate (Helbing and Balietti, 2011). It might be more suitable to abandon weak models that fail to explain the variance that exists in the data and instead seek stronger models that have the ability to mine bigger and complex real world datasets maintaining their performance stable, in order to provide more accurate and reliable answers to the fundamental research questions of Consumer Debt Analysis.

2.6.2 Potential of Data Mining models

As the need to develop fairly accurate quantitative prediction models becomes apparent (Atiya, 2001), the field of economics can benefit from the variety of techniques and models Data Mining has to offer. Accurate and powerful models like Random Forests and Neural Networks that can handle non-linearities in the data (Refenes et al., 1994; Gromping, 2009) can pose as strong candidates to analyse real world data like socio-economic data and provide meaningful answers to the research questions of Consumer Debt Analysis. Since both models are suitable for both regression and classification, they can handle the task of separating debtors from non debtors, which can be rephrased as a classification task and the tasks of predicting the level of debt and debt repayment, which can be rephrased as classification/regression tasks. Therefore, these two models with all the advantages they possess can be used to create a clear, complete and conceptual model of consumer indebtedness that has not yet emerged, despite the fact that many factors influencing consumer debt have been proposed in literature (Livingstone and Lunt, 1992). On the contrary, all the efforts so far have focused on a few or a subset of demographic, economic, psychological and situational factors limiting their predictive ability and their generalizability (Stone and Maury, 2006). As defined in (Stone and Maury, 2006) the purpose of this financial indebtedness model would be to accurately identify individuals who are at risk for developing personal financial management problems.

Both models have started to be used in field of economics replacing traditional statistical modelling. More accuretely, Random Forests have been utilised in marketing applications like (Ghose and Ipeirotis, 2011) where a model measuring the impact of the reviews of products in sales and perceived usefulness was constructed. On the other hand Neural Networks have been used in stock performance modelling (Nicholas Refenes et al., 1994) and for credit risk assessment (Atiya, 2001) where banks need to predict the possibility of default of a potential counterpart before they extend a loan. For credit risk assessment any improvement in default prediction accuracy not only can lead to increased savings and assist in estimating a fair value of interest rate but also can help in accurately assessing the credit risk of bank loan portfolios. For that reason, Neural Neuworks seem to replace linear regression in default prediction (Sousa et al., 2007) as part of an ongoing trend that maximises the utilisation of Neural Networks for credit risk assessment. A characteristic example of this emerging trend is the Moody's *Public Firm and Risk* model that is now based on Neural Networks (Atiya, 2001).

Another model that is very popular in Machine Learning and Data Mining applications is the Support Vector Machines (SVM's) (Cortes and Vapnik, 1995). Support Vector Machines have gained a momentum recently because of their strong mathematical background which guarantees that the solution will reach a global optimum. They exhibit good generelisation

2.6.2.1 Random Forests

Random forests (Breiman, 2001) are an ensemble learning method that operate by constructing a multitude of decision trees on random samples of the training data based on the ideas of *bootstrap* sampling, *bagging* and the *random selection of features*. They have gained a lot of popularity recently because of their ability to handle large number of variables with relatively small number of instances and they provide a mechanism to assess variable importance (Gromping, 2009; Segal, 2004) in contrast with Neural Networks. They allow for non-linearities to be learned from the data without any need to be explicitly modelled (Gromping, 2009) as linear regression requires and manage to achieve exceptional performance (Segal, 2004). In fact, research has shown that a large number of trees can be particularly important when the interest is in diagnostic quantities like variable importance (Gromping, 2009), while their averaging is responsible for variance reduction which is enhanced by the reduction in correlation between the averaged results of the trees caused by the injection of randomness. Hence, random forests correct for decision trees' habit of *overfitting* to the training set.

2.6.2.2 Neural Networks

The development of artificial neural networks (ANN's) arose from the attempt to simulate biological nervous systems by combining many simple computing elements (neurons) into a highly interconnected system and hoping that complex phenomena as "intelligence" would emerge as the result of self-organisation or learning (Sarle, 1994). Neural Networks are capable of processing vast amounts of data making accurate predictions. In fact they can serve as universal approximators as they can approximate any function to any desired degree of accuracy when given enough hidden neurons and enough data, a fact that is confirmed in (Hornik et al., 1989).

Their strength lies on their ability to handle non-linearities in the data (Sarle, 1994; Refenes et al., 1994), to allow for extrapolation (Sousa et al., 2007) that makes them suitable for generalisation (Sousa et al., 2007; Refenes et al., 1994) and to deal with the problem of *structural ability* which refers to the situation when the relationship between dependent and independent variables changes over time. In addition to this, a very interesting ability they possess is the ability to fully parametrise the topology of the network introducing a concept of logical structure among the neurons that compose the network. This gives the ability to design a network that will incorporate the knowledge extracted from the Behavioural Extraction phase into a Behavioural Modelling suitable for the purposes of our analysis. The same idea has been exploited in (Shifei et al., 2011) where factor analysis is utilised in order to define the topology of the network and although their result has shown not to actually improve the precision of the existing neural network, it manages to speed up the convergence of the algorithm.

In contrast with Linear Regression and Random Forests, their biggest disadvantage stems from their inability to provide transparent results. The mechanisms that occur inside the Neural Network are hidden and ignored and thus they are usually characterised as "black boxes" (Gevrey et al., 2003; Harrington and Wan, 1998). Because of this difficulty, the practical use of Neural Networks is limited in real world applications (Harrington and Wan, 1998) and it is prohibited in the social sciences where the interpretation of the models is very significant. But in literature there have been a series of techniques proposed to assess the importance of input variables in Neural Networks. In (Harrington and Wan, 1998) Sensitivity Analysis offers an approach to identify important variables in Neural Networks that can be achieved by perturbing the input variables and measuring the impact on the outcome variable. In (Gevrey et al., 2003) a comparison of different techniques measuring the contribution of input variables is provided. Seven techniques of contribution analysis are examined and analysed. From them, the Partial Derivatives method and the Profile method offer the most complete results.

2.6.2.3 Support Vector Machines

Another model that is very popular in Machine Learning and Data Mining applications is the Support Vector Machines (SVM's) (Cortes and Vapnik, 1995). Support Vector Machines have gained a momentum recently because of their strong mathematical background which guarantees that the solution will reach a global optimum. They embody the Structural Risk Minimisation (SRM) principle which tries to minimise an upper bound on the expected risk in contrast with the majority of learning algorithms that try to minimise the error on the training data. This grants them the ability to generalise in unseen data and avoid data over-fitting which is a valuable property in real world applications (Dibike et al., 2000). Also by performing the Kernel "trick" they SVM's can extend to handle non-linearities in the data.

However, in this work we are going to use the first two models. That is because the parameters of the Support Vector Machines are not interpretable as Random Forests can be and they don't exhibit the same level of flexibility Neural Networks do when building the model. In addition to this, as show in the comparative review of (Meyer et al., 2003) SVM's were not able to outperform Random Forests and Neural Networks in many cases despite the fact that they yielded a very good performance.

2.6.3 The special case of class imbalance

When it comes to classification it requires extra attention to build a classifier on an imbalanced dataset. A dataset is considered imbalanced if the classification categories are not equally represented (Chawla, 2005). Class imbalance problem has become an important issue in Data Mining (Longadge and Dongre, 2013) as it may produce an important deterioration of the performance of the model (Mollineda et al., 2007). Classification in Random forests for example has been shown to perform poorly in instances of class imbalance (Segal, 2004). The reason behind this deterioration in performance stems from the fact that under class imbalance occasions the minority classes do not have enough examples to train the classifier properly in order to discriminate them against the majority class. Therefore the accuracy for those classes remains poor.

The solutions proposed in literature fall into two main categories, to fix the class balance or modify the learning procedure in algorithmic level. When fixing the balance one can choose to oversample the minority classes or to undersample the majority class in order to create evenly represented classes. The utilisation of resampling or focused resampling that focuses on selecting objects close to the decision region since these are more likely to be misscassified, like SMOTE (Chawla, 2005) and injecting synthetic instances in the data, are among the solutions presented under oversampling, but they may may cause overfitting. Undersampling on the other hand involves techniques that focus on removing the most redundant objects, but they can result in loss of information (Chawla, 2005; Longadge and Dongre, 2013). Finally, at the algorithmic level a cost sensitive learning that penalises missclassification has been proposed together with building one class classifiers for each class or using classifier ensembles (Mollineda et al., 2007). However, all of these solutions work mainly for two-class classification tasks (Mollineda et al., 2007; Chawla, 2005).

Evaluating classification performance upon imbalanced classes requires additional care since predictive accuracy, a popular choice for evaluating performance of a classifier might not be appropriate when the data is imbalanced and the costs of different errors vary markedly (Chawla, 2005). The main goal for learning in this case is to improve *recall* without hurting the *precision*. That is assessed by the *F-measure* or by the *ROC curve* (Powers, 2011).

2.6.4 Modelling Consumer Indebtedness within a Data Mining Framework

All the findings in literature sketch the beneficial role Data Mining can play in modelling consumer indebtedness. Strong and powerful models possess the ability to replace traditional statistical models and improve the accuracy of level of debt prediction. At the same time the reliable and transparent models can provide reliable insight regarding important research questions of Consumer Debt Analysis.

Drawing upon this great potential Data Mining seems to offer, in this work we try to develop a Data Mining framework of consumer indebtedness in order to understand if it has the potential to replace traditional statistical methods, answering the third research question of our work. We then use the same framework to provide reliable answers regarding the beneficial nature of behavioural data and profiles and whether they improve the modelling of consumer indebtedness. This will provide an answer to our fifth research question as defined in the Chapter 1 in the subsection 1.4.1.

2.7 Conclusions

Our research in the literature has revealed the potential of Data Mining methods to analyse problems of complex “nature” like consumer indebtedness. Powerful and accurate predictive models pose as strong candidates to replace classical statistical modelling which is dominating the field and provide meaningful and reliable answers to the fundamental questions that have been the focus of the research so far. Unlike classical statistical modelling Data Mining offers models that can successfully deal with the complexities of real world data and they exhibit a great flexibility to adapt to more complicated and sensitive tasks in order to successfully mine behaviours of consumers. As consumer indebtedness is shown to be a phenomenon of complex “nature” and with many disctinct facets, this flexibility provides the means to incorporate extracted knowledge from the behavioural profiling and the behavioural transformations, forming a complete Data Mining framework to analyse consumer indebtedness.

Chapter 3

Methods and Techniques

3.1 Introduction

In this chapter we present the methods and techniques we use to develop a complete Data Mining framework in order to analyse consumer indebtedness. More accurately, we present a series of data transformations that transform the original socio-economic data into behavioural data together with the clustering algorithms for the purposes of behavioural profiling. These techniques implement the Behavioural Extraction phase of our framework. Finally we study the technical aspects of classification/regression algorithms from Data Mining that implement the Behavioural Modelling phase of our framework and provide meaningful answers to the fundamental research questions of Consumer Debt Analysis.

3.2 Data Transformations

3.2.1 Data Preparation

Data transformations are an essential step of every Data Mining procedure as they deal with the inconsistencies of the data. This prepares the data for the later stages of the Knowledge Discovery process as seen in Fig 1.1 in Chapter 1. The two pre-processing techniques presented here, namely Homogeneity Analysis (Homals) and Factor Analysis, are very powerful transformations that not only can they serve as summarisation and dimensionality reduction techniques but they also have the ability to extract behavioural elements hidden in the data and represent them clearly in a new behavioural space.

3.2.2 Homogeneity Analysis

As shown in Chapter 2 Homogeneity Analysis creates a low dimensional space to represent the similarities among objects and the similarities between categories. Given a dataset of n objects and m categorical variables, it tries to create a representation of the object scores and the categories quantification (levels of the categorical variables) in a joint p-dimensional space (p¡¡m) based on the criterion of minimising the departure from homogeneity which is expressed by the loss function:

$$\sigma(X; Y_1, Y_2......, Y_m) = \frac{1}{m}\sum_{j=1}^{m} tr(X - G_jY_j)M_j(X - G_jY_j) \tag{3.1}$$

where X is a $(n \times p)$ matrix containing the object scores, Y_j is a $(K_j \times p)$ matrix containing the categories quantification of the j_{th} categorical variable, G_j is a $(n \times K_j)$ indicator matrix with ones in the cells where the objects have the corresponding category level and zeros in all the rest of cells, K_j is the number of the categorical levels of the j_{th} categorical variable, M_j is a $(K_j \times K_j)$ diagonal matrix with the row sums of G_j representing the missing values and tr represents the trace function of linear algebra that sums the elements of the diagonal.

As it can be understood the loss function measures the sum of squared distances between the object points and their corresponding categories and therefore it measures the departure from homogeneity. Homegeinity describes the optimal state of the representation, where all object points coincide with their category points it is stated by the following definitions of perfection:

1. Y_j are perfectly *homogeneous* if $G_1Y_1 = G_2Y_2 = ... = G_mY_m$
2. X is perfectly *discriminated* if $X = P_1X = P_2X = ... = P_mX$, where P_j is the orthogonal projector defined in $P_j = G_jD_j^{-1}G_j^{-1}$ where D_j is a $(K_j \times K_j)$ diagonal matrix with the relative frequencies of each level in the diagonal.
3. X and Y_j are perfectly *consistent* if $X = G_1Y_1 = G_2Y_2 = ... = G_mY_m$

The loss function sits in the heart of Homogeneity Analysis and is subject to the following constraints in order to avoid the trivial solution corresponding to X=0 and $Y_j = 0$.

$$u^{'}M_{\bullet}X = 0 \tag{3.2}$$

that centers the graph plot around the origin and

$$X^{'}M_{\bullet}X = I \tag{3.3}$$

that makes the columns of the object score matrix orthogonal.

In order to achieve the maximum homogeneity Homals utilises the Alternating Least Squares (ALS) algorithm to minimise simultaneously over the X's and Y_j's and therefore minimise the loss function. It is an iterative algorithm that each iteration consists of the three steps:

1. Randomise objects scores X
2. Update categories quantification $Y_j = D_j^{-1}G_j^{'}X$, where D_j is a $K_j \times K_j$ diagonal matrix containing the relative frequencies of the categories of variable j on its diagonal.
3. Update object scores $\tilde{X} = M_{\star}^{-1} \sum_{j=1}^{m} G_j Y_j$, where $M_{\star}$ is the average of matrix M
4. Normalise object scores X
5. repeat steps 2 to 4 until it converges

The second step of the algorithm expresses the *first centroid principle*, according to which a category quantification is in the centroid of the object scores that belong to it. The third step of the algorithm shows that an object score is the average of quantifications of the categories it belongs to. Hence, this solutions guarantees that the resulting representation places objects close to the categories they fall in and categories close to the objects belonging in them. The fourth step ensures that the normalisation constraints 3.2 and 3.3.

The resulting representation possesses some very interesting properties that are very useful for interpretation:

- The distance between two objects indicates their similarity.
- A variable discriminates better if its categories are further apart.
- Objects with identical profiles will receive identical object scores.
- Category points with high marginal frequencies will tend to locate closer to the origin (center of the representation).
- Objects with a profile similar to the average profile will tend to locate closer to the origin.

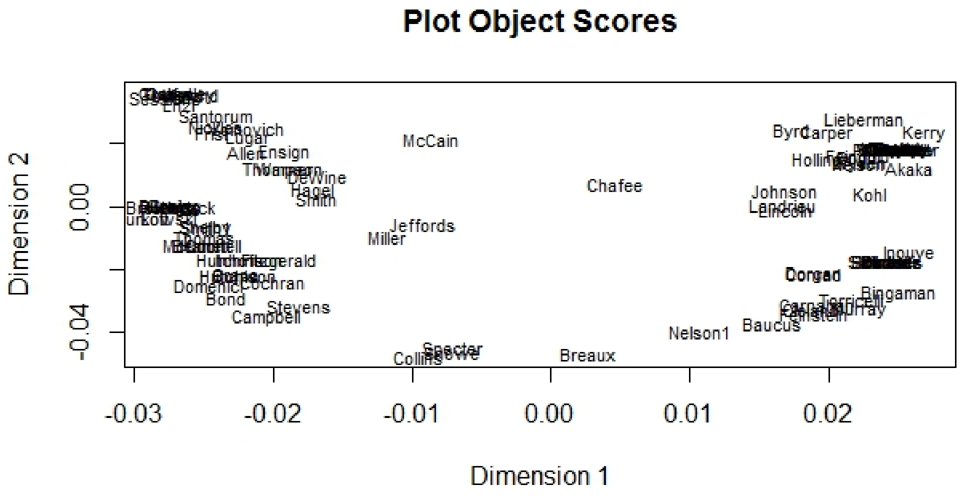


Figure 3.1: An object plot of senators based on their votes on twenty issues

- Categories of binary variables are placed on a straight line through the origin and their distance is determined by the marginal frequency.
- The category quantifications of each variable have a weighted sum over categories equal to zero.

In Fig. 3.1 you can see an example of the resulting representation when Homogeneity Analysis is applied on dataset containing the votes of senators on twenty different topics. The example is analysed properly in (De Leeuw and Mair, 2009) but here we can demonstrate the power of Homals to place similar objects together. The objects in this case are U.S senators and they are split in two groups as we can see from the plot. The left part represents the Republican senators and the right part the Democratic senators. Assuming that senators of the same party have similar voting patterns, the Homogeneity Analysis manages to identify these patterns and represent them in clear and interpretable way.

This summarises the functionality of Homals under its strict definitions. Possible extensions of the described functionality to numerical and ordinal variables can be found here (De Leeuw and Mair, 2007).

In the R environment Homals has been implemented in the “homals” package (De Leeuw and Mair, 2007) and is available for download in the CRAN repository: `https://cran.r-project.org/web/packages/homals/index.html`

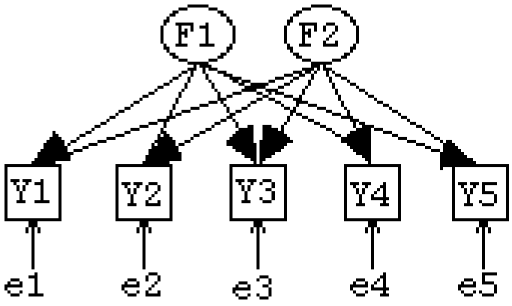


Figure 3.2: A diagram of Factor Analysis model

3.2.3 Exploratory Factor Analysis

Factor Analysis is a generic term for a family of statistical techniques concerned with the reduction of a set of observable variables in terms of a small number of latent factors. The latent factors exert linear influences on the variables and therefore the latent factors determine the values of the observed variables. As a result each observed variable can be expressed as a weighted composite of a set of latent variables. This can be seen clearly in Fig 3.2 that depicts a factor model. There, F1 and F2 are two common factors, Y1, Y2, Y3, Y4, and Y5 are observed variables and e1, e2, e3, e4, and e5 represent residuals or unique factors, which are assumed to be uncorrelated with each other. Thus according to this model every variable is a result of a linear combination between the common factors and a unique factor.

In order to find the latent factors that form the underlying structure of the data, Factor Analysis tries to appoint the appropriate values to the loadings of the factors in such a way so that their linear combination can approximate the correlation matrix of the measured variables. More formally this is expressed in the following equation:

$$P = \Lambda\Phi\Lambda^T + D_\psi \tag{3.4}$$

where P is the the correlation matrix of the observed variables, Λ is the factor loading matrix, Φ is the correlation matrix among common factors, which is usually equal to the identity matrix I as orthogonality is generally assumed, and D_ψ is the covariance matrix of the unique factors, which is usually a diagonal matrix with the variance of the unique factors on the diagonal when orthogonality is assumed again.

The most common algorithm for Factor Analysis tries to estimate the reduced correlation matrix $P - D_\psi$ in an iterative manner that tries to minimise the residual sum of squared differences. The resulting factors are ordered by the proportion of variance they explain which quarantees the uniqueness of the solution.

In Explanatory Factor Analysis, which makes no "a priori" assumptions about relationships among factors, it is important to determine the number of factors that best describe the relationships in the data. Two of the most common techniques for finding the number of factors in applied EFA are the scree test and parallel analysis (Fabrigar et al., 1999). In the Scree test the eigenvalues of the correlation matrix of the variables are plotted in descending order. The number of optimal factors is then chosen by the number of eigenvalues that precede the last substantial drop in the graph. In Parallel Analysis the eigenvalues of the same correlation matrix are compared to eigenvalues of the correlation matrices of randomly generated datasets with same size as the original. The number of factors is chosen by the number of eigenvalues that is bigger than the number of eigenvalues of random data.

3.3 Clustering

3.3.1 Implementing Behavioural Profiling

Consensus Clustering has been proved to be able to handle the complexities of real world data overcoming the limitations of traditional clustering approaches. For this reason it is used in our framework for the purposes of behavioural profiling. Thus, in this section we present a complete and sophisticated consensus clustering framework that finds representative and robust clusters. We then improve this framework by proposing a novel method to obtain the consensus clusters serving as the consensus function of this framework.

3.3.2 A framework to elucidate core classes in a dataset

This complete and sophisticated clustering framework, defined in (Soria and Garibaldi, 2010), formulates a consensus clustering solution that emphasises building a cluster ensemble that is characterised by quality and diversity complying perfectly with the findings of (Fern and Lin, 2008; Kuncheva et al., 2006) and then uses the agreement of the different clustering results to define the consensus clusters, which in this case they are named as core classes. In more detail the framework consists of five steps:

1. **Data pre-processing:** In the first step, if necessary the data is cleaned, normalised and inconsistencies are fixed.

2. **Clustering:** Two clustering methods are utilised, K-means (KM) and Partitioning Around Medoids (PAM)(Kaufman and Rousseeuw, 1987). These techniques differ from each other as they have different measurements to define proximity of the data instances to establish the clusters. This means that they group the data considering different characteristics present in the data. They were chosen as they are among the most widely used clustering methods in data mining.

3. **Determining the number of clusters:** In this step, validity indices are applied to clustering results. These indices indicate the appropriate number of groups to consider in the analysis.

4. **Data Visualisation:** Graphs like box plots and biplots are employed in order to obtain a general characterisation of the clusters obtained.

5. **Consensus:** In this step, the clusters found by the different techniques are aligned and the core classes are established on those samples assigned to the same group by distinct techniques, while those that do not co-occur in the same cluster are considered unclassified.

After the first essential step of pre-processing, the framework ensures the desired diversity in the second step by obtaining different clusterings from K-means and PAM while the third one guarantees the quality of the cluster ensemble as six validation indices, namely Calinski and Harabasz (Caliński and Harabasz, 1974), Hartigan (Hartigan, 1975), Scott and Symons (Scott and Symons, 1971), Marriot (Marriott, 1971), traceW and Friedman (Friedman and Rubin, 1967), are utilised in order to find the optimal number of clusters for each clustering algorithm, iterating them for different number of clusters each time and following the rules dictated in (Dimitriadou et al., 2002). According to these rules the number of cluster for each index is selected if it produces a much better value from the index it produces for one cluster less and if it is not much worse than the index it produces for one more cluster. In general, this logic is very similar to the Elbow method that finds the optimal number of factors in the case of Factor Analysis. In case of disagreement among the indices the groupings are being ranked for each index and the best one is chosen based on the minimum sum of ranks.

The level of agreement between the algorithms can serve as validation criterion of the quality of the clustering results. The bigger the agreement between the two clustering algorithms the bigger is the chance that these algorithms identified true patterns within the data. This stems from the fact that that patterns belonging to a natural cluster are likely to be in the same cluster in different data partitions (Duarte et al., 2010). The level of agreement can be measured by calculating the Cohen's kappa (Cohen et al., 1960) between the different clusterings. The Cohen's kappa is defined as:

$$\frac{p_0 - p_c}{1 - p_c} \tag{3.5}$$

where p_0 is the observed proportion of agreement and p_c is the proportion of agreement expected by chance. Kappa takes negative values when there is less observed agreement than is expected by chance, zero when observed agreement can be exactly accounted for by chance and one when there is complete agreement.

In addition to this, we are going to use the Silhouette criterion (Rousseeuw, 1987) in order to check how well separated, distinct and compact the clustering of the two algorithms are. The Silhouette coefficient of an object o is defined as:

$$s(o) = \frac{b(o) - a(o)}{max\{a(o), b(o)\}} \tag{3.6}$$

where a(o) is the average distance between o and all other objects in the cluster to which o belongs and b(o) is the minimum average distance from o to all clusters to which does not belong. The silhouette value of a clustering is calculated as the average silhouette of all objects and it takes value from -1 to 1, with 1 being the best case. Therefore it is a standardised metric for measuring the goodness of clustering, a fact that makes it appropriate for comparing the results of clusterings on different datasets.

Apart from determining the number of clusters of the respective clustering solutions based on the values of validation metrics, characterisation of clusters is also a very significant task. Characterisation of clusters can reveal the desired patterns that can enhance the Knowledge Discovery on the data, establishing the importance of the clusters for the domain of application. Visual inspection can be utilised as the framework dictated, producing a low dimensionality plot to depict the clusters or alternative ANOVA techniques to observe the differences in the numerical values. However considering the big size of the data some of these techniques can potentially lead to wrong conclusions. A pattern that will characterise a cluster should hold two important qualities. First it should be common in the cluster being expressed by the majority of instances and secondly it should be uncommon in the whole population. For this reason we propose a different technique to characterise the returned clusters that we apply in the fifth chapter.

The numerical variables are discretised in three classes (Low, Average, High) depending on which side of the interquartile range they fall. So the Low class describes the bottom 25%, the High Class the top 25% and the Average Class the middle 50%. Then after this transformation is being applied, each cluster will be expressed by the mode of each numerical variable. As the Average class is the most common class, a potential occurrence of the Low or the High class would be significant enough to characterise the cluster. For the categorical clusters we borrow the rules of Confidence and Lift from association rule mining (Agrawal et al., 1993) in order to assess the impact of the clustering on the initial distribution of categorical levels. Clustering is seen as an event and we measure the Lift and the Confidence of the rule (Cl → L) where Cl is the cluster examined and L the level of the categorical variable under examination. The Lift of a rule (Cl → L) is defined as

$$P(Cl \rightarrow L) = \frac{P(Cl \cap L}{P(Cl) * P(L)} \tag{3.7}$$

and it measures how bigger or smaller the probability of the level L is after the event of clustering Cl take place. High values can be used to characterise clusters only if the level of the categorical variables is possessed by the majority of the members of the cluster. On the other hand the Confidence of the same rule is defined as

$$P(Cl \rightarrow L) = \frac{P(Cl \cap L}{P(Cl)} \tag{3.8}$$

and shows how many instances of the initial population that express the categorical level L fall into the cluster CL. High values of confidence can reveal patterns that are expressed exclusively on this cluster even if they are not representative of the majority of the members of the clusters.

Finally the fifth step defines the core classes of the consensus, modelling strictly the agreement of the cluster ensemble.

3.3.3 Our proposed Homals Consensus

The above approach results in clear and good clusters but it focuses only on a subset of observations, meaning the ones that co-occur in the same clusters. That means that it ignores all the objects that fall into different clusters, a fact that may lead to a significant loss of information. It is important for the consensus clustering to be able to capture the information that might exist in these cases since if this missclassification occurs systematically, this might lead to the creation of a new cluster that describes these controversial objects.

Therefore it is important to improve this consensus clustering framework by incorporating a different consensus function that can lead to more representative clustering that takes into account all the objects even the ones of controversial nature. For this reason we propose a novel consensus clustering framework that uses Homogeneity Analysis

(Homals) to create a new representation of the cluster ensemble, called the agreement space, which reflects all the agreements and disagreements of the clusterings that form the cluster ensemble, taking into account both types of relationships either between clusters or between clusters and data points. This representation can enable traditional clustering methods like hierarchical clustering to uncover the desired consensus clusters.

More formally, given a cluster ensemble, the Homals Consensus can obtain a consensus partition in the following steps:

1. Use Homogeneity analysis on the cluster ensemble by viewing each clustering as a categorical variable in order to create the agreement space of the cluster ensemble.
2. Perform Hierarchical Clustering on the agreement space
3. Obtain the final consensus clustering by specifying a cut in the cluster tree that maximises the silhouette criterion in the original data representation and a second cut that maximises the silhouette criterion in the agreement space.

The last step creates two consensus partitions but since both partitions are part of the same tree then always the one will be a partition of the other. Thus our method creates a two-tier hierarchical clustering that is organised in two levels that optimise different criteria. The choice of the criteria agrees with the two evaluation criteria for consensus clustering proposed in (Duarte et al., 2010). The first one, which embraces the notions of *Compactness* and *Separation* and tries to produce well defined clusters, remains to the best of our knowledge a clear objective for every clustering algorithm. It should be the aim for every clustering application despite all the arguments against it presented in the previous chapter, as it provides a natural view on the underlying structure of the data, even if that disagrees with other classifications on the data, which actually consist of more subjective views on the data. For the same reason it seems unreasonable to compare the clustering results to a predefined classification, as (Fred and Jain, 2005) dictates in the definitions of a good consensus clustering especially when the predefined classes are not well defined. Moreover, Homals Consensus completes an existing framework that is biased towards producing compact and distinct clusters as this occurs from the utilisation of internal validation criteria in the procedure, which share the assumptions of *Compactness* and *Separation* and thus it would be more logical to enhance towards this direction.

The second criterion aims to maximise the agreement of the consensus partition to the original ensemble and it is very important for consensus clustering applications to be able to capture all the diversity provided by the cluster ensemble as they might eventually approach the optimal solution. Our intuition is that a value that will maximise the silhouette criterion in the agreement space will maximise the agreement between the resulting clusters and the initial cluster ensemble as the final consensus cluster are obtained by hierarchical clustering on the agreement space. Well defined clusters in the agreement space potentially means that the consensus clusters are representative of the cluster ensemble and that all the information regarding the agreements and disagreements of the clusters are expressed in the final consensus partition.

Homals as a typical example of Correspondence Analysis tries to find a numerical representation for categorical variables that places as close as possible objects that have similar categorical levels across the categories and categories that are common across objects. That means that based on the co-occurrence of categorical levels across objects a spatial representation can be formulated with similar objects and categories being placed close together. Therefore our idea is that we can view the different clusterings as categorical variables. Then performing Homogeneity Analysis on this categorical table can result in a graphical representation where the similarity of clusters and the similarity of objects based on co-occurrence will be reflected in the distances between the objects and between the categories. This highlights a way to create a combined clustering from the initial set of partitions as the relationships among all the clusters of all the partitions and the relationships of objects can be quantified. Thus a potential hierarchical clustering on this representation can discover the consensus clusters and provide a good Consensus Clustering.

3.3.4 Related Work to Homals Consensus

This Consunsus Clustering solution is based on a very similar idea to the work of (He et al., 2005) with the only difference being that it can work for the more than two clusters and also capitalises the idea to examine both the similarities between objects and the similarities of the clusters for the creation of representative consensus clusters (Fern and Brodley, 2004). However, in contrast with most of Consensus Clustering solutions presented in literature, this proposed solution does not require a cluster alignment or relabeling before it takes place, as Homogeneity Analysis is a statistical method that solves the correspondence problem implicitly by representing the relationships in a joint p-dimensional space where the clusters are aligned based on the frequencies of co-occurrence. Moreover, our proposed solution can work on diverse clusterings adapting to the requirement of diversity and it does not need to know the number of consensus clusters a-priori as it contains mechanisms to find the optimal number of clusters.

3.4 Classification/Regression

3.4.1 Implementing Behavioural Modelling

Three classification and regression algorithms are analysed here that can serve as the necessary tools for answering fundamental questions of Consumer Debt Analysis and implement the Behavioural Modelling phase of our framework. In the end we take advantage of the flexibility Neural Networks provide in designing their network topology and we propose a novel technique that can incorporate the extracted knowledge from exploratory techniques inside their topology. This way our novel method respects the semantics of the data and uses this knowledge to improve its performance.

3.4.2 Linear Regression

Linear Regression (Freedman, 2009) is the simplest of the statistical models and it tries to model the relationship between a dependent variable and one or more explanatory variables. As one can infer from the name, Linear Regression assumes a linear relationship between the dependent variable and the explanatory variables and tries to fit a straight line in the data. More formally Linear Regression is defined as:

$$Y = \beta_0 + X_1\beta_1 + ... + X_p\beta_p + \epsilon, \tag{3.9}$$

where β_0, β_1, ..., β_p are the coefficients and X_j,j=1,....p denote p regressor variables. Finally ϵ denotes the error term which is assumed to be uncorrelated to the regressors and have mean equal to zero and a constant variance. The model takes as input the observations and tries to fit the straight line by estimating the parameters (coefficients and error term). A widely used algorithm for estimating the parameters is the Ordinary Least Squares(OLS) which tries to minimise the sum of squared residuals.

Its strength lies on its simplicity and explanatory power. In fact by examining the partial regression coefficients the influence of the regressor variables upon the the response variable can be assessed. The coefficients reveal the strength and the direction of the relationship between predictors and the response variable. However this can only be done on the variables that have proven to be statistically significant (Gevrey et al., 2003). Its main drawback on the other hand is the assumption of a linear relationship between the dependent variable and the regressor variables, which is often unrealistic.

3.4.3 Random Forests

Random Forest is an example of ensemble learning that generates many classifiers and aggregate the results (Breiman, 2001). The Random Forest method creates a large number of decision trees for the case of classification or regression trees for the case of regression from different random samples of the data. The samples are drawn based on bootstrap techniques that allow resampling of instances. The appropriate tree is constructed based on each sample and its accuracy is estimated on the rest of the samples. The difference from the common decision tree is that when a split on a node is to be decided, a randomly selected subset of the attributes can participate as candidates and not all of them. When the random forest is built, the prediction is made by aggregating the votes of all the trees for the case of classification and by averaging the results of all the trees for the case of regression. It needs the specification of only two parameters, the size of the forest and the number of predictors that are candidates for each node split and for this reason its success it has the advantage of simplicity. The notion of randomness it adopts in its process allows the model to be robust against data overfitting.

Random Forests also offer a way to assess the variable importance by using as measure the mean decrease in Gini Index The second measure is the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index (Gini, 1997). For regression, it is measured by residual sum of squares. Another measure of variable importance that is used in Random Forests is the prediction error on the out-of-bag portion. For each tree the prediction on the data it is not build and is recorded. Then the difference in the prediction error is calculated after permuting each input variable and is averaged across all the trees and normalised by the standard deviation of the differences.

An example of a regression tree can be seen in Fig. 3.3 where the predictions are saved on the leaves of the tree. Traversing the tree in a top-down approach and examining the values each observation has on the appropriate predictors that are saved on the nodes, a prediction is made reaching the appropriate leaf. For example if an observation has a value of less than 19.5 in the attribute X_{11} and more or equal than 0.605 in the attribute X_4 then the tree assigns the value 4.952 for the prediction.

In R environment, a Random Forest implementation is provided in the package *RandomForest* available in the CRAN repository: https://cran.r-project.org/web/packages/randomForest/index.html.

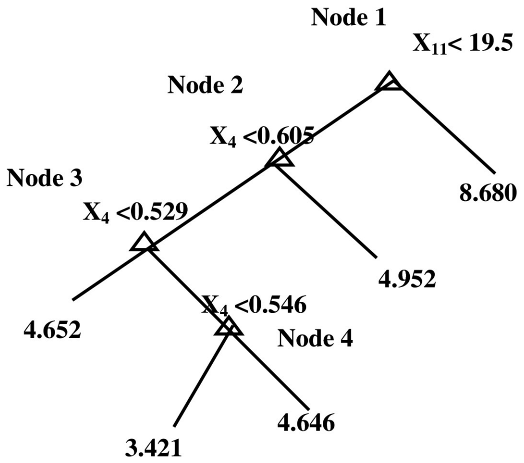


FIGURE 3.3: Example of a Regression Tree

3.4.4 Neural Networks

A Neural Network is a directed graph consisting of nodes and edges that are organised in layers. As it models a relationship between the predictors and the response variables, the input layer consists of nodes that represent the predictors and the output layer of nodes that represent the response variables if there are more than one. One or more hidden layers with a certain number of nodes connect these two layers. Typically, each layer is fully connected with the next layer and each edge assigns a weight to the value it takes as input and passes it on the next node. Thus in each node the weighted sum of the output values of all the nodes that belong to the previous layer is calculated adding the intercept and the result is fed into an activation function and passed to the next layer. The activation function is usually a non-linear activation function like the sigmoid function or the hyperbolic tangent. The simplest Neural Network (Perceptron) has n inputs and one output and it is analogous to the logistic regression as it is a non-linear function of the linear aggregation of the input. With this in mind we can easily conclude that a Neural Network with more than one node in the hidden layer is an extension of the Generalised Linear Models.

An example of a simple Neural Network can be seen in Fig. 3.4 where the relationship between the response variable Y and the two predictors, A and B, is modelled. The network also includes a hidden layer of three nodes. The two upper nodes represent the intercept being added to the output and the hidden layer.

A Neural Network takes as parameters the starting weights of the edges that are usually initialised randomly and the network topology meaning the organisation of the nodes in the hidden layers. Then the model tries to find the optimal weights of the edges by using a learning algorithm like back propagation on the data. Back propagation (Rumelhart

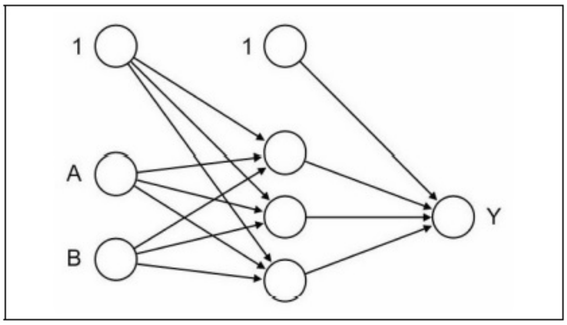


Figure 3.4: Example of a Neural Network with a two inputs, one output and hidden layer of three nodes. The image is being taken from Günther and Fritsch (2010)

et al., 1988) tries to minimise the difference between the predicted value calculated by the model and the actual value. It does that by calculating this difference and then following the chain rule, it moves from the output to the input adapting all the appropriate weights according to a specific learning rate. Resilient back propagation, which is argued to be more suitable for regression purposes (Günther and Fritsch, 2010), is similar to back propagation but instead of subtracting a ratio of the gradient of the error function like back propagation does, it increases the weight if the gradient is negative and reduces it if its positive. It updates the weights by using only the sign of the gradient and some predefined values. The value of the update is bigger if the gradient changes sign from the previous update and smaller if it keeps the same sign. This way it ensures that a potential global minimum won't be missed.

The Neural Networks tend to overfit the data, a fact that raises a concern of how they can be properly used. A common technique for avoiding data overfitting is to train the model on a subset of the data and validate it on the rest of the data. A very popular technique in supervised learning for this, is the 10-fold cross validation where the data is divided in ten folds and then a model is trained for each fold and gets validated on the rest of the folds. This is the way to evaluate the accuracy of the model and thus to choose the appropriate number of hidden layers and hidden nodes since this is not known beforehand. Usually different topologies are being tested and the one that minimises the error between the predicted and the actual values on the test set is selected.

In R environment, a Neural Networks implementation that allows the design of the topology of the neurons is provided in the package *neuralnet* available in the CRAN repository: `https://cran.r-project.org/web/packages/neuralnet/index.html`.

3.4.4.1 Profile Method

The biggest disadvantage of Neural Networks is that they are often considered as “black boxes” due to their inability to provide transparent modelling. Among the several explanatory methods developed (Gevrey et al., 2003; Harrington and Wan, 1998) of assessing variable importance for Neural Networks, a very interesting method is the Profile method. The idea of Profile method (Gevrey et al., 2003) is to study each input variable successively when the others are blocked at fixed values. In more detail each variable is divided into a certain number of equal intervals between its minimum and its maximum value. This number of intervals is called scale. All variables except one are set initially (as many times as required for each scale) at their minimum values and then successively at their 1st quartile, median, 3rd quartile and maximum value. For each variable studied, five values for each of the scale’s points are obtained. These five values are reduced to their median value. Then the profile of each variable can be plotted for the scale’s values of the variable considered and the same calculations can be repeated for each of the variables. For each variable a sketch is then obtained. This gives a set of profiles of the variation of the dependent variable according to the increase of the input variables.

3.4.5 Topology Defined Neural Network (TopDNN)

The flexibility that Neural Networks provide in designing their topology can be exploited to incorporate knowledge extracted by unsupervised learning performed on the data. Thus, in this work we tried to organise the neurons in the hidden layers based on the knowledge extracted by Factor Analysis and clustering and propose a new Neural Network solution we named TopDNN. The idea behind this was based on the striking resemblance Neural Networks have with Latent Factor Models, like Factor Analysis, and on the assumption that the classes of debtors identified by clustering define different relationships between the response variable and the predictors.

Factor Analysis is a common Latent Factor Model that organises the variables of a dataset into a smaller number of hidden factors that would still contain most of the information from the initial variables. This way neurons in the first hidden layer can be depicted as latent factors that summarise the input. The main difference with Factor Analysis, a widely used Latent Factor Model, is that the relationship between the input variables and the factors is non-linear. This non-linear relationship would also be able to model the linear relationships between the input variables and the neurons identified by Factor Analysis. This idea has been incorporated in the algorithm proposed in (Shifei et al., 2011).

Clustering on the other hand divides the debtors into classes with distinct characteristics. As these classes may model different relationships between the response variable and the explanatory variables, this could be introduced in the neural network as an extra hidden layer with as many neurons as the classes. This would create different functions for each class that will be combined in a more complex relationship in order to produce the final modelling. The intuition is something similar to Clusterwise Regression but the combination of different functions for each class is more fuzzy since they are included in a neural network.

These two ideas are central to our proposed methodology of building Neural Network models. Our aim is to test TopDNN in the socio-economic context but its disciplines can be extended in creating Neural Networks models for any real world application of complex nature.

3.4.6 Related work to TopDNN

Determining the topology of a Neural Network is usually carried by an optimisation procedure in Data Mining community. As a consequence the resulting topologies might be optimal in performance but they lack interpretation. An effort to design typologies of Neurons based on Knowledge extracted by exploratory research can improve the performance of the model but it can also lead to a potentially interpretable Neural Network architecture.

A similar idea is implemented by in (Shifei et al., 2011), where the authors uses Factor Analysis in an effort to design a Neural Network in a similar fashion with TopDNN. However they limit their idea only to Factor Analysis and not clustering. Their results show that this organisation of neurons based on Factor Analysis does not improve the performance but it speeds up the convergence of the algorithm. In other pieces of work Factor Analysis or more likely Principal Component Analysis (PCA) has been extensively used in combination with Neural Networks (Hoyle, 2008; Zekić-Sušac et al., 2013; Jiang, 1996), but in order to transform the input data not to organise the architecture of the Neural Network.

Chapter 4

Behavioural Profiling on CCCS

4.1 Introduction

In this chapter we proceed with the implementation of the Behavioural Extraction phase of our framework. In doing so, we attempt to extract meaningful Behavioural Profiles of consumers from a socio-economic dataset that as any other real world dataset is characterised by several inconsistencies in the data. For this reason it is important to use data transformations that can improve the data quality and prepare them for the clustering procedure. These transformations also aim to extract behavioural elements hidden in the data in order to enhance the process of Knowledge Discovery in Consumer Debt Analysis. Therefore our goal here is to assess the contribution of transformations in extracting Behavioural Profiles of consumers that can provide a better understanding of the complex problem of consumer indebtedness by revealing important patterns of consumer behaviour. This work will try to provide answers to the research question 1,2 and 4 as they are defined in the first Chapter.

4.2 The Consumer Credit Counselling Service (CCCS) Dataset

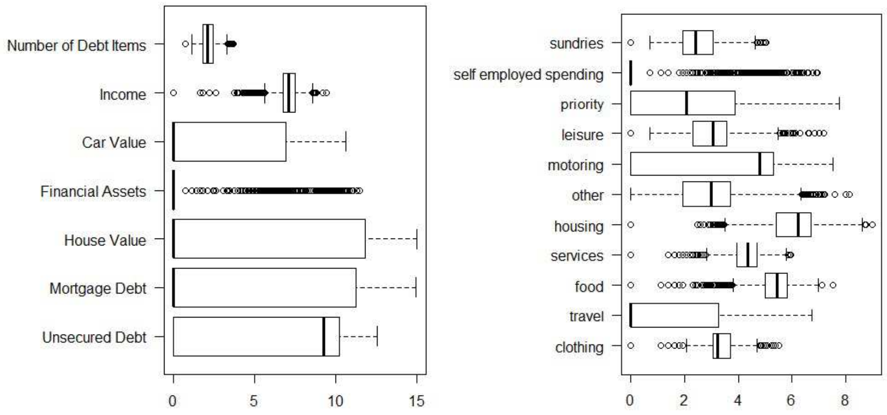
In an effort to implement the first part of our research plan we performed clustering on the CCCS dataset in order to detect Behavioural Profiles of debtors. The CCCS dataset, introduced in (Disney and Gathergood, 2009), is a socioeconomic crossectional dataset based on the data provided by the Consumer Credit Counseling Service. In its 58 attributes it contains information about approximately 70000 clients who contacted the service between the years 2004 and 2008 in order to require advice about how they can overcome their debts. The information was gathered through interviews when each client first contacted the service and it varies from standard demographics to financial

Table 4.1: Description of CCCS attributes

Attribute	*Description*
pid	individual identifier
Demographics	
age	age of person
mstat	marital status
empstat	employment status
male	sex of person
hstatus	housing status
ndep	number of dependants in household
nadults	number of adults in household
Financial Attributes	
udebt	total value of unsecured debt
mortdebt	total value of mortgage debt
hvalue	total value all housing owned
finasset	total value of financial assets
carvalue	resale value of car
income	total monthly income
Expenditure	
clothing	total monthly spending on clothing
travel	total monthly spending on travel
food	total monthly spending on food
services	total monthly spending on utilities
housing	total monthly spending on housing
motoring	total monthly spending on motoring
leisure	total monthly spending on leisure
priority	total monthly spending on priority debt
sundries	total monthly spending on sundries
sempspend	total monthly self-employed spending
other	total other spending
Debt Information	
ndebtitems	number of debt items

details, aggregated spending in categories and debt details. The attributes of interest for the purpose of Consumer Debt Analysis are limited to Demographics, Expenditure and Financial attributes as they can be seen in Table 4.1 together with their description.

In Fig. 4.1 we can see the boxplots of the numerical attributes of the dataset in log scale. We can instantly spot the much larger values of the financial attributes over the expenditure attributes. More precisely, *House Value*, *Unsecured Debt* and *Mortgage Debt* are expressed in significant bigger levels than the rest of the financial attributes as well. Moreover we can spot the existence of many outliers (bullets outside the boundaries of the boxplots) in attributes like *Income* and many spending attributes. If we add the fact that *Self Employed Spending* and *Financial Assets* are attributes mainly consisted of zeros, then we can easily understand the problematic nature of this dataset. Clustering



(a) Financial Attributes

(b) Expenditure Attributes

Figure 4.1: Boxplots of CCCS attributes in log scale

algorithms, especially K-means, will suffer from the outliers and their results will be shaped by the heavily expressed attributes which will produce the bigger differences in Euclidean distances. All the rest of the attributes will not be able to have an impact on the clustering procedure and therefore a potential analysis cannot reveal useful results for the understanding of Consumer Debt.

Our first simplistic approach is described in the published paper in UKCI 2012 (Ladas et al., 2012), where the CCCS dataset is presented analytically. Despite the fact that the results reveal patterns in the behaviour of debtors, they cannot be considered reliable since our clustering approach was not sophisticated enough. In more detail our clustering results were shaped by the aforementioned difficulties, revealing clusters characterised by misleading patterns. Also no effort was implemented to include the categorical attributes in the clustering procedure leaving out important demographic information Therefore, a more careful consensus clustering approach is required in order to produce robust behavioural profiles that reflect the differences of consumers in demographics and economic and spending behaviours given the complexity of this dataset.

4.3 Transformations of the CCCS dataset

4.3.1 Dealing with the inconsistencies of the CCCS dataset

From the above it is made clear that in order to successfully mine patterns in the CCCS dataset we have to perform a series of data transformations upon the socio economic dataset. More precisely, we use Homogeneity Analysis (Homals)(De Leeuw and Mair, 2009, 2007) in order to map the categorical demographic data, significant attributes concerning the Consumer Debt Analysis, into two-dimensional coordinates so they can be taken into account in the clustering procedure. Moreover we perform a Factor Analysis on the financial attributes and a clustering on the correlation of the spending items in order to reduce the dimensionality and also to provide interpretability to the results with meaningful financial factors and spending behaviours. By these transformations not only we can overcome the technical difficulties this dataset imposes on Data Mining approaches but we also aim to construct behavioural data that could help the mining procedure to extract behavioural patterns.

4.3.2 Transforming Demographics

Here we try to represent the categorical demographic variables of the dataset in a two-dimensional space. In order to do that we discretise the age attribute into three categories, H (high), M (medium) and L (low) depending on whether the values of the attribute fall over the interquartile range (IQR), inside or below respectively. Then we apply Homals to obtain the category plot of the variables as seen in Fig. 4.2.

We can notice how well the two dimensions separate the levels of all the categorical variables providing an appropriate numerical representation for clustering purposes. The examination of the values of the two dimensions reveals that it places the category points in different quandrants and so it can assist in identifying a person's sex, age, employment and marital status, determining the number of adults and dependants in the household and whether he owns his residence or he rents it.

4.3.3 Financial Attributes

In an effort to uncover potential relations between attributes that reflect the income, the assets and the level of debt, we perform a Factor Analysis on the financial attributes. With a latent factor model we can reduce the dimensionality of these attributes, uncover associations between them and deal with the problems these attributes impose on the dataset and were described in the previous section.

In order to determine the number of factors, we examine the scree plot and use parallel analysis. Our inspection shows that these attributes can be explained by a three factor model, that can explain 64% of the variance of the financial attributes. This is because some of the attributes tend to not co-vary with other attributes, like *Car Value* and *Financial Assets*, and therefore they cannot be explained by a latent common factor. The factor loadings can be seen in Table 4.2, where we can see that the first factor has a physical interpretation as it is associated with the house value and the level of mortgage

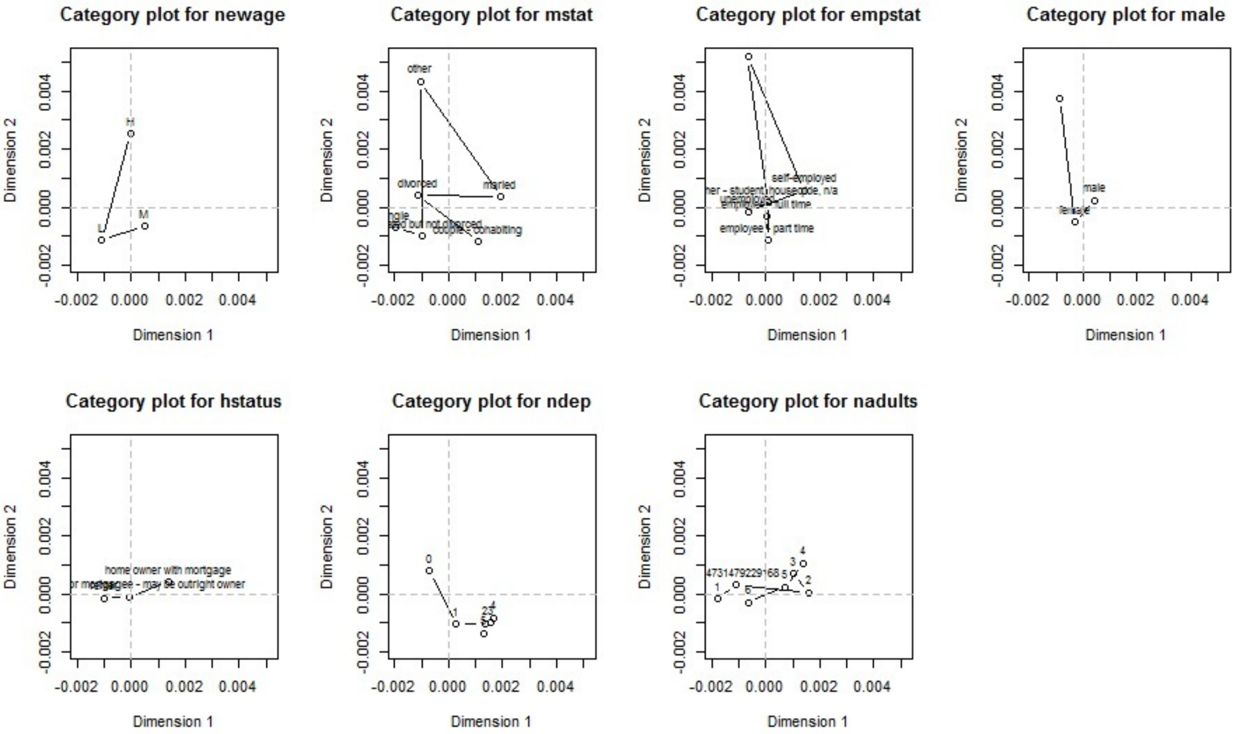


FIGURE 4.2: The representation of the labels of categorical variables in 2- dimensional space for each attribute, newage is the new discretized *Age*, mstat is *Marital Status*, empstat is *Employment Status*, male is a binary value for *Sex*, hstatus is *Housing Status*, ndep is *Number of Dependants* in household and nadults is *Number of Adults* in household

TABLE 4.2: Loadings of the 3 factors on financial attributes

	$Factor1$	$Factor2$	$Factor3$
Unsecured Debt		0.353	
Mortgage Debt	0.902	0.127	0.221
House Value	0.948	0.107	0.292
Financial Assets			
Car Value			0.292
Mortgage Terms	0.151		0.506
Income		0.875	0.477

debt. The second factor relates income with the level of unsecured debt, a relationship that is evident in the literature of Economics but the relationship is not clear and is not similar to the debt-to-income ratio. The third factor is not interpretable as it is unable to reflect sensible associations in the data, but it can be useful for the purpose of dimensionality reduction. As the Financial Assets attribute is not being explained by any factor and since it is a problematic attribute, it is dropped from our further analysis.

4.3.4 Expenditure

The expenditure of the debtors is aggregated in 11 categories in (Disney and Gathergood, 2009). This aggregation is based on common logic but it does not represent common spending behaviours. In this work we wanted to group the items debtors have spent upon based on the correlations between the items. For this reason we perform the framework of consensus clustering on the correlation matrix of the 100 most frequent items which identified four behavioural clusters. These clusters are depicted in the Fig. 4.3 as part of the dendrogram returned by hierarchical clustering on the same correlation matrix. For the hierarchical clustering complete linkage was used as it defines the similarity of two clusters as the similarity of their most dissimilar members and thus it merges the two clusters that minimise this distance resulting in more compact clusters. On the contrary single linkage defines the similiraty between two clusters as the similarity between their two most similar members. This method might result in very disperse clusters and for this reason is not chosen. A middle point between those two methods, called average linkage tries to minimise the average of these two distances but it does not lead to clear interpretations.

We can see that the four clusters returned as optimal from our framework are not in perfect agreement with the clusters returned from hierarchical clustering. In particular, the Cluster 2 is comprised by a small cluster that belongs to the left subtree of the dendrogram and another cluster that resides in the right subtree and the same applies for Cluster 4. However they reflect different spending behaviours of the debtors and in an attempt to characterise them we label them as *Necessary Spending*, *Household Spending*, *Excessive Spending* and *Leisure Spending*. While this labelling is arbitrary it is useful for separating the different spending behaviours. It is a technique usually referred as customer segmentation and it is widely used in marketing industry. The idea is based on (Otto et al., 2009) where this segmentation is used to improve the marketing research. Each debtor's score in these four different spendings defined by the clusters is calculated as the sum of the items included in the cluster that the debtor purchased.

After this final transformation we performed on the CCCS dataset, we obtain nine new numerical attributes. Two coordinates for representing demographic variables, three factor scores for explaining the financial attributes and four behavioural spending clusters which replace the original aggregated spending attributes. No further pre-processing is performed as we choose to experiment with a random sample of 10000 debtors which is representative of the whole dataset and contains no missing values.

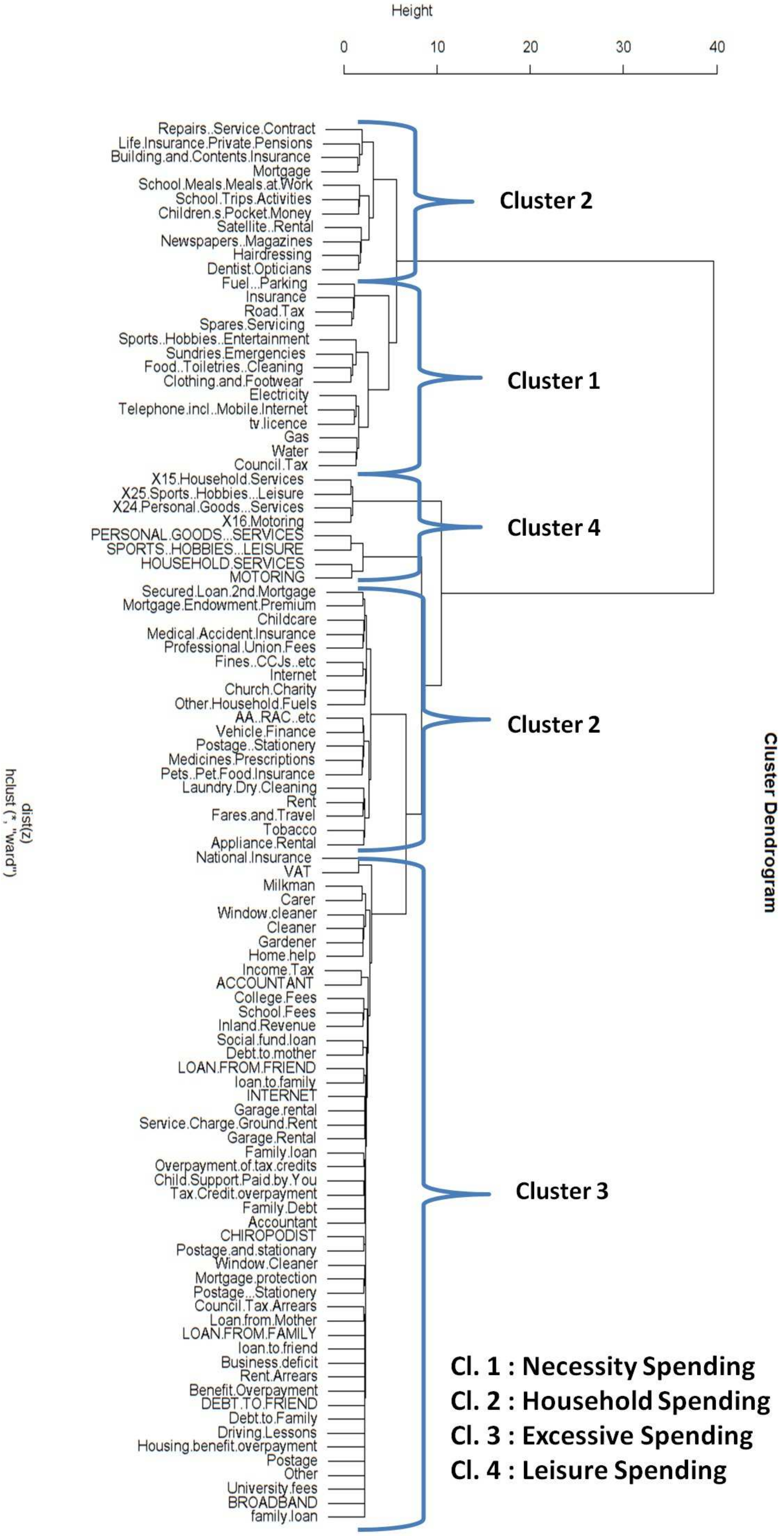


Figure 4.3: The 4 behavioural clusters as they are depicted on the dendrogram of Hierarchical Clustering

4.4 Methodology

For evaluating the impact of the described transformations on the quality of the clustering of socio-economic data we use a framework of consensus clustering (Soria and Garibaldi, 2010). As described, the framework utilizes six validation indices, namely Calinski and Harabasz, Hartigan, Scott and Symons, Marriot, traceW and Friedman, in order to find the optimal number of clusters for each clustering algorithm, iterating them for different number of clusters each time and following the rules dictated in (Dimitriadou et al., 2002). In case of disagreement among the indices the groupings are being ranked for each index and the best one is chosen based on the minimum sum of ranks. Then the level of agreement between the algorithms is measured by calculating the Cohen's kappa (Cohen et al., 1960) between the different clusterings.

This framework is suitable for our purposes as we don't know the correct number of clusters a priori and Cohen's kappa serves as a validation measure of the quality of the clustering result. We decided not to use any external validation because external information is not available. In addition, neither of the internal validation measures presented in literature were selected, they suffer from their own limitations and assumptions (Liu et al., 2010; Vendramin et al., 2010) and it is not known if they can provide meaningful comparisons between clusterings of different datasets. On the other hand, the intuition behind the choice of Cohen's kappa is that if patterns among the objects truly exist in the data then these should be revealed by any clustering algorithm. So the bigger the agreement between two clustering algorithms the bigger is the chance that these algorithms identified true patterns within the data.

For this reason we are going to use this framework for two clustering algorithms widely used in Data Mining applications, K-means and PAM for a series of experiments described in Table 4.3. Since K-means is sensitive to cluster initialisation it is initialised with cluster assignments obtained by hierarchical clustering as dictated in (Eshghi et al., 2011). For each one of the experiments, the level of agreement between the two clusterings obtained by K-means and PAM is computed and it is used as validation criterion of the quality of the clustering. Then the core classes are defined by grouping together clusters of the two algorithms that share the most objects.

In addition to this, we use the Silhouette criterion (Rousseeuw, 1987) in order to check how well separated, distinct and compact the clusterings of the two algorithms are.

Finally, the resulting core classes from each experiment are inspected in order to identify the patterns that were discovered in each experiment and provide a suitable characterisation of the classes. For this purpose we used ANOVA tests to find which of the numerical

Table 4.3: Experiments on the CCCS dataset

Experiment	*Attributes*
A	Financial and Expenditure attributes of CCCS
B	Financial and Expenditure attributes of CCCS, plus the demographic coordinates
C	3 Financial factors and 4 Behavioural Spending clusters
D	3 Financial Factors and 4 Behavioural Spending clusters, plus the demographic coordinates

attributes of the dataset produced significant differences, and for these attributes further pairwise t-tests and Z-tests for all the clusters were used in order to understand the expression of each attribute in each class compared with the expression of the attribute in other classes and in the whole population. For the categorical attributes their distribution among their nominal values is examined in order to find alterations in these distributions among the classes.

4.5 Results

4.5.1 Experiments Non Scaled Data

Examining the results in Table 4.4, we can see that the agreement between PAM and K-means is very poor when they have to cluster attributes from the original CCCS dataset (A). The level of disagreement is reflected on the small value of kappa index (0.189) and on the big difference between the optimal number of clusters returned for PAM and K-means. When the framework is applied on the processed attributes (C), the agreement between the Cohen's kappa increases substantially. This comes in contrast with the silhouette values for both algorithms which decrease in a great extent, indicating that the clusters returned from K-means and PAM are not distinct and well separated. In addition to this, when the demographic coordinates are included in the clustering framework together with either the original attributes of CCCS (B) or the preprocessed (D), they fail to produce any effect on the clustering procedure, producing identical clusterings each time. This is explained by the big difference between the values of the demographic coordinates and the values of the rest of the attributes. This verifies our concern that the clustering procedure is guided exclusively by attributes whose values vary in higher levels than the rest as they are able to produce significant differences in the Euclidean distance.

Checking the description of classes returned in Table 4.5, we can see that in Experiment A, the 4th class contains only four members with extremely high values in *udebt*

Table 4.4: Results of experiments on non scaled data. S_K is the average silhouette for Kmeans and S_P for PAM

$Experiment$	$Kmeansclusters$	$PAMclusters$	S_K	S_P	$kappa$	$CoreClasses$
A	5	16	0.7	0.42	0.189	5
B	5	16	0.7	0.42	0.189	5
C	10	4	0.17	0.15	0.663	4
D	10	4	0.17	0.15	0.663	4

Table 4.5: Characterisation of clusters returned in experiments with non-scaled data

	Experiment A		Experiment C	
Class	Size	Characterisation	Size	Characterisation
1	6180	hvalue-, single (38%), renters (41%)	4466	income-, udebt-, assets-, spending-, renters (34%), unemployed (24%), single (40%)
2	2802		2466	assets+, services+, housing+
3	568	income+,udebt+, assets+, spending+ (except leisure), self employed (18%)	411	income+, udebt+, ndebtitems+, spending+, married(74%)
4	4	income+, udebt+, assets+, spending+, leisure-	1588	udebt+, ndebtitems
5	66	income+, udebt+, assets+, spending+, married (73%), self employed (30%), mortgagees (67%)		

(unsecured debt) and *hvalue (house value)* indicating the presence of outliers. In addition to this, the first two classes describe debtors with average expression in almost all numerical attributes except the *hvalue (housevalue)* in the first class which is underrepresented and classes 3,4 and 5 exhibit larger expression in all numerical attributes except *leisure* in classes 3 and 4. The only thing that is different between the first two classes is the magnitude of expression of the numerical variables, which is higher in the 2nd, and the concentration of significant proportions of singles and renters in the 1st class. Similarly there are differences in the expression of numerical variables in the other three classes, with the 3rd class exhibiting the lowest expression in numerical variables from all three, class 4 exhibiting the highest expression in *udebt (unsecured debt)* and *hvalue (house value)* and the 5th class exhibiting the highest expression in all the rest of the attributes. Also the presence of large proportions of married and mortgagees in class 5 and the difference in the proportions of self employed between classes 3 and 5 separates more these three classes.

Moving to the Experiment C we can notice that the financial factors were able to produce different financial behaviours among the classes. In particular in class 2 we identify debtors with average levels of unsecured debt, income and spending but expensive assets, in class 3 debtors with high levels of unsecured debt, income and spending but average values of assets and finally in class 4 debtors with average levels of income, spending and assets but high levels of unsecured debts. In the same way, behavioural clusters despite the fact that couldn't produce differences in the behavioural spending, they managed to produce differences in the expression of *services* and *housing* in class 2. In demographics we can notice the large concentrations of renters, unemployed and singles in the low income and low spending 1st class and the large concentrations of married in the high income and high spending 3rd class.

In conclusion, the financial factors and behavioural clusters have proved to be beneficial for the clustering of this dataset as this is indicated by the increase in the agreement, which is measured by Cohen's kappa, and as this can be seen in the patterns returned from the two clusterings. More specifically in the experiment A, a strong association between spending, income, level of debt and total value of assets is demonstrated. This association is consistent with findings in economic psychology (Livingstone and Lunt, 1992; Kim and DeVaney, 2001; Wang et al., 2011), where the importance of income as a predictor of the level of debt is highlighted. However, in Experiment C the financial factors produce different financial behaviours and especially in class 4 we can see debtors with average spending but high levels of debt, a finding that comes in contradiction with the previous findings. Also the behavioural clusters were able to produce significant differences in the expression of some spending attributes but not the majority of them, showing that the relationship between income and spending still holds. Moreover the transformation managed to reveal a large concentration of self employed debtors with high levels of unsecured debt, which is an association that has been observed in the literature (Ottaviani and Vandone, 2011). Finally, the proposed transformation managed to eradicate the effect of outliers in the dataset which is evident in experiment A at the formation of the 4th Class, but they could not decrease the impact of the heavily expressed attributes as the transformed demographics couldn't alter the results in either case (B or D) and the patterns discovered in the demographics were the same in both the experiments (A and C), revealing large proportions in singles and renters in the low spending classes and large proportions of married in the high spending classes. This signifies the importance of scaling in order for the transformed demographic data to be taken into consideration in the clustering process.

4.5.2 Experiments with Scaled Data

At first we experimented with standardising the data using Z-scores. The results can be seen in Table 4.6. We notice that the silhouette values of both algorithms and the

Table 4.6: Results of experiments on scaled data with z-scores. S_K is the average silhouette for Kmeans and S_P for PAM

$Experiment$	$Kmeansclusters$	$PAMclusters$	S_K	S_P	$kappa$	$CoreClasses$
A	6	14	0.25	0.3	0.134	4
B	3	14	0.23	0.3	0.117	3
C	4	8	0.38	0.14	0.14	2
D	4	6	0.28	0.13	0.284	2

Table 4.7: Results of experiments on scaled data with 0 - 1 normalization. S_K is the average silhouette for Kmeans and S_P for PAM

$Experiment$	$Kmeansclusters$	$PAMclusters$	S_K	S_P	$kappa$	$CoreClasses$
A	4	4	0.14	0.08	0.509	4
B	4	5	0.21	0.15	0.614	4
C	9	4	0.25	0.27	0.622	4
D	9	8	0.2	0.18	0.775	7

level of agreement drop to a great extent. This is probably explained by the fact that standardisation produces a dataset that is normalised and very dense around central areas, which hardens the work of clustering algorithms, that are now unable to produce distinct and compact clusters easily. In addition to this we can see that financial factors and Spending behavioural clusters increase the quality of the clusters in all criteria. The inclusion of demographics seem to beneficial only together with other transformations.

Hence, results of standardisation with Z-scores do not seem to improve the quality of the clustering, a fact that comes in contradiction with the work of (Hennig and Liao, 2013) where standardisation is being picked as the appropriate method of scaling the data. For this reason we performed another scaling method.

In Table 4.7, the results of the same experiments can be seen after we scaled all the numerical attributes with 0 to 1 normalisation. Here the impact of the demographic coordinates is evident in both cases (B) and (D), and their inclusion increases the agreement between the two clustering algorithms. In addition to this, when the data are scaled we can verify how beneficial were the transformations we performed as the Cohen's kappa increases steadily across the experiments, meaning that each of the transformations improved the quality of the clustering. Moreover we can see that the agreement is generally higher when data is scaled, indicating the importance of scaling in real world data. As it is shown, not only it improves the quality of the clustering but it makes the clustering more fair as it takes into account attributes with small values. Thus, it becomes apparent that scaling is an essential preprocessing step when clustering is performed on real world data.

Table 4.8: Characterisation of clusters returned in experiments with linearly transformed data

	Experiment D	
Class	Size	Characterisation
1	2301	income-, udebt-, hvalue-, spending- , unempoyed(23%), renters(45%), singles(82%)
2	1440	hvalue-, clothing+, food+, p/t employment(20%), cohabiting (34%)
3	1033	income+, udebt+, hvalue+, spending+, self employed (18%), mortgagees (54%)
4	948	udebt-, hvalue+, clothing-, low f/t employment (32%), retired(21%)
5	507	income+, udebt+, hvalue-, spending+
6	1588	udebt-, p/t employment (22%), divorced(14%), singles (42%), separated (26%)
7	553	income-, udebt-, hvalue-, spending- , low f/t employment (17%), retired(53%), other marital status (65%)

As the beneficial nature of the financial factors and behavioural clusters has been established already from the experiments in the non-scaled clusters, and here it still shows significant increase in the level of agreement between the two clustering algorithms we will proceed with the analysis of the results of the Experiment D where all the proposed transformations are included in the clustering procedure and all the data are linearly transformed. The seven resulting classes can be seen in Table 4.8 and we can easily notice the emergence of a lot of patterns in demographics. More specifically, the transformed demographics not only provide detailed descriptions of every class of debtors but manage to differentiate the low income-spending classes 1 and 7 by identifying different groups of debtors that exhibit the same low financial and spending behaviours. In the 1st class there are included debtors that have a large possibility to be unemployed and more likely to be singles and rent their residence, whereas in the 7th class debtors which are more likely to be retired or be involved in other marital status are described. In a similar way, by presenting large proportions of self employed and mortgagees it manages to separate better class 3 from 5 where in both classes debtors are characterised by high spending, income and levels of debt.

In consistency with the results of experiment C in the non-scaled data, the financial factors managed again to produce different financial behaviour, which are evident in the average income and spending classes 2,4 and 6, where class 4 is characterised by low levels of debt but expensive *hvalue* and class 6 by low levels of unsecured debt and in the high income and spending classes where class 3 is characterised by expensive *hvalue* and class 5 by low *hvalue*. As noted before the strong relationship between income and spending remains intact. Concerning the spending, interestingly enough the overexpression of spending in clothing and food in a general average spending class like 2 and the underrepresentation of spending in clothing in the average spending class 4 provides some insight into different spending behaviours.

From all these, we can easily conclude that the proposed transformations managed to improve the quality of the clustering as this is verified not only by the increment in the level of agreement but also by the emergence of numerous patterns that can provide detailed descriptions of debtors based on the differences in demographics, financial and spending behaviour. In particular, the proposed transformations highlighted a way for incorporating categorical demographic attributes in the process of clustering and processed data that were able to reflect different behaviours among the debtors. This means that the quality of clustering is quantified by the level of agreement but at the same time is also observed in the patterns extracted that can introduce new findings in the field of Consumer Debt Analysis.

The resulting behavioural profiles describe distinct groups of consumers based on the expression of meaningful behavioural patterns that provide a better understanding of consumer indebtedness. More specifically, in the observed patterns we notice that the widely referred association between income and unsecured debt seems not to be stable across the clusters. The description of clusters can help us understand when this relation holds and when it does not. Besides this, the verified association between a large percentage of self-employed debtors in clusters with high levels of debt is worth extra attention as it has been briefly covered in the literature.

Furthermore, the significance of scaling as an essential pre-preprocessing step is emphasised as it allows the inclusion of the transformed demographics to be taken into account and treats each attribute fairly, removing the bad effects of the heavily expressed numerical attributes and the outliers. More importantly these results justified the choice of Cohen's kappa as an evaluation criterion, since it was shown that an increase in the level of agreement was accompanied by the observation of meaningful patterns.

4.5.3 Low Silhouette values

However, we have to notice that the Silhouette values are quite small not only in the cases of clustering scaled data but also in the cases (C,D) in clustering non-scaled data. This comes in agreement with the observation in (Caruana et al., 2006), where it is stated that compact clusters do not ensure that they are meaningful for the application. We are unsure if we should regard these silhouette values as low, because we do not know what is considered to be an accepted silhouette value for clustering real world data and the high values of the first experiment in the non-scaled data can be a result of the presence of outliers, since outliers lie far away from the areas that most of the objects of the dataset fall and therefore their *Compactness* and *Separation* increases, ensuring them a good silhouette value. This could explain the fact that these small silhouette values coincide with the large levels of agreement, meaning that despite the fact that two different clustering algorithms produce numerous similar patterns these are considered not so compact and not so well separated.

This leaves a question mark about whether Knowledge Discovery in real world data through the clustering process should share exclusively the theoretical assumptions of *Compactness* and *Separation*, on which most of the internal validation indices are based. Perhaps in real world data, patterns are not compact and well separated. This is shown by the emergence of numerous and interesting patterns and by the increase of the level of agreement that is not followed by increase in the silhouette values. That signifies the utilisation of Cohen's kappa as an alternative validation criterion in clustering since the agreement of the cluster ensemble seems to lead to the discovery of true patterns hidden in the data and therefore it should be considered by a consensus clustering solution.

4.6 Conclusions

In this chapter we developed a framework that provides answers regarding how to process the data in order to extract psychological features of consumers and how to use this information in order to build behavioural profiles of consumers. In more detail, we performed a series of transformations on CCCS, a socio-economic dataset, in order to deal with the inconsistencies this dataset exhibited but also to identify behavioural elements in the data. The resulting behavioural data proved to be beneficial for the goodness of clustering not only because they improve the quality of the data but also because they increase the level of agreement between two different clustering algorithms, assisting this way the clustering process to reveal interesting and meaningful patterns in the results as the inspection of the classes revealed. In fact, the successfully extracted behavioural profiles are characterised by differences in the financial behaviours of debtors, the demographics and few of the spending behaviours of the consumers, and consist a meaningful classification of debtors which represents interesting patterns of consumer behaviour. Thus, the behavioural data enhanced the process of Knowledge Discovery for the purposes of Consumer Debt Analysis.

In addition to this, a way for including categorical data in the clustering procedure is shown by using Homogeneity Analysis(Homals) and scaling was proved to be an essential preprocessing step for the clustering procedure. These results not only highlight the way to mine this socio-economic dataset and extract patterns that will become useful in the analysis of consumer debt, but they also provide useful guidelines for clustering applications on any real world dataset. The resulting patterns sketch alternative debtor profiles which highlight differences in the demographics, and financial and spending behaviour.

Moreover, the utilisation of Cohen's kappa as an alternative method of clustering validation proved successful. As the maximisation of this criterion was followed by the emergence of numerous patterns that lead to the definition of distinct and meaningful behavioural profiles of consumers, it seems that it manages to uncover patterns that transcend the ideas of *Compactness* and *Separation*. This leaves open questions regarding whether a clustering application in the real data should rely exclusively on the notions of Compactness and Separation.

Chapter 5

Homals Consensus

5.1 Introduction

In this chapter we evaluate the performance of our proposed Homals Consensus as the last part of a consensus clustering framework defined in (Soria and Garibaldi, 2010). In doing so, we apply our method on three different datasets from the UCI repository that have different qualities. One that can be easily clustered, one dense dataset that is very difficult to be partitioned in well separated clusters and one more of moderate difficulty to be clustered. Then we evaluate the results of the proposed two tier clustering structure in terms of optimising *Separation* and *Compactness* and in terms of agreement to the initial cluster ensemble. The first objective tries to enhance the current functionality of the framework, whereas the second tries to find interesting patterns in the data that can occur by successfully modelling the agreement and disagreement of the cluster ensemble in an effort to build on the success of the utilisation of Cohen's kappa as evaluation metric in the previous chapter, where it achieved the discovery of meaningful patterns only by considering the level of agreement in the cluster ensemble. As the objective of Homals Consensus is to improve the existing consensus clustering framework, it is also compared to the original consensus framework and to the initial cluster ensemble. After the performance of the Homals Consensus in these datasets is established the proposed solution is used to extract behavioural profiles of consumers from the DebtTrack dataset, a survey with socio-economic and psychological items. These profiles should define meaningful classes of debtors that describe different consumer behaviours and should be able to provide a deeper insight to the complex problem of consumer indebtedness. This work will provide further answers for the second research question stated in Chapter 1 in an effort to improve the quality of our methodology.

Table 5.1: Attributes of User Knowledge Modelling dataset

Attribute	**Description**
STG	(The degree of study time for goal object materials)
SCG	(The degree of repetition number of user for goal object materials)
STR	(The degree of study time of user for related objects with goal object)
LPR	(The exam performance of user for related objects with goal object)
PEG	(The exam performance of user for goal objects)
UNS	(The knowledge level of user) (target value)

Table 5.2: UNS categorical levels and their frequencies

UNS levels			
High	**Low**	**Middle**	**very_low**
63	83	88	24

5.2 Datasets

5.2.1 User Knowledge Modelling

This dataset includes information regarding the students' knowledge status about the subject of Electrical DC Machines and is available in the UCI repository: https://archive.ics.uci.edu/ml/datasets/User+Knowledge+Modeling. It represents the knowledge level of 258 users measured by five numerical attributes and classified in one class attribute (UNS) which has four categorical levels. In Table 5.1 we can see the information the dataset contains and in table 5.2 we can see the four categorical levels together with their frequencies.

In order to get a visual representation of the UNS classification we draw bitplot of five numerical attributes coloured by their class in UNS. In the graph of Fig 5.1 we can see that the High class (turquoise) and the Very low (gold) are distinct and well separated with the first one positioned on the negative side of the first component and the other on the positive side of the same component. However the two bigger UNS classes, Low (Red) and Middle (Blue) respectively, are positioned in the center of the graph and they are not well separated. This causes a very low Silhouette value for this classification (0.05). From this we can understand that the UNS classes are not well separated in the n-dimensional space (n=5) which means that it is very difficult for clustering algorithms to discover them.

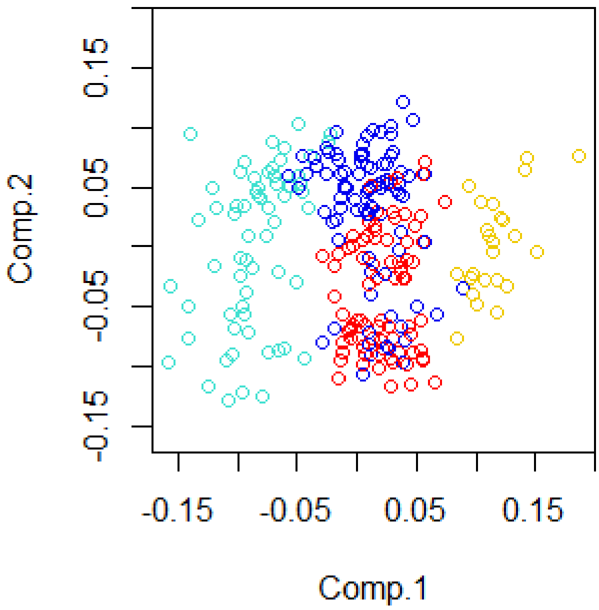


Figure 5.1: Biplot of the User Knowledge Modelling dataset by UNS

Table 5.3: Attributes of seeds dataset

Attribute	**Description**
Area	area A,
Perimeter	perimeter P,
Compactness	compactness C = 4*pi*A/P2̂,
Kernel_length	length of kernel,
Kernel_width	width of kernel,
Assymetry	asymmetry coefficient
Kernel_groove	length of kernel groove.
Class	wheat variety

5.2.2 Seeds Dataset

This dataset is available in the UCI repository: https://archive.ics.uci.edu/ml/datasets/seeds and summarises the measurements of geometrical properties of kernels belonging to three different varieties of wheat with each variety having 70 members. The size of the dataset has 210 members and a description of its attributes is provided in Table 5.3.

A visual representation of the varieties of wheat in the seeds dataset is depicted in Fig 5.2. As we can see the three wheat varieties are well defined being distinct and compact. Only very few occassions of the second class (red) and the third class (blue) are found to reside inside the outer area of the first class. In this dataset the clustering algorithms

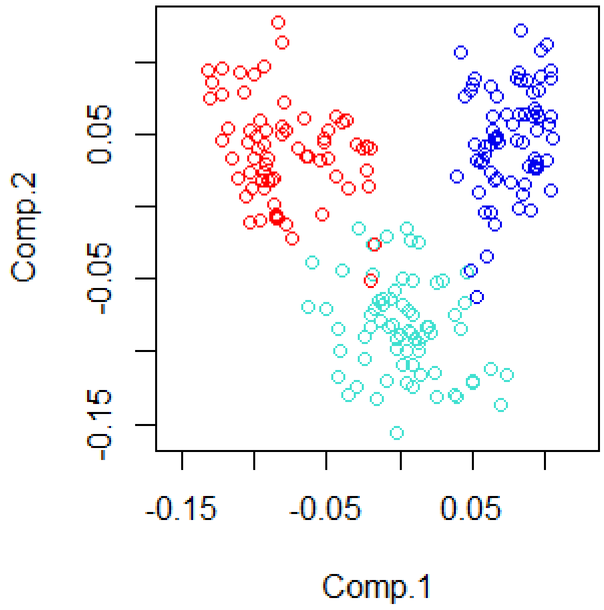


FIGURE 5.2: Biplot of the seeds dataset classified by wheat variety

TABLE 5.4: Attributes Wholesale customers dataset

Attribute	**Description**
FRESH	annual spending on fresh products
MILK	annual spending on milk products
GROCERY	annual spending on grocery products
FROZEN	annual spending on frozen products
DETERGENTS_PAPER	annual spending on detergents and paper products
DELICATESSEN	annual spending on and delicatessen products
CHANNEL	Horeca (Hotel/Restaurant/Cafe) or Retail
REGION	Lisboa, Oporto or Other

can achieve a very good performance with the only difficulty being in these few boundary objects.

5.2.3 Wholesale Customers

The dataset refers to clients of a wholesale distributor in Portugal. It includes the annual spending in monetary units (m.u.) on diverse product categories and is available in the UCI repository https://archive.ics.uci.edu/ml/datasets/Wholesale+customers

The visual representations of this dataset depicted in the biplots in the Fig. 5.3 and Fig. 5.4 reveal a very dense dataset. In fact most of the points seem to be concentrated in a very small region of the representation with a few outliers placed a bit further away. In the case of Region the classes are visually incomprehensible as they seem not to be

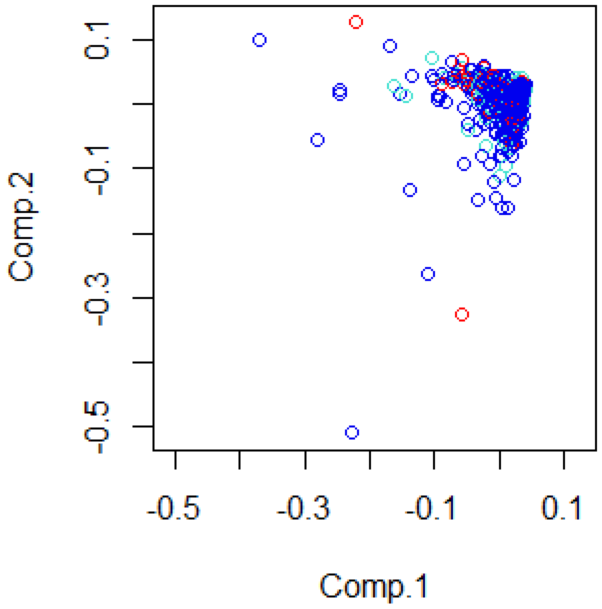


Figure 5.3: Biplot of the Wholesale customers dataset classified by region

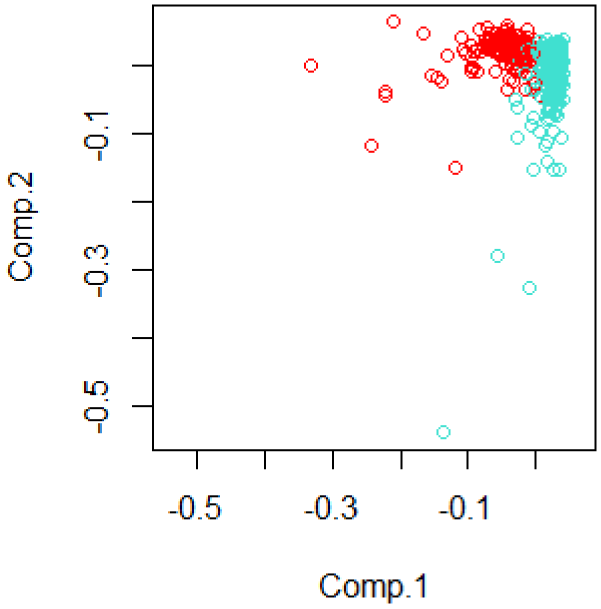


Figure 5.4: Biplot of the Wholesale customers dataset classified by channel

distinct. However in the case of Channel the clusters seem to be distinct but not well separated.

5.3 Experimental Evaluation

5.3.1 Experimental Setup

As the Homals Consensus defines the consensus cluster in the final phase of the consensus clustering framework, specified in (Soria and Garibaldi, 2010), we retain the first steps of the framework for the formation of the cluster ensemble on all three datasets. Then Homogeneity Analysis performed on this cluster ensemble treating each clustering as a categorical variable creates an agreement space upon which hierarchical clustering produces the final consensus partition.

In an effort to evaluate the impact of our proposed consensus approach we examine the resulting representation in the agreement space to understand whether Homals Consensus achieves to solve implicitly the correspondence problem without explicit relabeling. Then after the hierarchical clustering organises the clusters of the data, we examine whether the two proposed optimisation criteria of Silhouette and Silhouette of Agreement, as they are specified in Chapter 3 in section 3.3.3, achieve the optimal score. Finally, we compare our results with the clusterings that form the initial cluster ensemble and with the original consensus clustering framework, both visually and by the silhouette and average rand index scores. These are applied to all the three datasets described above.

5.3.2 Correspondence

After forming the cluster ensemble with the results of the K-means and PAM, it is essential to examine the level of agreement between those two algorithms and understand whether Homals Analysis can represent the underlying structure of associations between those clusters in a meaningful way that can lead to the creation of consensus clusters.

In Table 5.5 we can see the cross tabulation for the clusters returned by PAM and K-means performed on the User Knowledge Modelling dataset. K-means returned four clusters whereas PAM returned six. The difference in the number of clusters returned signifies the differences between the two clustering algorithms and provides the necessary diversity in the results. As we check the number of instances that co-occur in the same cluster between the clusterings produced by the two algorithms we can understand that there is a substantial agreement between the two algorithms. More accurately, it is safe to assume that the first cluster of K-means aligns with the first cluster of PAM as they both share their larger number of instances with each other and in the same way the third cluster of K-means can be aligned with the fourth cluster of PAM and the fourth cluster

Table 5.5: Cross tabulation for clustering results of PAM and K-means for User Knowledge Modelling Dataset

		PAM					
		1	**2**	**3**	**4**	**5**	**6**
K-means	**1**	**44**	4	8	0	3	0
	2	0	37	**48**	0	25	0
	3	13	0	4	**42**	1	0
	4	0	2	1	4	0	**22**

of K-means with the sixth cluster of PAM. Also there is a clear association between the second cluster of K-means and the clusters 2, 3 and 5 of PAM as the instances of the this cluster are divided in these three clusters of PAM. The alignment is highlighted with bold expression in Table 5.5. But even after this alignment there is a significant number of instances that are missclassified to other clusters, as 40 out of the 258 instances of the dataset do not follow this alignment. That means that the previous crude method of cluster alignment can result in significant loss of information.

In Fig 5.5 we can see that the relationships between clusters analysed from the cross tabulation are represented accordingly in the agreement space created by Homogeneity Analysis. In fact, the second cluster of K-means and the clusters 2,3 and 5 of PAM form a distinct cluster in the negative sides of both dimensions. The same applies to the associations between the fourth cluster of K-means and the sixth cluster of PAM that fall very close together in the lower right area of the representation that is specified by the positive values of the first dimension and the negative values of the second dimension, the association between the first clusters of K-means and PAM that are placed close together and finally the association between the third cluster of K-means and the fourth cluster of PAM that coincide almost in that same spot. It is important to point out that since Homogeneity Analysis defines the similarity of clusters as the frequencies of co-occurences between the cluster, the fact that the last four clusters are placed in a relatively close distance from each other means that they share objects between them. Checking Table 5.5 again we can see that the relationship of the four clusters uncovered by Homogeneity Analysis is caused by the fact that the first cluster of PAM also shares a significant number of objects with the third cluster of K-means. These four clusters placed close together can lead to the formation of one bigger cluster that classifies all the instances of these clusters.

The last observations highlight the strength of Homogeneity Analysis in creating a meaningful representation that reflects all the possible relations among the clusters returned by different clustering algorithms. In this agreement space, Homogeneity Analysis not only can verify the associations defined by a simple one to one alignment that occurs

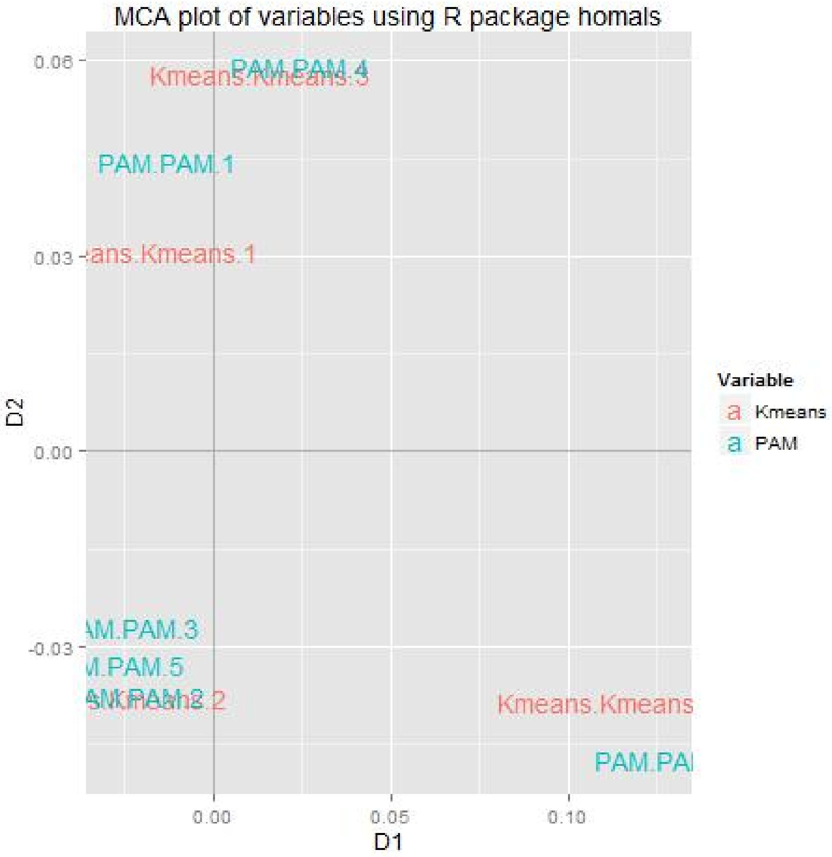


Figure 5.5: Plot of K-means and PAM clusters in the two dimensions of agreement space for User Knowledge Modelling Dataset

in the consensus clustering framework in (Soria and Garibaldi, 2010), but also manages to take into account all the relations between the clusters and represent them in a very meaningful way that can lead to the formation of well defined consensus clusters. In addition to this, it takes into account both types of relationships found in the cluster ensemble, either between data points or between clusters and seems to be able to solve the correspondence problem even when the clusters returned from the clustering algorithms differ both in number and in cluster membership.

Moving to the Seeds dataset we can see in Table 5.6 that the agreement between the clusters produced by K-means and PAM is particularly high. The agreement between the clusters of both algorithms is highlighted with bold numbers in Table 5.6. That is reasonable according to the biplot in Fig. 5.2, which revealed three well defined clusters. So, both K-means and PAM managed to return these three clusters that agree with each other to a great extent with only five instances being misclassified. These five instances are five members of first K-means cluster that are classified in the second cluster by PAM.

Table 5.6: Cross tabulation for clustering results of PAM and K-means for Seeds Dataset

		PAM		
		1	**2**	**3**
K-means	**1**	**67**	5	0
	2	0	**77**	0
	3	0	0	**61**

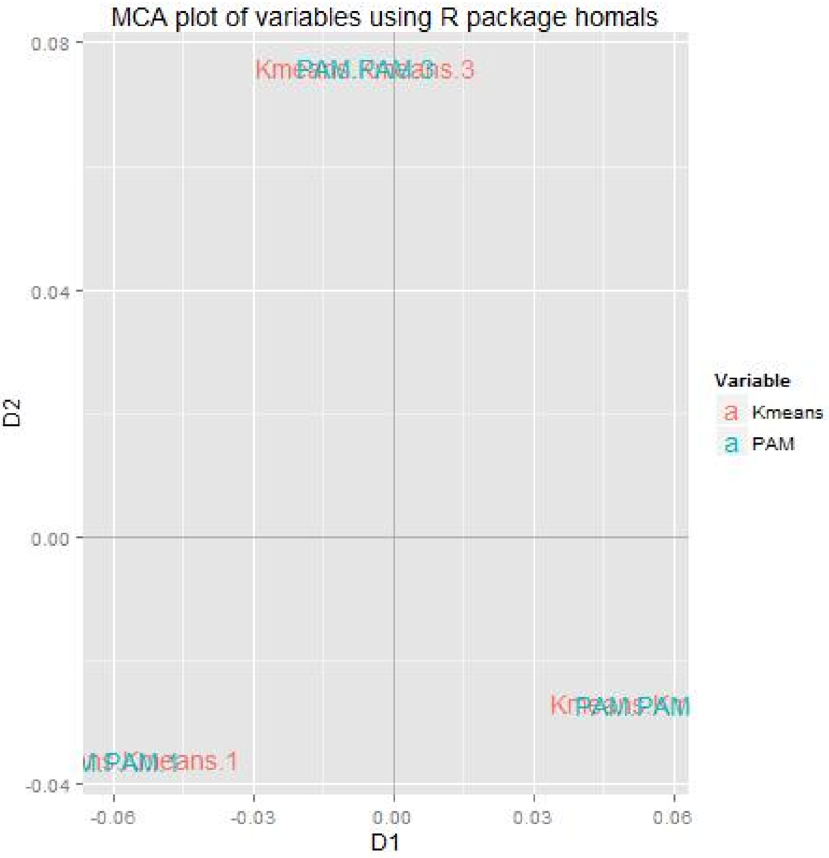


Figure 5.6: Plot of K-means and PAM clusters in the two dimensions of agreement space for Seeds Dataset

Inspecting the position of the clusters in the representation created by Homogeneity Analysis in Fig 5.6, we can observe the nearly perfect agreement of clusters as the corresponding clusters of K-means and PAM coincide in the same locations being as distant as possible from the other clusters at the same time. The small association between the first cluster of K-means and the second cluster of PAM causes the cluster 1 and 2 from K-means and PAM to be placed in the negative side of the second dimension of the graph and a bit closer to each other. As expected Homogeneity Analysis in this dataset proved to be able to capture the high level of agreement between the results of the two clustering algorithms and reflect it in the most emphatic way.

Table 5.7: Cross tabulation for clustering results of PAM and K-means for Wholesale Customers Dataset

		PAM						
		1	**2**	**3**	**4**	**5**	**6**	**7**
K-means	**1**	29	27	0	**129**	0	0	0
	2	0	**53**	0	0	34	0	0
	3	0	0	1	0	0	1	0
	4	0	0	0	0	0	**4**	0
	5	0	0	16	0	0	0	16
	6	0	0	0	0	0	0	**3**
	7	**77**	0	29	0	0	0	0
	8	0	0	0	0	**14**	7	0

Moving to the most difficult dataset of the three, the Wholesale Customers dataset, we can see in the cross tabulations of the cluster returned by K-means and PAM in the Table 5.6 that both clustering algorithms returned a large number of clusters, eight and seven respectively. This means that both algorithms failed to capture the underlying structure of a dataset with most of its data points residing in a very small and dense area and very few data points being further away as it is revealed in Fig. 5.3 and Fig 5.4. Despite the fact that a simple alignment strategy in this cross tabulations is still possible, as then non-zero elements are relatively few, there is a number of occasions that is very hard to decide. For example, it is controversial if one could align the third and the fifth clusters of K-means as the both divide equally the members they share with two clusters of PAM, cluster 3 and 6 for the third and 3 and 7 for the fifth. Other relations stand out a bit more clearly, like the one between the fourth cluster of K-means and the sixth cluster of PAM, the one between the eight cluster of K-means and the fifth cluster of PAM. Others are possible to define, but they still share a lot of common members with multiple clusters. All the relations that follow the alignment are expressed with bold fonts in the Table 5.6.

In Fig. 5.7 we can see that the resulting representation of relationships between the clusters in the cluster ensemble is not as clear as in the previous cases. While all the relationships are depicted quite clearly in the graph, the multiple relationships of strong nature among the majority of the clusters cause the graph to not produce distinct patterns of clusters. However it is very interesting to point out that two clusters seem to form according to the visual representation, with one being very close to the center of the representation where all the big clusters with many objects reside and the other one consisting of all the rest of the small clusters that are placed in the bottom region of the graph and represent the uncommon points in the dataset, possibly the outliers. The last cluster could be split in more than one clusters as it does not seem very well defined and distinct.

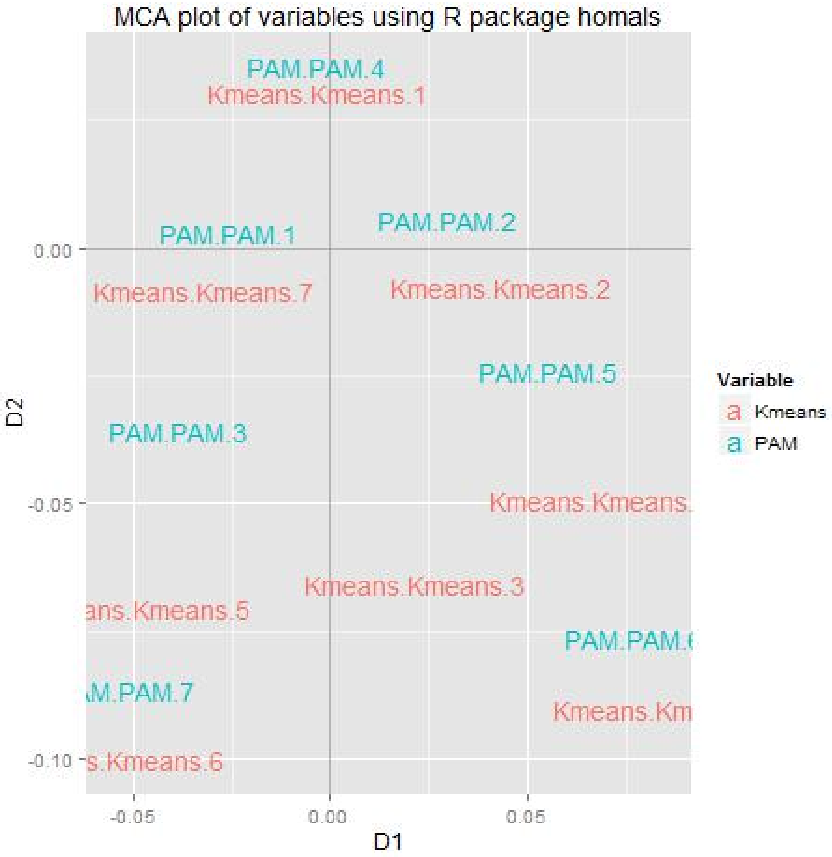


Figure 5.7: Plot of K-means and PAM clusters in the two dimensions of agreement space for Wholesale Customers Dataset

From these visual examinations of the agreement spaces the Homals creates for each cluster ensemble, we can conclude that Homogeneity Analysis manages to reflect all the associations between the clusters of different clusterings and the between the data points themselves. Not only it can identify the simple aligning by placing very close the clusters of the two algorithms that have a lot of common members but also it manages to represent the lesser relationships of the ensemble among clusters that share few common members, placing those clusters in a respectively close distance.

5.3.3 Determining Consensus Clusters

As Homogeneity Analysis creates this representation by examining both the similarity between clusters and the similarity between objects, a Hierarchical Clustering on this representation can uncover the consensus clusters. It is essential however to determine the optimal cut of the cluster tree (dendrogram) in order to obtain clusters that maximise the the Silhouette criterion in the original representation and the Silhouette criterion in the agreement space. The second criterion is based on our intuition that well defined, compact and distinct clusters in the agreement space will maximise the agreement between the consensus partition and the initial cluster ensemble. For that reason we will examine also which cut of the cluster tree maximises this agreement in order to make appropriate comparison. Finally we will also try to find which cut maximises the agreement with the ground truth since the latter is provided in these datasets, despite our disagreements that such a criterion should not be considered in evaluating clustering results.

In particular, we produce cuts of the cluster tree starting from two up to 20 clusters and we obtain the scores of each of the four criteria mentioned above for each of the 19 clusters returned. The silhouette criterion is calculated as it is specified in the equation 3.6 in the section 3.3.1, the agreement with the cluster ensemble by the mean Rand Index between the consensus partition and each of the clusterings of the cluster ensemble, and the agreement with ground truth by the Rand Index of the consensus partition with the classification provided by the dataset after these two are aligned.

In Fig. 5.8 we can see the level of disagreement between the four different criteria for the User Knowledge Modelling dataset. The Silhouette criterion is maximised in the case of the three clusters which is also the number of natural clusters that seem to form naturally inspecting the agreement space in Fig. 5.5. The Silhouette of agreement reaches the optimal point at the 15 clusters whereas the Agreement to the cluster ensemble identifies eight as the optimal number of clusters. Despite the fact that the Silhouette of agreement fails to agree with the criterion of maximising the similarity with the cluster ensemble, its optimal point achieves a score of agreement to the cluster ensemble that is very close to the optimal. The 17 clusters identified by the comparison to the ground truth shows how incompatible the discovery of the original classifications by clustering can be, especially when the classes are not well defined, like in the case of UNS. However we can observe that the value of similarity to the ground truth obtained by the point indicated by the Silhouette of agreement is again very close to the optimal.

Moving to the results of Seeds dataset in Fig. 5.9, the Silhouette criterion identifies two clusters as optimal, whereas the agreement to consensus three, which is the number of wheat varieties specified in the dataset. Again Silhouette of agreement fails to agree with the agreement to consensus but it still presents a point that is very close to the maximum agreement between the consensus partition and the cluster ensemble. It also manages to agree with the agreement to the ground truth which similarly identifies four clusters as optimal.

In Fig. 5.10 and Fig. 5.11 we can see that Silhoutte identifies two as the optimal number of clusters. The agreement to ground truth is incomprehensible in the case of the Region class, a result that complies with the visual representation of the class in the biplot of

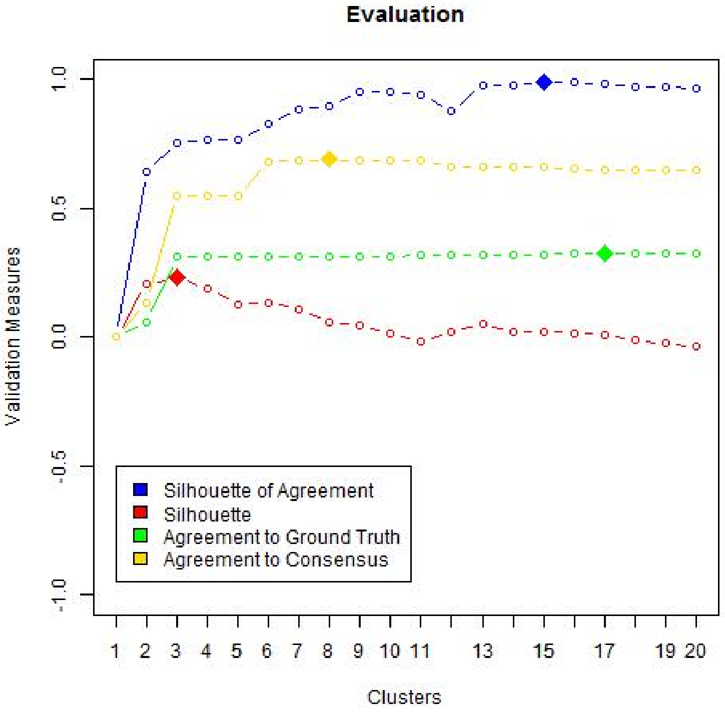


FIGURE 5.8: Plot of the four criteria against the number of consensus clusters obtained by hierarchical clustering for the User Knowledge Modelling dataset

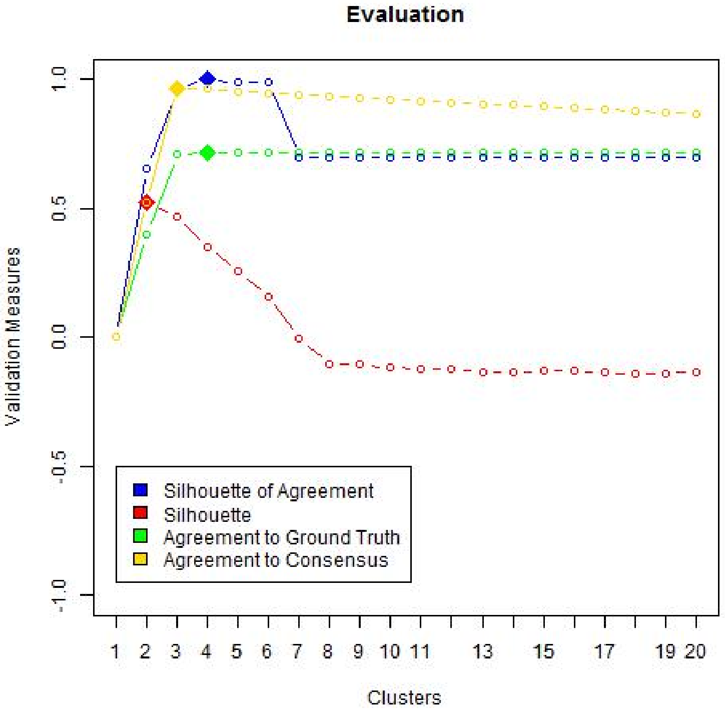


FIGURE 5.9: Plot of the four criteria against the number of consensus clusters obtained by hierarchical clustering for the Seeds dataset

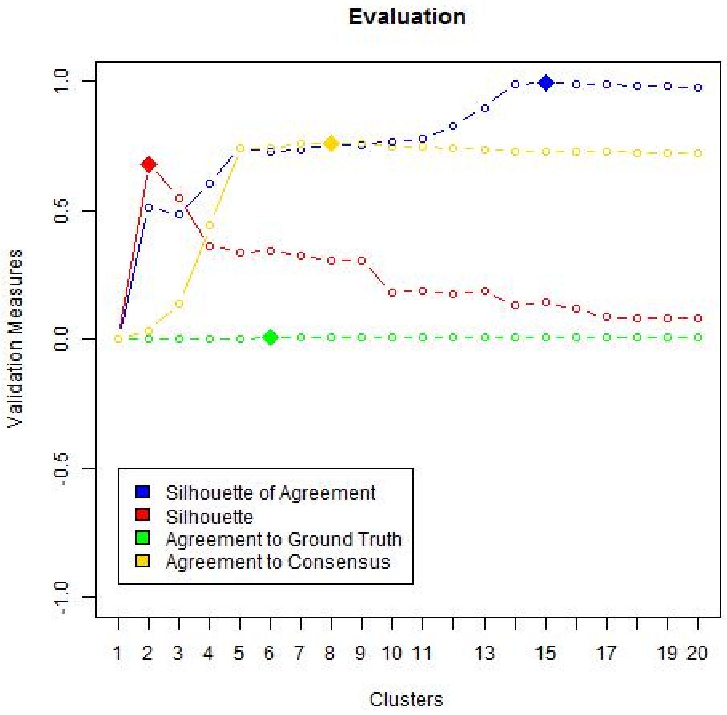


Figure 5.10: Plot of the four criteria against the number of consensus clusters obtained by hierarchical clustering for the Wholesale Customers dataset for the class of Region

Fig 5.3. However, it reaches its maximum in the 16 clusters for the case of Channel class. Again the Silhouette of agreement finds 15 clusters as optimal, which is a clustering that provides scores of agreement with the consensus and with original classification that are very close to being optimal.

In all cases the Silhouette criterion finds a number of well defined clusters that is different from the number of clusters specified by the original classifications. This is justified in two of these cases as the biplots in an earlier section showed that these classes are not compact or well separated, whereas in the case of the Seeds dataset where the clusters seem to be well defined, the few instances that lie beyond the boundaries of the clusters cause a small deterioration in the Silhouette of the three clusters. A two cluster classification seems to overcome this issue.

The Silhouette of agreement was not able to maximise the agreement of the consensus partition with the initial cluster ensemble, but in all cases gave values that are very close to optimal. This small difference from the optimal value is caused by the small size of the cluster ensemble. We can observe that the number of clusters that maximises the silhouette criterion in the agreement space is exactly equal to non-zero elements of all the cross tabulations. That means that in the agreement space Homogeneity Analysis places together all the members of each of these non-zero elements. This can be seen

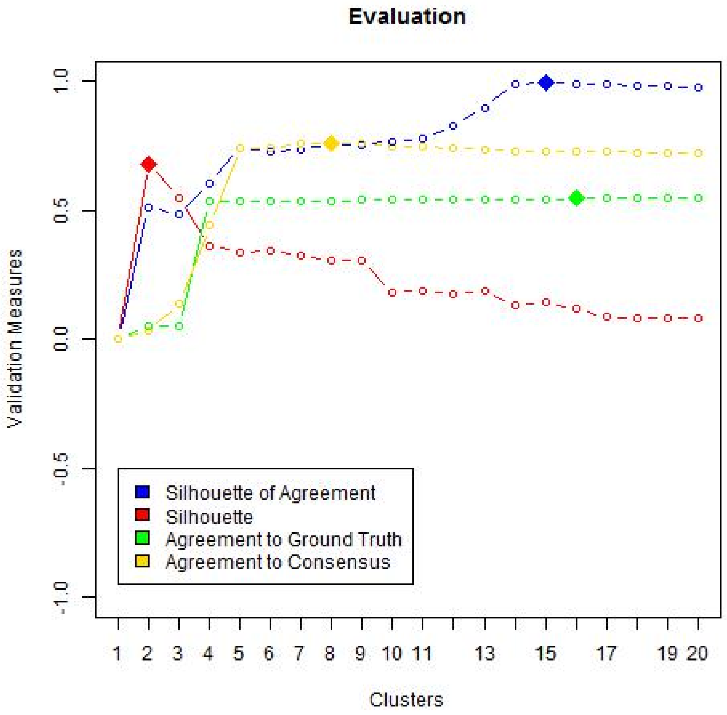


FIGURE 5.11: Plot of the four criteria against the number of consensus clusters obtained by hierarchical clustering for the Wholesale Customers dataset for the class of Channel

for example in the case of the Seeds dataset in the joint plot of Fig. 5.12, where the five members of the first cluster of K-means fall in the second cluster of PAM and therefore they are placed in the middle of their distance. This causes the Silhouette criterion in the agreement space to be maximises with four clusters and not the three which is indicated by the agreement to the consensus. If we used a bigger cluster ensemble this phenomenon would not be that distinct and it could lead to a more meaningful organisation of the disagreements among the clusterings of the ensemble. As desirable as bigger ensemble might be, this simplistic modelling of the disagreement remains particularly valuable as it reveals the patterns or occasions where the two clustering algorithms fail to agree. So the proposed two tier structure specified by the two different cuts on the hierarchical based on the silhouette values in the original representation and in the agreement space, can still present significant and meaningful information.

5.3.4 Comparison of Results

Comparing the silhouette values of the Homals Consensus against the values of the initial K-means and PAM clusterings that form the cluster ensemble and the traditional method of consensus clustering defined in (Soria and Garibaldi, 2010), in Table 5.8 we can see that our method achieves to maximise the *Compactness* and *Separation* of

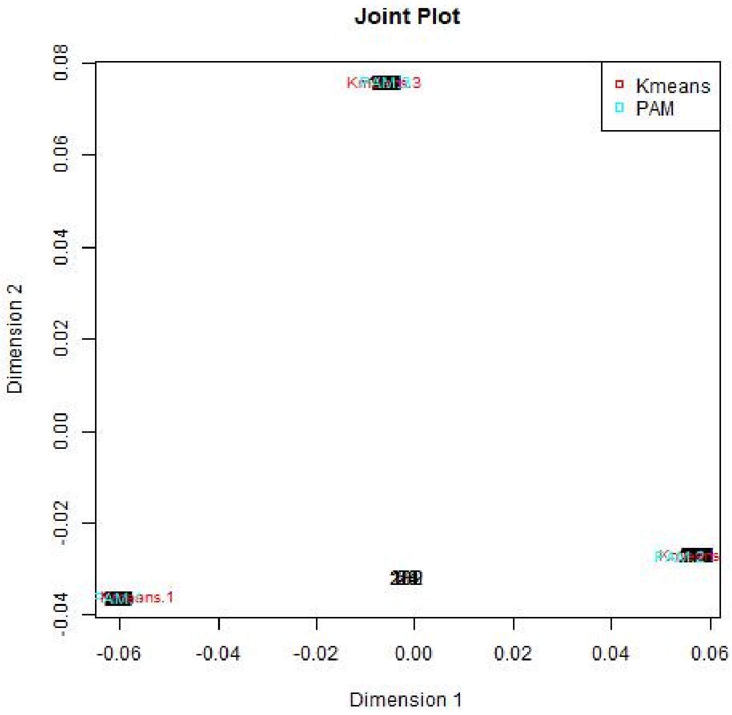


FIGURE 5.12: Plot of the four criteria against the number of consensus clusters obtained by hierarchical clustering for the Seeds dataset

the initial cluster ensemble and produce more well defined clusters than the traditional consensus approach in all three cases. The difference in the performance in the silhouette criterion becomes even greater in the difficult dataset of Wholesale Consumers, where our Homals-based consensus solution manages to produce clusters that almost double the best silhouette scores of the best clustering it compares to. On the other hand the traditional consensus approach of the original consensus clustering framework produces clusters that are less well defined compared to the clusters produced by K-means and PAM that form the initial cluster ensemble, a fact that confirms the significant loss of information that this simplistic approach imposes on the clustering.

The same can be concluded when we compare our Homals-based consensus solution to the original consensus method of the original consensus clustering framework in terms of agreement to the initial cluster ensemble in Table 5.9. Our proposed consensus solution achieves a better agreement to the cluster ensemble than the traditional approach, although in two of the cases the level of agreement is comparable. In fact, in the case of Seeds dataset both approaches achieve the same score of agreement and in the case of User Knowledge Modelling dataset our solution is only slightly better. However, in the case of Wholesale Customers our method achieves a much greater level of agreement.

The superior performance of our Homals-based consensus solution is confirmed visually

Table 5.8: Silhouette values of clusterings

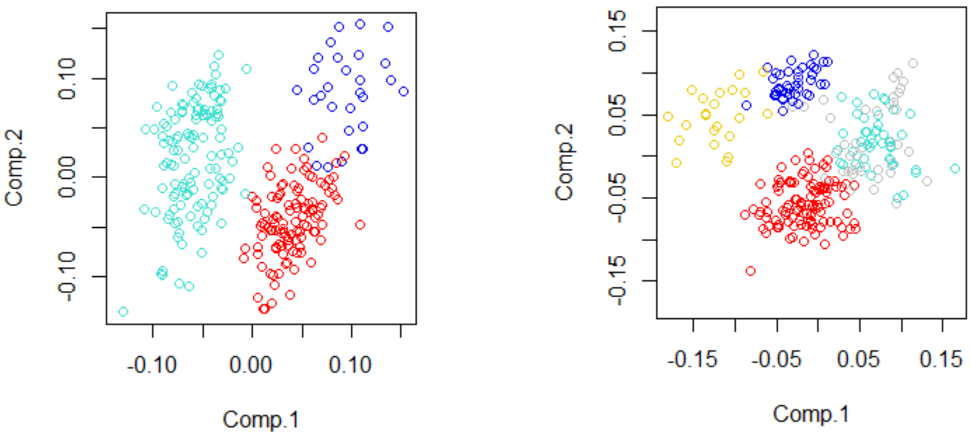
	Silhouette
User Knowledge Modelling	
K-means	0.205042966
PAM	0.185297385
Original Consensus	0.170237
Homals Consensus	**0.2319754**
Seeds	
K-means	0.4719337
PAM	0.4681391
Original Consensus	0.3519592
Homals Consensus	**0.5233003**
Wholesale Customers	
K-means	0.3464764
PAM	0.2817748
Original Consensus	0.1310384
Homals Consensus	**0.6804235**

Table 5.9: Agreement to cluster ensembles of the consensus partitions

	Average Rand Index to cluster ensemble
User Knowledge Modelling	
Original Consensus	0.656948
Homals Consensus	**0.6578813**
Seeds	
Original Consensus	0.9626916
Homals Consensus	0.9626916
Wholesale Customers	
Original Consensus	0.546224
Homals Consensus	**0.730767**

when we inspect the biplots of the returned clusters. In Fig. 5.13 we can see that the three returned clusters defined from Homals Consensus are more compact and well separated than the four clusters returned by the Original Consensus and the four UNS classes depicted in Fig 5.1. It seems that Homals Consensus manages to take into account the boundary instances that cause the clustering to be less well defined and incorporate it in a more clear clustering, whereas in the case of the original consensus approach these instances are part of the disagreement between the two algorithms and they are left unclassified (grey). The three resulting clusters seem to capture the natural underlying structure of the User Knowledge Modelling Dataset.

Inspecting the same biplots in the Fig. 5.14 for the Seeds dataset, it seems that the original consensus approach distinguish one clear cluster (blue) and it produces two more relatively distinct clusters that are not entirely separated. Homals consensus on



(a) Homals Consensus

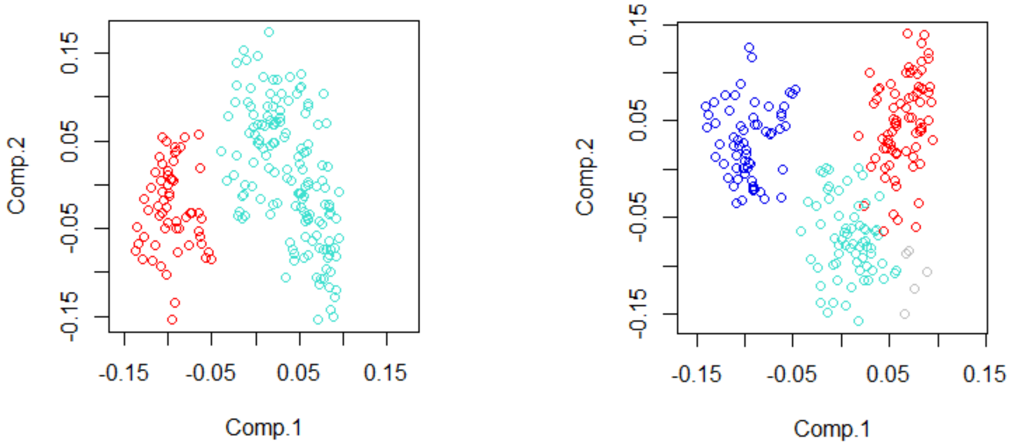
(b) Original Consensus

FIGURE 5.13: Biplots of consensus clustering solutions on the User Knowledge Modelling dataset

the other hand produces two well separated clusters that seem to represent more clearly the natural underlying structure of the data than the original three classes of the dataset.

Finally, inspecting the biplots of the consensus approaches in the Fig. 5.15 for the Wholesale Customers, reveals clearly the strength of the Homals Consensus. In a very dense and difficult for clustering dataset, Homals Consenus manages to cluster together almost all the outliers of the data in one cluster while it classifies the rest of the data as one big cluster, something that sees reasonable concerning the shape of the data in the biplot. The original consensus approach on the other hand produces multiple clusters of the data that do not seem to produce any patterns in the data and leaves a large number of clusters unclassified.

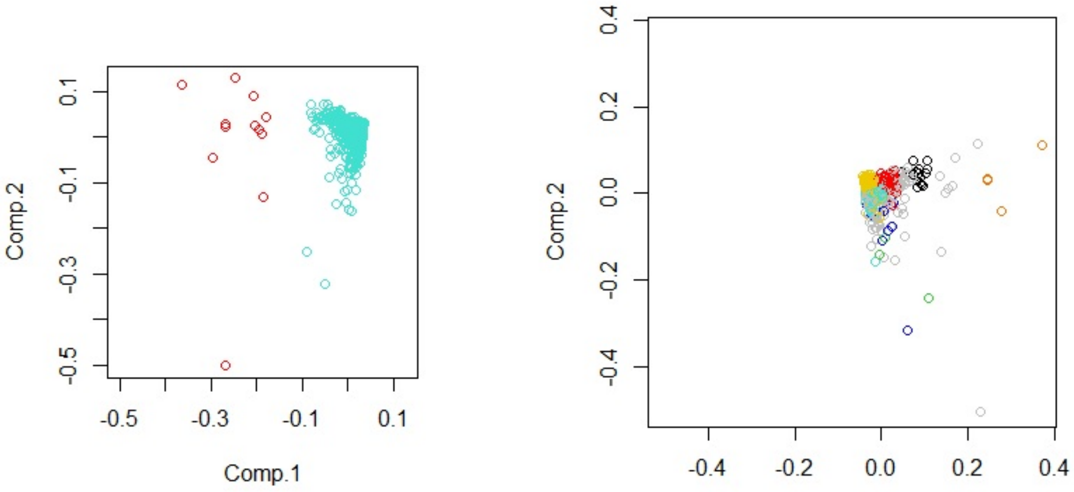
As a conclusion our method manages to model the agreement between clusterings in a similar way with the original consensus approach, while at the same time manages to model the disagreement between the clustering algorithms and it uses this modelling to classify unclassified objects and find the clusters that maximise the silhouette criterion in the original representation, resulting in compact and distinct clusters. At the same time important information that can be noticed in a potential organisation of the disagreements between clusterings reveals interesting patterns regarding objects of controversial nature. This information is being captured by the second tier of our hierarchical cluster that maximises the Silhouette of the agreement that also forms clusters for the objects that co-occur. However, given the small size of this ensemble, these clusters cannot reveal systematic patterns of disagreement, but a bigger ensemble can create more variety in the disagreement, which might be used to reveal systematic patterns that can lead to



(a) Homals Consensus

(b) Original Consensus

FIGURE 5.14: Biplots of consensus clustering solutions on the Seeds dataset



(a) Homals Consensus

(b) Original Consensus

FIGURE 5.15: Biplots of consensus clustering solutions on the Wholesale Consumers dataset

a better understanding of the reasons behind the disagreement and the data themselves. Thus, our two-tier clustering structure can provide two views on the data. One that reveals well defined clusters and classifies the data in more natural sense and a second that produces more fuzzy clusters not that distinct that model the agreement and the disagreement of the cluster ensemble.

Its superiority over single clustering algorithms and the original consensus is more clear in very difficult datasets, which makes it appropriate for real world data applications. It capitalises the strength of Homals to model both types of relationships and it manages to represent them accordingly, taking into account all the agreements and disagreements of the cluster ensemble at the same time and solving the correspondence problem implicitly, without the need of relabeling, and it can work with clusterings that are not comparable. Finally the second tier of the hierarchy provides a more fuzzy view on the data ignoring the notions of Compactness and Separation that dominate the first tier. But since both clustering are provided by different cuts of the same cluster tree, one tier is always a partition of the other tier.

5.4 Applying Homals Consensus for Behavioural Profiling in the DebtTrack Dataset

The DebtTrack Survey (Gathergood, 2012) is a quarterly repeated cross-section survey of a representative sample of UK households. It is been carried out by the market research company YouGov and it focuses on the consumer-credit information of households. It consists of 2084 households and in its 85 questions it covers household demographics, labour market information, income and balance sheet details. The consumer credit data are particularly detailed, providing information about the number and type of consumer credit products, outstanding balances for each debt product , monthly payments and whether they are one month or three months in arrears on the product, as well as the value of arrears. The most interesting aspect of this dataset is that it contains 28 questions that measure psychological characteristics like *Impulsivity* and *Risk aversion*, which gives us the opportunity to explore the multifaceted "nature" of consumer indebtedness on a complete dataset by the standards specified in (Stone and Maury, 2006)

The variables of interest from demographics and financial attributes can be seen in Table 5.10. The survey contains a much bigger number of demographic and financial questions, but they had a big number of uncertain answers that is why they were excluded from our analysis. By uncertain answers we refer to answers that were given to questions, like "I don't know" or "Prefer not to answer", which while they cannot be regarded as

Table 5.10: Demographic and Financial attributes of DebtTrack

Attribute	**Unknown Answers**
Demographic	
Marital_Status	17
Emp_Status	32
Age_Sex	
Social_Grade	
Education	32
Guardian	
Financial	
Household Income	492
Income	548
Liquid Assets	340
House_Status	30
Insurance	

missing values they definitely do not provide any valuable information. In the variables included in Table 5.10 we can see that some of the financial attributes exhibit a big number of uncertain answers that can be regarded as noisy data. It is worth noted that the *Guardian* and *Insurance* variables are actually groups of smaller Boolean variables, five for the case of *Guardian* and 11 for the *Insurance*.

Before applying the Homals Consensus on this dataset a lot of instances containing missing values were removed, resulting in a dataset of 1253 households. Factor Analysis and Homogeneity analysis transformed the data producing two demographic dimensions, two financial dimensions and five psychological factors.

On these transformed behavioural data, Homals Consensus produced six profiles of consumers classified in two types of consumers in its two-tier hierarchical structure which can be seen in Table 5.11. In order to obtain the characterisation of the clusters for the categorical variables, the Lift and Confidence rules were utilised with the threshold being set to 1.5 and 0.85 respectively. The numerical variables were discretised to High, Average and Medium in a similar fashion described in Chapter 3, and the mode of each cluster was examined.

As we can see in the resulting descriptions of the clusters, the two types are mainly differentiated by the likelihood to have children, whereas the two big clusters that fall into Type 1 (cluster 1 and 2) describe two very different profiles of consumers based on differences on demographic variables like *age*, *marital status*, *employment status* and *education*. At the same time differences are expressed in financial variables like *house status* and *income*. The most interesting behavioural profile though, is emerging in the fifth cluster, where psychological factors like *organisational responsibility*, *risk management belief* and *planful saving* exhibit significant differences in their scores when

Table 5.11: Description of consumer profiles in DebtTrack

		Size	Description
Type 1		940	Households without children
Cluster	**1**	456	Older retired individuals, potentially widowed with big certainty with lower education and income that own their property and have no children or high levels of debt
	2	427	Young f/t professionals mainly single or co-living with higher education that tend to rent their residence or live in rent-free housing
	4	36	Middle aged individuals with lower education that have Self Invested Personal Pension
	5	18	Middle aged females working full time low scores in organisational responsibility but high scores risk in management belief and planful saving with average education that have liquid assets, a big variety of insurances, higher income and exhibit higher level of unsecured debt while the are under mortgage
	6	3	Lower income lower social class middle aged individuals with lower income and education that are under mortgage and exhibit low scores in risk management belief but high scores in planful saving
Type 2		313	Households with children
Cluster	**3**	313	Part time employed middle aged individuals that have children and critical illness insurance under mortgage or renting on housing association

compared to the whole population. This small and unigue cluster also produces the first distinction in the gender of consumers. The described females exhibit low scores in *organisational responsibility*, a fact that in combination with the high income they earn potentially explains the big concentration of unsecured debt or the numerous insurances they acquired. High scores of *Planful saving* do not seem to prevent the accumulation of debt as someone would expect, as they manage to keep the unsecured debt to the usual levels in cluster 6. Finally, the third cluster, which defines exclusively the second type of consumers, also describes a big cluster of consumers that have a distinct behavioural profile with a lot of differences in economic and demographic variables.

As meaningful behavioural profiles are needed to prove the goodness of our clustering

Table 5.12: Performance of Homals Consensus in DebtTrack

	Silhouette
K-means	0.18622
PAM	0.18197
Original Consensus	0.16538
Homals Consensus	**0.209091**

for the purposes of Consumer Debt Analyisis according to the ideas of (Hennig and Liao, 2013; Caruana et al., 2006), we conclude that our two-tier structure clustering revealed important behavioural profiles of consumers. In more detail, our clustering identified three big profiles of consumers that reflect standard socio-economic classes of consumers in different life stages, expressing differences in economic behaviours and in the expression of demographis as the *life-cycle theory* (Webley and Nyhus, 2001) would dictate. Apart from different socio-economic behaviours, our clustering also produces smaller clusters that reveal psychological patterns. The most interesting of these clusters isolates a small group of consumers with high *income* and poor *organisational responsibility* that exhibit high levels of debt. This confirms the commonly mentioned association between income and debt (Stone and Maury, 2006; Ottaviani and Vandone, 2011; Kim and DeVaney, 2001; Lea et al., 1993). While the psychological factor of *organisational responsibility* also involves items that capture *money management practices*, as it can be seen in the Appendix A, we can also identify the link mentioned in literature between poor *money management practices* and the accumulation of debt (Lea et al., 1995). However, we should be careful not to generalise the patterns of behaviours observed in this cluster as they characterise only a small group of consumers. Several other interesting psychological patterns also emerge in other small clusters as well, but their actual impact on Consumer Debt Analysis necessitates further research.

Focusing on traditional evaluation criteria, Homals Consensus as before achieves a bigger silhouette score than the cluster ensemble and the original consensus partition, which can be seen in Table 5.12.

As a conclusion, the application of Homals Consensus on DebtTrack dataset results in the creation of meaningful and important behavioural profiles of consumers that can provide a better understanding of the landscape of consumer indebtedness. Expected socio-economic patterns of consumer behaviour are verified in the bigger clusters, whereas novel and interesting patterns of psychological behaviour emerge in the smaller clusters. In more detail, the successfully modelled agreements and disagreements of the cluster ensemble leads to the identification of two distinct, well separated and compact types of consumers that are separated based on the presence of children, whereas the smaller clusters that emerge from the agreement of ensemble reveals interesting patterns of behaviour that possibly transcend the ideas of *Compactness* and *Separation*.

5.5 Conclusions

In this chapter we enhanced the process of *Behavioural Profiling* of consumers for the purposes of Consumer Debt Analysis by improving the existing consensus clustering framework. The utilisation of Homogeneity Analysis for implementing the consensus function of the framework proved to be particularly beneficial as it managed to provide a representation that models all the relationships between the clusters and between the data points in the cluster ensemble and enabled the formation of well defined profiles that are consistent with the initial cluster ensemble at the same time. In comparison with other consensus solutions in literature, our method achieved to solve implicitly the correspondence problem without relabeling and it showed to provide meaningful consensus partitions even in cases where the clusterings forming the cluster ensemble did not provide comparable clusters.

When applied to a socio-economic dataset, Homals Consensus produced types of consumers that follow the ideas of *Compactness* and *Separation* and six classes of consumers describe patterns of consumer behaviour of great interest for the purposes of Consumer Debt Analysis, providing a better understanding of the “nature” of consumer indebtedness.

The enhanced *Behavioural Profiling* method presented here completed the process of *Behavioural Extraction* for Consumer Debt Analysis. Together with the beneficial behavioural data it passes on to the next stage all the extracted insight to help modelling answer responsively the important research of this field.

Chapter 6

Evaluating TopDNN Regression for Behavioural Modelling of Consumer Indebtedness

6.1 Introduction

In this chapter we propose a novel Neural Network approach that incorporates the knowledge extracted from exploratory analysis in the topology of its network in order to develop a strong and accurate behavioural model to predict the level of indebtedness of consumers. Neural Networks, coming from Data Mining, possess a strong predictive ability as they can deal successfully with the complexities real world data exhibit. At the same time, methods of Sensitivity Analysis can provide the desired interpretability assessing the variable importance within a Neural Network, which is of great significance for economic and social sciences. In combination with their flexibility in designing their architecture, Neural Networks offer the means to implement the process of Behavioural Modelling for the purposes of Consumer Debt Analysis.

However, before we proceed with the process of Behavioural Modelling of consumer indebtedness on a socio-economic dataset (CCCS) it is essential to establish the impact Neural Networks and Data Mining models in general can have on modelling consumer indebtedness and whether they have the potential to replace the traditional statistical modelling that dominates the area, answering the third research question stated in the first chapter in 1.4.1. Also, it is of great importance to establish the beneficial role behavioural data and profiles can play in this modelling in order to understand whether they should be included in it, providing answers on the fifth research question in 1.4.1.

Then, the performance of TopDNN, a neural network that designs its topology of neurons based on the behavioural profiles and exploratory search on the data, is assessed.

6.2 The Consumer Credit Counelling Service (CCCS) Dataset

6.2.1 Description

The CCCS dataset, introduced in (Disney and Gathergood, 2009), is a socio-economic crossectional dataset based on the data provided by the Consumer Credit Counseling Service. Its 58 attributes contain information about approximately 70000 clients who contacted the service between the years 2004 and 2008 in order to require advice about how they can overcome their debts. The information was gathered through interviews when each client first contacted the service and it varies from standard demographics to financial details, aggregated spending in categories and debt details. The attributes of interest for the purpose of Consumer Debt Analysis are limited to demographics, expenditure and financial attributes, as they can be seen in Table 6.1 together with their description.

6.2.2 Behavioural Transformations

Like any other real world dataset, CCCS contains noise and outliers, while at the same time it suffers from high dimensionality. In order to tackle the aforementioned difficulties a series of transformations steps were performed in an earlier work described in Chapter 4, which proved to be beneficial for the unsupervised learning approach used in this dataset. More precisely, Homogeneity Analysis (Homals) (De Leeuw and Mair, 2009) was utilised in order to map the categorical demographic data, significant attributes concerning the Consumer Debt Analysis, into two-dimensional coordinates together with a Factor Analysis on the financial attributes and a clustering on the correlation of the spending items. These transformations reduced the dimensionality to more compact attributes, removed noise and outliers, provided a sense of interpretability and improved the quality of the clustering. More importantly, these transformations managed to capture behavioural elements in the data and represent it in a behavioural space which enables clustering approaches to discover behavioural profiles of consumers that describe different economic and spending behavious. A summary of the transformations can be seen in Fig. 6.1, whereas the new nine transformed attributes include two spatial coordinates that discriminate the demographic variables, three financial factors that summarise different economic behaviours expressed in the financial attributes

Table 6.1: Description of CCCS attributes

Attribute	*Description*
pid	individual identifier
Demographics	
age	age of person
mstat	marital status
empstat	employment status
male	sex of person
hstatus	housing status
ndep	number of dependants in household
nadults	number of adults in household
Financial Attributes	
udebt	total value of unsecured debt
mortdebt	total value of mortgage debt
hvalue	total value all housing owned
finasset	total value of financial assets
carvalue	resale value of car
income	total monthly income
Expenditure	
clothing	total monthly spending on clothing
travel	total monthly spending on travel
food	total monthly spending on food
services	total monthly spending on utilities
housing	total monthly spending on housing
motoring	total monthly spending on motoring
leisure	total monthly spending on leisure
priority	total monthly spending on priority debt
sundries	total monthly spending on sundries
sempspend	total monthly self-employed spending
other	total other spending
Debt Details	
ndebtitems	number of debt items

and four behavioural spending clusters that are characterised by spending in Necessity, Household, Excessive and Leisure.

6.2.3 Behavioural Profiles of consumers

The behavioural transformations were proved to be useful for the clustering of a random sample of 10000 debtors from the CCCS dataset that managed to classify 8370 debtors in seven behavioural profiles with distinct characteristics in Chapter 4. The characteristics of these profiles can be seen in Table 6.2, which also includes the 1630 debtors that remained unclassified.

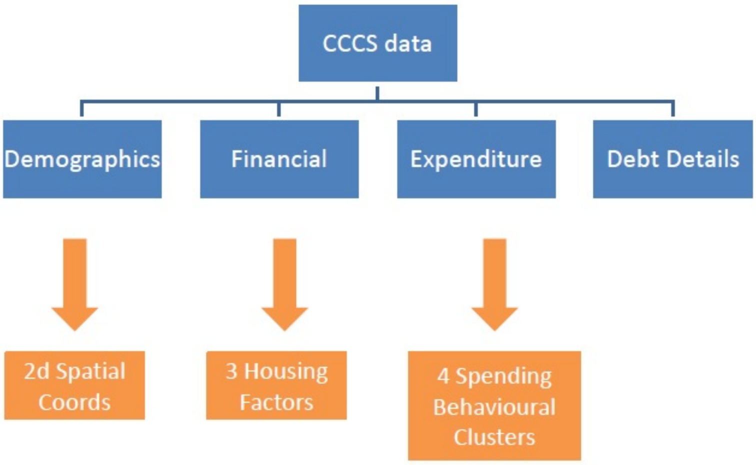


FIGURE 6.1: Transformations of CCCS attributes

TABLE 6.2: Description of behavioural profiles of consumers

Class	*Size*	*Characterisation*
1	2301	Young single unemployed debtors with low income, debt and spending
2	1440	Average Income-spending- debt debtors usually p/t employed and cohabiting with high spending in clothing and food
3	1033	High Income-Debt-Spending Debtors, usually self-employed and with expensive houses
4	948	Older and retired debtors with average income-spending and low levels of debt
5	507	High Income-Debt-Spending Debtors with cheap houses
6	1588	Average Income-spending-debt debtors usually p/t employed but single, divorced or separated
7	553	Old and retired Debtors with low income, debt and spending, other marital status
8	1630	Unclassified

6.3 Experimental Setup

The aim of this work is to evaluate the performance of Neural Networks as a regression model that can predict the amount of unsecured debts (*udebt*) a debtor in the CCCS dataset has by using the rest of the variables as predictors. For this reason we compare its performance against different regression models with different characteristics, like Linear Regression and Random Forests' regression. Furthermore we check whether a series of behavioural transformations we performed in Chapter 4 and behavioural profiles of consumers we provided in the same work can improve the performance of the regression, so that they be incorporated in the final Neural Network we aim to develop.

Since these models try to optimise different criteria and they are internally validated on different measures when they are fitted into data, we needed to test all these models under a common framework. So we use the 10-fold cross validation as the method to compare the different models and we select RMSE and R^2 as the evaluation criteria. 10-fold cross validation is a standard method for evaluating models in Supervised Learning and it also allows Neural Networks to avoid data overfitting providing more representative results for their case.

R^2 measures the percentage of variance that is explained by the model and it is a standardised measure taking values from 0 to 1 with 1 being a perfect fit. The Root Mean Square Error (RMSE) measures the difference between the predicted values from the model and the actual values. It is defined as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{obs,i} - y_{model}, i)^2}{n}} \tag{6.1}$$

where n is the number of observations, $y_{obs,i}$ is the observed value of the target variable in observation i and $y_{model,i}$ is the calculated value of the the target variable observation i. The best model will minimise the RMSE.

For model training we use a random sample of 10000 debtors from the CCCS dataset, a subset of the dataset that contains no missing values and on which we already have performed the transformations described in Chapter 4. All the models are built in R using the *caret* package `https://cran.r-project.org/web/packages/caret/index.html`. For Linear Regression we calculate the weights using the Ordinary Least Squares (OLS) algorithm and for Random Forests we create 500 trees and initialise the number of potential candidates for a node split as m/3, where m equals the number of predictors. For Neural Networks the initial weights are randomly assigned and a hidden layer is chosen. In order to choose the optimal number of hidden nodes, we produce ten neural

Table 6.3: Description of Datasets

Dataset	*Attributes*
A	Original CCCS variables
B	Transformed Variables
C	Original CCCS variables and clustering classification
D	Transformed variables and clustering classification

networks for each case with the number of neurons varying from 1 to 10. Because of the necessary parameter tuning of neural networks we perform a two way 10-fold cross validation. The first validation is to pick out the optimal parameters of the neural networks and the second cross validation to evaluate the performance of the final model. For this reason a sample 1000 records is separated from the training data and kept as the test set upon which the 2nd 10 cross fold validation evaluates the final model. The final model is selected for the number of neurons that minimise RMSE. We also use both back propagation and Resilient back propagation for making the appropriate comparisons. All models are built using both the actual data and the transformed data and the classification is introduced as an additional categorical variable. For all of the above we have to create four different datasets that all the regression models will be built upon. These necessary datasets, which are used to establish the contribution of the behavioural transformation and behavioural profiles provided by clustering together with the performance of the regression models, are summarised in Table 6.3.

Finally, we construct a Neural Network based on our intuition to utilise behavioural profiles and Factor Analysis for designing its topology and we evaluate its performance on the same datasets.

6.4 Results

6.4.1 Comparison of Regression Models

From a quick look at Table 6.4, which presents the performance of the models built on the four datasets, with the numbers in the brackets indicating the optimal number of neurons in the hidden layer found, we can see that Neural Networks and Random Forests clearly outperform Linear Regression on almost all datasets, with the only exception being the Neural Network models built on the C dataset. In all the rest of the cases Neural Networks and Random Forests produce smaller RMSE and bigger R^2. In addition to this, we can identify the beneficial nature of the behavioural transformations performed on CCCS attributes, since all four different regression models seem to improve their performance when they are built on the transformed data. More specifically, the models built on datasets containing the transformed attributes (B and D) reduce the RMSE and increase R^2 when compared with models built on datasets A and C respectively. Especially in the cases of Neural Networks trained with resilient back propagation and the Random Forests' regression, the improvement in the performance is significantly big, reducing the RMSE to around 0.06 for the case of Random Forests and to around 0.05 for the case of Neural Networks trained with resilient back propagation. Similarly R^2 was raised to around 0.5 for Random Forests and around 0.6 for the Neural Networks trained with resilient back propagation. For the cases of Linear Regression and Neural Networks trained with back propagation the improvement was significant but much smaller.

On the other hand, the contribution of the behavioural profiles provided from clustering remains less clear. It manages to provide a rather small improvement in the Linear Regression, but it decreases the performance of Random Forests when it is combined with the original variables. In the case of Neural Networks it seems that it leaves unchanged the performance of the models trained with back propagation but it is beneficial for resilient back propagation Neural Networks. This can be seen when you compare models built on datasets C and D that contain the additional categorical variable of classification with the models build on datasets A and B respectively. Interestingly, the Random Forest Regression model build on dataset C has an increased RMSE and a bigger proportion of variance explained at the same time.

Looking at the performance of the models, the best performance was achieved by the resilient back propagation models followed closely by the Random Forests Regression, whereas the performance of back propagation Neural Networks and Linear Regression remained comparable with the first one being better though. The model that exhibits the minimum RMSE and the biggest R^2 is the resilient back propagation Neural Network built on the transformed variables together with the classification of debtors (D). This verified the argument of (Günther and Fritsch, 2010) that resilient back propagation is more suitable for regression purposes. It also strengthens the argument regarding the potential of using Neural Networks in applications of economics, traditionally dominated by statistical models. Data Mining and Computational Intelligence in a broader sense introduce a holistic approach in order to extract knowledge from the data as it offers a large number of tools to pre-process the data, techniques to explore the relationships with unsupervised learning algorithms, like clustering and accurate models to be used for prediction, that when combined in a sophisticated framework can build models which achieve impressive results. In our case this was verified not only by the better performance of Neural Networks and Random Forests but also from the beneficial nature of the behavioural transformations performed as part of pre-processing the data, which

Table 6.4: Results of Regression Models

Dataset	**RMSE**	R^2	**RMSESD**	R^2 **SD**
Linear Regression				
A	0,078	0,235	0,003	0,026
B	0,0731	0,328	0,004	0,042
C	0,0769	0,257	0,004	0,033
D	0,0727	0,336	0,004	0,056
Random Forests				
A	0,0727	0,293	0,003	0,035
B	0,0672	0,572	0,004	0,045
C	0,0741	0,311	0,004	0,026
D	0,0626	0,5	0,002	0,044
Neural Networks back propagation				
A (4 Neurons)	0,0781	0,238	0,002	0,048
B (2 Neurons)	0,0671	0,442	0,003	0,067
C (6 Neurons)	0,0781	0,236	0,004	0,048
D (2 Neurons)	0,0672	0,442	0,003	0,051
Neural Networks resilient back propagation				
A (2 Neurons)	0,0764	0,240	0,001	0,042
B (3 Neurons)	0,0542	0,605	0,003	0,052
C (2 Neurons)	0,0749	0,23	0,005	0,132
D (4 Neurons)	0,0522	0,647	0,002	0,029

improved all the models. Despite the fact that the contribution of the classification of debtors returned from clustering was not beneficial for all the cases tested, it managed to provide a small improvement in the performance of the Linear Regression models but its most significant impact is shown in the performance of resilient back propagation Neural Networks, especially when it is combined with the transformed data.

Proceeding with the examination the R^2 achieved by the models, we notice that the best model has the ability to explain approximately two times the proportion of variance explained by the best Linear Regression model. When these models are compared to the ones found in literature, Linear Regression performance seems to be comparable to the one presented in (Kim and DeVaney, 2001) but better than the rest of the models whereas the performance of the Neural Networks trained by resilient back propagation is significantly higher and can only be compared with the Linear Regression model in (Livingstone and Lunt, 1992), although this performance was not considered representative enough due to the limited number of observations the model was built upon. In fact, a more realistic value of R^2 for this model given in (Stone and Maury, 2006) was arround 30%, meaning that the performance of the best model found here is still significantly higher than the ones found in literature.

Table 6.5: Factor Analysis on transformed variables

	Factor1	**Factor2**	**Factor3**
x	0.298	0.487	
y			
housingfactor		0.385	-0.477
financialfactor1	0.280	0.574	0.766
financialfactor2	0.118	0.792	
Necessity.Spending	0.983		0.167
Household.Spending	0.728	0.286	0.232
Excessive.Spending	0.217		0.128
Leisure.Spending			

6.4.2 Results for TopDNN

Since the beneficial nature of the behavioural data is experimentally verified in all cases we are encouraged to test our novel approach on dataset B using resilient back propagation. Therefore we begin with performing a Factor Analysis on the attributes of the dataset B. Three was the number of factors that is found to be optimal for summarising the nine attributes of the dataset after examining the scree plot of the eigenvalues and performing a parallel analysis. In the scree plot the eigenvalues of the correlation matrix are plotted in order of descending values. The last substantial drop in the graph indicates the number of factors. In parallel analysis the same eigenvalues are compared to eigenvalues derived from random data. The number of cases where they are bigger indicate the number of factors in the model. These methods for determining the number of factors are two of the most popular and effective methods in the bibliography and they are preferred from others as dictated in (Fabrigar et al., 1999). Interestingly, three is also the number of neurons that was found to be optimal for the case of building Neural Networks on dataset B using resilient back propagation, indicating the agreement between two different techniques in designing the network topology of a neural network. The three factors and their loadings can be seen in Table 6.5.

Then we train two Neural Networks, one with one hidden layer of three neurons and one with an additional hidden layer of eight neurons representing the classes of debtors, as shown in Table 6.2 in order to test in a stepwise fashion the two main ideas of our approach. Again we utilise 10-fold cross validation and RMSE and R^2 as evaluation criteria in order to get comparable results with the rest of the experiments. The results can be seen in Table 6.6. We can see that designing the network topology according to the knowledge extracted by Factor Analysis is beneficial for the performance of the model. The inclusion of the hidden layer of three nodes as dictated by Factor Analysis

Table 6.6: Results of TopDNN

	RMSE	R^2	**RMSESD**	R^2 **SD**
TopDNN				
NN with Factor Analysis	0,054	0,623	0,006	0,048
NN with Factor Analysis and clustering	0,053	0,633	0,001	0,074

improves the performance of the model when compared with the Neural Network build on dataset B, but has worse performance from the best model of the previous experiments. The additional layer of eight neurons on the other hand improves the performance of the model in a further extend but it is still lower than the performance of the resilient back propagation Neural Networks built on D dataset. However, the performance of TopDNN using Factor Analusis and clustering is comparable to the best performance seen for these data. This verifies our intuition that the flexibility Neural Networks offer in designing their topology can be exploited properly in order to include knowledge that stems from the unsupervised learning approaches performed on the data. Thus our model manages to achieve a very good performance, indicating the ability of Neural Networks to incorporate in their modelling results from previous steps of the Data Mining process that produced important patterns of consumer behaviour.

6.5 Analysis of Regression Models

6.5.1 Analysing Linear Regression

The low performance of Linear Regression comparing to the Data Mining methods can be explained easily if we take a careful look at the diagnostics plots of the best linear model for the CCCS dataset in Fig. 6.2. The plot of the residuals against the fitted values indicates that the error terms are not independent and that their variance is not constant as they are not randomly scattered throughout the zero point. Besides this, the normal probability reveals that the error terms are not normally distributed as there is a strong deviation from the line with two big curves in the beginning and the end of the plot. Furthermore in Fig. 6.3, where the partial residuals plot for *Housing Factor, FinanceFactor1* and *FinanceFactor2* are depicted, we can identify the non-linear relationship these input variables have with the response variable. Partial residuals are utilised instead of normal residuals because in a multiple regression they account for the effect the rest of the independent variables have on this relationship. These observations come in contrast with almost all the assumptions of linear regression, degrading the quality of the linear model. A series of transformations on the response

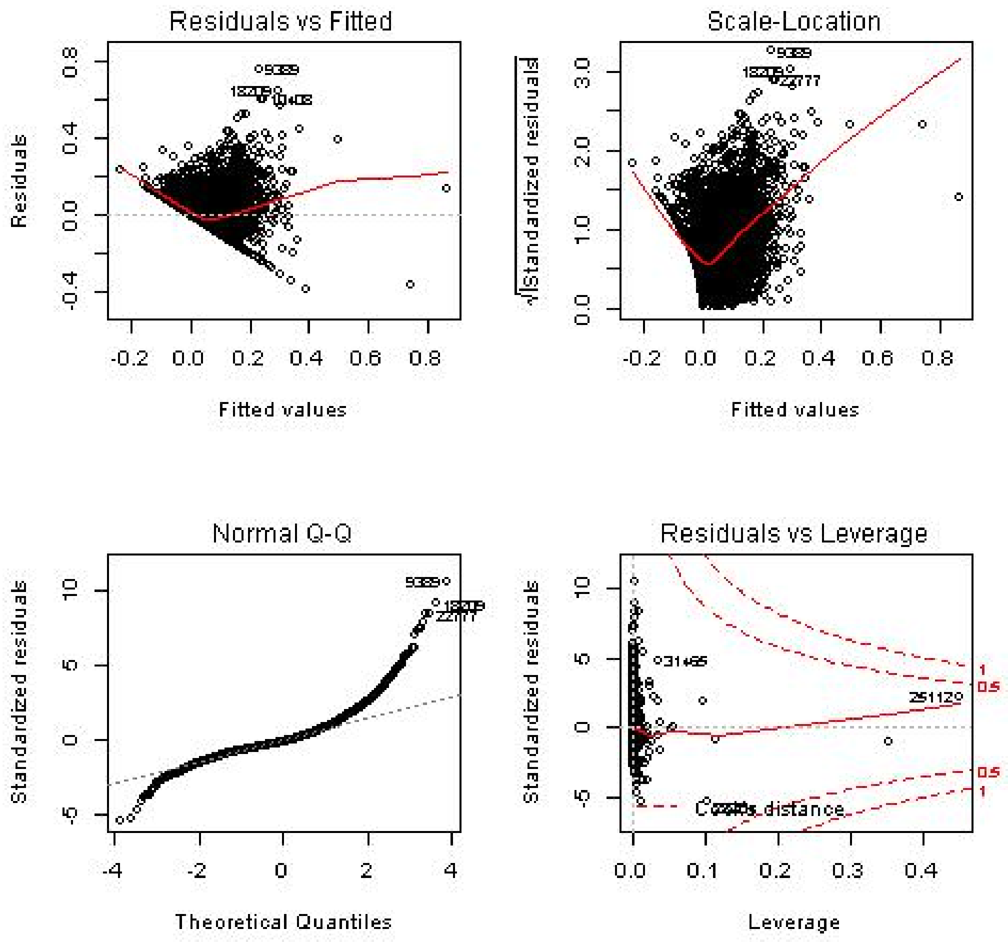


Figure 6.2: Diagnostic plots of Linear Regression model built on Dataset D

variable or the explanatory variables, following established techniques like power and log transformations as dictated in (Draper et al., 1966) were not able to improve the quality of the model as the R^2 remained low and the assumptions were still violated.

Trying to identify the significant predictors of the linear model we take a closer look at the significant tests performed on the coefficients of each predictor. This can be seen in Table 6.7, which shows the importance of the variables included in dataset D, the dataset upon which the best linear model is built. Surprisingly we can see that all the attributes are considered statistically significant by the model, with the only exception being some classes of the classification of debtors. However, some of them like *HousingFactor*,*FinanceFactor1*,*FinanceFactor2* and y (coordinate) seem more important than the others based on their high t-value, while almost all the behavioural clusters are among the weakest predictors but still statistically significant. Also *FinanceFactor2* and clusters 4, 6, 7 are the only significant variables that have a negative impact on accumulating unsecured debt.

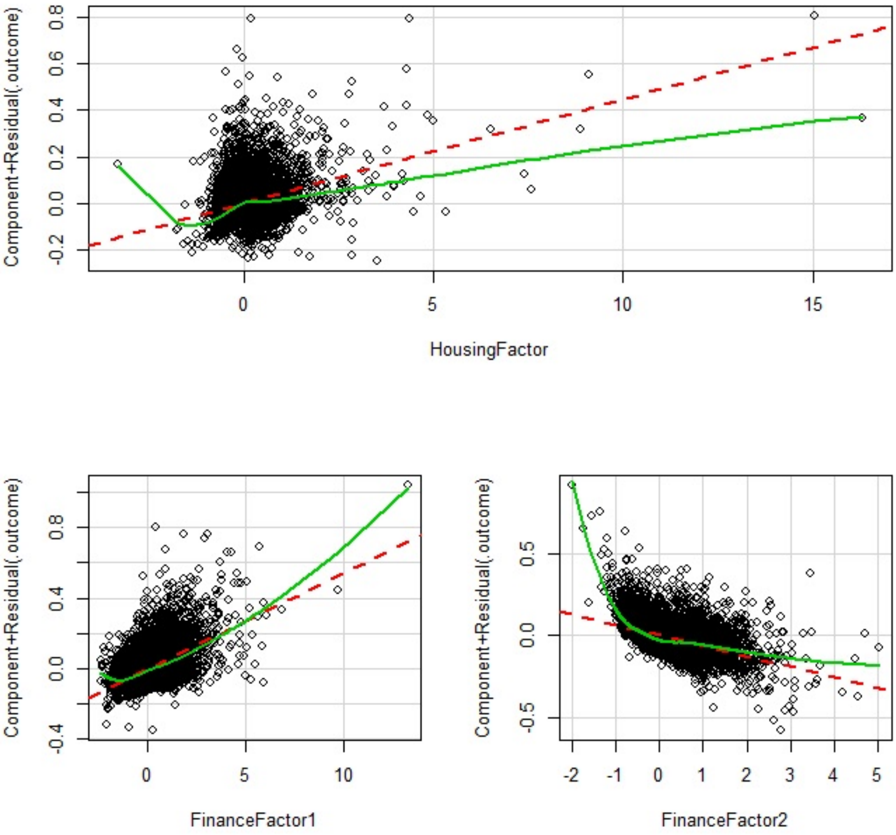


Figure 6.3: Partial Residuals plot of *Housing Factor*

Table 6.7: Importance of attributes in Linear Regression model

	t value	$Pr(>\|t\|)$
(Intercept)	27.171	2e-16 ***
x	5.486	4.22e-08 ***
y	11.316	2e-16 ***
HousingFactor	29.498	2e-16 ***
FinanceFactor1	45.801	2e-16 ***
FinanceFactor2	-42.815	2e-16 ***
Necessity.Spending	2.108	0.03505 *
Household.Spending	8.139	4.46e-16 ***
Excessive.Spending	2.101	0.03569 *
Leisure.Spending	-2.839	0.00454 **
newexperiment15.clustering1	-1.634	0.10230
newexperiment15.clustering2	-1.343	0.17941
newexperiment15.clustering3	2.322	0.02023 *
newexperiment15.clustering4	-8.683	2e-16 ***
newexperiment15.clustering5	-1.804	0.07125 .
newexperiment15.clustering6	-7.039	2.07e-12 ***
newexperiment15.clustering7	-4.990	6.14e-07 ***

Table 6.8: Variable Importance of Random Forest Regression in dataset C

	% MSE Increase	**Node Impurity**
x	4,46E+02	5,E+09
y	2,44E+02	5,E+09
HousingFactor	1,83E+03	9,E+09
FinanceFactor1	1,56E+04	2,E+10
FinancialFactor2	1,13E+04	2,E+10
Necessity.Spending	1,25E+03	6,E+09
Household.Spending	1,22E+03	7,E+09
Excessive.Spending	5,36E+01	2,E+09
Leisure.Spending	5,33E-03	0.009396788

6.5.2 Analysing Random Forests

Random Forests also have the ability to measure the importance of variables. In particular, they have two measures of importance. The first is the total decrease in node impurities from splitting on the variable, averaged over all trees and the second is the mean increase in Mean Squared Error (MSE) after permuting each variable. The idea is that if the variable is not important then rearranging the values of that variable will not degrade prediction accuracy. Thus the bigger the increase in the error the bigger the importance of the variable, and the bigger the increase in impurity the bigger the importance. In Table 6.8 we can see the values of these measures for the Random Forests' regression model built on dataset C which demonstrated the best performance. Here the significant variables identified from Random Forests signify the importance of the financial factors which are considered the most significant predictors, followed by two types of spending (Necessity and Household) and the *HousingFactor*. Among the weakest predictors are the spatial coordinates of the Demographic Variables x and y and the *Excessive Spending*. The fact that *leisure spending* is shown to be a very weak predictor with very low values in both measurements comes in contrast with the Linear Regression modelling that considered the variable as significant. Spatial coordinates x and y are also identified as very significant predictors by the linear model whereas here are considered weak, a fact that indicates another big difference between the two models. The aforementioned differences observed from the two models in detecting the significance of the predictors raises questions whether they can be attributed to the weak modelling of Linear Regression or not representative measurements.

6.5.3 Analysing TopDNN

A simple first step for understanding the modelling of Neural Networks is by examining the plot of the network. The plot of the Neural Network built with the TopDNN approach, which exhibited the best results, can be seen in Fig. 6.4. The weights of the edges have been omitted for clarity but the lines have been modified accordingly to depict the magnitude and the sign of the weights, with thinner lines representing small weights and thicker lines larger weights, whereas the grey colour of the lines indicate a negative sign and the black a positive. We can notice that the interpretation of Neural Networks is not a trivial task, especially when the network is complicated. That is their main drawback comparing to Linear Regression and Random Forests, which have mechanisms to assess the variable importance of their models. However, tracing the very thick black lines of the plot we can immediately detect the strong influence *FinancialFactor1* has on the final outcome as it influences heavily the first neuron of the first hidden layer, which influenced strongly the sixth neuron of the second hidden layer which belongs to the four neurons of the second layer that affect moderately the final outcome. This relationship between the *FinancialFactor1* and *udebt* cannot be quantified or defined from this plot but it can be signified.

Applying the profile method of the nine predictors we can get a better understanding of the contribution of each predictor in accumulating unsecured debt. In Fig 6.5. we can see the profiles of the input variables. These depict how the response variable is altered by increasing each input variable from its 1st percentile to its last. As we can see most of the predictors remain stable at zero indicating that the value of unsecured debt is not affected by the differences in their input. However there are three variables that seem to affect the level of unsecured debt in a great extent. In particular, *HousingFactor* and *FinanceFactor1* seem to cause a non-linear increase in the unsecured debt as they increase their values. As we can see until the 70th percentile approximately in the case of *FinanceFactor1* the response variable retains a linear relationship with the predictor, but then the response variable increases suddenly. The same happens in the case of *HousingFactor*, with the only difference being that the sudden increase happens close to the last percentile. Finally, *FinanceFactor2* causes the response variable to decrease as this increases, exhibiting the negative relationhsip it has with the response variable.

Comparing the results with the other models we can see that in a similar way with the Random Forests the importance of all the financial factors, *HousingFactor*, *FinanceFactor1* and *FinanceFactor2* were identified, but now the exact relationship between *FinanceFactor2* and unsecured debt has been demonstrated revealing its strength and negative sign. However, the Profile method revealed that the rest of the input variables have no effect on the outcome. This contradicts the findings of Random Forests, which

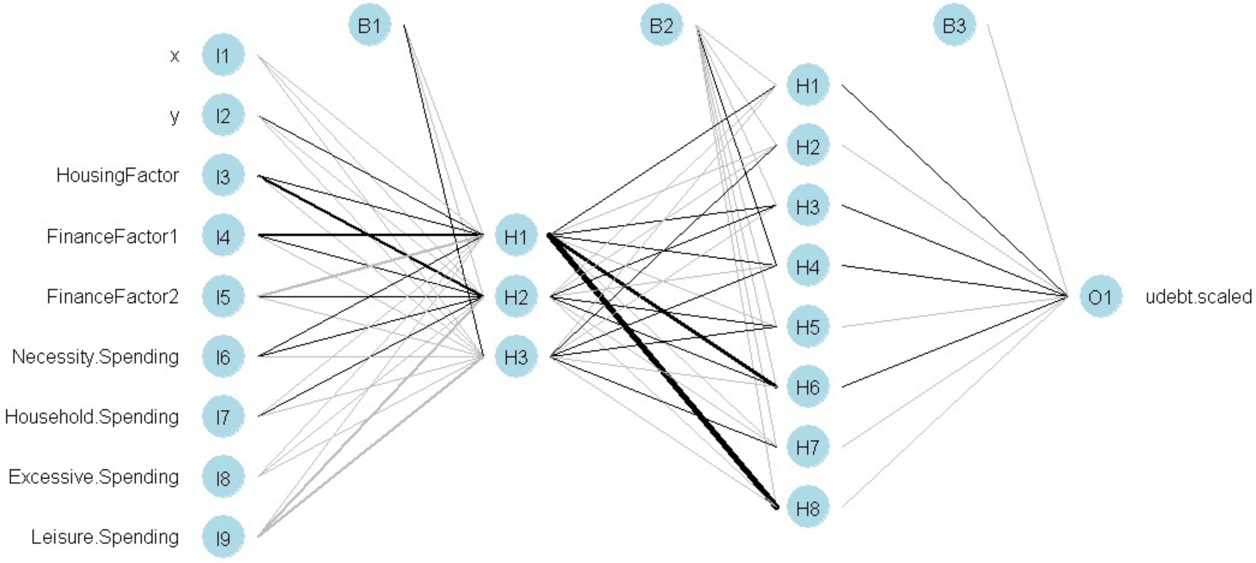


Figure 6.4: TopDNN with two hidden layers. The first one represents the number of factors and the second one the number of classes of debtors

assign some importance to some more attributes and especially the findings of the Linear Regression, which considers almost all the predictors as significant.

All models exhibited differences in the identification of significant predictors, but they all agree on the three financial factors. As Linear Regression exhibits the biggest differences with the other two Data Mining methods we can understand the challenge of identifying predictors based on weak modelling. That is why Neural Networks pose as suitable models for Consumer Debt Analysis, since their strong predictive accuracy can now be combined with a better understanding of their modelling, offering reliable results that can provide deeper knowledge to this complex problem. In this work the role of financial factors on predicting the level of unsecured debt was signified by Neural Networks.

6.6 Conclusions

In this work we tried to construct an accurate regression model for the level of debt prediction, a significant task for Consumer Debt Analysis, utilising a widely used Data Mining method, Neural Networks. For this reason we compared their performance against Linear Regression and Random Forests. Our results show that Neural Networks clearly outperform Linear Regression whereas Random Forests achieve comparable performance. The results also proved that all regression models can benefit from the behavioural transformations and from the unsupervised learning approaches on the data, if

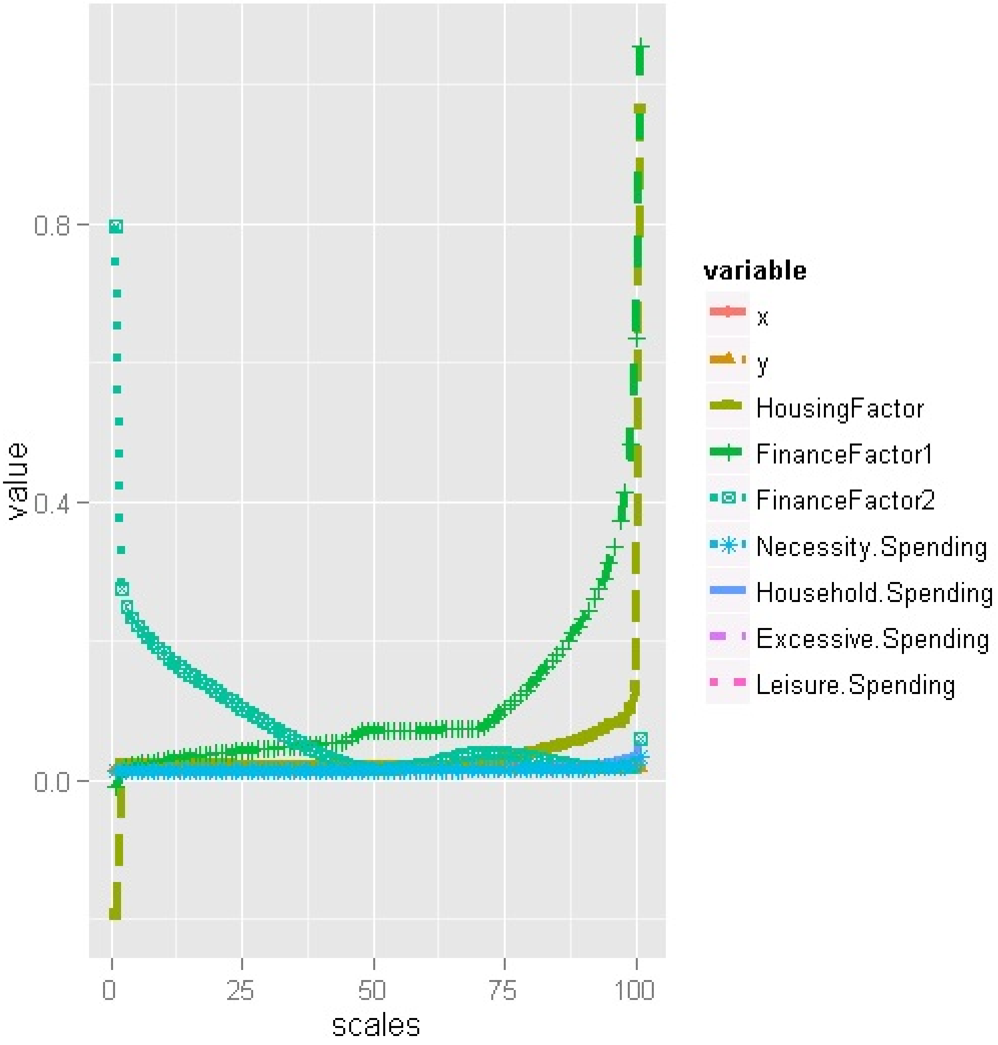


Figure 6.5: Sensitivity Analysis on TopDNN

these are incorporated properly in the data. Trying the latter we devised a novel method for designing the topology of the Neural Networks utilising information that stems from Factor Analysis and clustering performed on the data. Topology Defined Neural Networks (TopDNN), as our method is named, achieves a very good performance of the models that is comparable to the optimal performance witnessed in our experiments and signifies the ability Neural Networks possess in adopting in their design results from previous steps of explanatory research conducted on the dataset. Finally, we applied a method of measuring the variable importance in Neural Networks called profile method that reveals the strongest predictors of unsecured debt together with the nature of their relationship. This proves the value of Neural Networks for real world applications as their strong predictive modelling can be combined with a better understanding of the nature of the problem.

Therefore, our work forms a data mining framework with the pre-processing of data, clustering to uncover important relationships, a strong and reliable regression model that is suitable for the purposes of Consumer Data Analysis and provides transparent results. This framework exhibits much better performance than the existing statistical methods that dominate the field of economics (namely Linear Regression) and it signifies the further utilisation of Data Mining models for economic sciences and Consumer Debt Analysis. In other words, the developed framework achieves a step towards behavioural modelling as it demonstrates the ability to incorporate significant extracted behavioural knowledge in its procedure enhancing its performance, while at the same time it maintains the desired quality of transparent results that can help the process of Knowledge Discovery in Consumer Debt Analysis.

Chapter 7

Assessing the importance of Psychological Factors for Consumer Debt Analysis using Classification

7.1 Introduction

In this chapter we explore the multifaceted nature of consumer indebtedness within a Data Mining framework, trying to provide an answer to our hypothesis regarding the potential of psychological information to achieve Knowledge Discovery for the purposes of Consumer Debt Analysis. The beneficial role of behavioural data and profiles was established in Chapters 4 and 5, where it was demonstrated how to extract psychological information from a socio-economic dataset and how this information can provide a deeper understanding of complex "nature" of consumer indebtedness. In Chapter 6, it was shown that Data Mining provides a strong and powerful modelling approach that takes into consideration a series of transformations and exploratory models that successfully mine and represent behaviours in such manner that can enhance the process of Knowledge Discovery, replacing the traditional statistical modelling that dominates economic and social sciences. Its reliability and transparency on the other hand can provide the necessary means to assess the impact of the extracted psychological information upon modelling consumer indebtedness.

Therefore, drawing upon the complete nature of the DebtTrack survey, a very detailed socio-economic dataset that contains items that measure psychological factors like *Impulsivity* and *Risk Aversion*, we are given the opportunity to implement a holistic Data Mining model of financial indebtedness that takes into account a series of diverse factors simultaneously and it can provide reliable answers regarding the contribution of each of these factors. In more detail, we pre-process the data to reduce dimensionality and remove the noisy data as well as to extract and verify the psychological factors of the data, building new behavioural data with our established methods. Afterwards, we utilise three Data Mining methods from different families of predictive modelling, namely Logistic Regression, Random Forests, and Neural Networks to assess the contribution of psychological factors in Consumer Debt Analysis in an extensive number of experiments.

7.2 The DebtTrack Survey

7.2.1 General Overview

The DebtTrack Survey (Gathergood, 2012) is a quarterly repeated cross-section survey of a representative sample of UK households. It has been carried out by the market research company YouGov and it focuses on the consumer-credit information of households. It consists of 2084 households and in its 85 questions it covers household demographics, labour market information, income and balance sheet details. The consumer credit data are particularly detailed, providing information about the number and type of consumer credit products, outstanding balances for each debt product, monthly payments and whether they are one month or three months in arrears on the product as well as the value of arrears. The most interesting aspect of this dataset is that it contains 28 questions that measure psychological characteristics like *Impulsivity* and *Risk aversion*, which gives us the opportunity to explore the multifaceted “nature” of consumer indebtedness on a complete dataset by the standards specified in (Stone and Maury, 2006).

The variables of interest from demographics and financial attributes can be seen in Table 7.1. The survey contains a much larger number of demographic and financial questions but they had a big number of uncertain answers, and that is why they were excluded from our analysis. By uncertain answers we refer to answers that were given to questions, like “I don’t know” or “Prefer not to answer”, which while they cannot be regarded as missing values they definitely do not provide any valuable information. In the variables included in Table 7.1 we can see that some of the financial attributes exhibit a big number of uncertain answers that can be regarded as noisy data. It is

Table 7.1: Demographic and financial attributes of DebtTrack

Attribute	**Unknown Answers**
Demographic	
Marital_Status	17
Emp_Status	32
Age_Sex	
Social_Grade	
Education	32
Guardian	
financial	
Household Income	492
Income	548
Liquid Assets	340
House_Status	30
Insurance	

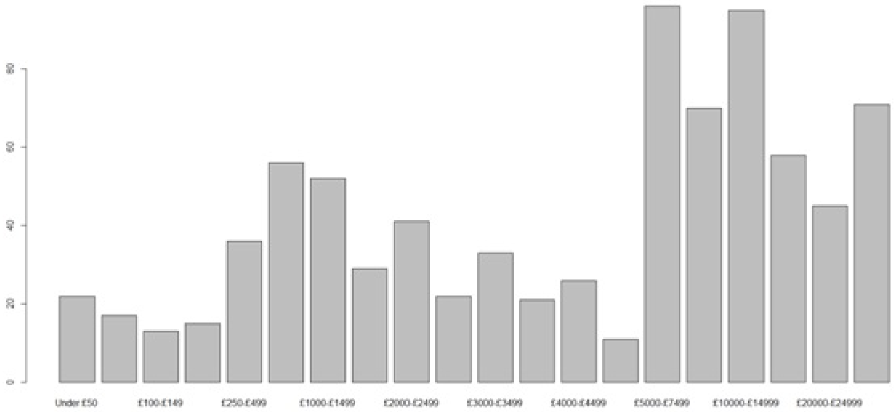


Figure 7.1: Histogram of Unsecured Debt

worth to point out that the *Guardian* and *Insurance* variables are actually groups of smaller boolean variables, five for the case of *Guardian* and 11 for the *Insurance*.

Consumer indebtedness can be modelled in very different ways, as it can be seen throughout the literature. In this work we are going to use the unsecured debt as dependent variable of the models. The choice is based on the well established relationship between unsecured debt and *Impulsivity* (Ottaviani and Vandone, 2011), which is a psychological factor that is verified in our dataset. Unsecured debt is a categorical variable with 21 levels. As we can see in Fig 7.1. the class variable is heavily imbalanced and most of the classes have very few members so they will be under-represented. For this reason we decided to move to a two-class classification (In Debt, No Debt) which is more balanced with 652 consumers being in debt and 601 not in debt and therefore we focus on building models that can discriminate the debtors from non-debtors, a fundamental research question of Consumer Debt Analysis (Wang et al., 2011). As the level of debt

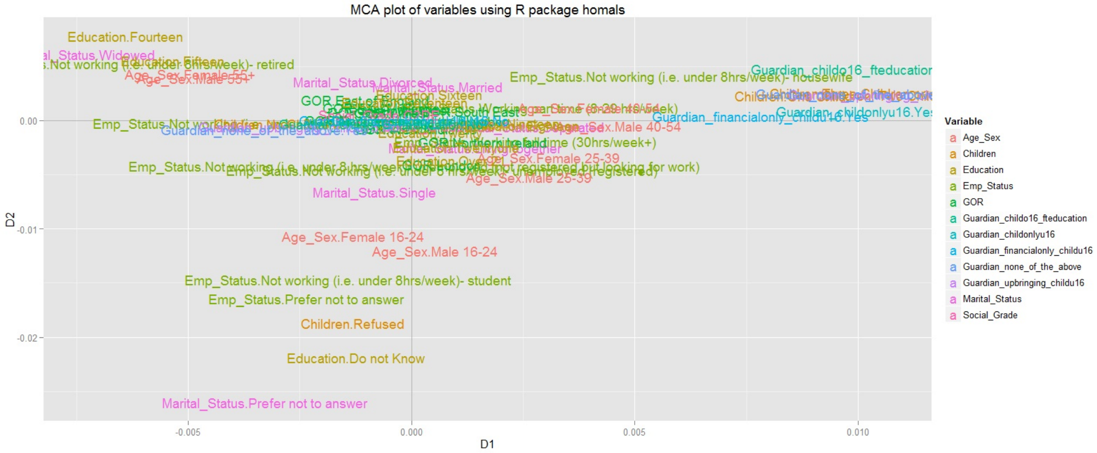


FIGURE 7.2: Category plots of demographic variables in the 2-dimensional space created by Homogeneity Analysis

prediction is another important direction of this research we also attempt to build a three-class classification model (No Debt, High, Low) for predicting the level of debt, but that is also imbalanced. No Debt class has 600 members whereas the other two class around 300. Imbalanced classes is a traditional problem in Data Mining (Chawla, 2005; Mollineda et al., 2007) but while a lot of sophisticated methods have been presented in literature they are usually not suitable for a multiclass model (Mollineda et al., 2007). For this reason we decided to use a simple under-sampling of the "No Debt" class and we chose 300 instances randomly to be the new "No Debt" which despite its simplicity under-sampling is considered a very reliable method (Chawla, 2005; Mollineda et al., 2007). The new sample is representative of the whole "No Debt" class as statistical tests showed, so there is no loss of information.

7.3 Data Pre-processing

7.3.1 Handling Noise

As we were unsure whether to include in our models the instances that contained a large number of uncertain answers, we performed a Homogeneity Analysis (Homals) on the demographic and financial attributes in order to get a better insight of the data.

A view of the 2-dimensional space created by Homals can be seen in Fig 7.2. and reveals associations between categorical levels. We can notice that the categorical levels of "Don't know" and " Prefer not to answer" form an outlier cluster away from the other categories which seem to be centered. In the case of the demographic variables in Fig. 7.2 the cluster is positioned below the center and in the case of the financial attributes

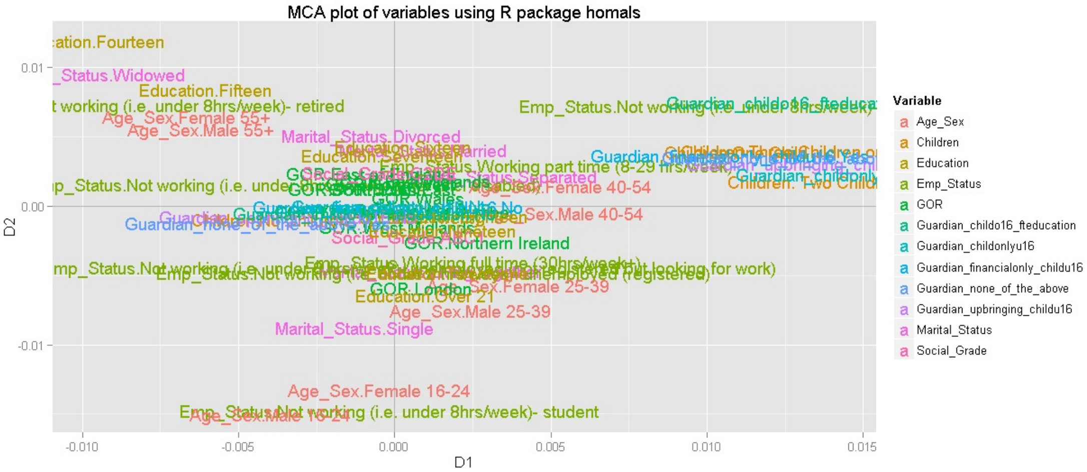


Figure 7.3: Category plots of demographic variables in the 2-dimensional space created by Homogeneity Analysis on the reduced dataset

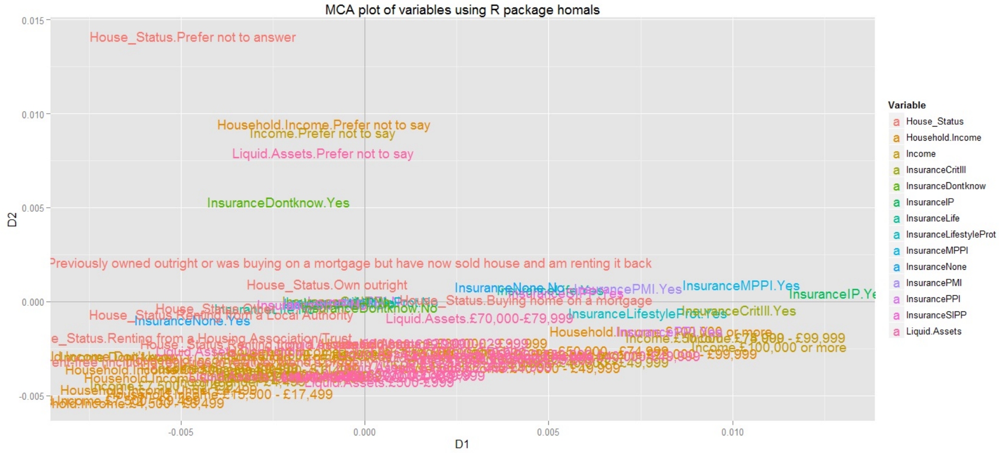


Figure 7.4: Category plots of financial variables in the 2-dimensional space created by Homogeneity Analysis

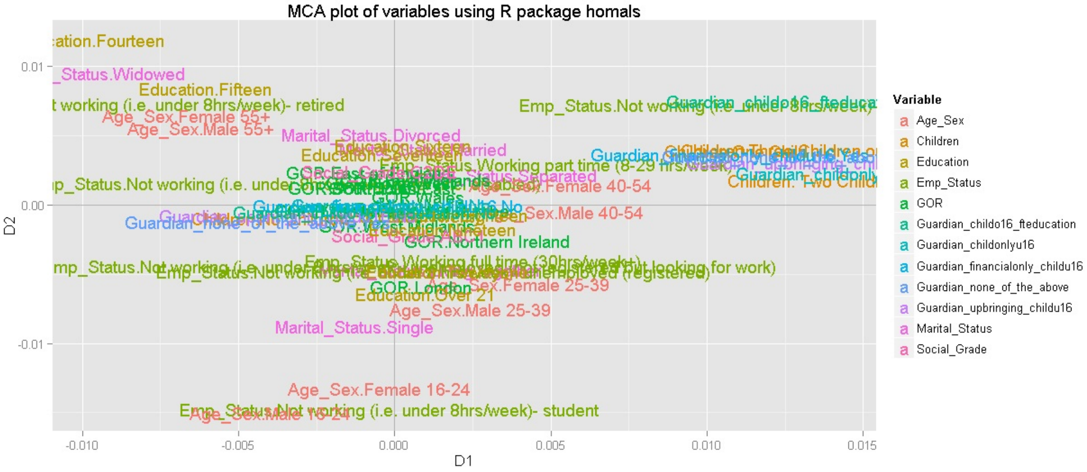


Figure 7.5: Category plots of financial variables in the 2-dimensional space created by Homogeneity Analysis on the reduced dataset

in Fig. 7.4 above the center. That means that the people who preferred not to reveal any information did so systematically to a series of questions. So, these instances cannot contribute any valuable information to the modelling and therefore it was decided to remove them from the further analysis, moving to a reduced data size.

After the removal of these noisy instances the data size was reduced to 1253 instances from the 2084 the data originally had. Chi-square tests were utilised to test the differences between the expression of the same categorical variables in the sample and in the original population. All the results showed that the reduced dataset holds the same kind of information as the original dataset as no significant differences in the distribution of categorical levels were noticed, with the bigger one being 5%. That is a very small proportion of change considering that almost half of the dataset is removed. The category plots of the smaller dataset can be seen in Fig. 7.3 and Fig 7.5. Removing these noisy instances caused both category plots to be more dispersed giving more discriminative power to the dimensions of Homogeneity Analysis. This way the representation can differentiate more easily among categories by assigning a more meaningful score to the representation. The biggest impact of our strategy can be seen in the biplot of the financial attributes in Fig 7.5, where the small and dense cluster in the bottom of the plot in Fig. 7.4 has changed to a bigger cluster centered in the plot that occupies most of the space. Regarding the demographics in Fig 7.3 we can see that the deletion of the noisy instances caused the biplot to be more understandable since it seems that a formation of clusters starts to appear. This cannot be noticed in the case of the financial dimensions where the categories of financial attributes appear to be very close to each other.

As this representation of Homals proved to be successful in our previous work on modelling consumer indebtedness (Ladas et al., 2014), as they increased the predictive power of all our models, we decided to keep these transformed variables, obtaining two demographic dimensions and two financial dimensions.

7.3.2 Factor Analysis on Psychological Items

In an effort to understand the structure of the psychological items of the survey and to validate the theoretical constructs they were trying to assess, we performed an Exploratory Factor Analysis on them (EFA). The 28 psychological Items in the survey cover psychological aspects like *Self Control* or *Impulsivity*, *Risk Management/Knowledge* and *Risk Aversion*. Since the number of theoretical constructs was unclear in the original data, we chose EFA instead of CFA in order to find the underlying latent structure that

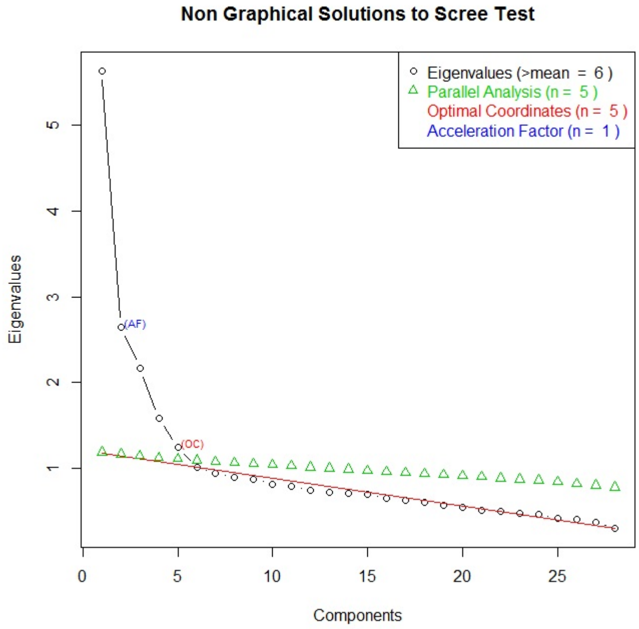


Figure 7.6: Scree plot of the eigenvalues of the psychological Items and random data

can represent the 28 items. After the factors have been specified we can verify which of the theoretical constructs are actually being assessed in this survey.

In order to define the optimal number of factors that describe best the psychological items, we utilised the scree test and parallel analysis, two widely used techniques in applied EFA for determining the number of latent factors (Fabrigar et al., 1999). In Fig. 7.6 you can see the results of the scree test and parallel analysis suggesting that five should be the optimal number of latent factors.

So we proceed with a 5-factor model to represent the psychological items. The factor loadings of the five factors on the psychological items can be seen in Table 7.2. We chose an orthogonal rotation of the factors despite the many advantages of oblique rotation (Fabrigar et al., 1999) because uncorrelated predictors is much desired in Data Mining models. The 5-factor model explains 36% of the total variance and the first factor explains almost the half of this amount.

After looking at how the psychological items are grouped together based on the 5-factor model we can see that the first factor loads significantly on items measuring *Impulsivity* and *Self Control*, the 2nd factor on *Risk Aversion* items, the 3rd on *Organisational Responsibility*, the 4th on *Risk Management Belief* and finally the 5th one on *Planful Saving*. The renamed factors and their Cronbach's alpha can be seen in Table 7.3. Cronbach's alpha (Cronbach, 1951) measures the internal consistency of each factor, meaning how intercorrelated are the items that the factor loads on. In other words, it tests if items loaded on a factor measure the same construct the latent factor represents.

Table 7.2: Factor loadings of 5-factor model on psychological items

	factor 1	**factor 2**	**factor 3**	**factor 4**	**factor5**
Q70r1	0.75				-0.195
Q70r2	0.689				-0.123
Q70r3	0.732		-0.112		
Q70r4	-0.168		0.201		0.391
Q70r5	0.681				-0.426
Q70r6	0.583				-0.179
Q70r7	-0.428	0.133			0.656
Q70r8	0.522	-0.138		0.128	
Q70r9	0.646		-0.209	0.123	
Q70r10			0.275	0.412	
Q70r11	-0.154		0.592	0.156	
Q70r12		-0.102	0.408	0.334	
Q70r13	0.11		0.305	0.215	
Q70r14	0.107			0.648	
Q70r15				0.592	
Q70r16			0.428		
Q70r17	-0.183	0.118	0.46	-0.118	
Q71r1	0.194	-0.389	0.165	0.129	0.174
Q71r2		0.601			
Q71r3		0.643			0.126
Q71r4	-0.383	0.425			0.328
Q71r5	0.326				
Q71r6	-0.129	0.209	0.414		
Q71r7		0.373	0.109		-0.161
Q71r8	0.636	-0.154			-0.107
Q71r9	-0.409	0.479	0.222		0.121
Q71r10		-0.122	0.395	0.104	0.205
Q71r11	0.428	0.146			-0.113

Table 7.3: Identified psychological factors and their Internal Consistency

factors	**Cronbach's alpha**
Impulsivity	0.86
Risk Aversion	0.64
Organisational Responsibility	0.61
Risk Management Belief	0.57
Planful Saving	0.57

Cronbach's alpha takes values from 0 to 1 with values above 0.7 being considered good, between 0.6 and 0.7 as acceptable and between 0.5 and 0.6 as poor. In similar fashion, as we can see in Table 7.3 only *Impulsivity* can be considered a reliable measure. *Risk aversion* and *Organisational Responsibility* can be regarded as acceptable and the last two as poor.

7.4 Experimental Evaluation of Psychological Factors

7.4.1 Experimental Setup

After having identified the psychological factors in the dataset we can now evaluate their contribution in modelling consumer indebtedness. For this reason we group the variables into three groups, demographics, financial and psychological factors. For the demographics and financial groups we also have obtained their transformed representations from the Homogeneity Analysis, whereas the psychological factors include the five factors identified by Exploratory Factor Analysis. Then we check each of the five groups of variables individually in order to understand their predictive power. Then we proceed in a stepwise fashion starting from financial variables (Step 1) and we continue adding demographics (Step 2) and psychological factors (Step 3) in the process, to examine carefully the accuracy of the multifaceted model as this develops gradually. This is repeated for the transformed variables. Finally we compare the differences in the performance of the models between Step 3 (financial and demographics and psychological) and Step 2 (financial and demographics only) in order to assess the impact of psychological factors on modelling consumer indebtedness. Since in Step 2 the models are based on socio-economic variables only and in Step 3 psychological factors are also added, this comparison points out the importance of including psychological factors in the traditional economic modelling of consumer indebtedness. The significance of the impact is measured by statistical significant testing.

As classifiers, three different Data Mining methods with different characteristics are used, Multinomial Logistic Regression, Random Forests and Neural Networks. Multinomial Logistic Regression belongs to the family of linear classifiers and measures the relationship between a categorical dependent variable and one or more independent variables, by using probability scores as the predicted values of the dependent variable. Random Forests is an example of ensemble learning that generates a large number of decision trees, built on different samples with bootstrap methods that allow re-sampling of instances, and aggregate the results. The difference from being an ensemble of decision trees is that when a split on a node is to be decided, a specific number of the attributes are chosen randomly to participate as candidates and not all of them. The 3rd method is Neural Networks which is considered a non-linear classifier. Neural Networks connect the input variables (predictors) to the output variable through a network of neurons organised in layers, where each neuron of every layer is fully connected to the output of all the neurons in the previous layer. The output of each neuron is a non-linear transformation of the sum of all its inputs. The three different classifiers can test different aspects of modelling consumer indebtedness and reveal important characteristics of this dataset.

Finally, for evaluating the performance of all the models we build and estimating their accuracy, we use a repeated 10-fold cross validation method, which is a widely used evaluation method in Data Mining. 10-fold cross validation gives a generalised and reliable evaluation of the model as it doesn't allow over-fitting, which means that the model will exhibit the same performance on unseen data. The repeated version of 10-fold cross validation gives an even more reliable estimate of performance of the model as it is less affected by the partitioning of the folds.

All of our models are built in R using the caret package (Kuhn and Johnson, 2013), and for Neural Networks we use one hidden layer of neurons and we tune the number of neurons in this layer by keeping the model with the best performance from all the possible models with neurons varying from one to ten. Also ten is the number of repetitions for 10-cross fold validation.

7.4.2 Results for Two-Class Models

In Fig. 7.7 we can see the performance of the models that are built on a single group of variables. We can see that financial variables hold the strongest predictive ability among the five different groups, while psychological factors come 2nd and demographic variables 3rd. It is clear that psychological factors hold significant predictive ability and they exhibit better performance than the demographic variables which were traditionally included in economic models of consumer indebtedness. However, we can also see that the transformed variables show the worst performance of all five groups and especially the biggest drop is manifested in the case of financial dimensions, a fact that signifies a significant loss of information. In case of demographics the performance of demographic dimensions is worse but comparable to the original demographic variables. Considering the performance of the classifiers we notice that Neural Networks and Multinomial Logistic Regression exhibit similar performance with Random Forests being slightly worse.

The superior performance of the original variables over the transformed ones can be seen more clearly in Fig. 7.8, where the performance of the model is presented as the groups of variables are added step by step. All the models exhibit similar behaviour and performance as in each step the performance improves when a group of variables is inserted into the model. That means that every group of variable is beneficial for modelling consumer indebtedness. However, the increase in the accuracy of the model is not very big, around 10% in almost models except the case of Random Forests for transformed variables where the final model (Step 3) increases its accuracy around 20%.

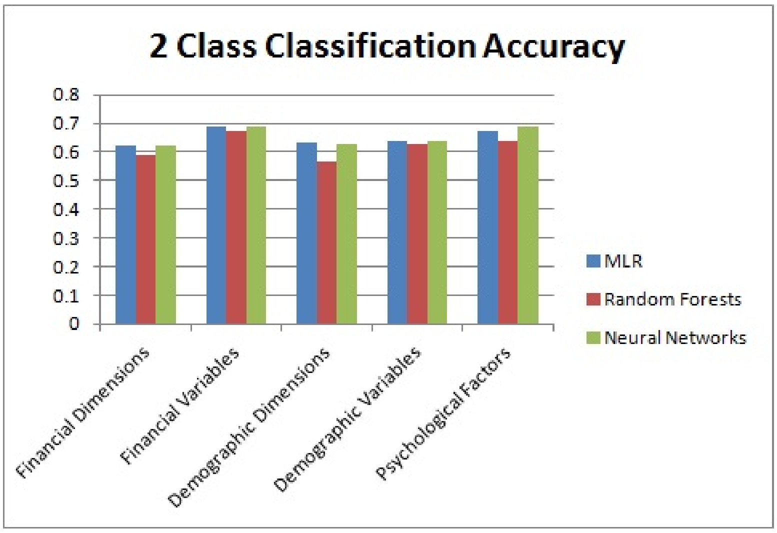


FIGURE 7.7: Performance of groups of variables in 2-class classification

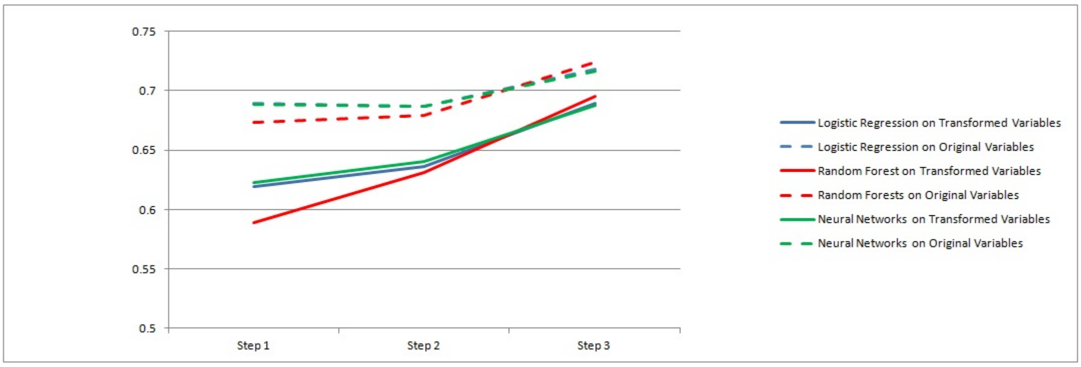


FIGURE 7.8: Performance of the models in stepwise fashion, Step 1: financial Variables, Step 2: demographics Added, Step 3: psychological factors Added

This behaviour of Random Forests is evident also in the case of the original variables and interestingly enough, while they exhibit the worst performance in the beginning, as it was also seen in Fig. 7.7 when examined for groups of variables individually, soon their performance increases and in the end they exhibit the best performance at Step 3.

As psychological factors seem to increase the performance from all the models improving their predictive ability, their impact was assessed by statistical significant testing. In more detail, all the 100 folds from the ten repetitions of 10-fold cross validation of each model in Step 3 were compared to the corresponding 100 folds of Step 2 with t-tests. The results can be seen in Table 7.4, where the importance of psychological factors in modelling consumer indebtedness is pointed out as the results of all the tests were statistically significant with their p-value being much smaller than 0.025. That means that the traditional economic modelling of consumer indebtedness which was exclusively based on socio-economic variables like in Step 2 can benefit from the incorporation of psychological factors in Step 3.

TABLE 7.4: Statistical Significance between Step 2 and 3 for 2-class Classification

Classifier	**p value**
Original	
Multinomial Logistic Regression	$5.153e^{-6}$
Random Forests	$6.312e^{-11}$
Neural Networks	$3.918e^{-6}$
Transformed	
Multinomial Logistic Regression	$6.558e^{-15}$
Random Forests	$< 2.2e^{-16}$
Neural Networks	$2.361e^{-15}$

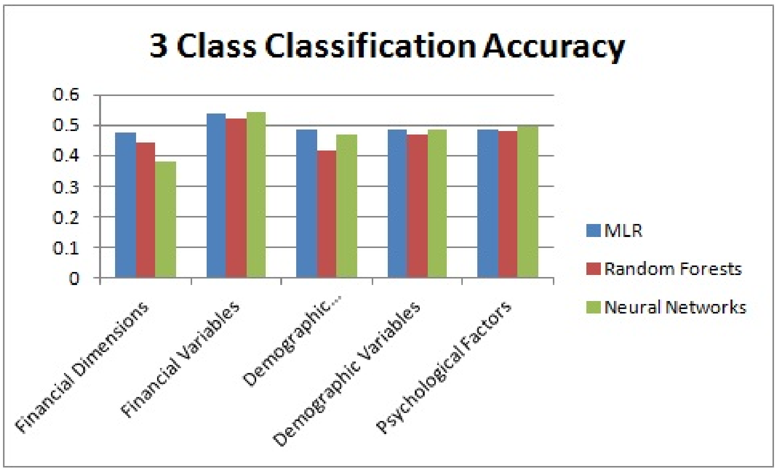


FIGURE 7.9: Performance of groups of variables in 3-class classification

7.4.3 Results of Three-CAlass Models

Moving to a three-class classification we can see in Fig. 7.9, that the predictive ability of the variables has dropped around 20% when compared with the performance of the same variables in the two-class classification. However, the relative predictive ability of the five groups of variables remained approximately the same. Similarly with the two class classification, financial variables appear to be the strongest predictors whereas all the transformed variables continue to exhibit the worst performance of all five groups. The predictive ability of psychological factors is still better than the one of original demographic variables but now their difference is much smaller, while transformed demographic dimensions achieve again comparable performance with the original demographic variables. Considering the models, Random Forests continue to exhibit the worst performance when they are examined for single group of variables in almost all groups, except the transformed financial dimensions where Neural Networks are substantially worse.

Checking the performance of the models in the stepwise fashion in Fig. 7.10 verifies

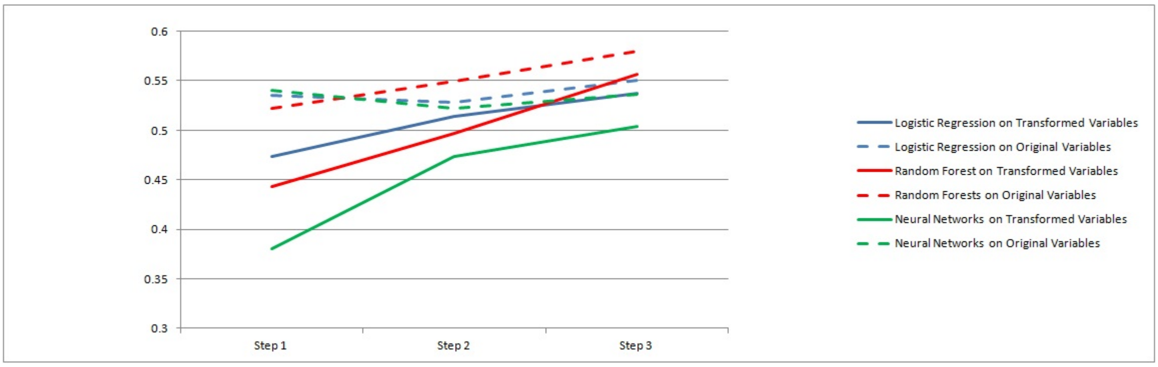


FIGURE 7.10: Performance of the models in stepwise fashion, Step 1: financial Variables, Step 2: demographics Added, Step 3: psychological factors Added

TABLE 7.5: Statistical Significance between Step 2 and 3 for 3-class Classification

Classifier	**p value**
Original	
Multinomial Logistic Regression	0.007919
Random Forests	0.0001202
Neural Networks	$9.47e^{-7}$
Transformed	
Multinomial Logistic Regression	0.0005011
Random Forests	$8.524e^{-14}$
Neural Networks	$4.149e^{-6}$

the overall worse accuracy of the three-class classification models comparing to the corresponding two-class classification models. Now the performance of the classifiers is not similar. Random Forests are clearly superior and the Neural Networks exhibit the worse predictive Accuracy. In a similar way as before, Random Forests' accuracy picks up as more variables are added in the modelling and they are the only classifier that improve the performance in Step 2 in case of original variables. Neural Networks and Multinomial Logistic Regression exhibit a drop in the performance when demographic variables are added in Step 2. On the other hand the inclusion of psychological factors is beneficial for all the models, both for original and transformed variables. Now psychological factors seem to be more important for modelling consumer indebtedness than the demographics.

Looking at the statistical significance of the increase between Step 2 and 3 in Table 7.5, where the same tests as in the case of two-class classification were repeated, we see again the very small p-values. Again all the tests produced statistically significant results, showing the importance of psychological factors in modelling consumer indebtedness.

It is now evident from both the 2-class classification and the 3-class classification approaches to model consumer indebtedness that psychological factors are important predictors. This has been shown from all the results and statistical tests for all the models for both original and transformed variables. Their contribution to Consumer Debt Analysis seems to be vital and they pose as strong candidates to supplement the existing traditional socio-economic modelling of consumer indebtedness. On the other hand the demographic variables seem to exhibit worse performance than psychological factors not only when they are examined separately but also when they cause a drop in the performance of the models for the cases of Neural Networks and Multinomial Logistic Regression for original variables in the three-class classification, a fact that raises some questions regarding their contribution to the level of debt prediction. In addition to this, the performance of the models in the two-class classification is superior to the performance of the three-class classification. More accurately, the two-class classification achieves a respectable performance from all the models, whereas the performance of the models in the three-class classification is below average. While this gives better chances to build models to separate debtors from non-debtors, trying to predict the level of debt remains one of the important questions of Consumer Debt Analysis (Wang et al., 2011). Given that the multifaceted "nature" of the level of debt prediction was confirmed by showing the importance of the psychological aspect of the problem, perhaps more research needs to be done in order to produce models with better predictive ability.

Looking closer at the performance of the classifiers, Random Forests manifested the best results especially in the case of three-classification, whereas in two-class classification the performance of the classifiers is comparable. That means that solving the problem of discriminating debtors from non-debtors does not depend on the characteristics of the three different classifiers. On the contrary, trying to predict the level of debt seems to benefit from the usage of ensemble learning.

Finally, the transformations were not able to improve the classifications in any model. That is more clear in the case of financial dimensions, where it seems that the loss of information in this transformed representation is significant. This verifies the absence of discriminative ability of the financial dimensions, as this was noticed on the visual representation of the financial categories in Fig. 7.5. On the other hand the transformed demographic dimensions, which seem to have more discriminative ability in Fig. 7.3, appear as a better representation of the original demographic variables since the performance of both sets of variables is comparable, and that means that they can be used in an effort to reduce dimensionality.

7.4.4 Random Forests Analysis

In our results, it is clear that we can build a good model for discriminating debtors from non-debtors, but building a more reliable and accurate model that predicts the level of

Table 7.6: Descriptive Statistics of Variable Importance

Mean Decrease in Gini Index					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.2091	1.406	2.479	4.179	3.338	50.83

Debt requires a deeper and more sophisticated research. From the two-class models, Random Forests exhibited the best performance. More accurately, the model achieved a 72% accuracy, 70% sensitivity (true positive rate) and 74% specificity (true negative rate). Random Forests also offer a way to assess the variable importance by using as measure the mean decrease in Gini Index. Gini index measures the impurity of data, and therefore the attribute that causes the biggest decrease in the impurity is considered a significant predictor. Having a look at the values of this measure which are summarised in the descriptive statistics of Table 7.6, we can notice that more than the 75% of the attributes have a measured importance that is less than the mean importance of all the attributes. That means that there are few attributes that seem to be very important for the accuracy of the model.

Taking a closer look at the values of the mean decrease in Gini Index for all the attributes we can spot seven attributes that have much bigger values than the rest of the dataset. These attributes are presented in Table 7.7 where we can see that the values of these attributes in the decrease of Gini Index are many times bigger than the average decrease in Gini index of all the attributes of the dataset. Interestingly enough all the psychological factors are included as predictors, especially *Impulsivity* and *Planful Saving*. Finding that *Impulsivity* is the strongest predictor among of all the variables confirms the findings of (Gathergood, 2012; Ottaviani and Vandone, 2011), where it was mentioned to be strongly associated with Consumer Debt, especially its unsecured forms of debt. From the demographic variables, *Employment Status* seems to hold some significance for separating debtors from non-debtors, and from economic variables *House Status* appears to be an important predictor.

Table 7.7: Important Variables in Random Forest Model

Variable Importance	
Variable Importance	**Mean Decrease in Gini Index**
Impulsivity	50.8253
Risk Aversion	28.6954
Organisational Responsibility	28.1296
Risk Management Belief	27.8619
Planful Saving	35.2052
Emp_Status_retired	11.3329
House_StatusOwn.outright	24.1645

Checking which attributes seem to have a measured variable importance above the mean, we can also find among the important predictors additional attributes like *Marital Status*, *Age/Sex*, *Social Grade*, *Education*, *Liquid Assets*, *Life Insurance* and whether the households an *Insurance* at all. Most of these socio-economic attributes have been well researched in the literature (Wang et al., 2011; Kamleitner et al., 2012; Stone and Maury, 2006; Kamleitner and Kirchler, 2007) but some attributes like *Liquid Assets* and *Life Insurance* appear for the first time to be associated with consumer debt. It is of great surprise, however that *Income* and *Household Income* are not included in the important predictors. Income's ranking in variable importance is in the top 50% but still below the average in this model, whereas throughout the literature it considered one of the most important predictors. Having in mind that the mean decrease in Gini Index might not be the appropriate method to identify the importance of each variable, we think that deeper and more careful analysis of the results is required in order to validate our findings from this model.

7.5 Conclusions

In this work, we explored the multifaceted nature of consumer indebtedness in a complete and detailed socio-economic dataset that contains psychological information. After we identified and verified the psychological factors in the dataset we proceeded to the assessment of their contribution on modelling consumer indebtedness within a complete Data Mining Framework, with powerful classification models and reliable evaluation techniques. It was shown that all the psychological factors increase the performance of the models in a statistically significant sense in all cases. Especially Impulsivity, which is a well researched psychological factor (Gathergood, 2012; Ottaviani and Vandone, 2011) seems to be the strongest predictor in our modelling. That verifies the potential psychological information has to achieve Knowledge Discovery in Consumer Debt Analysis and confirms our hypothesis, at is stated in 1.4.1. The latter has a lot to gain by embracing the studying of psychological factors and by adopting Data Mining approaches in their methods and practices. Data Mining offers a powerful and transparent modelling that is suitable for answering any research question, especially of complex and sensitive nature whereas the psychological information has the ability to explain better the behaviours of consumers and supplement the existing economic models.

Chapter 8

Conclusions

8.1 An Overview of the Data Mining Framework to model Consumer Indebtedness

In this work we utilised Data Mining techniques and methods to develop a framework for modelling consumer indebtedness. Considering the complex and sensitive "nature" of this emerging social phenomenon, a careful analysis of the data is required in order to extract safe results that can lead to a deeper insight and better understanding of consumer indebtedness and a better modelling in a further extent. Data Mining offers a complete set of techniques and methods that vary from processing the data in order to improve their quality and using unsupervised learning to explore the relationships hidden in the data to stronger and more accurate predictive modelling that can be utilised to provide reliable answers to the fundamental research questions of Consumer Debt Analysis (Wang et al., 2011).

We begin with the essential data transformations of the socio-economic variables of the Consumer Credit Counseling Service (CCCS) dataset in Chapter 3. The transformations are able to deal with the inconsistencies commonly found in real world data and squeeze out behavioural elements hidden in the data. This transformed behavioural representation enabled clustering techniques and more specifically consensus clustering to mine meaningful behaviours of consumers as these were reflected in seven behavioural profiles. The quality of the clusters was assessed by different evaluation methods that identified the importance of the agreement between the cluster ensemble to uncover novelties hidden in the data and achieve Knowledge Discovery within a socio-economic context. The resulting profiles are shown also to provide a better understanding of consumer indebtedness, describing diverse socio-economic behaviours and thus the importance of behavioural transformations for Consumer Debt Analysis was verified as it improved the quality of the clustering results.

The emergence of the agreement of cluster ensemble as an alternative validation criterion of the clustering inspired us to devise a new consensus function that can improve the current consensus clustering framework we used in Chapter 3. Thus in Chapter 4 we utilise Homogeneity analysis to create a new representation that models successfully all the relationships of cluster agreement including the disagreements of the cluster ensemble, information that was absent in the previous consensus clustering approach. This informative representation enhanced clustering to improve its performance and quality of results by producing well defined and distinct consensus clusters that follow the assumptions of *Compactness* and *Separation*. Since agreement can uncover different patterns that transcend the aforementioned assumptions, our two-tier clustering result also includes clusters that are defined based on the agreement of the cluster ensemble. The superiority of our approach was confirmed in the datasets both visually and by metrics. Then it was applied to a socio-economic dataset that contained psychological information to discover six behavioural profiles of consumers organised in two broader types. The analysis of the clustering result has revealed significant socio-economic patterns for the purposes of Consumer Debt Analysis and also patterns regarding the personality of consumers that have not been thoroughly researched in literature.

These two chapters describe the implementation of the Behavioural Feature Extraction phase of our framework and it shows how we can utilise Data Mining techniques to construct Behavioural Data and Profiles for achieving a better Knowledge Discovery within socio-economic context. In the Chapters 5 and 6 we proceed with the Behavioural Modelling phase of our framework, which describes our efforts to include the extracted Knowledge from the previous chapters into the modelling of consumer indebtedness for more accurate and better results.

In doing so, in Chapter 6 we first establish the superiority of Random Forests and Neural Networks models against some of the traditional statistical modelling widely used in economics and social sciences in a series of experiments. Throughout this experimentation the importance of Behavioural Data for the modelling was also verified. For this reason we exploited the flexibility Neural Networks offer in designing the topology of their networks to propose a novel technique that can incorporate the extracted knowledge into the modelling. Our novel Neural Networks approach achieved the best performance in predicting the level of consumer debt highlighting the importance of Behavioural Extraction for the modelling when this is incorporated carefully in the Data Mining process. The powerful modelling of Data Mining methods was accompanied with techniques of assessing the variable importance, further supporting the utilisation of these methods for the purposes of economics and social sciences.

Therefore in Chapter 7 we use Data Mining models to explore the multifaceted nature of consumer indebtedness within a socio-economic dataset that contains psychological information. Based on our Behavioural Extraction techniques we extract psychological factors from the data and then, relying on the strong and accurate performance of Data Mining modelling, we show that psychological factors improve the accuracy of the models when separating debtors from non-debtors and when predicting the level of consumer indebtedness. More accurately, the analysis of the models showed that the psychological factors were the most significant predictors of consumer indebtedness, challenging the existing rational economic modelling.

8.2 Conclusions

8.2.1 Importance of Psychological Information

The last chapter demonstrated the beneficial role psychological information can play in modelling consumer indebtedness. In its more direct form as it is found in the DebtTrack survey, psychological information shows a great potential to enhance the predictive accuracy of the applied models. Even in its undirect form where psychological information has to be constructed through Behavioural Extraction from socio-economic datasets that do not include measured psychological items, it proved particularly beneficial as it helped both the supervised and unsupervised models to achieve meaninghful knowledge discovery for the purposes of Consumer Debt Analysis. Therefore we can safely conclude that psychological information possesses a great potential to enhance the analysis of consumer indebtedness, a fact that verifies our initial research hypothesis, stated in 1.4.1.

8.2.2 Importance of Data Mining

Numerous techniques and methods from the diverse toolbox of Data Mining were utilised to guarantee the reliability of this conclusion. In more detail Data Mining exhibited a great potential in analysing a problem of complex and sensitive nature, which is evident in the careful processing of the data that respects the semantics of the data, in the exploratory techniques that uncover informative patterns in the data and provide a meaningful representation of this knowledge and in the powerful and accurate modelling that can provide reliable and transparent answers to the significant research questions. The greatest contribution of Data Mining methods however is the flexibility they provides to incorporate all the extracted knowledge from the different stages in the Knowledge Discovery process in order to achieve even greater performance sketching a complete and sophisticated framework. For this reason we were able to implement the Behavioural Extraction and Behavioural Modelling phases of our framework and analyse the multifaceted and complex nature of consumer indebtedness.

8.3 Contributions

Our resulting framework highlights a series of novelties in several technical aspects. The use of Homegeneity Analysis was identified as an important pre-processing technique as it manages to transform the categorical variables of a dataset in a meaningful numerical representation which enables the application of clustering algorithms on mixed data. Besides this, the representation can be used to solve the correspondence problem of a cluster ensemble implicitly without relabeling as it models successfully all the possible relationships among the clusters and the data points of the cluster ensemble, providing this way a novel consensus clustering solution that demonstrates superior performance. Together with Homogeneity Analysis several other commonly used pre-processing techniques like scaling or Factor Analysis were shown to be essential for analysing a complex problem of sensitive nature.

In unsupervised learning and more specifically in clustering, we evaluated the potential of the agreement of the cluster ensemble to uncover meaningful patterns that do not comply with the traditional assumptions of *Separation* and *Compactness*. As its importance was verified by the numerous patterns that followed the optimisation of this new criterion, it showed that Knowledge is not always included in well defined clusters as it is generally assumed. For this reason the patterns that follow the optimisation of this criterion were included in our novel Homals Consensus solution within a two-tier structure of the clustering results in Chapter 5.

Finally, in supervised learning the flexibility of Neural Networks in designing their network topology provided the means to incorporate the extracted Knowledge from the other stages of our framework into the modelling process in a novel Neural Network solution that we named TopDNN (Topology Defined Neural Networks). TopDNN utilises exploratory techniques like Factor Analysis and clustering to gain a better understanding of the data and uses the extracted information to design its network of neurons. When applied on the CCCS dataset in order to predict the level of debt, it manifested superior performance from all the other models in Chapter 6.

All these novelties together with the existing powerful methods and techniques from Data Mining sketch a complete Data Mining framework that is suitable for the purposes of Consumer Debt Analysis. As the proposed framework respects the sensitive nature of the analysed problem it becomes suitable for further applications in economic and social sciences but also any real world complex problem that shares the same characteristics (noise, non-linearities, etc) of socio-economic data.

8.4 Future Work

8.4.1 Extending our framework for Credit Risk Assessment with Personality Profiles

A possible extension of our framework can be provided for modelling credit score. Credit Score is a measure of trustworthiness for the consumers and it reflects the likelihood that the person will repay his debt (Baesens and van Gestel, 2008). Credit score modelling refers to the process of calculating such score for each individual and traditionally it is based on demographics and financial/economic data as it can be seen in Fig 8.1 in the blue boxes. Our purpose is to include in this process the personality profiles of consumers that will capture their psychological characteristics, as these are expressed by personality traits, behaviours and attitudes and to check whether this additional information can create a more fair credit score that fits best the consumers. Throughout the literature the significance of psychological factors for analysing the behaviours of debtors is highlighted but whether this can be beneficial for the Credit Score modelling is a question that remains to be answered. This modification in this Credit Scoring modelling is indicated by the orange box in the Fig 8.1. and depicts the purpose of the future research.

Similarly with modelling the consumer indebtedness our alternative Credit Score modelling will consist of the same Behavioural Extraction and Behavioural Modelling phases and it will provide an alternative route to calculate the credit score of consumers. This is of great significance to the industry as sometimes the required information of consumers needed from the existing Credit Score modelling to assess their Credit Risk is absent. Thus exploiting alternative data sources to sketch the Behavioural Profiles of consumers can provide a way to assess reliably their Credir Score.

However while this potential extension of research was part of my studies was never applied as data for this task never became available, leaving this interesting path of research for the future.

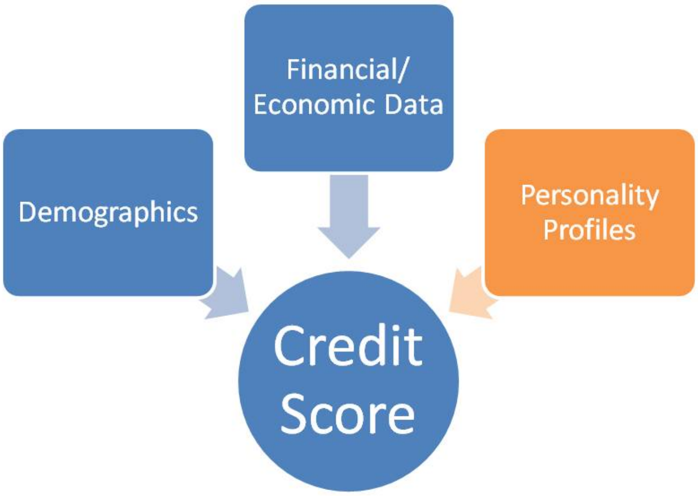


Figure 8.1: Incorporation of Personality Profiles into Credit Scoring models

8.4.2 Improving Homals Consensus

As the results in our research showed Consensus Clustering proved very useful for partitioning difficult real world data and extracting useful patterns that give a better insight of the data. Building a better consensus function based on Homegeneity Analysis that manages to model the relationships between cluster ensemble proved to be a vital part of this process. Homals Consensus showed several interesting properties in modelling the cluster ensemble and representing it with consensus clusters and it managed to uncover very interesting patterns of behaviour.

However, Homals Consensus was tested under a specific consensus clustering framework. It will interesting to evaluate Homals Consensus under other frameworks that build differently the cluster ensemble. It might be more interesting to test the performance of this consensus function in a bigger and more diverse cluster ensemble that is not necessarily characterised only by its quality. Different clustering algorithms with diverse properties and assumptions can be utilised for this reason. This way the ability of Homals Consensus can be tested in a more difficult cluster ensemble and its current strengths can be verified.

8.4.3 Verifying TopDNN

TopDNN managed to show very good performance when analysed the CCCS data but it was never tested in different problems with different data. Despite the promising characteristics TopDNN possesses it is essential to verify the performance of this approach on multiple cases. Alternatively it might be interesting to extend the central idea of

TopDNN to incorporate information extracted from other exploratory techniques or improve the way clustering is incorporated into the network. Extensive experimentation can highlight the way to optimise the inclusion of clustering results in the design of a neural network.

Appendix A

Appendix: YouGov Psychological Items

A.1 YouGov Psychological Items

- **Q70r1** If I want something I am prepared to buy it on credit and think about how I will repay the money afterwards : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r2** I am prepared to spend now and let the future take care of itself : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r3** If lenders offer to lend me money I will take it : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r4** I would rather cut back than put everyday spending on a credit card : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r5** I would rather buy things on credit than save up : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r6** Borrowing has become a way of life : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r7** 'I would always save up for something I want, rather than borrow money to buy it' : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r8** Buying things on credit does not feel like spending : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r9** I would miss a payment on an existing financial commitment if it meant I could have what I wanted now : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r10** Companies lending money have only themselves to blame if people stop repaying : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r11** Home owners shouldnt be offered a mortgage thats difficult to pay each month : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r12** Mortgage lenders are more interested in making sales than providing a mortgage that I can afford : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r13** Getting an agreement from your creditors to only repay a part of what you owe over time is a sensible thing to do to reduce your debts : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r14** Bankruptcy is now regarded as being socially acceptable : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r15** Bankruptcy is an easy way to escape from your money problems : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r16** It is difficult to build up your credit history after bankruptcy : Can you tell us how much you agree, or disagree, with the following statements?

- **Q70r17** I would feel ashamed if I had to go through bankruptcy : Can you tell us how much you agree, or disagree, with the following statements?

- **Q71r1** Financial services are complicated and confusing to me : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r2** I regularly read the personal finance pages in the press : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r3** Friends and family often come to me for advice on financial matters : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r4** I am more of a saver than a spender : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r5** I will buy more things with cash or a debit card in the next 6 months than I did before : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r6** I would only buy financial products from a company I have heard of and trust : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r7** Buying things on a credit card and paying everything back each month is a smart way to manage your money : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r8** I am impulsive and tend to buy things even when I cant really afford them. : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r9** I am organised when it comes to managing my money on a day to day basis : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r10** Buying things with cash makes me realise how much I am spending : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

- **Q71r11** I often move money from one credit card to another to take advantage of 0% interest deals : Here are some statements that have been made about managing money and your financial affairs, how much you agree or disagree with each statement?

Appendix B

Appendix: Publications

B.1 Publications

- 2015: *Indebted households profiling: a knowledge discovery from database approach* Scarpel Rodrigo Ladas Alexandros Aickelin Uwe, Annals of Data Science, 2 (1), pp. 43-59, 2015.

- 2014: *A Data Mining framework to model Consumer Indebtedness with Psychological Factors*
  Ladas Alexandros, Ferguson Eamonn, Garibaldi Jon, Aickelin, Uwe, IEEE International Conference of Data Mining: The Seventh International Workshop on Domain Driven Data Mining 2014, Shenzhen, China.

- 2014: *Augmented Neural Networks for Modelling Consumer Indebtness*
  Alexandros Ladas, Jon Garibaldi, Rodrigo Scarpel, Uwe Aickelin, IEEE IJCNN 2014, Beijing, China.

- 2012: *Biomarker Clustering of Colorectal Cancer Data to Complement Clinical Classification*
  Chris M. Roadknight, Uwe Aickelin, Alex Ladas, Daniele Soria, John Scholefield, Lindy Durrant, FEDCSIS 2012, Wroclaw, Poland.

- 2012: *Using Clustering to extract Personality Information from socio economic data*
  Alexandros Ladas, Uwe Aickelin, Jonathan M. Garibaldi, Eamonn Ferguson, UKCI 2012, Edinburgh, UK.

Bibliography

Agrawal, R., Imieliński, T. and Swami, A. (1993), Mining association rules between sets of items in large databases, *in* 'ACM SIGMOD Record', Vol. 22, ACM, pp. 207–216.

Ahmad, A. and Dey, L. (2007), 'A k-mean clustering algorithm for mixed numeric and categorical data', *Data & Knowledge Engineering* **63**(2), 503–527.

Almlund, M., Duckworth, A., Heckman, J. and Kautz, T. (2011), Personality psychology and economics, Technical report, National Bureau of Economic Research.

Atiya, A. F. (2001), 'Bankruptcy prediction for credit risk using neural networks: A survey and new results', *Neural Networks, IEEE Transactions on* **12**(4), 929–935.

Baesens, B. and van Gestel, T. (2008), *Credit Risk Management: Basic Concepts*, Oxford university Press.

Betti, G., Dourmashkin, N., Rossi, M. C., Verma, V. and Yin, Y. (2001), 'Study of the problem of consumer indebtedness: Statistical aspects contract n: B5-1000/00/000197 final report'.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Caliński, T. and Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics-theory and Methods* **3**(1), 1–27.

Cao, L. (2010), 'In-depth behavior understanding and use: The behavior informatics approach', *Information Sciences* **180**(17), 3067–3085.

Caruana, R., Elhawary, M., Nguyen, N. and Smith, C. (2006), Meta clustering, *in* 'Data Mining, 2006. ICDM'06. Sixth International Conference on', IEEE, pp. 107–118.

Chawla, N. V. (2005), Data mining for imbalanced datasets: An overview, *in* 'Data mining and knowledge discovery handbook', Springer, pp. 853–867.

Chen, Z. (2006), 'From data mining to behavior mining', *International Journal of Information Technology and Decision Making* **5**(4), 703–711.

Cohen, J. et al. (1960), 'A coefficient of agreement for nominal scales', *Educational and psychological measurement* **20**(1), 37–46.

Consulting, C. (2013), The overindebtedness of european households: updated mapping of the situation, nature and causes, effects and initiatives for alleviating its impact, *in* 'Summary document for stakeholder workshop'.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Cortinovis, I., Vella, V. and Ndiku, J. (1993), 'Construction of a socio-economic index to facilitate analysis of health data in developing countries', *Social science & medicine* **36**(8), 1087–1097.

Costa, P. T. and MacCrae, R. R. (1992), *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*, Psychological Assessment Resources.

Cronbach, L. J. (1951), 'Coefficient alpha and the internal structure of tests', *psychometrika* **16**(3), 297–334.

D'Alessio, G. and Iezzi, S. (2013), 'Household over-indebtedness: definition and measurement with italian data', *Bank of Italy Occasional Paper* (149).

De Leeuw, J. and Mair, P. (2007), 'Homogeneity analysis in r: The package homals'.

De Leeuw, J. and Mair, P. (2009), 'Gifi methods for optimal scaling in r: The package homals', *Journal of Statistical Software, forthcoming* pp. 1–30.

Dibike, Y. B., Velickov, S. and Solomatine, D. (2000), Support vector machines: Review and applications in civil engineering, *in* 'Proceedings of the 2nd Joint Workshop on Application of AI in Civil Engineering', Citeseer, pp. 215–218.

Dimitriadou, E., Dolničar, S. and Weingessel, A. (2002), 'An examination of indexes for determining the number of clusters in binary data sets', *Psychometrika* **67**(1), 137–159.

Disney, R. and Gathergood, J. (2009), 'Understanding consumer over-indebtedness using counselling sector data: Scoping study', *Report to the Department for Business, Innovation and Skills (BIS)* .

Dittmar, H. and Drury, J. (2000), 'Self-image–is it in the bag? a qualitative comparison between ordinary and excessive consumers', *Journal of Economic Psychology* **21**(2), 109–142.

Draper, N. R., Smith, H. and Pownell, E. (1966), *Applied regression analysis*, Vol. 3, Wiley New York.

Duarte, J. M., Fred, A. L., Lourenço, A. and Duarte, F. J. F. (2010), On consensus clustering validation, *in* 'Structural, Syntactic, and Statistical Pattern Recognition', Springer, pp. 385–394.

Eshghi, A., Haughton, D., Legrand, P., Skaletsky, M. and Woolford, S. (2011), 'Identifying groups: A comparison of methodologies', *J Data Sci* **9**, 271–291.

Fabrigar, L. R. and Wegener, D. T. (2011), *Exploratory factor analysis*, Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C. and Strahan, E. J. (1999), 'Evaluating the use of exploratory factor analysis in psychological research.', *Psychological methods* **4**(3), 272.

Ferguson, E. and Cox, T. (1993), 'Exploratory factor analysis: A users guide', *International Journal of Selection and Assessment* **1**(2), 84–94.

Ferguson, E., Heckman, J. J. and Corr, P. (2011), 'Personality and economics: Overview and proposed framework', *Personality and Individual Differences* **51**(3), 201 – 209. Special Issue on Personality and Economics.
URL: *http://www.sciencedirect.com/science/article/pii/S0191886911001516*

Fern, X. Z. and Brodley, C. E. (2004), Solving cluster ensemble problems by bipartite graph partitioning, *in* 'Proceedings of the twenty-first international conference on Machine learning', ACM, p. 36.

Fern, X. Z. and Lin, W. (2008), 'Cluster ensemble selection', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **1**(3), 128–141.

Filkov, V. and Skiena, S. (2004), 'Integrating microarray data by consensus clustering', *International Journal on Artificial Intelligence Tools* **13**(04), 863–880.

Fred, A. L. and Jain, A. K. (2005), 'Combining multiple clusterings using evidence accumulation', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(6), 835–850.

Freedman, D. A. (2009), *Statistical models: theory and practice*, cambridge university press.

Friedman, H. P. and Rubin, J. (1967), 'On some invariant criteria for grouping data', *Journal of the American Statistical Association* **62**(320), 1159–1178.

Gardharsdottir, R. B. and Dittmar, H. (2012), ‘The relationship of materialism to debt and financial well-being: The case of icelands perceived prosperity’, *Journal of Economic Psychology* **33**(3), 471–481.

Gathergood, J. (2012), ‘Self-control, financial literacy and consumer over-indebtedness’, *Journal of Economic Psychology* **33**(3), 590–602.

Gevrey, M., Dimopoulos, I. and Lek, S. (2003), ‘Review and comparison of methods to study the contribution of variables in artificial neural network models’, *Ecological Modelling* **160**(3), 249–264.

Ghose, A. and Ipeirotis, P. G. (2011), ‘Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics’, *Knowledge and Data Engineering, IEEE Transactions on* **23**(10), 1498–1512.

Gini, C. (1997), ‘Concentration and dependency ratios’, *Rivista di Politica Economica* **87**, 769–792.

Goder, A. and Filkov, V. (2008), Consensus clustering algorithms: Comparison and refinement., *in* ‘ALENEX’, Vol. 8, pp. 109–117.

Gromping, U. (2009), ‘Variable importance assessment in regression: linear regression versus random forest’, *The American Statistician* **63**(4).

Günther, F. and Fritsch, S. (2010), ‘neuralnet: Training of neural networks’, *The R Journal* **2**(1), 30–38.

Han, J., Kamber, M. and Pei, J. (2006), *Data mining, southeast asia edition: Concepts and techniques*, Morgan kaufmann.

Harrington, P. d. B. and Wan, C. (1998), ‘Sensitivity analysis applied to artificial neural networks: What has my neural network actually learned?’, *Anal. Chem* **70**, 2983–2990.

Hartigan, J. A. (1975), *Clustering algorithms*, John Wiley & Sons, Inc.

He, Z., Xu, X. and Deng, S. (2005), ‘A cluster ensemble method for clustering categorical data’, *Information Fusion* **6**(2), 143–151.

Helbing, D. and Balietti, S. (2011), ‘From social data mining to forecasting socio-economic crises’, *The European Physical Journal-Special Topics* **195**(1), 3–68.

Hennig, C. and Liao, T. F. (2013), ‘How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(3), 309–369.
URL: *http://dx.doi.org/10.1111/j.1467-9876.2012.01066.x*

Hornik, K., Stinchcombe, M. and White, H. (1989), 'Multilayer feedforward networks are universal approximators', *Neural networks* **2**(5), 359–366.

Hotelling, H. (1933), 'Analysis of a complex of statistical variables into principal components.', *Journal of educational psychology* **24**(6), 417.

Hoyle, B. S. (2008), 'Improved neural network performance using principal component analysis on matlab'.

Huang, Z. (1998), 'Extensions to the k-means algorithm for clustering large data sets with categorical values', *Data mining and knowledge discovery* **2**(3), 283–304.

Hunt, C. et al. (2015), 'Economic implications of high and rising household indebtedness', *Reserve Bank of New Zealand Bulletin* **78**, 1–12.

Jiang, Q. (1996), 'Principal component analysis and neural network based face recognition', *University of Chicago* .

Kamleitner, B., Hoelzl, E. and Kirchler, E. (2012), 'Credit use: Psychological perspectives on a multifaceted phenomenon', *International Journal of Psychology* **47**(1), 1–27.

Kamleitner, B. and Kirchler, E. (2007), 'Consumer credit use: A process model and literature review', *Revue Europeenne de Psychologie Appliquee/European Review of Applied Psychology* **57**(4), 267–283.

Kaufman, L. and Rousseeuw, P. (1987), *Clustering by means of medoids*, North-Holland.

Kim, H. and DeVaney, S. A. (2001), 'The determinants of outstanding balances among credit card revolvers', *Financial Counseling and Planning* **12**(1), 67–77.

Kuhn, M. and Johnson, K. (2013), *Applied predictive modeling*, Springer New York.

Kuncheva, L. I., Hadjitodorov, S. T. and Todorova, L. P. (2006), Experimental comparison of cluster ensemble methods, *in* 'Information Fusion, 2006 9th International Conference on', IEEE, pp. 1–7.

Ladas, A., Aickelin, U., Garibaldi, J. and Ferguson, E. (2012), 'Using clustering to extract personality information from socio economic data', *UKCI, 12th Annual Workshop on Computer Intelligence* .
URL: *http://ima.ac.uk/papers/ladas2012a.pdf*

Ladas, A., Garibaldi, J. and Aickelin, U. (2014), Augmented neural networks for modelling consumer indebtness, *in* 'International Joint Conference on Neural Networks', IEEE.

Larsen, R. J. and Buss, D. M. (2008), 'Personality psychology', *New York, mc grew-hill* .

Lea, S. E., Webley, P. and Levine, R. M. (1993), 'The economic psychology of consumer debt', *Journal of economic psychology* **14**(1), 85–119.

Lea, S. E., Webley, P. and Walker, C. M. (1995), 'Psychological factors in consumer debt: Money management, economic socialization, and credit use', *Journal of economic psychology* **16**(4), 681–701.

Lebart, L. and Salem, A. (1988), 'Analyse statistiques des donnees textuelles', *Paris, Dunod* .

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010), Understanding of internal clustering validation measures, *in* 'Data Mining (ICDM), 2010 IEEE 10th International Conference on', IEEE, pp. 911–916.

Livingstone, S. M. and Lunt, P. K. (1992), 'Predicting personal debt and debt repayment: Psychological, social and economic determinants', *Journal of Economic Psychology* **13**(1), 111–134.

Lloyd, S. P. (1982), 'Least squares quantization in pcm', *Information Theory, IEEE Transactions on* **28**(2), 129–137.

Longadge, R. and Dongre, S. (2013), 'Class imbalance problem in data mining review', *arXiv preprint arXiv:1305.1707* .

Marriott, F. (1971), 'Practical problems in a method of cluster analysis', *Biometrics* pp. 501–514.

Meyer, D., Leisch, F. and Hornik, K. (2003), 'The support vector machine under test', *Neurocomputing* **55**(1), 169–186.

Michailidis, G. and de Leeuw, J. (1998), 'The gifi system of descriptive multivariate analysis', *Statistical Science* pp. 307–336.

Mollineda, R., Alejo, R. and Sotoca, J. (2007), The class imbalance problem in pattern classification and learning, *in* 'II Congreso Español de Informática (CEDI 2007). ISBN', Citeseer, pp. 978–84.

Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003), 'Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data', *Machine learning* **52**(1-2), 91–118.

Nicholas Refenes, A., Zapranis, A. and Francis, G. (1994), 'Stock performance modeling using neural networks: a comparative study with regression models', *Neural Networks* **7**(2), 375–388.

Noone, J., O'Loughlin, K. and Kendig, H. (2012), 'Socioeconomic, psychological and demographic determinants of australian baby boomers' financial planning for retirement', *Australasian journal on ageing* **31**(3), 194–197.

Norvilitis, J. M., Szablicki, P. B. and Wilson, S. D. (2003), 'Factors influencing levels of credit-card debt in college students1', *Journal of Applied Social Psychology* **33**(5), 935–947.

Ottaviani, C. and Vandone, D. (2011), 'Impulsivity and household indebtedness: Evidence from real life', *Journal of economic psychology* **32**(5), 754–761.

Otto, P. E., Davies, G. B., Chater, N. and Stott, H. (2009), 'From spending to understanding: Analyzing customers by their spending behavior', *Journal of Retailing and Consumer Services* **16**(1), 10–18.

Pham, T. H., Yap, K. and Dowling, N. A. (2012), 'The impact of financial management practices and financial attitudes on the relationship between materialism and compulsive buying', *Journal of Economic Psychology* **33**(3), 461 – 470.

Powers, D. M. (2011), 'Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation'.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Refenes, A. N., Zapranis, A. and Francis, G. (1994), 'Stock performance modeling using neural networks: a comparative study with regression models', *Neural networks* **7**(2), 375–388.

Robb, C. A. and Sharpe, D. L. (2009), 'Effect of personal financial knowledge on college students credit card behavior', *Journal of Financial Counseling and Planning* **20**(1).

Rousseeuw, P. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1988), 'Learning representations by back-propagating errors', *Cognitive modeling* **5**, 3.

Sarle, W. S. (1994), 'Neural networks and statistical models'.

Scott, A. J. and Symons, M. J. (1971), 'Clustering methods based on likelihood ratio criteria', *Biometrics* pp. 387–397.

Segal, M. R. (2004), 'Machine learning benchmarks and random forest regression', *Center for Bioinformatics & Molecular Biostatistics* .

Shifei, D., Liwen, Zhang abd Weikuan, J., Lili, L. and Chunyang, S. (2011), 'Research of neural network algorithm based on factor analysis and cluster analysis', *NEURAL COMPUTING AND APPLICATIONS* **20**(2), 297–302.

Soria, D. and Garibaldi, J. (2010), 'A novel framework to elucidate core classes in a dataset', pp. 1–8.

Sousa, S., Martins, F., Alvim-Ferraz, M. and Pereira, M. C. (2007), 'Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations', *Environmental Modelling & Software* **22**(1), 97–103.

Stone, B. and Maury, R. V. (2006), 'Indicators of personal financial debt using a multidisciplinary behavioral model', *Journal of Economic Psychology* **27**(4), 543–556.

Strehl, A. and Ghosh, J. (2003), 'Cluster ensembles—a knowledge reuse framework for combining multiple partitions', *The Journal of Machine Learning Research* **3**, 583–617.

Thorndike, R. L. (1953), 'Who belongs in the family?', *Psychometrika* **18**(4), 267–276.

Topchy, A. P., Law, M. H., Jain, A. K. and Fred, A. L. (2004), Analysis of consensus partition in cluster ensemble, *in* 'Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on', IEEE, pp. 225–232.

Vendramin, L., Campello, R. J. and Hruschka, E. R. (2010), 'Relative clustering validity criteria: A comparative overview', *Statistical Analysis and Data Mining* **3**(4), 209–235.

Vissing-Jorgensen, A. (2011), 'Consumer credit: Learning your customer's default risk from what (s) he buys', *Available at SSRN 2023238* .

Vyas, S. and Kumaranayake, L. (2006), 'Constructing socio-economic status indices: how to use principal components analysis', *Health Policy and Planning* **21**(6), 459–468.

Wang, L., Lu, W. and Malhotra, N. K. (2011), 'Demographics, attitude, personality and credit card features correlate with credit card debt: A view from china', *Journal of economic psychology* **32**(1), 179–193.

Watson, J. J. (2003), 'The relationship of materialism to spending tendencies, saving, and debt', *Journal of economic psychology* **24**(6), 723–739.

Webley, P. and Nyhus, E. K. (2001), 'Life-cycle and dispositional routes into problem debt', *British Journal of Psychology* **92**(3), 423–446.

Wikipedia (2011), 'Consumer debt', http://en.wikipedia.org/wiki/Consumer_debt.

Zekić-Sušac, M., Šarlija, N. and Pfeifer, S. (2013), 'Combining pca analysis and artificial neural networks in modelling entrepreneurial intentions of students', *Croatian Operational Research Review* **4**(1), 306–317.

Zhang, Y., Fu, A. W.-c., Cai, C. H. and Heng, P. A. (2000), Clustering categorical data, *in* 'icde', IEEE, p. 305.