DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

APPLICATION OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT MARKET BASKET ANALYSIS IN A RETAILER STORE

by

Mine DURDU

July, 2012
İZMİR

APPLICATION OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT MARKET BASKET ANALYSIS IN A RETAILER STORE

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science of Industrial Engineering, Applied Industrial Engineering Program

by
Mine DURDU

July, 2012
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"APPLICATION OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT MARKET BASKET ANALYSIS IN A RETAILER STORE"** completed by **MİNE DURDU** under supervision of **ASSOC. PROF. HASAN SELİM** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Hasan SELIM

Supervisor

Doç.Dr.İpek Deveci KOCAKOĞ

(Jury Member)

Yrd.Doç.Dr.Gökalp YILDIZ

(Jury Member)

Prof.Dr. Mustafa SABUNCU

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank my supervisor Assoc. Prof. Hasan Selim who has dedicated his time and effort for this study. I would like to express my sincere gratitude to him for his professional support, guidance and encouragements from the beginning of this research.

I would like to thank to my grateful friend Handan Aldemir who was always with me with an endless patience, support and friendship during writing this thesis. I am also greatly thankful to my special friend Erdem Kullep for his kindly friendship and contribution to getting the best data mining software, SPSS Clementine for this study.

Finally, I would like to thank and acknowledge my parents for their support. And special thanks to my sister Başak Durdu for her endless support, love and also everything from the beginning of this long story.

APPLICATION OF DATA MINING IN CUSTOMER RELATIONSHIP MANAGEMENT: MARKET BASKET ANALYSIS IN A RETAILER STORE

ABSTRACT

In today's world, hard conditions in the market lead the companies to find new ways to compete better. With the intensive global competition and rapidly changing technological environments, meeting customers' various needs and maximizing the value of profitable customers are becoming the only viable option for many contemporary companies. Customer Relationship Management (CRM) provides organizations with the platform to obtain a competitive advantage by embracing customer needs and building value driven long-term relationships.

CRM is an iterative process that turns customer data into customer loyalty. Analyzing the customer database and convert the data into information that will help company develop programs for building customer loyalty. In the analysis of this data, data mining techniques are essentially used. Association rules are one of the most frequently used methods which are the special application areas of the data mining. Association rules are the rules that include which items commonly occur together in the same transactions. The Apriori algorithm is the most popular association rule algorithm which discovers all frequent itemsets in large database of transactions. This algorithm uses iterative approach to count the frequent itemsets. Using this algorithm, candidate patterns which receive sufficient support from the database and the algorithm uses aprior gen actions join and prune to find all frequent itemsets.

The aim of this study is to propose a base for the customer relationship management activities by using data mining tools and applications for a firm in retail sector. Customer master data and sales transactions of customers are converted to meaningful information that can be used for customer relationship management activities. In this concern, a market basket analysis is performed, and an application was conducted to find association rules from market datasets by using apriori algorithm.

Keywords: Customer relationship management, data mining, market basket analysis.

MÜŞTERİ İLİŞKİLERİ YÖNETİMİNDE VERİ MADENCİLİĞİ UYGULAMASI: BİR PERAKENDE MAĞAZASINDA MARKET SEPET ANALİZİ

ÖZ

Günümüz dünyasında, pazarda yaşanan yoğun rekabet şirketleri daha iyi rekabet edebilmek için yeni arayışlara itmektedir. Yoğun küresel rekabet ve hızla değişen teknolojik ortamlarda müşterilerin çeşitli ihtiyaçlarını karşılamak ve karlı müşterilerinin değerini maksimize etmek birçok çağdaş şirket için tek uygun seçenek haline gelmektedir. Müşteri ilişkileri yönetimi organizasyonlara, müşteri ihtiyaçlarını karşılayarak ve değer odaklı uzun vadeli ilişkiler kurarak rekabet avantajı elde etmek için bir platform sağlar.

Müşteri ilişkileri yönetimi, müşteri verilerini müşteri sadakatine dönüştüren tekrarlı bir süreçtir. Müşteri veri tabanının analiz edilmesi ve verilerin bilgiye dönüştürülmesi, şirketin müşteri sadakatini oluşturması için programlar geliştirmesine yardımcı olacaktır. Bu veri kümelerinin çok büyük hacimde olması nedeniyle analizlerde kaçınılmaz olarak veri madenciliği tekniklerinin kullanılması gerekmektedir. Veri madenciliğinde en sık kullanılan yöntemlerden biri ise birliktelik kurallarıdır. Birliktelik kuralları, aynı işlem içinde çoğunlukla beraber görülen nesneleri içeren kurallardır. Apriori algoritması, veri madenciliğinde sık geçen öğelerin keşfedilmesinde en çok kullanılan birliktelik kuralı algoritmasıdır. Sık geçen öğeleri bulmak için veritabanını birçok kez taramak gerekir ve bu taramalar aşamasında Apriori algoritmasının birleştirme, budama işlemleri ve minimum destek ölçütü yardımı ile birliktelik ilişkisi olan öğeler bulunur.

Bu çalışmanın amacı, veri madenciliği araçları ve uygulamalarını kullanarak perakende sektöründe yer alan bir firma için, müşteri ilişkileri yönetimi aktivitelerine temel olabilecek bir yapı geliştirmektir. Bu amaca yönelik olarak, müşteri ana verisi ve satış işlemleri, müşteri ilişkileri yönetimi için kullanılabilecek anlamlı verilere dönüştürülmüştür. Bu kapsamda, bir market sepet analizi gerçekleştirilmiş ve market veri setinden, apriori algoritması kullanılarak birliktelik kurallarını bulan bir uygulama geliştirilmiştir.

Anahtar sözcükler: Müşteri ilişkileri yönetimi, veri madenciliği, market sepet analizi.

CONTENTS

CHAPTER ONE
CUSTOMER RELATIONSHIP MANAGEMENT

1.1 Foundation of CRM

CRM is a logical step in the series of major commercial and IT initiatives that have been implemented since 1980s, beginning with downsizing. Most of these early initiatives had a cost-cutting focus on the internal workings of the business, concentrating on employees, working methods, or technology. Increased profitability was the desired result, which was to be engineered through cost savings. All of these initiatives were based on decreasing costs through increasing efficiency, which is one of the key benefits of a successful CRM strategy in addition to its significant impact on the customer (Sharp, 2001). CRM has been one of the most popular terms throughout the world since the early 2000s. However, how did this term appear and where did it come from?

Despite the recent birth of CRM, which stands in the nineties, since then it has become a key tool for business management (Ngai, 2005). Similarly, research on CRM has increased significantly over the past few years (Romano & Fjermestad, 2003), but there are still research needs in different areas: search for a definition or a generally accepted conceptual framework, analysis of its key dimensions, study of CRM impact on business results, barriers to its successful implementation, development of valid and reliable scales to study the degree of implementation and success and rigorous empirical studies on the subject (Colgate & Danaher, 2000; Parvatiyar & Sheth, 2001; Sin, Tse, & Yim, 2005).

In 1990s, the roles of buyer and provider/supplier changed as mentioned. With the supply getting greater than demand, customers’ role for the suppliers has changed from “hunted” to “special”. Before, the leaders of global brands were deciding who the customer would be, the basic idea being “The public wants what the public gets”. But the days of Henry Ford, telling everyone can get any color car that he wants as long as it is black has expired when someone decided to listen to the customers and offer them a second colour (Swift, 2000).

In 1990s, companies started making person-to-person marketing activities, being customer focused, listening and understanding customers. Activities likes ending happy birthday card to customers started in these years. Banks began to offer education credit to their customers who have children. These activities were the first steps of CRM implementation.

The benefits of CRM not only can assist the enterprise to locate the profitable market, but it also improves the competitive advantage, through lowering cost and gaining higher customer value, in comparison with the competitors. However, a real successful CRM should integrate information technology such as basic installation, applicable system, etc., information resource such as customer data base, interview record of salesman, well interaction with customer, and so on, as well as organizational resource, for example, customer-oriented business culture, etc. All these can actually exert the best effectiveness (Pushkala, Michael Wittmann, & Rauseo, 2006).

From the report of Spengler (1999), one can find out that extended functions of ‘‘Contact Management” are: Customer data collection, as well as gathering and application of useful information. It further developed to be the call center, representing the unit or research tool to analyze customer data. To understand CRM system from the aspect of marketing, its ultimate target also involves of how to fit the customer’s requirement; with quest to achieving the objective of establishing the ‘‘Relationship Marketing”, in other words, a long-term customer relationship. The only differentiation is in the application of information technology enhancing its effectiveness (Ryals & Payne, 2001). Kalakota and Robinson (1999) considered that CRM can be seen as the consistent organizational activity under usage of integrated selling, marketing and service strategy. That is, trying to define the real need of the customer, by the enterprise integrating various process and technology, in asking internal product and service improvement, in order to dawn effort of enhancing customer satisfaction and loyalty. Additionally, Kalakota & Robinson (2001) offered the concept of CRM system to synthesize with functions of sales, customer service, and marketing activity, all based on customer orientation. The same idea also served as the developmental foundation of CRM system upgrades in the present.

After reviewing the literature on the concept of CRM (i.e., Paas & Kuijlen, 2001; Parvatiyar & Sheth, 2001; Plakoyiannaki & Tzokas, 2002), we can say that there is not yet a consensus about a clear conceptual framework of the concept of CRM (Zablah, Bellenger, & Johnston, 2004). At the theoretical level CRM clearly offers numerous advantages, but a large number of studies indicate a high failure rate in the implementation of this type of strategy (Xu & Walton, 2005). When examining the various causes of these negative results, several authors (Rigby et al., 2002; Starkey & Woodcock, 2002) suggest that one of the main causes of failure is not integrating CRM into the firm’s overall strategy, in other words, considering CRM as an exclusively technological tool and not assuming the various organizational and cultural changes it entails. Additionally, Sin et al. (2005) argue that there is no integrative conceptual framework that translates the CRM concept into specific organizational activities and guides firms in how to implement the strategy successfully.

As a result, the number of implemented CRM systems, generally in the form of IT databases and communications systems has grown markedly during the past ten years (DeSisto, 2005). In today's competitive business environment, the success of firm increasingly hinges on the ability to operate customer relationship management (CRM) that enables the development and implementation of more efficient and effective customer-focused strategies. Based on this belief, many companies have made enormous investment in CRM technology as a means to actualize CRM efficiently. Despite conceptual underpinnings of CRM technology and substantial financial implications, empirical research examining the CRM technology- performance link has met with equivocal results. Recent studies demonstrate that only 30% of the organizations introducing CRM technology achieved improvements in their organizational performance (Bull, 2003; Cornerand Hinton, 2002).

1.2 Definition of CRM

A general definition of CRM could not be achieved in the literature. CRM has various definitions by different researchers. It has been described as a business tool, a technology component, customer data management, call center or only customized e-mails.

CRM contribute to business excellence and enables a business to keep in tune with the requirements of customers and enhance customer relations and satisfaction. Peters (1988) points out that being close to customers and listening to them are important for a business when it would like to manage change and pursue excellence. Waterman (1987) also emphasizes the importance of regarding information such as customer knowledge as a business's main strategic advantage, and also looking at the business itself from a different perspective, such as that of its customers, for the pursuit of businesses excellence. Kanji (1998) and Kanji & Wallace (2000) argue that customer satisfaction is a critical success factor for business excellence. Therefore, CRM that may create value for customers, inform further quality improvement and enhance customer satisfaction plays an important role in the pursuit of business excellence and a close examination of the CRM strategies of a business is very important for that reason.

There exist many definitions of CRM in the literature. Among them, Chablo (1999) defines CRM as "a comprehensive approach which provides seamless integration of every area of business that affects the customer, namely, marketing, sales, customer service and field support through the integration of people, process and technology, taking advantage of the revolutionary impact of the Internet." Although most of the others are similar to this comprehensive definition, it is necessary to examine a few others.

Peppers, Rogers, & Dorf (1999) describe CRM as a concept that makes it possible to an organization to customize specific products or services to each individual customer. They have focused on four steps, identify, differentiate, interact and customize, for one-to-one marketing.

As the business world shifts from product focus to customer focus, managers are discovering that the enhancement of existing customer relations will be of benefit for profitable and sustainable revenue growth. Brown (2000) defines CRM as 'the key competitive strategy you need to stay focused on the needs of your customers and to integrate a customer-facing approach throughout your organization'. He states that CRM is neither a concept nor a project. CRM is a business strategy, which aims to understand, anticipate and manage the needs of an organization's existing and potential customers. He presents the strategic customer care 5-pillar model to build a CRM model for enterprises. These are strategic, process, organizational and technical change and management of enterprise around customer behavior.

Chatterjee (2000) points out that CRM is a discipline which focuses on automating and improving the business processes associated with managing customer relationships in the area of sales, management, customer service, and support. Actually, CRM is very important because acquiring customers is much more expensive than keeping them. Srivastava et al. (1999) develop a framework for understanding the integration of marketing with business processes and shareholder value. They also emphasize that the CRM process addresses all aspects of identifying customers, creating customer knowledge, building customer values, and shaping customers' perceptions of an organization and its products.

CRM may enable a business to understand better the stated and especially the implied requirements of its customers. With this understanding, a business may have a better opportunity to provide its customers with products or services that are more in tune to their requirements and their view of quality. Russell (1999) argues that it is important for a business to understand not only the view of its internal stakeholders but also that of its external stakeholders such as customers in order to have a clearer sense of direction and prevent changes in the wrong direction.

More simply, Handen (2000) defines CRM as the process of acquiring, retaining and growing profitable customers. Actually, this definition summarizes the core of CRM thought. In order to represent value to the customers and create loyalty, CRM requires an obvious focus on the service attributes. According to Handen (2000), to implement a CRM project effectively, five dimensions are considered important: strategy, organization, technology, segmentation, and process.

In a similar approach, Findlay (2000) mentions that CRM contemplates on the retention of customers by collecting all data from every interaction and from all access points whether they are phone, mail, web or field. The organization can then use this data for specific business purposes, which could be marketing, service, support or sales while concentrating on a customer-centric approach rather than a product-centric approach.

According to Swift (2001), CRM is “an enterprise approach to understanding and influencing customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability”. Parvatiyar and Sheth (2001) state that CRM is a comprehensive strategy and process of acquiring, retaining, and partnering with selective customers to create superior value for the company and the customer. It involves the integration of marketing, sales, customer service, and the supply chain functions of the organization to achieve greater efficiencies and effectiveness in delivering customer value.

CRM is the technology that forms relation between a company and its clients through the customer relationship cycle (Butler Group, 2001). Another definition states that CRM is the variety of methods and contact strategies that the companies use to build lasting and profitable relationship in order to retain the best customers and generate profitable revenue. According to Chen & Popovich (2003), CRM is an enterprise wide customer centric business model that must be built around the customer. Kincaid (2003) defines CRM as “the strategic use of information, processes, technology, and people to manage the customer's relationship with the company across the whole customer life cycle.” According to Reinartz et al. (2004), CRM is the “systematic process to manage customer relationship initiation, maintenance and termination across all customer contact points to maximize the value of the relationship portfolio.” According to Ko et al. (2004), CRM is the integrated customer management strategy of a firm to efficiently manage customers by providing customized goods and services and maximizing customers' lifetime values.

Peppers & Rogers (2004) describe CRM as a set of business practices designed simply, to put an enterprise into closer and closer touch with its customers, in order to learn more about each one and to deliver greater and greater value to each one, with the overall goal of making each one more valuable to the company. It is both an evolution and a revolution. It is evolution in the respect that the permanent change has been experienced in the marketing environment. It is revolution because the merging of change and technology has created an opportunity for the marketers to enter a new period of sophistication in understanding the customers, exactly who purchases their products (Vaura, 1992; 10).

CRM is the technique or set of processes for collecting information from prospects and customers about their needs, and for providing information that helps customer evaluate and purchase products that deliver the best possible value to them. It is a process for managing the company’s resources to create the best possible experience and value for customers while generating the highest possible revenue and profit for the company (Doole et. al., 2005; 280). It is an overall process of building and maintaining profitable customer relationships by delivering superior customer value and satisfaction (Kotler & Armstrong, 2006; 13).

As the definitions reveal, CRM is an interactive process that turns customer information into positive customer relationship. The technology used for data transformation and analyses is very important. With a highly improved technology data, analyses can be made more rapidly and healthy, and as a result, it accelerates the decision making speed of the management. It empowers tight customer contacts, more efficient and useful marketing activities. Thus, the company becomes more informed about their customers. Figure 1.1 shows the CRM process cycle.

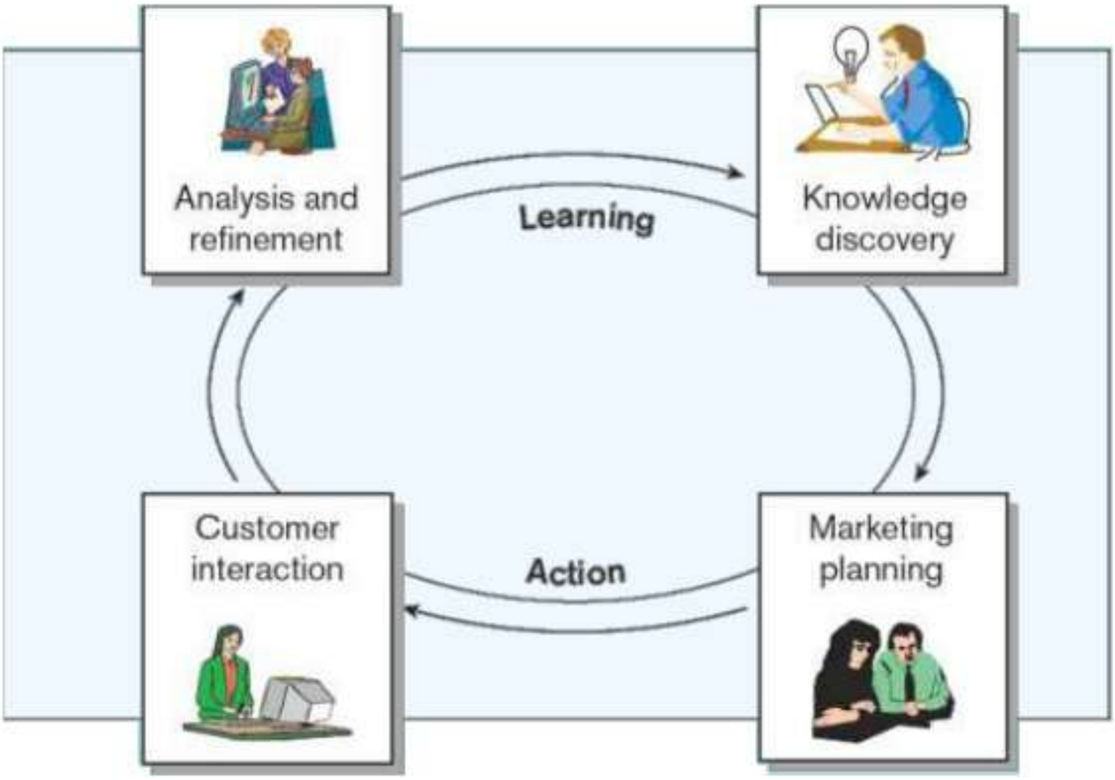


Figure 1.1 CRM process cycle (Swift, 2001)

Note that CRM needs to be associated with everything a company does, everyone it employs, and everywhere it transacts. When a company claims its goal as to implement CRM and to form good customer relationship with good customer services, it should be talking about the whole company.

In fact, the most useful definition of CRM is the term itself: the management of relationships with the customers. The keyword of the term is relationship. However, it is very important to comprehend the meaning of this word "relationship" clearly, because many companies believe that they have relationship with their customer although such a relationship does not exist. These companies conceive of a transaction done by the customer or a sale done as a relationship. In fact, the communication must be bilateral, integrated, recorded and managed in order to call it a "relationship". Without historical data, transaction details and complete customer information it is impossible to talk about forming a stable and lasting relationship. Finally, each company should determine what CRM means for its organization and for the future of it.

As these definitions clearly show, CRM has been defined as a corporate wide approach to understanding customer behavior, influencing it through continuous relevant communication, and developing long-term relationships to enhance customer loyalty, retention, acquisition, and profitability. CRM is often perceived by senior management with mixed feelings on the one hand, it is a great opportunity to enhance customer relationships and to increase revenues and profitability at the same time and on the other hand, it is a costly and time-consuming process that will alter fundamentally the corporate culture. CRM is also fraught with the numerous potential pitfalls that confront any major corporate project involving people, processes, and technologies (Sharp, 2003).

CRM is not a technology or even a group of technologies; it is a continually evolving process that requires a shift in attitude away from the traditional internal focus of a business and defines the approach a company takes toward its customers, backed up by a thoughtful investment in people, technology and business processes (Sharp, 2003).

1.3 Goals of CRM

The goal of CRM is to achieve a competitive advantage in customer management and ultimately increase profit levels (Gartner Group, 2005; 2006).

As CRM is the way of managing relations with the right customers, CRM’s goal is to increase the opportunity of communicating with the right customer, and offering the right product at right price, through the right channel, at the right time. Understanding the historical behavior of customers, information about their buying habits and their ideas, a company could catch the right customers.

The goals stated above can be characterized as follows:

Right Customers: Customer relationships must be managed throughout their life cycle and the customer potential must be realized by increasing “share of wallet”.

Right Offer: The company and its products must be introduced to the customers efficiently, and offers must be customized for each customer.

Right Channel(s): Communications must be coordinated across every customer touch point, each customer must communicate through the channel that he or she prefers, and the channel information must be captured and analyzed for continuous learning.

Right Time: Marketing during the communications with the customers must be, as near to real time marketing as possible and the customer communications must relate.

1.4 The Components of CRM

There have different views about the components of CRM among the researchers. For example, Hansotia (2002; 122) states that there are three distinct components of CRM; strategy design and organizational readiness, planning and analysis, and execution of customer interaction. Sin et. al. (2005) assert that there are four components of CRM; key customer focus, CRM organization, knowledge management and technology-based CRM. When the general viewpoint is analyzed, CRM system is constructed around three components (Rajola, 2003; 26, Karimi et al., 2001; 128).

The analytical component of CRM is the information that the company has to gather to make the customer more valuable and the tools that are used to analyze this information. Data warehouse and data marts play the main part of the analytical component (Rajola, 2003; 26). The other tools are vertical application tools (e.g. data mining, OLAP etc.), marketing automation and campaign manager system that use a data warehouse to plan and execute targeted marketing campaigns to respond to customer behavior (Pan & Lee, 2003; 97). The analytical component contains building an analysis system on the operational system and by this way discovering the potential customers, segmentation, and giving one-to-one marketing services. At the formation of marketing and company strategies, analyzing data correctly is very important. Marketing, analysis of sales and service operations and customer behavior type, customer value and customer portfolio analysis are included in this category.

The operational component of CRM is the process of achieving a long-term relationship with customers, across all available touch points through customized products, so that the contribution from each customer to overall profitability of the company is maximized (Ramaseshan et. al., 2006; 196). Operational CRM focuses on the software installations and the changes in process affecting the day-to-day operations of a company (Peppers & Rogers, 2004; 8).

It is the first category remembered which is defined as an automation system that helps to see the customer contact points, channels and work processes as a whole. Supply chain management, post sales service, marketing automation, sales automation and mobile sales are included in this category.

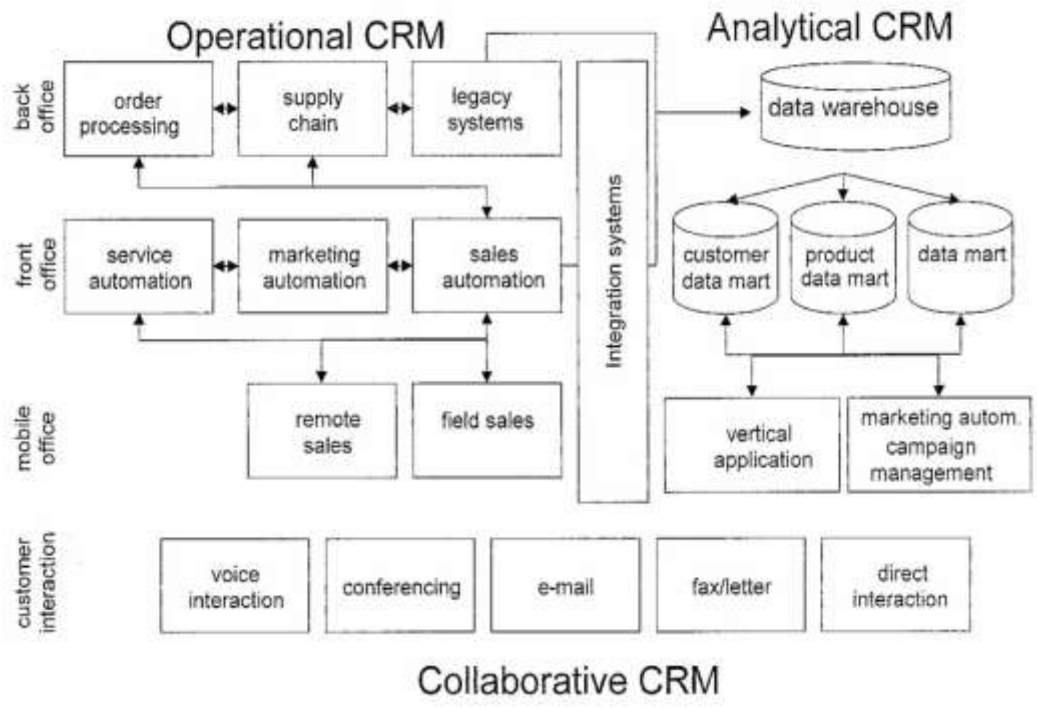


Figure 1.2 Components of CRM.

The collaborative component of CRM is the collaboration of the customer and the company for a mutually beneficial relationship (Peppers & Rogers, 2004; 22) through direct interaction, e-mail, fax/letter, conferencing and voice interaction (Rajola, 2003; 28).

These are the functions that are built on the logic of detailing special services for the customers more by sharing the data of customers with the work partners, channel and suppliers. Direct connection, telephone (call center), web and letter/fax are included in this category.

CHAPTER TWO
DATA MINING

2.1 Introduction

A simple definition of data mining in marketing is: extraction of previously unknown, comprehensible and actionable information from large repositories of data, and using it to make crucial business decisions and support their implementation, including formulating tactical and strategic marketing initiatives and measuring their success (Stone & Foss, 2001; 67).

Data mining method is widely used around the world for processing data via usage of many classifying, clustering, associating tests on data attributes and instances. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cut costs, or both.

Data Mining is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Berry & Linoff, 2004). In other words, data mining is a business process for maximizing the value of data collected by the business.

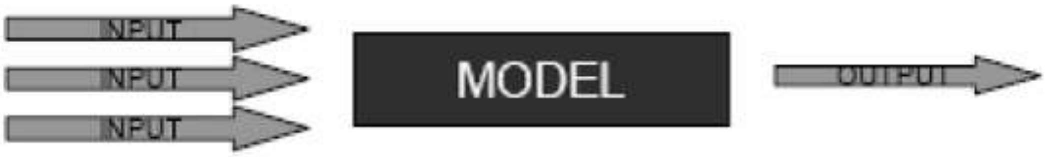


Figure 2.1 Simple diagram of data mining.

As shown at the figure above, data that is prepared to analyze is put into the model and prediction or pattern that is used to build strategies is got. Data mining is an iterative learning process and requires long-term hard work and commitment. As everything in life, data mining needs effort but the successful result transforms the company from being reactive to being proactive.

2.2 The Process of Data Mining

In general, the process of data mining can be summarized by the following four steps:

- ✓ Defining the objectives of the analysis.
- ✓ Collecting and preprocessing of the data.
- ✓ Using data mining techniques to transform the data into valuable information.
- ✓ Interpreting the model and drawing conclusions.

The basic steps are illustrated in Figure 2.2 (Hui and Jha, 2000):

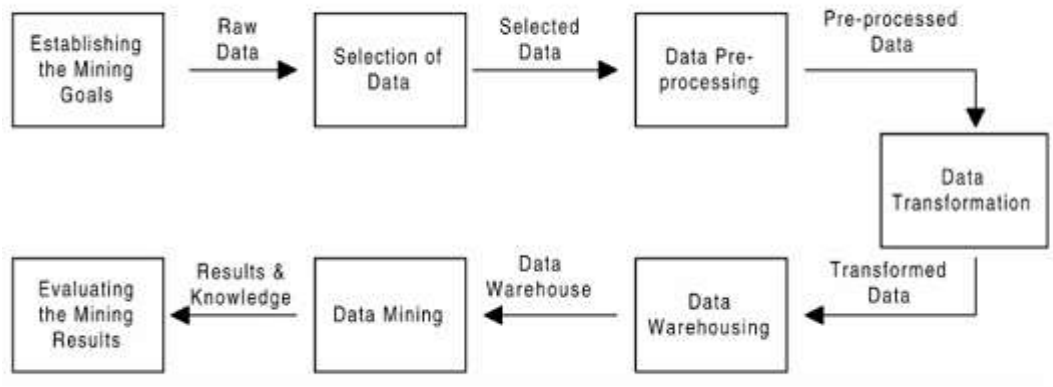


Figure 2.2 Steps of a data mining process.

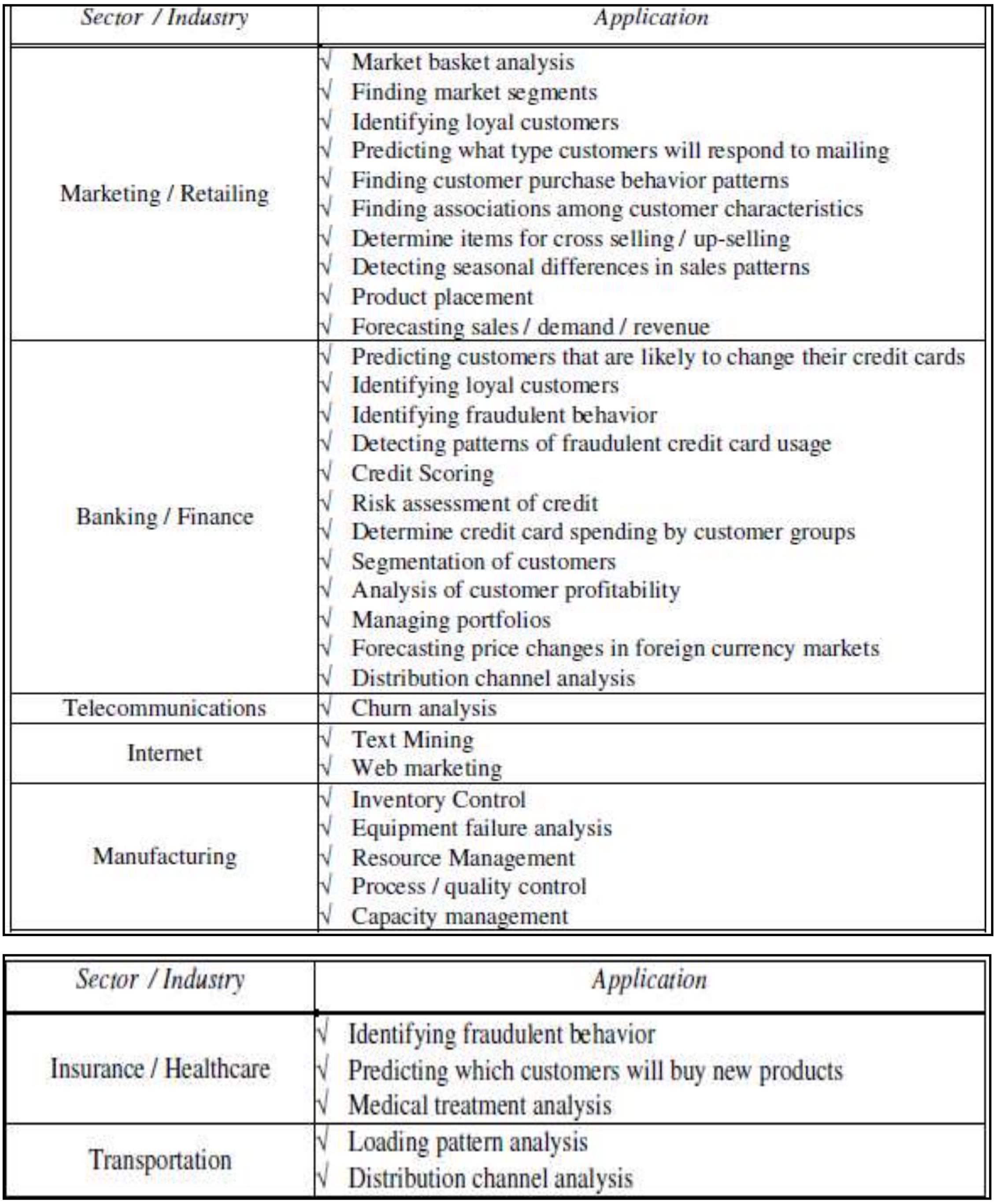
Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid prediction.

The first and simplest analytical step in data mining is to describe the data summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables (such as values that often occur together). But the data description alone cannot provide an action plan. You must build a predictive model based on patterns determined from results, and then test that model on results outside the original sample. The final step is to empirically verify the model. For example, from a database of customers who have already responded to a particular offer, you've built a model predicting which prospects are likeliest to respond to the same offer.

Data mining is primarily used today by companies with a strong consumer focus-retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning or staff skills, and "external" factors such as economic indicators, competition and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

Data mining is a broad technology that can potentially benefit any functional areas within a business where there is a major need or opportunity for improved performance and where data is available for analysis that can impact the performance improvement. Table 2.1 shows examples of business applications in various sectors and industries that can most benefit from data mining (Musaoğlu, 2003).

Table 2.1 Examples of data mining business applications in various sectors (Musaoğlu, 2003)



Sector / Industry	*Application*
Marketing / Retailing	√ Market basket analysis √ Finding market segments √ Identifying loyal customers √ Predicting what type customers will respond to mailing √ Finding customer purchase behavior patterns √ Finding associations among customer characteristics √ Determine items for cross selling / up-selling √ Detecting seasonal differences in sales patterns √ Product placement √ Forecasting sales / demand / revenue
Banking / Finance	√ Predicting customers that are likely to change their credit cards √ Identifying loyal customers √ Identifying fraudulent behavior √ Detecting patterns of fraudulent credit card usage √ Credit Scoring √ Risk assessment of credit √ Determine credit card spending by customer groups √ Segmentation of customers √ Analysis of customer profitability √ Managing portfolios √ Forecasting price changes in foreign currency markets √ Distribution channel analysis
Telecommunications	√ Churn analysis
Internet	√ Text Mining √ Web marketing
Manufacturing	√ Inventory Control √ Equipment failure analysis √ Resource Management √ Process / quality control √ Capacity management
Insurance / Healthcare	√ Identifying fraudulent behavior √ Predicting which customers will buy new products √ Medical treatment analysis
Transportation	√ Loading pattern analysis √ Distribution channel analysis

Databases today can range in size into the terabytes (over 1.000.000.000.000 bytes). For example AT&T, one of the leading telecommunication companies, handles over 250 million long distance calls daily. Within these masses of data lies hidden information of strategic importance. But, how meaningful conclusions can be drawn? The newest answer to the question is data mining. In simple terms, data mining can be defined as the automated extraction of predictive information from databases.

According to Berry and Linoff, data mining is the exploration and the analysis, by automatic and semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. But, in a more business-eyed way data mining can be defined as the process of extracting valid, useful, unknown, and comprehensible information from data and using it to make business decisions.

There are several tools in the data mining market, of which more than 50 listed in the KDNuggets web site (www.kdnuggets.com). Hence, choosing a tool for the company's needs may seem to be a big problem and must be done in a systematic manner. According to the market shares, the top three market leaders are: Clementine of SPSS, Enterprise Miner of SAS Institute, and Intelligent Miner of IBM (Groth, 1999), so their perspectives about data mining will be summarized in the following chapter. Data mining is an interactive and iterative process. So, concentration on understanding relationships and features that underlie in the data before using any data mining technique would be valuable.

Companies are experiencing a fast changing environment and also almost everybody knows that enterprises who adapt to this changing environment will survive and the others most likely will die. There is a tool, which has been either not discovered or not totally utilized. That is the power of the data. But the data itself doesn't make sense without exploitation. In other words, as Davenport, Harris, and Kohli (2001) state "Companies may know more about their customers, but most of them don't know the customers themselves or how to attract new ones." For that reason those who utilize this power which is come from several internal and external sources will survive and others will be out of business.

It is a well known fact that a customer database moves a company from a reactive to a proactive context in business building. Chye and Gerry (2002) point out that "Examining and analyzing the data can turn raw data into valuable information about customer's needs. By predicting customer needs in advance, businesses can then market the right products to the right segments at the right time through the right delivery channels."

Data mining tools give companies the ability to predict what will happen next based on the past experiences. That wisdom makes data mining a valuable step through customer relationship management because if the responses or behaviors of customers were known formerly then necessary precautions could have been taken to retain them in hand.

As Berson, Smith, and Thearling (1999) state, “we can define information as that which resolves uncertainty. We can further say that the decision-making is the progressive resolution of uncertainty and is a key to a purposeful behavior by any mechanism (or organism).”

Data mining describes a collection of techniques that aim to find useful but undiscovered patterns in collected data (Berson, Smith, and Thearling, 1999, Preface). Data mining gives a company the wisdom of decision making when it is used to address some strategic business objectives.

As Berson, Smith, and Thearling (1999) emphasize in their study, “Companies should give their customers what they exactly need.” Data mining helps companies achieve their goals of customer retention because with good exploitation of data mining companies can determine customers who has propensity to churn and may prepare a marketing campaign tailored to them to convince them stay with them. So data mining firstly helps segmentation and customer segmentation makes custom-tailored marketing possible.

Data mining enables the analysis of large quantities of data to discover meaningful patterns and relationships (Payne & Frow, 2005) and to discover insights of customer needs (Paas & Kuijlen, 2001). Knowing each customer through data mining techniques and a customer-centric business strategy helps the organization to proactively and steadily offer more products and services for improved long-term customer retention and loyalty (Chen & Popovich, 2003). According to Stone and Foss (2001), data mining helps marketing managers in the following ways:

It helps them understand and predict future customer actions (Dyche, 2002) in a complex (multifactor) world. It helps them discover customer groupings which would be hard to discover using theory-based hypothetical world, for profiling, needs analysis and focused actions.

2.3 Development of Data Mining

Data mining takes advantages in the fields of Artificial Intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification. Both communities have made great contributions to the understanding and application of neural nets and decision trees. Data mining does not replace traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed (Larose, 1999). The increased power of computers and their lower cost, coupled with the need to analyze enormous data sets with millions of rows have allowed the development of new techniques based on a brute force exploration of possible solutions.

Data mining is a tool for increasing the productivity of people trying to build predictive models. Innovative organizations worldwide are already using data mining to locate and appeal to higher value customers, to reconfigure their product offering to increase sales, and to minimize losses due to errors or fraud.

Many organization are using data mining to help manage all phase of customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who have bought particular product it can focus attention on similar customers who have not bought that product (cross-selling). By profiling customers who have left, a company can act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one.

2.4 Data Mining Techniques

Building a model to represent the data set is at the heart of data mining process. Implementing methods to the models, there exists lots of algorithms stemming from statistics, data mining and artificial intelligence. While some techniques belong to completely different approaches, also techniques may vary with analyzer approach differences and combinatory usage in models. Because of the wide spectrum of techniques, only some popular algorithms will be mentioned in this section.

2.4.1 Decision Trees

Decision Tree technique develops a classification model from a set of records. Each record in the training set is assigned to one of the many predefined classes which represent the records best as a general concept description. After the model is set, the model can be used for automatic prediction of the class of unclassified records. If each node of decision tree has two branches at most the tree is called as Binary Tree, if node can have more than two branches the tree is called n-way (Multi-way) tree.

The representations of this technique are easier to understand, and their implementation is more efficient than those of neural network or genetic algorithms. This technique can be used in data exploration and modeling phases and especially useful when there are many ways to reach the target. For example a profitable customer can be a high premium paying customer from the commission point of view or low loss ratio customers from the profit sharing point of view. A simple decision tree structure is illustrated in Figure 2.3.

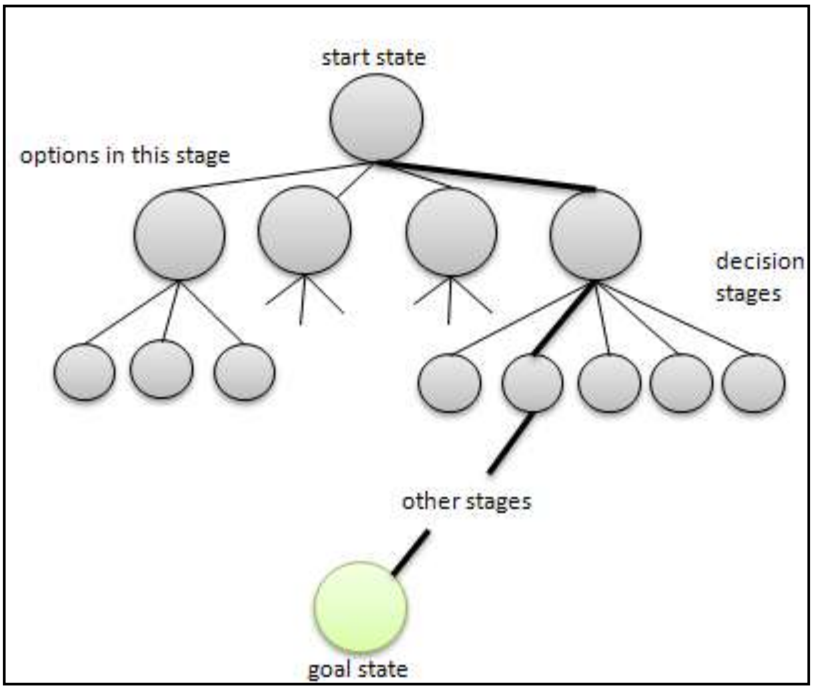


Figure 2.3 Sample decision tree structure.

Decision tree consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The best split is defined as one that does the best job of separating the data into groups where a single class predominates in each group. Each node has more homogeneous data set and similar size for best split. To test the data sets are homogeneous or not, some tests like Gini, Entropy, Chi-Square are used. These tests are also used for another primary consideration when developing a tree, deciding on how large to grow the tree or what nodes to prune off the tree, in other words limit the tree (Berry & Linoff, 2004).

Specific decision tree methods include Classification and Regression Trees (CART), Chi-squared Automatic Interaction Detection (CHAID) algorithm, C5.0 and Quest. Although decision trees are popular, easy and visually powerful, they are limited to one output variable that must be categorical, and processing numeric values can be complex. Moreover, separation of categories with strict rules like customer having less than 60% loss ratio is profitable but 60.01% loss ratio is unprofitable may lead biases in decision (Berry & Linoff, 2004; Edelstein, 2000; Larose, 1999; Guo, 2003; Hand, Mannila & Smyth, 2001).

2.4.2 Neural Networks

Neural Networks is a class of powerful, general-purpose tools applicable in prediction, classification, link analysis and clustering models. Neural networks are popular since they enable efficient modeling of large and complex problems in which there may be hundreds of input attributes having many interactions like biological neural networks. This widely used technique imitates the way the human brain learns and uses rules coming from data patterns to construct hidden layers of logic for analysis. A neural net technique represents its model in the form of nodes arranged in layers with weighted links between the nodes. The technique can be divided into two as directed and undirected.

Directed neural net algorithms such as Back Propagation and Perceptron require predefined output values to develop a classification model. Among the many algorithms, Back propagation is the most popular directed neural net algorithm which is used since 1980's. Back propagation can be used to develop not only a classification model, but also a regression model.

Undirected neural net algorithms such as ART do not require predefined output values for input data in the training set and employ self-organizing learning schemes to segment the target data set. Such self-organizing networks divide data set into clusters depending on similarity and each cluster represents an unlabeled category. Kohonen's Feature Map is a well-known method in undirected neural networks (Cerny, 2001).

In the process of a neural network, inputs with weights go through the activation function that consists of the combination function that combines the inputs into a value and the transfer function that calculates the output using combination function output. The cycle is repeated iteratively to optimize the target value. Each iteration passing through all nodes is called an epoch.

Neural network supports us to develop a model by using historical data that are able to learn just as people. Actually, the neural networks can be simple and also very complex in nature. A simple network may consist of a couple of inputs and one output equals to linear regression statistical technique when combination is weighted sum and transfer function is linear. Moreover, in most of the Neural networks, there are one or more additional layer between the input and output layer which are called ―hidden layers. The size of this layer increases the efficiency of the network but also raises the risk of over fitting. These networks can produce more than one output. If the hidden layer has certain non-linear activation functions, more specifically, the combination function is weighted sum and transfer function is logistic, these nets are called logistic regression.

Perceptron is the simplest NN. In this architecture, there is a single neuron with multiple inputs and one output. A network of perceptrons is called a multilayer perceptron (MLP). MLP is the simple feedforward NN and it has multiple layers (Dunham, 2003). Multilayer perceptron is a bit more complicated, that is, the inputs are combined in each hidden node by weighted sum and a transfer function (hyperbolic tangent). The outputs of the hidden nodes are combined again inside the output node.

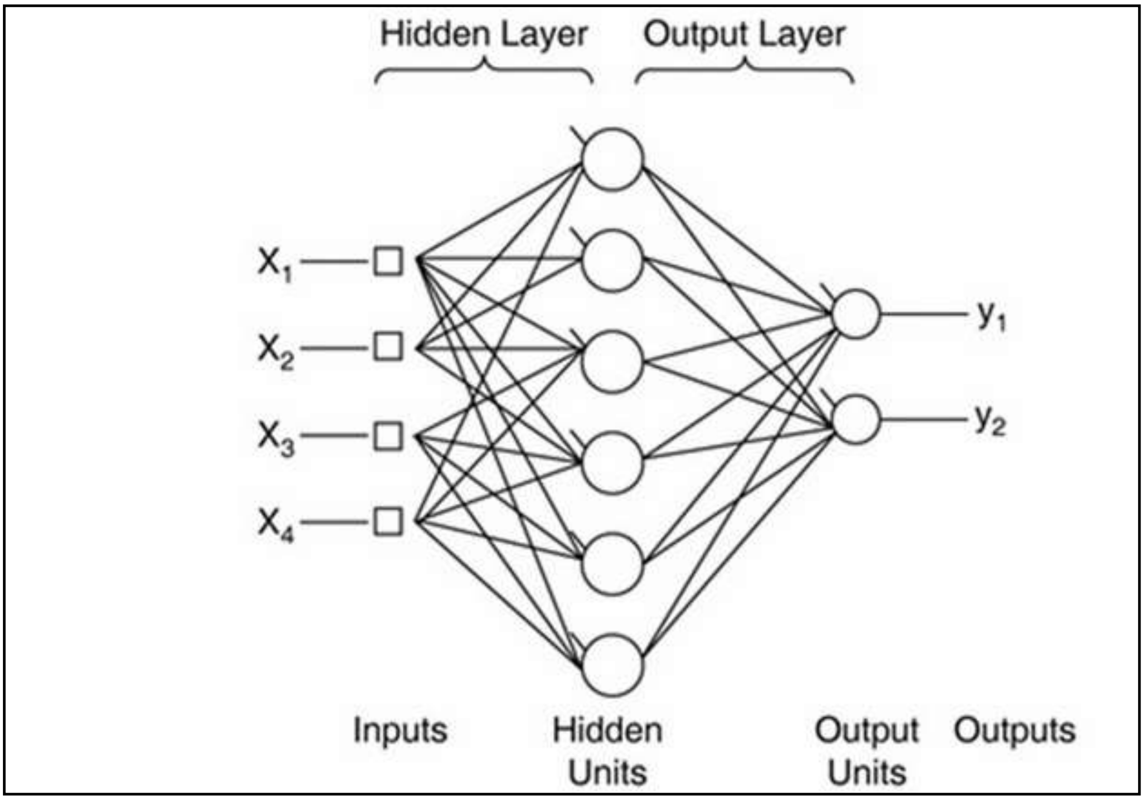


Figure 2.4 Schematic diagram of a neural network

The weights of node connections are estimated by a training method. However, finding the best set of weights for the network in many alternatives under time limitation is a difficult problem that affects the success of the model to a great extend. Backpropagation is the most commonly used learning technique. It is easily understood and applicable. ―It adjusts the weights in the NN by propagating weight changes backward from the sink to the source nodes (Dunham, 2003). Some of other methods are conjugate gradient, quickprop, quasi-Newton, Levenberg-Marquardt and genetic algorithms. Each training method has a set of parameters that control various aspects of training such as avoiding local optima or adjusting the speed of conversion. As an example, back propagation begins a first training then calculates the error by taking the difference between the calculated result and the expected actual result. The error feedback is used to make adjustments to minimize the error. The name back propagation comes from the iterative method of sending the errors back through the network. However, quickprop uses some partial derivatives to fit a multidimensional parabola and converges minimum error in short time.

The greatest strength of neural net models is their ability to approximate any continuous function without making any assumptions about the underlying form of the function to be approximated. The linear combinations of sigmoid surfaces generated increases in the capability of estimation but also complexity.

2.4.3 Nearest Neighbor Techniques

General tendency in finding solutions to new problems is often by the review of the similar situations faced before. Memory Based Reasoning or in other words, Nearest Neighbor classification techniques operate on the same principle that the results are based on analogous situations in the past. Nearest neighbor algorithm predicts unknown values for records in a data set based on a combination of values for the records most similar to it in an historical dataset. These most similar data are the neighbors giving the name nearest neighbor. For example, if the patient has a circular rash and has recently been bitten by a tick, the patient possibly has Lyme disease because circular rash is the first symptom many patients notice. If the patient later develops fever and joint pain, the diagnosis becomes more certain because these are symptoms that often follow the initial rash in Lyme disease.

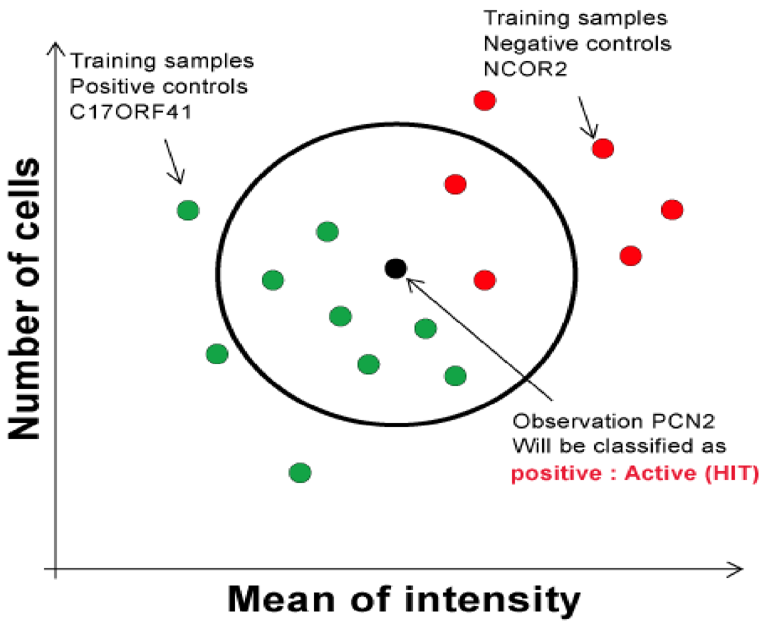


Figure 2.5 Groups of similar records Nearest-neighbour.

The nearest-neighbour supervised method first involves the construction of hypothetical siRNAs that best fit the desired patterns. The technique then finds individual siRNAs that are most similar to the hypothetical genes.

The flow of the technique is finding the neighbors of data and combining the records to form a new class. Calculating and deciding on the distance between the records is the first step. The classification goes on using distances to desired level of neighborhood level. Some distant neighborhoods can be weighted lower in the model with respect to the closely related classes. The results of the neighbors are combined to a result set.

The decisions on the distance are difficult when multi dimensions and nonnumeric fields exist and in many sets this is the case. Multi dimensional sets are treated attribute by attribute and the categorical attributes are turned computable numeric values. Memory based reasoning also requires long computations while operating on each record.

Some applications of memory based reasoning are fraud detection, customer response prediction and classification, medical diagnosis and treatments and some interesting applications like face recognition. E.g. the technique can be used to recommend movies based on the votes and neighborhood of the other clients in a movie store, called collaborative filtering method (Berry & Linoff, 2004; Larose, 1999).

2.4.4 Regression

Regression algorithms predict the value of a continuous attribute of an instance using the given values for other attributes. In the process, a function is found to return the value of the continuous attribute given the values of other attributes. The function is used for prediction of missing continuous value of an incoming instance. As an example a regression algorithm can be used to predict a safe credit card limit value for a customer. Regression is a common technique easily used in combination with other data mining techniques like decision trees that is for example, each split in a decision tree is chosen to minimize the error of a simple regression model on the data at that node.

2.4.4.1 Linear Regression

Regression is based on correlated attributes used to predict from one to the other. The most common and easiest type of regression is linear regression. Linear regression tries to fit a straight line to explain behavior and this line is used to estimate a value for one variable given that the other.

Simply, the process of the linear regression is to fit the data to the equation $Y = \beta_0 + \beta_1 X_1$. The fitting aims to minimize the distance between the observed data points and the equation line. The linear regression can be easily applied and explained. However, linear regressions are not generally sufficient enough for real world cases. In real life, interactions are so complex that multiple variable techniques have to be utilized. Hence, other techniques like multiple regression, logistic regression, decision trees and neural nets are necessary for these cases (Berry & Linoff, 2004; Larose, 1999).

2.4.4.2 Multivariate Adaptive Regression

The Multivariate Adaptive Regression Splines (MARS) is developed by the inventors of classification and regression tree (CART). One of the developments in MARS compared to CART is replacing the hard splits to a continuous transition. This is modeled by a pair of straight lines in each node and leading to a smooth function (spline) at the end. Furthermore, dependence to the predecessors in a tree can be eliminated with MARS but the tree structure of CART is not used in MARS to produce rules. Instead, MARS algorithm can determine most important variables for prediction, associations between these variables and dependence to these variables. MARS is an automatic non-linear step-wise regression tool. Over fitting is also a problem of MARS like neural nets and decision trees. Cross validations with validation set or test set are useful for this problem (Larose, 1999).

2.4.4.3 Logistic Regression

Logistic regression tries to fit a curve to observed data instead of a line like in linear regression. The technique is a special case of generalized linear modeling using odds ratios. The algorithm compares the odds of the event of one category to the odds of the event in another category. Odds ratio is simply the ratio of the probability of being in that class to the probability of not being in that class. For example, if the probability of having exam is 20%, then odds ratio is 20% / (1- 20%), 25%. Odds function is a non-symmetric function going infinity. However, using logistic function adds some advantages like a symmetric function having negative values to positive values. As a result, log odds are the basis of logistic regressions.

Logistic regression formulation becomes as in $(Y/(1-Y)) = \beta_0 + \beta_1 X_1$. The logistic function itself has a characteristic *S* shape. The parameters on the model shift the curve left or right and stretch or compress the curve.

This function has useful properties like being around 0, having the slope about 45% and moving about the region of -1 to 1. Beyond this range, it gradually flattens out, saturating at 100% or 0%. These properties make the logistic a natural curve for expressing probabilities.

Ease of interpretation is one advantage of modeling with logistic regression. It is useful in binary and discrete variables. However, in large data sets, high dimensionality makes the detection of nonlinearities and interactions difficult. In addition, it is very likely that some segments of the data space have more records than other segments. When the data is not evenly distributed, the model that fits the whole data space might not be the best choice depending on the intended application. Although there are many existing methods such as backward elimination and forward selection that can help data analyst to build logistic regression model, judgment should be exercised regardless of the method selected (Berry & Linoff, 2004; Guo, 2003; Hand, Mannila & Smyth, 2001).

2.4.5 Clustering

Clustering techniques are employed to segment a database into groups, each of which shares common properties. The purpose of segmenting a database is often to summarize the contents of the target database by considering the common characteristics shared in a cluster. Clusters are also created to support the other types of data mining operations, generally cluster analysis are followed by other data mining techniques. Clustering is a tool used primarily for undirected data mining with no pre-classified training data set and no distinction between independent and dependent variables, but can be used for directed data mining for forming marketing segments which is a popular application of clustering.

A database can be clustered by traditional methods of Gaussian mixture models, Agglomerative clustering, Divisive clustering, undirected neural nets such as ART and Kohonen's Feature Map, conceptual clustering techniques such as COBWEB and UNIMEM, or Bayesian approach like AutoClass.

Conceptual clustering algorithms consider all the attributes that characterize each record and identify the subset of the attributes that will describe each created cluster to form concepts. The concepts in a conceptual clustering algorithm can be represented as relationships of attributes and their values. Bayesian clustering algorithms automatically discover a clustering that is maximally probable with respect to the data by a Bayesian approach. The various clustering algorithms can be characterized by the type of acceptable attribute values such as continuous, discrete or qualitative; by the presentation methods of each cluster; and by the methods of organizing the set of clusters, either hierarchically or into flat files. K-means algorithm, which is the most popular method, Gaussian mixture models, Agglomerative clustering and Divisive clustering will be explained in the following sections.

Since the method needs some evaluation like deciding on *K* in *K*-means or ending level of Agglomerative clustering, best tool for this decision is variance because of the similarity is the concern. Best cluster is the one with the lowest variance. If the cluster size is big, average variance is an evaluation technique. However, since agglomerative and divisive clustering begins or ends with zero variance, the time that elapses between when the cluster is formed and when it is merged into another, larger cluster is a more suitable way to evaluate. Another measure that works for all clustering techniques is to compare the average distance between cluster members and the cluster centroid (seed) with the average distance between cluster centroids (using the distance metric that is used to create the clusters in the first place).

Clustering can be utilized to understand large amounts of data, customer segmentation, reducing records and to break up large data sets into smaller homogeneous pieces (Berry & Linoff, 2004; Edelstein, 2000; Guo, 2003).

2.4.5.1 K-means Clustering

K-means clustering is an undirected method that looks for a number of clusters which are defined in terms of proximity of data points to each other. In other words, the method depends on a measure of distance or similarity between points. Different distance metrics used in k-means clustering can result in different clusters.

K-means clustering assumes a geometric interpretation of the data that is, the records are points in an n-dimensional data space. Algorithm begins with determination of number of clusters, which is presented with *K*. Selection of *K* data points (seeds) randomly is the second step. Remaining data are assigned to the closest cluster, bind to one of the *K* seeds. The mean of the each cluster is calculated and the seed of each cluster becomes this calculated mean. Since the place of each cluster's seed has changed, reassignment is needed to find new closest records. The cycle goes on until no change occurs by the recalculations and seeds and clusters are fixed. Due to the iterative process, beginning seeds are important for the solution time performance. The method is efficient provided the initial cluster seeds are intelligently placed. The distance is Euclidean, so the K-means algorithm attempts to minimize the sum-of-squares. The beginning assumption of the number of clusters may change the results if there isn't any reliable reason of this selection. In this case, it is possible to run the algorithm for different *K* values and determine the result by providing lowest variance etc. (Berry & Linoff, 2004; Guo, 2003).

An example to K-Means clustering is sizing military clothes (Edelstein, 2000). The standard sizing of women's clothes where all dimensions increase together and further diversified by body measures causes many different sizes that is difficult to manage and causes high costs. The analyst came with a radical approach and a clustering algorithm, e.g. K-Means, is applied to solve the problem. The detailed body measures are analyzed and the clusters by just a few variables were formed. The study results in less variety of sizes and better fitting uniforms.

2.4.5.2 Gaussian Mixture

Gaussian mixture is very similar to K-Means clustering adding model a probabilistic approach. Gaussian distribution, a probability distribution often assumed for high-dimensional problems, has given method the name Gauss. The algorithm starts by choosing *K* seeds with a difference however, the seeds are considered to be the means of Gaussian distributions. Remaining work is to optimize the parameters of each Gaussian and the weights used to combine them to maximize the likelihood of the observed points.

The algorithm iterates over two steps called the *estimation step* and the *maximization step*. The estimation step calculates the responsibility that each Gaussian has for each data point. Each Gaussian has strong responsibility (weight) for points that are close to it and weak responsibility for points that are distant. In the maximization step, the mean of each Gaussian is moved towards the centroid (seed) of the entire data set, weighted by the responsibilities. These steps are repeated until the Gaussians are no longer moving. The reason this is called a “mixture model” is that the overall probability distribution is the sum of a mixture of several distributions (Berry & Linoff, 2004).

2.4.5.3 Agglomerative Clustering

Agglomerative clustering starts with each record forming a cluster. A similarity matrix is created and the closest clusters are merged to reduce the number of clusters and increase the members of clusters. The similarity matrix is renewed with the new clusters and this process continues until all records are in one big cluster. For the distance calculations among clusters in multidimensional cases three different ways can be utilized. In the single linkage method, the distance between two clusters is given by the distance between the closest members. This method produces clusters with the property that every member of a cluster is more closely related to at least one member of its cluster than to any point outside it. In the complete linkage method, the distance between two clusters is given by the distance between their most distant members. This method produces clusters with the property that all members lay within some known maximum distance of one another. In the third method, the distance between two clusters is measured between the centroids of each. The centroid of a cluster is its average element.

2.4.5.4 Divisive Clustering

Opposite approach to Agglomerative clustering is utilized that all data set is one cluster at the beginning to be split into smaller parts. The set is divided into two clusters minimizing the variance and the process goes further by dividing the new clusters to converge zero variance in clusters. As it can be understood from the process, the method is similar to decision trees aiming pure clusters (Berry & Linoff, 2004).

2.4.5.5 Kohonen Maps

Kohonen map, in other words self-organizing map, is a neural network-based approach to clustering. The basic map has an input layer which is connected to the inputs like neural networks and an output layer. As in other neural networks, each unit in the Kohonen map has an independent weight associated with each incoming connection. In contrast to the multilayer perceptrons, the output layer consists of many units instead of just a handful. Each of the units in the output layer is connected to all of the units in the input layer, but not to each other. The output layer is laid out like a grid.

Competitive learning is an adaptive process in which the neurons in a neural network gradually become sensitive to different input categories, sets of samples. A division of labor occurs in the network when different neurons form specialization on different types of inputs. The specialization is enforced by competition among the neurons, that is when an input is feed to the network, the neuron that is best able to represent the input wins the competition and is allowed to learn it even better.

When a record of the training set is presented to the network, the values of the record flow forward through the network to the units in the output layer. The units in the output layer compete with each other to be output of the network and the one with the highest value wins. The reward is not only to adjust the weights leading up to the winning unit to strengthen its response to the input pattern but also the paths to its neighbors in the grid are strengthened as well. This way, the number of clusters found is not determined by the number of output units because several output units may together represent the cluster. Clusters similar to each other should be placed closer than more dissimilar clusters (Berry & Linoff, 2004; Kaski, 1997).

2.4.6 Association Discovery

Association discovery techniques discover the rules to identify affinities among the collection of items. In other words, association rule extraction algorithms try to find some relationships or hidden patterns in data. Given data, these algorithms try to find rules in *if-then* form (Agrawal, Imielinski, & Swami, 1993). For example, given an insurer’s sales database, such an algorithm can produce a rule like if a customer buys a travel insurance, he also buys a health insurance”. As a data mining technique, association rules focus almost exclusively on categorical data rather than on numerical data.

The algorithms discover the affinity rules by sorting the data while counting occurrences to calculate confidence. There are a variety of algorithms to identify association rules such as *Apriori algorithm* and using random sampling. *Bayesian Net* can also be used to identify distinctions and relationships between variables. Association rules may be categorized as actionable rules that contain high quality usable information, trivial rules that tell what everyone already knows and inexplicable rules which cannot be explained and not actionable. Unfortunately, the results are generally trivial and inexplicable rules.

Association rules are a good technique for exploring item-based data to determine which things co-occur. The probabilities and joint probabilities in the co occurrence matrix are calculated and processed to determine rules. An association rule uses measures of support, confidence and lift to represent the strength of association in between. The support is the proportion of market baskets where the rule is true (e.g. if it is thought that who buys rubber also buys pencil, the support is the proportion of baskets containing rubber and pencil together). The confidence is the probability of record occurrences given the triggered part (e.g. the support ratio divided by the proportion of baskets having pencil).

The term Lift (improvement) is how much better a rule is at predicting the result than just assuming the result in the first place. Lift is the ratio of the records that support the entire rule to the number that would be expected, assuming there is no relationship between the products (e.g. ratio of baskets having rubber and pencil divided by ratio of baskets having rubber and the result divided by probability of pencil in the basket). If lift ratio is greater than 1, it means that the rule is better at predicting the result than just guessing. If lift ratio is less than 1, the rule is doing worse.

The sets having many unique items are challenging for the method. For example, a Supermarket having 10 thousand unique items is leading to more than 50 million 2-item combinations and 100 billion 3-item combinations. In order to solve the problems of this size, usually groups of items are utilized. To conclude, number of groups is important for an efficient process (Berry & Linoff, 2004; Dunham, 2002; Guo, 2003).

2.4.6.1 Market Basket Analysis

Market-Basket analysis is a popular technique of association rules. In retail markets, each customer purchases different set of products in different quantities, in different times. The relations among the products going together to the shopping basket are crucial information for a retailer.

The technique uses the sales information to identify who customers are, to understand why they make certain purchases, to understand products like together purchased products, promotionable products, fast-slow sold products, to take action like promotion decisions, store layouts, package products etc.

Actually market-basket analysis is a combinational technique mainly based on association techniques. The data consists of customer data, transactions and product data. Generally there exists a relational database containing these data (Berry, & Linoff, 2004). Market basket analysis will be explained in the next chapter in detail.

2.4.6.2 Sequence Discovery

Sequence discovery is very similar to association discovery except that the collection of items occurs over a period of time. A sequence is treated as an association in which the items are linked by time. When customer names are available, their purchase patterns over time can be analyzed. For example, it could be found that, if a customer buys a rubber, he will buy pencil within one month 30% of the time (Guo, 2003).

2.4.7 Genetic Algorithms

Genetic algorithms are based on processes in biological evolution like neural networks and nearest neighbor algorithms. The basic idea is that over time, evolution has selected the *fittest species*. The name of genetic algorithms comes from the imitation of biological evolution process in which the members of one generation compete to pass on their characteristics to the next generation, until the best generation (model) is found. For a genetic algorithm, one can start with a random group of data and maximize the fitness function. For example, in clustering analysis, a fitness function could be a function to determine the level of similarity between data sets within a group and optimization of the goodness of fit.

The process includes selection, leaving the weak out, crossover, split and change parts, mutation and random change of a part. Genetic algorithms have often been used with other data mining techniques like neural networks to model data. For example, genetic algorithms are utilized in training neural networks to produce weights and to find distance values in memory based reasoning. They have been used to solve complex problems that other techniques have a difficult time with. The genetic algorithms are generally used for complex scheduling, assignment, resource optimization and classification (Berry & Linoff, 2004; Guo, 2003; Larose, 1999).

2.4.8 Distance Evaluation

Although distance measuring is not directly a technique, it occupies a great place in the analysis of techniques like memory based reasoning, clustering etc. In order to calculate the similarities or differences, distances among the records should be well-defined. Even when the variables are more than one, generally this is the case, or variables are categorical, it is difficult to decide on the distance (Berry, & Linoff, 2004).

2.4.8.1 Distance Measuring

If there exists a multi variable problem that variables are numeric, the distance function is calculated separately for each attribute. The most popular distance functions for numeric fields are the absolute value of difference, square of the difference, normalized absolute value of difference (absolute value of difference divided by maximum difference) and absolute value of difference of standardized values (absolute value of difference divided by standard deviation). A problem that has categorical fields with two possible values is easier to interpret. If the records match distance is 0 whereas don't match value is 1.

However, possible values are much more than two in life. In this case, hierarchy leveling that is a tree with many branches can be utilized. The categories that are closely related placed near each other in the tree and that are not related at all are connected only through the root. The distances can be transferred to numeric values by the number of hierarchy levels reaching to common ancestor as an example.

2.4.8.2 Distance Combination

The distance between the fields are determined in the above topic, however in the case of a multidimensional problem, these distances should be combined to form record to record distance. Some of the ways of combination are Euclidean approach that is square root of the sum of the squares of each field distance, Manhattan approach that is sum of the field distance, weighted approach that is sum of the field distances times field-specific weights.

When there are many categorical fields, the ratio of matching to non-matching fields is a useful measure. When relationships within a record are more important than differences between records, the similarity is not based on size. The angle between vectors is a better similarity metric than straight distance. As an example, sine of the angle can be used as distance.

CHAPTER THREE
MARKET BASKET ANALYSIS

3.1 Introduction

Some people may argue that the basic purpose of CRM is to increase sales revenues. The cash-flow enhancing potential of customer relationships has led to the recognition that customers are market-based assets. Successful implementation of CRM sub-processes can contribute to greater switching costs and loyalty.

Anderson (2006) explains why the future of retail business involves selling smaller quantities of more products. He sums up his idea as the ‘‘98% rule’’ contrasting with the well-known ‘‘80/20 rule’’. The ‘‘98% rule’’ means that in a statistical distribution of the products, only 2% of the items are very frequent and 98% of them have very low frequencies, creating a long-tail distribution. The long-tail shape emerges in the new markets due to three factors: democratization of the tools of production, democratization of the tools of distribution and finally the connection between supply and demand based on online networks. The long-tail distribution depends on the type of retailer. The physical retailer is limited by the store’s size, corresponding to a short tail. Then the online retailers, like Amazon.com, expanded the number of products, creating a longer tail. Finally, the pure digital retailers, like Rhapsody that sells music on line, working with no physical goods, have expanded the long tail even further. The emergence of the ‘‘98% rule’’ in the retail sector has made the software that works with many low-frequency items more relevant and appealing (Cavique, 2007).

Market baskets arise from consumers’ shopping trips and include items from multiple categories that are frequently chosen interdependently from each other. Hence, for large retail assortments, the issue emerges of how to determine the composition of shopping baskets with a meaningful selection of categories.

Shocker, Bayus, and Kim (2004, p.29) call for a better understanding of the connectedness among products. Indeed, it is clear and well known that the purchase of one product can influence purchases of other products. The underlying dynamics of these processes, however, remain less clear. Market basket analysis is a generic term for methodologies that study the composition of a basket of products purchased by a household during a single shopping trip.

A market or shopping basket represents a set of items or product categories included in a retail assortment that a consumer purchases during one and the same shopping trip. Retail managers are interested in better understanding the interdependency structure among categories purchased jointly by their customers for several reasons. Traditionally, insights into cross-category dependencies and corresponding marketing-mix effects are of particular interest for optimizing the overall profitability of retail category management (Manchanda et al., 1999; Chen et al., 2005; Song and Chintagunta, 2006).

There are two main research traditions for analyzing market basket data, namely exploratory and explanatory types of models (Mild and Reutterer, 2003). Exploratory approaches are restricted to the task of discovering distinguished cross-category interrelationships based on observed patterns of jointly purchased items or product categories. In the marketing literature, this is also referred to as *affinity analysis* (Russell et al., 1999). The majority of attempts contributed to this research field so far, however, examine cross-category purchase effects on the aggregate level of demand only. This especially applies to methods aiming at a parsimonious representation of pairwise symmetric association measures derived from cross-tabulations of joint purchases across multiple categories (Dickinson et al., 1992; Julander, 1992; Lattin et al., 1996).

In marketing research practice, meaningful cross correlational structures are merely *determined* by visual inspection. Thus, the marketing analyst usually aims for a parsimonious representation of the cross-category associations in a compressed and meaningful fashion. Multidimensional scaling techniques or hierarchical clustering are typically employed to accomplish this task. The practical relevance of such attempts obviously suffers from their limitations to a relatively small number of categories with symmetric pair wise relationships.

These constraints are successfully resolved by a huge amount of research on association rule discovery stemming from the data mining literature (Agrawal et al., 1995; Anand et al., 1998; Brin et al., 1998; Hahsler et al., 2006), which have seen recent applications in the marketing-related literature (Brijs et al., 2004; Van den Poel et al., 2004; Chen et al., 2005). Following a probabilistic concept, rule-mining techniques derive asymmetric implications (rules) for disjoint subsets of items or categories based on aggregated co-occurrence frequencies (associations).

Rule-mining algorithms are capable of dealing with both very large numbers of categories (or even single items) and shopping baskets. However, the issue of an *average* (or aggregate) market view remains.

The idea of representing cross-category purchase effects at a more disaggregate level is not new to the marketing community but was introduced only recently by Decker and Monien (2003) and Decker (2005). The authors utilize neural networks with unsupervised learning rules as a data compression device which results in a mapping of category purchase incidence vectors onto a set of so-called basket prototypes. In empirical applications, they illustrate that each of these prototypes is responsible for a specific class of market baskets with internally more pronounced (complementary) cross category purchase interrelationships as compared to the aggregate case.

More recently, Reutterer et al. (2006) extend this approach towards a customer segmentation tool with campaign design options for target marketing selection and report encouraging findings from a controlled field experiment. Despite their usefulness for discovering meaningful cross-category interrelationship patterns, the managerial value of all these exploratory approaches to market basket analysis is limited. Since no priori assumptions are made regarding the distinction between *response* and *effect* category (that is between categories that are affected by purchases of other categories and categories that exert a purchase effect) and, more specifically, no marketing variables are directly incorporated in the analytical framework, they provide marketing managers with only very limited recommendations regarding decision-making.

By contrast, explanatory (or predictive) types of multi category choice models mainly focus on estimating the effects of marketing-mix variables on category purchase incidences by explicitly accounting for cross-category dependencies among the retail assortment. Most of these explanatory models for market basket analysis introduced so far are either conceptualized as logit- or probit-type specifications within the framework of random utility theory, excellent state-of-the field reviews are provided by Russell et al. (1997; 1999).

Approaches that contribute to the estimation of segment-specific or even individual level marketing-mix effect parameters are included in the works by Russell and Kamakura (1997), Manchanda et al. (1999), Andrews and Currim (2002).

One practical problem with explanatory models is that the set of categories to be incorporated and simultaneously analyzed for cross-category effects on the selected response category is rather limited (typically, up to four of five categories). Indeed, for multivariate logit or probit approaches, significant improvements of powerful Markov chain Monte Carlo simulation methodologies can help to successfully alleviate estimation problems when the number of product categories to be analyzed increases.

Nevertheless, real-world retail assortments typically consist of dozens or even hundreds of potentially relevant product categories, which cause severe computational problems unless constraints are placed on excessively large covariance matrices. Yet another problem concerns the rather ad hoc selection of relevant categories for basket creation, which often needs to be guided by managerial intuition or a priori knowledge within the respective problem context.

Both exploratory approaches to market basket analysis (lack of implications for managerial decision making) and explanatory multi category choice models (issue of proper category selection because of computational restrictions) are limited in meeting the initially mentioned information requirements of modern retail marketing.

To summarize, the market basket analysis is a powerful tool for the implementation of cross-selling strategies. The input for the market basket analysis is a data set of purchases. A market basket is composed of items bought together in a single trip to a store. The most significant attributes are the transaction identification and item identification. While ignoring the quantity bought and the price, each transaction represents a purchase, which occurred in a specific time and place. This purchase can be linked to an identified customer (usually carrying a card) or to a non-identified customer. For instance, if a specific customer's buying profile fits into an identified market basket, the next item will be proposed. The cross-selling strategies lead to the recommender systems. A traditional recommender system is designed to suggest new products to frequent customers based on previous purchase patterns (Wang et al., 2005).

3.2 Market Basket Analysis Algorithms

One of the first approaches in finding the market basket is Collaborative Filtering software that finds a *soul mate* for each customer (Hughes, 2000). A customer's *soul mate* is a customer who has identical tastes and therefore the same market basket. The software is installed in hundreds of companies. However, its disadvantage is that it merely compares two individuals and this does not allow an overall view. For example: customer X bought four books about data mining and a book about cooking, and customer Y bought the same four books about data mining. Therefore, the software will suggest the cooking book as the next item for customer Y, leading to possible mistakes.

Actually, an early algorithm for finding all association rules, referred to as the AIS algorithm, was first explored by Agrawal et al. (1994). The algorithm requires to repeatedly scanning the database. It uses the large itemsets discovered in the previous pass as the basis to generate new potentially large itemsets, called *candidate itemsets,* and counts their supports during the pass over the data. Specifically, after reading a transaction, it is determined which of the large itemsets found in the previous pass are contained in the transaction. New candidate itemsets are generated by extending these large itemsets with other items in the transaction. However, the performance study shows that AIS is not efficient since it generates too many candidate itemsets. Although, there are several efficient algorithms such as DHP, PSI, and a newer one LigCid (Tsai and Chen, 2004). Two well-known algorithms, Apriori and GRI, will be introduced in the following.

3.2.1 Apriori Algorithm

In opposition to *soul mate* algorithm, the Apriori algorithm (Agrawal and Srikan, 1994; Agrawal et al., 1996) takes all of the transactions in the database into account in order to define the market basket. The market basket can be represented with association rules, with a left and a right side (Berry and Linoff, 1997).

Apriori technique deals with primarily on itemsets that make up transactions. Items are flag-type conditions that indicate the presence or absence of a particular thing in a specific transaction. An itemset is a group of items which may or may not tend to co-occur within transactions.

There are three selection criteria for controlling the Apriori algorithm: minimum rule support, minimum rule confidence, and maximum rule preconditions.

Apriori proceeds in two stages: The first step identifies frequent itemsets. A frequent itemset is defined as an itemset with a support value that is greater than or equal to the user-specified minimum support threshold, S_{min}. The support of an itemset is the number of records in which the itemset is found divided by the total number of records. The algorithm begins by scanning the data and identifying the single-item itemsets that satisfy this criterion. Any single items that do not satisfy the criterion are not considered further, because adding an infrequent item to an itemset will always result in an infrequent itemset.

Apriori then generates larger itemsets recursively using the following steps :

Generate a candidate set of itemsets of length k by combining existing itemsets of length $(k-1)$. For every possible pair of frequent itemsets p and q with length $(k-1)$, compare the first $(k-2)$ items (in lexicographic order). If they are the same, and the last item in q is greater than the last item in p, add the last item in q to the end of p to create a new candidate itemset with length k.

Prune the candidate itemset by checking every $(k-1)$ length subset of each candidate itemset. All subsets must be frequent itemsets, or if the candidate itemset is infrequent then it is removed from further consideration.

Calculate the support of each itemset in the candidate set, as $\text{support} = N / N_i$ where N_i is the number of records that match the itemset and N is the number of records in the training data.

Itemsets with support > S_{min} are added to the list of frequent itemsets.

If any frequent itemsets of length k were found, and k is less than the user-specified maximum rule size $k_{\max}$, repeat the process to find frequent itemsets of length $(k+1)$.

When all frequent itemsets have been identified, the algorithm extracts rules from the frequent itemsets. For each frequent itemset L with length $(k > 1)$, the following procedure is applied:

Calculate all subsets A of length $(k-1)$ of the itemset such that all the fields in A are input fields and all the other fields in the itemset are output fields. Call the latter subset A'.

For each subset A, calculate the evaluation measure for the rule $A \to A'$. There are several measures that can be used by Apriori as an evaluation measure for determining which rules to retain. Values are calculated based on the prior confidence and the posterior confidence, defined as: $C_{prior} = N / c$ and $C_{posterior} = a / r$. Where c is the support of the consequent, a is the support of the antecedent, r is the support of the conjunction of the antecedent and the consequent, and N is the number of records in the training data.

If the rule confidence is greater than the user-specified threshold, add the rule to the rule table, and, if the length k' of A is greater than 1, test all possible subsets of A with length $(k'-1)$.

Let me give a simple example about Apriori algorithm. Firstly 0-1 matrix is generated. Assume the user-specified minimum support is 50%, and compare minimum candidate support value with minimum support value, then generate all frequent itemsets.

Slip no	**Items**
1	1,3,4
2	2,3,5
3	1,2,3,5
4	2,5

→

	1	**2**	**3**	**4**	**5**
1	1	0	1	1	0
2	0	1	1	0	1
3	1	1	1	0	1
4	0	1	0	0	1

C1

Items	Support
1	0,5
2	0,75
3	0,75
4	0,25
5	0,75

→

L1

Items	Support
1	0,5
2	0,75
3	0,75
5	0,75

C2

Items	Support
1,2	0,25
1,3	0,5
1,5	0,25
2,3	0,5
2,5	0,75
3,5	0,5

→

L2

Items	Support
1,3	0,5
2,3	0,5
2,5	0,75
3,5	0,5

C3

Items	Support
2,3,5	0,5

→

L3

Items	Support
2,3,5	0,5

3.2.2 GRI Technique

Generalized Rule Induction (GRI) searches for the most interesting independent rules and generally finds them quicker than Apriori. GRI generates associations based on the information content of a rule which is judged with J measure that is based on confidence, or probabilities. The measure maximizes the simplicity with goodness-of-fit tradeoff for any rule. It uses both the prior and posterior probabilities discussed above. If a potential rule in GRI is of the form:

$$(Y = y) \le (X = x) \text{ (with some support and confidence)} \qquad (3.1)$$

where x and y are values of X and Y, respectively, then J is defined as:

$$J = p(x)[p(y\mathrm{I}x)\log\frac{p(y\mathrm{I}x)}{p(y)} + (1 - p(y\mathrm{I}x))\log\frac{(1 - p(y\mathrm{I}x))}{(1 - p(y))}] \qquad (3.2)$$

Here, $p(x)$ is the probability (confidence) of the value of x in the data set, $p(y)$ is the prior confidence of y in the data set, and $p(\, y|x\,)$ is the conditional probability, or posterior confidence, of the occurrence of y given x in the data. The J measure is generalized to more than one antecedent by allowing x referring to a conjunction of several values.

GRI generates a simple rule under the restrictions of confidence and support values specified and calculates its interest level (J value). If this value is higher than the lowest value for a rule stored in the rule table, then the new rule is inserted into the table. If the rule table is not full then the new rule is inserted regardless of its J value. At this point, bounds are calculated on the J value with a slightly different formula to determine whether the rule should be further specialized, i.e., more antecedents added (SPPS Inc., 2001).

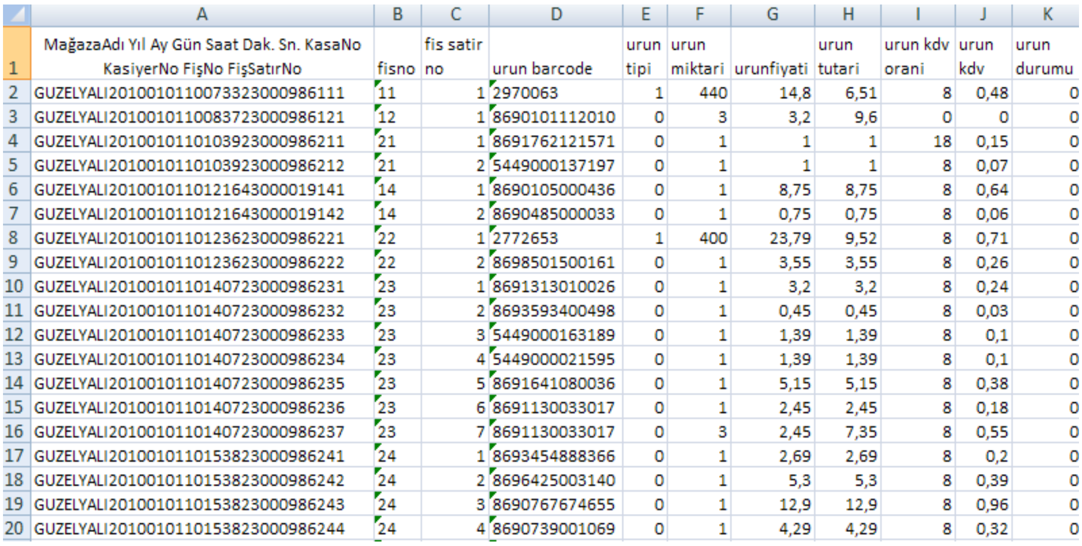
The GRI methodology can handle either categorical or numerical variables as inputs, but still requires categorical variables as outputs. Generalized rule induction was introduced by Smyth and Goodman in 1992. Rather than using frequent itemsets, GRI applies an information theoretic approach to determining the interestingness of a candidate association rule.

CHAPTER FOUR
APPLICATION

4.1 Introduction

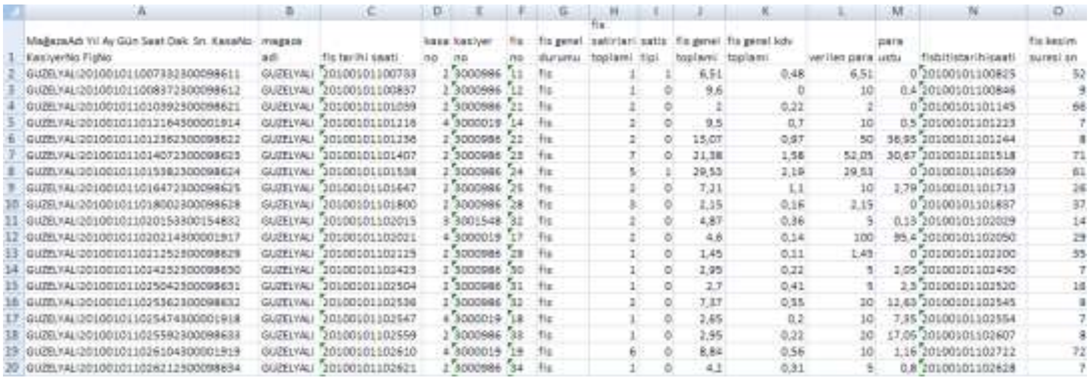
In this section, a market basket analysis application performed in a retail store is presented. The store is Güzelyalı branch of Pehlivanoğlu Marketçilik Gıda Paz. San ve Tic A.Ş. that is one of the market chains in retail sector in Turkey. In the analysis, relational products are determined by using association rules. Among the rules, Apriori Algorithm is preferred because of its advantage over other algorithms. The base data is analyzed and association rules between the products are set with Apriori algorithm.

The data includes sale records of January and February 2011. The records have totally 431218 purchases with the following data labels: branch name, date, time, cash box no, cashier no, slip no, barcodes and unit prices of bought products. As illustrated in the following figures,



	A	B	C	D	E	F	G	H	I	J	K
1	MağazaAdı Yıl Ay Gün Saat Dak. Sn. KasaNo KasiyerNo FişNo FişSatırNo	fisno	fis satir no	urun barcode	urun tipi	urun miktari	urunfiyati	urun tutari	urun kdv orani	urun kdv	urun durumu
2	GUZELYALI2010010110073323000986111	11	1	2970063	1	440	14,8	6,51	8	0,48	0
3	GUZELYALI2010010110083723000986121	12	1	8690101112010	0	3	3,2	9,6	0	0	0
4	GUZELYALI2010010110103923000986211	21	1	8691762121571	0	1	1	1	18	0,15	0
5	GUZELYALI2010010110103923000986212	21	2	5449000137197	0	1	1	1	8	0,07	0
6	GUZELYALI2010010110121643000019141	14	1	8690105000436	0	1	8,75	8,75	8	0,64	0
7	GUZELYALI2010010110121643000019142	14	2	8690485000033	0	1	0,75	0,75	8	0,06	0
8	GUZELYALI2010010110123623000986221	22	1	2772653	1	400	23,79	9,52	8	0,71	0
9	GUZELYALI2010010110123623000986222	22	2	8698501500161	0	1	3,55	3,55	8	0,26	0
10	GUZELYALI2010010110140723000986231	23	1	8691313010026	0	1	3,2	3,2	8	0,24	0
11	GUZELYALI2010010110140723000986232	23	2	8693593400498	0	1	0,45	0,45	8	0,03	0
12	GUZELYALI2010010110140723000986233	23	3	5449000163189	0	1	1,39	1,39	8	0,1	0
13	GUZELYALI2010010110140723000986234	23	4	5449000021595	0	1	1,39	1,39	8	0,1	0
14	GUZELYALI2010010110140723000986235	23	5	8691641080036	0	1	5,15	5,15	8	0,38	0
15	GUZELYALI2010010110140723000986236	23	6	8691130033017	0	1	2,45	2,45	8	0,18	0
16	GUZELYALI2010010110140723000986237	23	7	8691130033017	0	3	2,45	7,35	8	0,55	0
17	GUZELYALI2010010110153823000986241	24	1	8693454888366	0	1	2,69	2,69	8	0,2	0
18	GUZELYALI2010010110153823000986242	24	2	8696425003140	0	1	5,3	5,3	8	0,39	0
19	GUZELYALI2010010110153823000986243	24	3	8690767674655	0	1	12,9	12,9	8	0,96	0
20	GUZELYALI2010010110153823000986244	24	4	8690739001069	0	1	4,29	4,29	8	0,32	0

Figure 4.1 Details of sales records



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	MağazaAdı Yıl Ay Gün Saat Dak Sn KasaNo KasiyerNo FişNo	magaza adi	fis tarihi saati	kasa no	kasiyer no	fis no	fis genel durumu	fis satirlari toplami	satis tipi	fis genel toplami	fis genel kdv toplami	verilen para	para ustu	fisbitistarihisaati	fis kesim suresi sn
2	GUZELYALI2010010110073323000098611	GUZELYALI	20100101100733	2	3000986	11	fis	1	1	6,51	0,48	6,51	0	20100101100825	52
3	GUZELYALI2010010110083723000098612	GUZELYALI	20100101100837	2	3000986	12	fis	1	0	9,6	0	10	0,4	20100101100846	9
4	GUZELYALI2010010110103923000098621	GUZELYALI	20100101101039	2	3000986	21	fis	2	0	2	0,22	2	0	20100101101145	66
5	GUZELYALI201001011012164300001914	GUZELYALI	20100101101216	4	3000019	14	fis	2	0	9,5	0,7	10	0,5	20100101101223	7
6	GUZELYALI2010010110123623000098622	GUZELYALI	20100101101236	2	3000986	22	fis	2	0	15,07	0,97	50	34,93	20100101101244	8
7	GUZELYALI2010010110140723000098623	GUZELYALI	20100101101407	2	3000986	23	fis	7	0	21,38	1,58	52,05	30,67	20100101101518	71
8	GUZELYALI2010010110153823000098624	GUZELYALI	20100101101538	2	3000986	24	fis	5	1	29,53	2,19	29,53	0	20100101101639	61
9	GUZELYALI2010010110164723000098625	GUZELYALI	20100101101647	2	3000986	25	fis	2	0	7,21	1,1	10	2,79	20100101101713	26
10	GUZELYALI2010010110180023000098628	GUZELYALI	20100101101800	2	3000986	28	fis	3	0	2,15	0,16	2,15	0	20100101101837	37
11	GUZELYALI201001011020153300154832	GUZELYALI	20100101102015	3	3001548	32	fis	2	0	4,87	0,36	5	0,13	20100101102029	14
12	GUZELYALI201001011020214300001917	GUZELYALI	20100101102021	4	3000019	17	fis	2	0	4,6	0,14	100	95,4	20100101102050	29
13	GUZELYALI2010010110212523000098629	GUZELYALI	20100101102125	2	3000986	29	fis	1	0	1,45	0,11	1,45	0	20100101102200	35
14	GUZELYALI2010010110242523000098630	GUZELYALI	20100101102425	2	3000986	30	fis	1	0	2,95	0,22	5	2,05	20100101102450	7
15	GUZELYALI2010010110250423000098631	GUZELYALI	20100101102504	2	3000986	31	fis	1	0	2,7	0,41	5	2,3	20100101102520	16
16	GUZELYALI2010010110253623000098632	GUZELYALI	20100101102536	2	3000986	32	fis	2	0	7,37	0,55	20	12,63	20100101102545	8
17	GUZELYALI201001011025474300001918	GUZELYALI	20100101102547	4	3000019	18	fis	1	0	2,65	0,2	10	7,35	20100101102554	7
18	GUZELYALI2010010110255923000098633	GUZELYALI	20100101102559	2	3000986	33	fis	1	0	2,95	0,22	20	17,05	20100101102607	8
19	GUZELYALI201001011026104300001919	GUZELYALI	20100101102610	4	3000019	19	fis	6	0	8,84	0,56	10	1,16	20100101102712	72
20	GUZELYALI2010010110262123000098634	GUZELYALI	20100101102621	2	3000986	34	fis	1	0	4,1	0,31	5	0,9	20100101102628	7

Figure 4.2 Records of sales

Before starting the application with a base data, it should be analyzed to understand appropriateness of the data to the application. Therefore, first, product barcodes are matched with the products to arrange the data set for the analysis. Then, cleaning products, clothing, gift etc. purchases are disregarded since the focus of the analysis is food products.

The base data is composed of daily sales records. Hereby, number of customers shopping in a day from the market is examined. The daily number of customers in January and February 2011 are as follows.

Table 4.1 Daily counts of the customers

	Days	January	February		Days	January	February
1	Saturday	461	597	17	Monday	572	621
2	Sunday	568	584	18	Tuesday	613	628
3	Monday	520	587	19	Wednesday	561	597
4	Tuesday	656	549	20	Thursday	570	661
5	Wednesday	1000	593	21	Friday	578	525
6	Thursday	651	641	22	Saturday	469	656
7	Friday	711	388	23	Sunday	625	667
8	Saturday	615	650	24	Monday	515	563
9	Sunday	675	646	25	Tuesday	605	570
10	Monday	523	590	26	Wednesday	474	632
11	Tuesday	607	607	27	Thursday	562	623
12	Wednesday	529	526	28	Friday	614	539
13	Thursday	573	602	29	Saturday	537	
14	Friday	605	516	30	Sunday	617	
15	Saturday	590	657	31	Monday	429	
16	Sunday	638	670				

Table 4.2 Total number of customers in weekdays

Day	January	February
Monday	2559	2361
Tuesday	2481	2354
Wednesday	2564	2348
Thursday	2356	2527
Friday	2508	1968
Saturday	2672	2560
Sunday	3123	2567
Total	18263	16685

The data in Tables 4.1 and 4.2 reports that number of customers in weekdays is stable in general. However, as illustrated in Figure 4.3, the count of purchase increases significantly at the weekends.

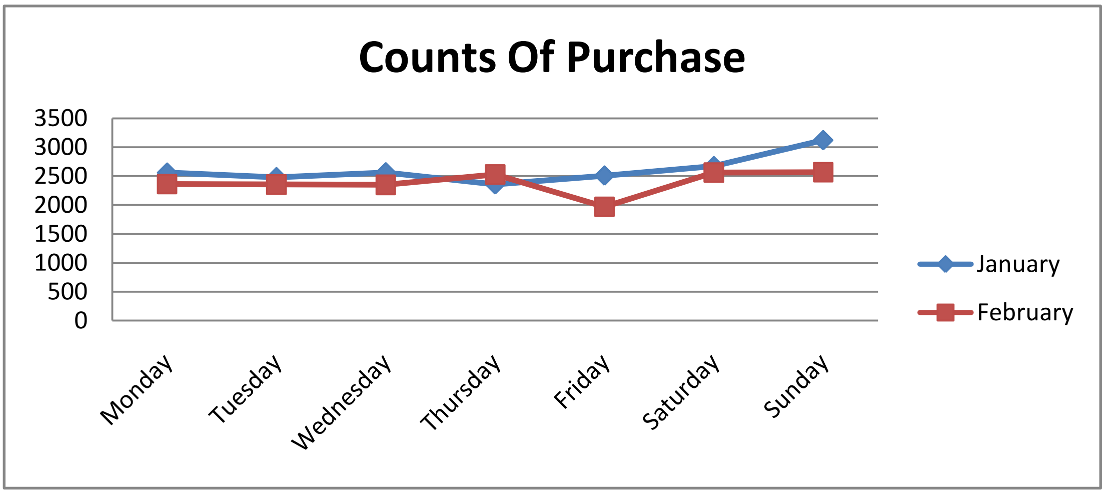


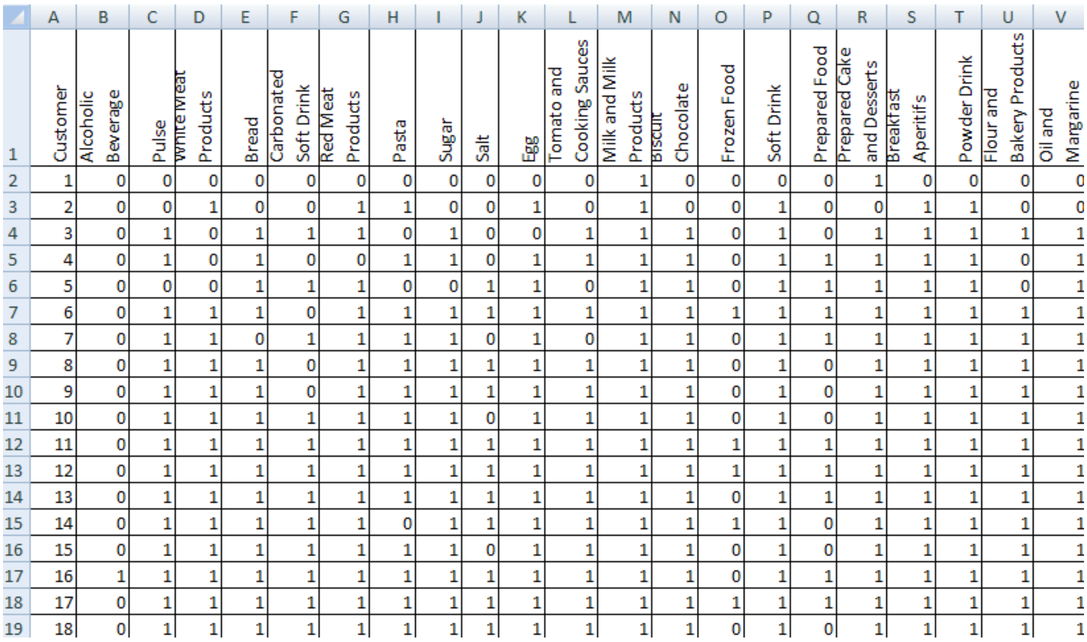
Figure 4.3 Counts of Customers

As stated previously, the purpose of *Market Basket Analysis* is to find relational product groups to utilize these relations in sales campaign and promotions, to set shelf array and to increase sales with revenue. The products which are bought by customer in one time are needed for the analysis. These products will be assessed in specific groups. Hence, they are grouped according to their specifications. In our application, the products are categorized into 21 groups as reported in Table 4.3.

Table 4.3 Product groups

1	Alcoholic Beverage	12	Milk and Milk Products
2	Pulse	13	Biscuit-Chocolate –Candy
3	White Meat Products	14	Frozen Food
4	Bread	15	Soft Drink
5	Carbonated Soft Drink	16	Prepared Food
6	Red Meat Products	17	Prepared Cake and Desserts
7	Pasta	18	Breakfast Aperitifs
8	Sugar	19	Powder Drink
9	Salt	20	Flour and Bakery Products
10	Egg	21	Oil and Margarine
11	Tomato and Cooking Sauces		

The products in slips are allocated to the above mentioned 21 groups and a 0-1 matrix is formed according to these groups. Herein, "1" represents the membership of the group under concern. A part of a 0-1 matrix is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Customer	Alcoholic Beverage	Pulse	White Meat Products	Bread	Carbonated Soft Drink	Red Meat Products	Pasta	Sugar	Salt	Egg	Tomato and Cooking Sauces	Milk and Milk Products	Biscuit Chocolate	Frozen Food	Soft Drink	Prepared Food	Prepared Cake and Desserts	Breakfast Aperitifs	Powder Drink	Flour and Bakery Products	Oil and Margarine
2	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
3	2	0	0	1	0	0	1	1	0	0	1	0	1	0	0	1	0	0	1	1	0	0
4	3	0	1	0	1	1	1	0	1	0	0	1	1	1	0	1	0	1	1	1	1	1
5	4	0	1	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1
6	5	0	0	0	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	0	1
7	6	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	7	0	1	1	0	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1	1
9	8	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
10	9	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
11	10	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1
12	11	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	12	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	13	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
15	14	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1
16	15	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1
17	16	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
18	17	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	18	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1

Figure 4.4 A part of 0-1 matrix

Multiple purchase of any product is disregarded in the analysis. After these arrangements, the data set consists of 1775 slips. Frequencies of the product groups in terms of purchase are presented in Table 4.4. and illustrated in Figure 4.5.

Table 4.4 Frequencies of the product groups

Product Groups	January	February	Total
Alcoholic Beverage	545	527	1072
Pulse	662	630	1292
White Meat Products	681	639	1320
Bread	690	646	1336
Carbonated Soft Drink	705	666	1371
Red Meat Products	721	654	1375
Pasta	654	619	1273
Sugar	652	623	1275
Salt	488	440	928
Egg	697	644	1341
Tomato and Cooking Sauces	667	641	1308
Milk and Milk Products	889	704	1593
Biscuit-Chocolate -Candy	797	690	1487
Frozen Food	571	574	1145
Soft Drink	706	670	1376
Prepared Food	565	571	1136
Prepared Cake and Desserts	693	655	1348
Breakfast Aperitifs	660	632	1292
Powder Drink	677	647	1324
Flour and Bakery Products	626	609	1235
Oil and Margarine	707	657	1364

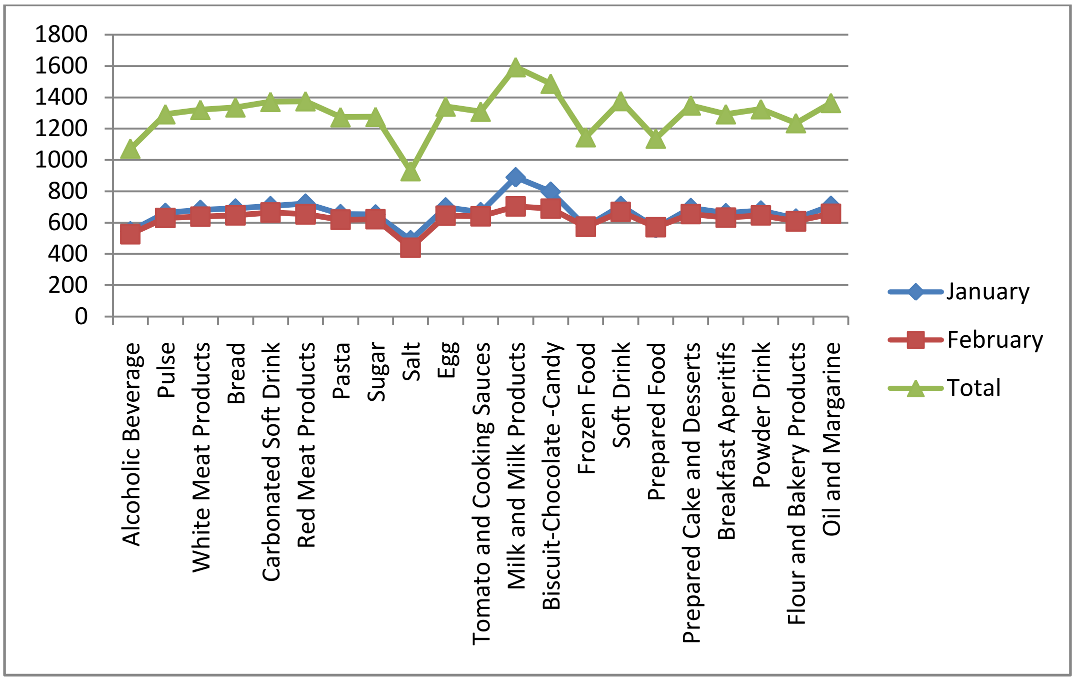


Figure 4.5 Frequencies of the product groups

As reported in Figure 4.5, *Milk and Milk Products* are in the first rank in purchase volume, while salt is the last product in purchase volume among the product group. As illustrated in figure, the product groups are slightly differs in purchase volume.

The purchase data is processed by employing SPSS Clementine software, and relationship between the products are determined using Apriori Algorithm. The rationales of using SPSS Clementine in our analysis are explained in the next section.

4.2 Selection of The Software

As known, several data mining tools exist in the market. Hence, choosing an appropriate tool for the company's needs may seem to be an important problem, and must be performed in a systematic way. According to the market shares, the top three market leaders are: *Clementine* of SPSS, *Enterprise Miner* of SAS Institute, and *Intelligent Miner* of IBM.

SPSS Clementine uses a methodology called CRISP-DM (Cross-Industry Standard Process for Data Mining). According to this methodology, the life cycle of a data mining project consists of following six phases.

Business Understanding: In this phase, our aim is to determine business and data mining objectives, and producing a project plan by assessing the situation.

Data Understanding: This phase's aim is to understand what data resources are. In this phase we collect, describe, explore data, and finally verify data quality.

Data Preparation: This phase's aim is to prepare data for data mining. Preparations include selecting, cleaning, constructing, integrating, and formatting data.

Modeling: In this phase sophisticated data analysis methods are used to extract information from the data. It involves selecting modeling techniques, generating test designs, and building and assessing the models.

Evaluation: In this phase, the data mining process is reviewed, evaluated the results in order to understand whether they are lead to achieve the business objectives or not.

Deployment: This phase focuses on integrating our new knowledge into business processes to solve our original business problem. This phase includes plan deployment, monitoring and maintenance, producing a final report, and reviewing the project.

The other market leader SAS Institute's *Enterprise Miner* uses the SEMMA (Sample, Explore, Modify, Model, and Assess) methodology (Groth, 1999).

Sample: Sampling of the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process.

Explore: The process of exploration of the data by searching for anticipated relationships, unanticipated trends, and anomalies in order to understand the ideas.

Modify: Making necessary modifications on the data by creating, selecting, and transforming the variables to focus the model-selection process.

Model: Constructing a model by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome.

Assess: The process of evaluating the usefulness and reliability of the findings from the data mining process.

In a data mining application, all of these steps may or may not be used and if needed one or more steps can be repeated until acceptable results are achieved.

On the other hand, IBM’s *Intelligent Miner* recommends the following steps as a generic data mining method:

- ✓ Definition of the business issue in precise statements.
- ✓ Definition of the data model and the data requirements.
- ✓ Using data from all available sources and preparation of the data.
- ✓ Assessment of the data quality.
- ✓ Choosing the mining function and defining the mining run.
- ✓ Interpretation of the results and identification of new information.
- ✓ Deploying the results and the new knowledge into the business.

As explained above, it is determined to use SPSS Clementine Program for this analysis because of common usage, being a preferred program, easy usage and data set integration from different sources easily.

4.3 Building The Model And Generating Associations

The associations corresponding to our data is generated by using SPSS Clementine. Diagram of the model in SPSS Clementine format is presented in Figure 4.6.

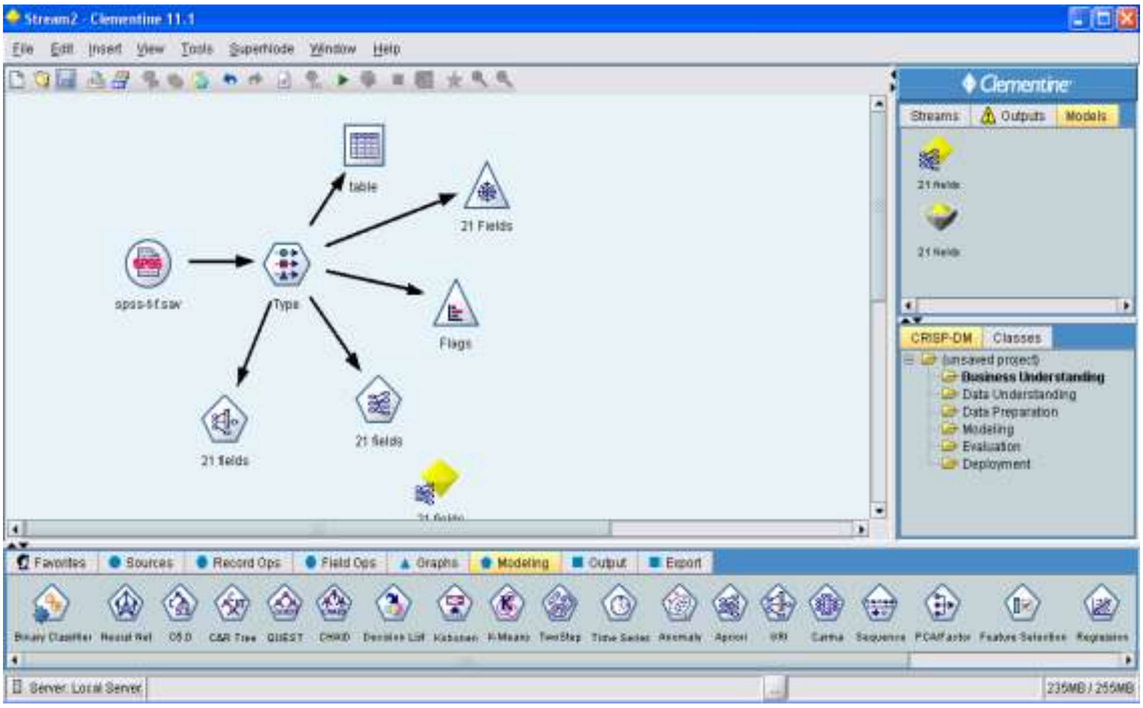
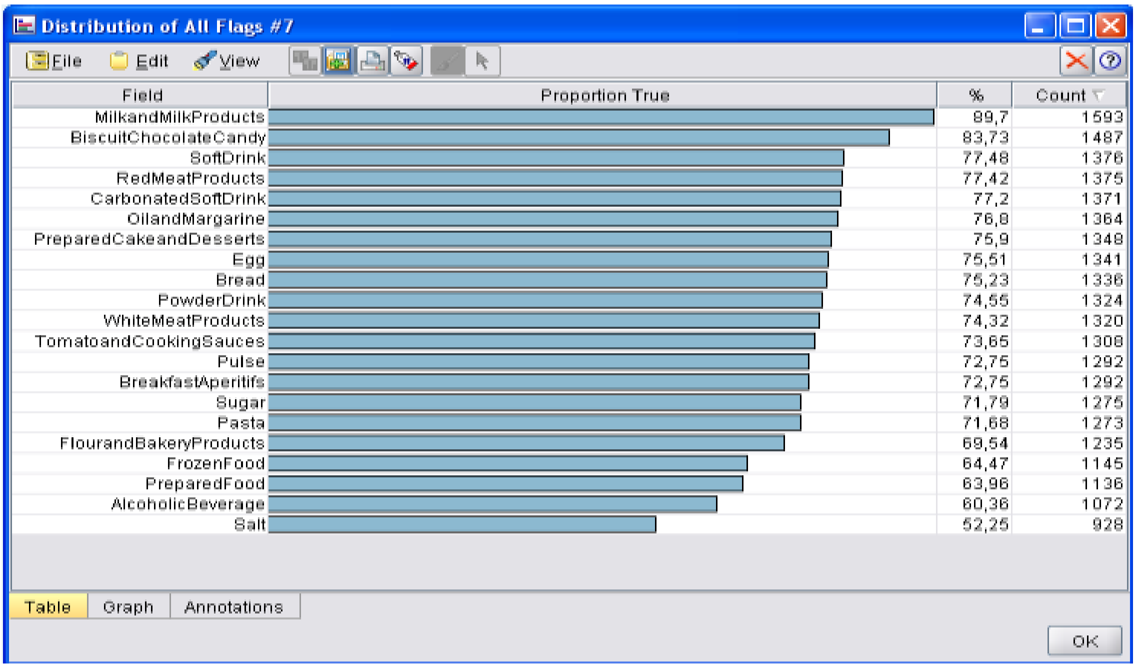

Stream2 - Clementine 11.1
File Edit Insert View Tools SuperNode Window Help
table
21 Fields
spss-fsf.sav
Type
Flags
21 fields
21 fields
Clementine
Streams Outputs Models
21 fields
21 fields
CRISP-DM Classes
(unsaved project)
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export
Binary Classifier Neural Net C5.0 C&R Tree QUEST CHAID Decision List Kohonen K-Means TwoStep Time Series Anomaly Apriori GRI Carma Sequence PCA/Factor Feature Selection Regression
Server: Local Server
235MB / 255MB

Figure 4.6 Model diagram in SPSS Clementine

In our analysis, first, distribution of the product groups has been determined. Figure 4.7 illustrates the rates and frequencies of the sales.



Field	Proportion True	%	Count
MilkandMilkProducts		89,7	1593
BiscuitChocolateCandy		83,73	1487
SoftDrink		77,48	1376
RedMeatProducts		77,42	1375
CarbonatedSoftDrink		77,2	1371
OilandMargarine		76,8	1364
PreparedCakeandDesserts		75,9	1348
Egg		75,51	1341
Bread		75,23	1336
PowderDrink		74,55	1324
WhiteMeatProducts		74,32	1320
TomatoandCookingSauces		73,65	1308
Pulse		72,75	1292
BreakfastAperitifs		72,75	1292
Sugar		71,79	1275
Pasta		71,68	1273
FlourandBakeryProducts		69,54	1235
FrozenFood		64,47	1145
PreparedFood		63,96	1136
AlcoholicBeverage		60,36	1072
Salt		52,25	928

Figure 4.7 Distribution of product groups

As reported in Figure 4.7, the product groups with the highest frequencies are *milk and milk products* (89.7%), *biscuit chocolate candy* (83.73%) and *soft drink* (77.48%). Actually, frequencies of all product groups are high because of the large data set. For instance, the product group with minimum frequency value is *salt*, 52.25%. Therefore, all product groups should be examined with the same sensibility.

As stated previously, Apriori Algorithm is employed in this study. SPSS Clementine necessitates that some parameters are entered by user for Apriori Algorithm. These parameters are *min support*, *min confidence* and *max number of antecedents*. The results are analyzed for max no of antecedent=1 and max no of antecedent=2, separately. *Min support* and *min confidence* are the most essential parameters which have been determined at the beginning of the market basket analysis.

Actually, market basket analysis is based on two metrics; *support* and *confidence*. These metrics are derived from the transactions record for the business. The first metric defined for market basket analysis is *support*, which is the probability of an association (probability of the two items being purchased together).

The chance of items being purchased together is governed by randomness. For small values of support, it is possible that the appearance of an affinity is the result of a small number of coincidences. To reduce the likelihood of such spurious associations, a *minimum support* value (e.g. $sup \geq 10\%$) is typically used in practice.

The next metric typically defined in the market basket analysis is the conditional probability of an item to be purchased, given that another one has already been purchased. A high *confidence* value may seem beneficial. Therefore, the confidence value is taken as 85% in this study. The association rules among the product groups obtained by SPSS Clementine software are presented in the Appendix 1. As reported, 327 rules have been obtained by the model for max no antecedent=1.

The results of the analysis can be summarized as in the following. Togetherness probability of Milk&milk products and frozen food in total slip moves is 64%. Therefore, it can be stated that customers who buy Milk&milk products buy Frozen food at the same time with 99% probability. Additionally, togetherness probability of Tomato&cooking sauces and flour&bakery products in total slip moves is 69%. Herein, it can be stated that the customers who buy Tomato&cooking sauces also buy flour&bakery products with 95% probability. The other rules can be interpreted in the same manner. On the other hand, the results reveal that 19 product groups are assessed as consequent. However, salt and alcoholic drinks are not consequent in any of the rules. It demonstrates that, salt and alcoholic drinks are not bought alone, they are generally bought with other product groups.

As the second case, the analysis is performed by taking the max no of antecedents as 2. In the analysis, 3890 rules are obtained. The results of the second case is the same with those of the first case. All of 3890 rules have been examined in this study and the top five rules are presented in the Appendix 2. At the same time association rules with highest confidence values are presented in the following table.

Table 4.5 Association rules of product groups with highest confidence level

Consequent	Antecedent	Support (%)	Confidence (%)
BiscuitChocolateCandy	Salt and BreakfastAperitifs	50,225	100
BiscuitChocolateCandy	Salt and Pulse	50,619	100
BiscuitChocolateCandy	Salt and Bread	50,732	100
BiscuitChocolateCandy	PreparedFood and Pulse	61,318	100
BiscuitChocolateCandy	FrozenFood and Sugar	60,529	100
MilkandMilkProducts	Salt and PreparedFood	46,791	100
MilkandMilkProducts	Salt and FrozenFood	46,453	100
MilkandMilkProducts	Salt and Pasta	50,113	100
MilkandMilkProducts	Salt and BreakfastAperitifs	50,225	100
MilkandMilkProducts	Salt and Pulse	50,619	100
RedMeatProducts	Salt and PreparedFood	46,791	100
RedMeatProducts	Salt and FrozenFood	46,453	100
RedMeatProducts	Salt and BreakfastAperitifs	50,225	100

The results reveal that togetherness probability of Bread, salt and prepared food in total slip moves is 47%. It can be stated here that customers who buy Bread, also buy salt and prepared food with 99.5% probability. Additionally, togetherness probability of Soft drink, prepared food and white meat products in total slip moves is %62. Then, it can be stated that customers who buy Soft drink also buy prepared food and white meat products with 99.5% probability.

CHAPTER FIVE
CONCLUSION

Generally, current business has the deficiency of past data usage in developing customer life cycle business strategies. Data mining is a set of techniques to bring information from large data sets to the surface. If one torture the data enough they will say what you want. Enough understanding of past, today and future leads the steps in the future.

Customer data mining is one of the strongest tools to derive information from data which contains the application of descriptive and predictive analytics (such as clustering, segmentation, estimation, prediction and affinity analysis) to support the marketing, sales and service functions. The wide spectrum tool set of data mining provides advantage of wide range of analysis according to the aim.

Many statistical and AI techniques are used in data mining. Apriori algorithm is one of the fastest and earliest tools for Association Mining. In this study, it is intended to use data mining methods to derive conclusions from a large set of real data to be used in strategy setting and decision making process in highly competitive environment. Apriori algorithm is employed for mining association rules in the large database of Pehlivanoğlu A.Ş..

The results of the analysis reveal that the item which is sold the most came out milk and milk products. Milk and milk products take place in 89.7% of the transactions. The second most sold item was biscuit chocolate candy. Then soft drink takes place. It can be concluded here that the customers of the given retail store consume milk and milk products in a high rate. Let me give examples of rules derived from these large itemsets. The people who bought milk and milk products also bought frozen food with confidence 99%. By saying “99% confidence” we mean that 99 % of people who bought milk and milk products also bought frozen food. It should be note that here that we do not have a rule saying that 99% of people who bought frozen food also bought milk and milk products because although the itemset frozen food, milk and milk products does not have enough confidence.

Another important point that must be addressed on the basis of the results is that salt and alcoholic beverage are not bought alone, they are generally bought with other product groups.

In this study, generally accepted values of support and confidence are used. However, in subsequent studies, a variety of comparisons can be made by using different support and confidence values on the same data set. Thus, the most appropriate support and confidence values can be obtained to explain the data set better.

Successfully competing in the new global economy requires immediate decision capability. This immediate decision capability requires quick analysis of both timely and relevant data to derive accurate and useful conclusions. Organizations are now on the stage of getting use of data in hand. To support this analysis, organizations give more effort to data mining.

In summary, increasing competition forces all companies to improve their relationship with their customers in order to increase their long term profit. This requires detecting valuable customers and retaining them instead of acquiring new ones. Data mining functionalities such as association mining can be used to detect these valuable customers. If the results of these functionalities are combined with a reporting environment that allows multiple level analyses, an effective base for CRM studies can be developed in future.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining Association Rules Between Sets of Items in Large Databases*. SIGMOD Conference, 207-216.

Agrawal, R. & R. Srikant. (1994). *Fast Algorithms for Mining Association Rules*, In Proc. 20th Int. Conf. Very Large Data Bases, VLDB, 487-499.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I. (1995). *Fast Discovery of Association Rules.* In: Fayyad, G., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in Knowledge Discovery and Data Mining. AAAI Press and The MIT Press, Menlo Park, CA, 307– 328.

Anand, S., Patrick, A.R., Hughes, J.G., Bell, D.A. (1998). *A Data Mining Methodology for Cross-sales.* Knowledge Based Systems, *10,* 449–461.

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More.* Hyperion Books.

Andrews, R.L., & Currim, I.S. (2002). Identifying Segments with Identical Choice Behaviors Across Product Categories: An Intercategory Logit Mixture Model. *International Journal of Research in Marketing*, *19*, 65–79.

Berry, M., Linoff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support.* Wiley.

Berry M.; Linoff G. (2004). *Data Mining Techniques for Marketing, Sales, and Customer Relationship Management,* Second Edition, John Wiley & Sons.

Berson, Alex, Smith Stephen, and Thearling Kurt. (1999). *Building Data Mining Applications for CRM.* New York: McGraw-Hill.

Brijs, T., Swinnen, G., Vanhoof, K., Wets, G. (2004). *Building An Association Rules Framework To Improve Product Assortment Decisions Data Mining and Knowledge Discovery*, *8*(1), 7–23.

Brin, S., Siverstein, C., Motwani, R. (1998). *Beyond Market Baskets: Generalizing Association Rules To Dependence Rules.* Data Mining and Knowledge Discovery, *2*, 39–68.

Brown, S.A. (2000). *A Case Study On CRM and Mass Customization*. In: S.A. Brown(Ed.) Customer Relationship Management: A Strategic Imperative in the World of e-Business (Canada, Wiley), 41–53.

Bull C. (2003). *Strategic Issues In Customer Relationship Management (CRM) Implementation.* Bus Process Manag J, *9*(5), 592–602.

Butler Group (2001). *Real CRM: Pitfalls and Potential*. Butler Group Report Series - June 2001.

Cavique, L., Rego, C., Themido, I. (2002). *A Scatter Search Algorithm For The Maximum Clique Problem.* In: Ribeiro, C., Hansen, P. (Eds.), Essays and Surveys in Meta-heuristics. Kluwer Academic Publishers, Dordrecht. 227–244.

Cerny P.A. (2001). *Data mining and Neural Networks from a Commercial Perspective.* Proceedings of the ORSNZ Conference Twenty Naught One, University of Canterbury, NZ.

Chablo, E. (1999). *The Importance of Marketing Data Intelligence in Delivering Successful CRM (Report).*

Chatterjee, J. (2000). Managing Customer Relationships in The E-business Economy, *Journal of Scientific & Industrial Research*. 749–752.

Chen, Injazz J. & Popovich, Karen (2003). Understanding Customer Relationship Management: People, Process and Technology. *Business Process Management Journal*, *9*(5), 672-688.

Chen, Y.L., Tang, K., & Hu, Y.H. (2005). *Market Basket Analysis In A Multiple Store Environment.* Decision Support Systems,*40* (2), 339–354.

Chye, Koh Hian, and Gerry Chan Kin Leong. (2002). 24(2), 1-28

Colgate, M. R., & Danaher, P. J. (2000). Implementing A Customer Relationship Strategy: The Asymmetric Impact Of Poor Versus Excellent Execution. *Journal of the Academy of Marketing Science*, *28*(3), 375–387.

Corner I, Hinton M. (2002) Customer Relationship Management Systems: Implementation Risks and Relationship Dynamics. *Qual Market Res: An Int J ,5*(4), 239–51.

Davenport, T. H., Harris J.G., & Kohli, A. K. (2001). *How Do They Know Their Customers So Well?* MIT Sloan Management Review, *42*(2), 63-74.

Decker, R. , Monien, K. (2003). Market Basket Analysis With Neural Gas Networks and Self-organising Maps. *Journal of Targeting, Measurement and Analysis for Marketing*,*11*(4), 373–386.

Decker, R. (2005). *Market Basket Analysis By means Of a Growing Neural Network.* The International Review of Retail, Distribution and Consumer Research, *15* (2), 151–169.

Dickinson, R., Harris, F., & Sircar, S.(1992). *Merchandise Compatibility: An Exploratory Study of Its Measurement and Effect On Department Store Performance.* International Review of Retail, Distribution and Consumer Research, *2* (4), 351–379.

Doole, I., Lancaster, P. & Lowe, R.(2005). *Understanding and Managing Customers*. UK: Pearson Education Limited

Dunham M.H. (2002). Data *Mining Introductory and Advanced Topics.* Prentice Hall, http://engr.smu.edu/~mhd/ , Appendix 1

Dyche, J. (2002). *The CRM Handbook* . Upper Saddle River, Addison-Wesley.

Edelstein H. (2000). *Building Profitable Customer Relationships With Data Mining.* SPSS White Paper.

Gartner Group. (2005). Gartner's top 54 CRM case studies.

Gartner Group. (2006). *Report Highlight For Market Trends*: CRM services, Asia/Pacific, 2006–2007.

Groth, R. (1999). *Data Mining: Building Competitive Advantage*. New Jersey: Prentice Hall.

Guo L. (2003). *Applying Data Mining Techniques in Property-Casualty Insurance.* Forums of the Casualty Actuarial Society.

Hahsler, M., Hornik, K., Reutterer, T. (2006). *Implications Of Probabilistic Data Modeling For Mining Association Rules*. In: Spiliopoulou, M., Kruse, R., Borgelt, C., Nu¨rnberger, A., Gaul, W. (Eds.), From Data and Information Analysis to Knowledge Engineering. Springer, Berlin, 598–605.

Hand D., Mannila H., Smyth P. (2001). *Principles of Data Mining.* The MIT Press, 352-353, 384-391

Handen, L. (2000). *The Three Ws Of Technology*. In: S.A. Brown (Ed.) Customer Relationship Management: A Strategic Imperative in the World of e-Business (Canada, Wiley), 219–225.

Hansotia, B. (2002). *Gearing Up For CRM:* Antecedents to Successful Implementation. *Journal of Database Marketing, 10*(2), 121-132.

Hughes, A.M. (2000). *Strategic Database Marketing*. McGraw-Hill, New York.

Julander, C.R.(1992). Basket Analysis. A New Way of Analyzing Scanner Data. *International Journal of Retail and Distribution Management,* 20, 10–18.

Kalakota, R., & Robinson, M. (1999). *E-business Roadmap for Success*. Boston, MA: Addison-Wesley.

Kalakota, R., & Robinson, M. (2001). *M-business: The Race to Mobility*. New York: McGraw-Hill.

Kanji, G.K. (1998). *Measurement of Business Excellence.* Total Quality Management, *7*, 633–643.

Kanji, G.K. & Wallace, W. (2000). *Business Excellence Through Customer Satisfaction.* Total Quality Management, *7*, 979-998.

Karimi, J., Somers, T.M. & Gupta, Y. P. (2001). Impact of Information Technology Management Practices on Customer Service. *Journal of Management Information Systems, 17*(4), 125-158.

Kaski S.(1997). *Data Exploration Using Self-organizing Maps*. Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No:82, Espoo, Finnish Academy of Technology, 57

Kincaid, J.W. (2003*). Customer Relationship Management: Getting It Right!* Upper Saddle River, NJ: Prentice-Hall.

Ko E, Lee SJ, Woo JY. (2004). *Current CRM Adoption in The Korean Apparel Industry*. Spring conference proceedings of Korean Society of Clothing & Textiles. Seoul.

Kotler, P. & Armstrong, G. (2006). *Principles of Marketing*. New Jersey: Pearson Prentice Hall.

Larose, D.T. (1999*). Introduction to Data Mining and Knowledge Discovery* (3). Two Crows Corporation

Manchanda, P., Ansari, A., & Gupta, S. (1999). *The Shopping Basket: A Model for Multi-category Purchase Incidence Decisions.* Marketing Science,*18,* 95–114.

Mild, A., Reutterer, T. (2003). An Improved Collaborative Filtering Approach for Predicting Cross-category Purchases Based On Binary Market Basket Data. *Journal of Retailing and Consumer Services, 10* (3), 123–133

Musaoglu, C. (2003). *Customer Acquisition and Retention Modeling In Consumer Finance Sector Using Data Mining*. Published master's thesis, University of Boğaziçi.

Ngai EWT. (2005). *Customer Relationship Management Research (1992–2002).* Mark Intell Plann , *23*(6), 582605.

Paas, L., & Kuijlen, T. (2001). Towards A General Definition of Customer Relationship Management. *Journal of Database Marketing, 9*(1), 51-60.

Pan, S.L. & Lee, J. (2003). *Using E-CRM for a Unified View of The Customer.* Communications of the ACM*, 46*(4), 95-99.

Parvatiyar, A. & Sheth, J. N. (2001). Customer Relationship Management: Emerging Practice, Process, and Discipline. *Journal of Economic and Social Research, 3*(2),1-34.

Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. *Journal of Marketing, 169*(4), 167-176.

Pepper, D. & Rogers, M. (1999). *Is Your Company Ready For One-to-One Marketing,* Harvard Business Review, January-February, 3–12.

Peppers, D. & Rogers, M. (2004). *Managing Customer Relationships: A Strategic Framework*. John Wiley & Sons.

Peters, T. (1988). *Thriving on Chaos.* London, Macmillan.

Pushkala, R., Michael Wittmann, C., & Rauseo, N. A. (2006). Leveraging CRM For Sales: The Role Of Organizational Capabilities In successful CRM Implementation. *Journal of Selling & Sales Management*, *26*(1), 39–53.

Rajola, F. (2003). *Customer Relationship Management*. Berlin: Springer-Verlag

Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The Customer Relationship Management Process: Its Measurement and Impact on Performance. *Journal of Marketing Research*, *41*, 293−305.

Russell, G.J., Bell, D., Bodapati, A., Brown, C., Chiang, J., Gaeth, G., Gupta, S., Manchanda, P. (1997). *Perspectives On Multiple Category Choice*. Marketing Letters *8* (3), 297–305.

Russell, G.J., Kamakura, W.A. (1997). Modeling Multiple Category Brand Preference with Household Basket Data. *Journal of Retailing*, *73* (4), 439–461.

Russel, S. (1999). *Business Excellence: From Outside In or Inside Out?*, Total Quality Management, *10*, 697–703.

Russell, G.J., Ratneshwar, S., Shocker, A.D., Bell, D., Bodapati, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., & Shankar, V.H. (1999). *Multiple-category Decision-making: Review and Synthesis.* Marketing Letters, *10*, 319–332.

Ryals, L., & Payne, A. (2001). Customer Relationship Management In Financial Services: Towards Information-enable Relationship Marketing. *Journal of Strategic Marketing*, *9*(1), 3–27.

Sharp D.E. (2003). *Customer Relationship Management Systems Handbook*. CRC Press, 114-115

Sin, Y.M., Tse, C.B. & Yim, H.K. (2005). CRM: Conceptualization and Scale Development. *European Journal of Marketing, 39*(11/12),1264-1290.

Smyth, P. & Goodman, R.M. (1992). *An Information Theoretic Approach To Rule Induction From Databases*. IEEE Trans. On Knowladge and Data Engineering, *4*(4), 301-316

Song, I., & Chintagunta, P.K.(2006). *Measuring Cross-category Price Effects With Aggregate Store Data*. Management Science, *52* (10), 1594–1609.

Spengler, B. (1999). *Eyes On The Customer*. Computerworld, 33, 60–62.

Srivastava, R.K., Shervani, T.A. & Fahey, L. (1999). Marketing, Business Processes, and Shareholder Value: An Organizationally Embedded View of Marketing Activities and The Discipline of Marketing. *Journal of Marketing*, *63*, 168–179.

Stone, M. & Foss, B. (2001). *Successful Customer Relationship Marketing*. Great Britain: The Bath Press

Swift, R.S. (2000). *Accelerating Customer Relationships: Using CRM and Relationship Technologies*. Prentice Hall PTR

Swift, Ronald (2001), *Accelerating Customer Relationship*. Printece Hall PTR

Tsai, P. M.; Chen, C. M. (2004). *Mining Interesting Association Rules From Customer Databases and Transaction Databases.* Information Systems, *29*(8), 2-3.

Vaura, T.G. (1992). *Aftermarketing: How to Keep Customers For Life Through Relationship Marketing*. USA

Wang, Y., Chuang, Y.L., Hsu, M.H., & Keh, H.C. (2005). *A Personalized Recommender System For The Cosmetic Business.* Expert Systems with Applications,*26*, 427–434.

Wang, F.-S., Shao, H.-M. (2005). *Effective Personalized Recommendation Based On Time-Framed Navigation Clustering and Association Mining*. Expert Systems with Applications, *27*(3), 365–377.

Waterman, R. (1987). *The Renewal Factor*. New York, Bantam.

APPENDIX

Table A1. Association rules of product groups for one antecedent

Consequent	Antecedent	Support (%)	Confidence (%)
MilkandMilkProducts	FrozenFood	64,471	99,476
MilkandMilkProducts	PreparedFood	63,964	99,472
MilkandMilkProducts	Sugar	71,791	99,451
MilkandMilkProducts	TomatoandCookingSauces	73,649	99,159
MilkandMilkProducts	Salt	52,252	99,138
MilkandMilkProducts	FlourandBakeryProducts	69,538	99,109
MilkandMilkProducts	PowderDrink	74,55	99,094
MilkandMilkProducts	Egg	75,507	99,031
MilkandMilkProducts	BreakfastAperitifs	72,748	98,994
MilkandMilkProducts	Pasta	71,678	98,9
MilkandMilkProducts	Pulse	72,748	98,839
MilkandMilkProducts	OilandMargarine	76,802	98,827
MilkandMilkProducts	PreparedCakeandDesserts	75,901	98,813
MilkandMilkProducts	Bread	75,225	98,728
BiscuitChocolateCandy	Salt	52,252	98,707
BiscuitChocolateCandy	FrozenFood	64,471	98,69
MilkandMilkProducts	AlcoholicBeverage	60,36	98,507
MilkandMilkProducts	WhiteMeatProducts	74,324	98,485
BiscuitChocolateCandy	PreparedFood	63,964	98,415
MilkandMilkProducts	RedMeatProducts	77,421	98,327
RedMeatProducts	Salt	52,252	98,276
MilkandMilkProducts	SoftDrink	77,477	98,183
MilkandMilkProducts	CarbonatedSoftDrink	77,196	98,177
OilandMargarine	Salt	52,252	98,168
BiscuitChocolateCandy	Sugar	71,791	98,118
RedMeatProducts	PreparedFood	63,964	98,063
PreparedCakeandDesserts	Salt	52,252	97,953
SoftDrink	Salt	52,252	97,953
BiscuitChocolateCandy	AlcoholicBeverage	60,36	97,948
BiscuitChocolateCandy	TomatoandCookingSauces	73,649	97,936
BiscuitChocolateCandy	Pulse	72,748	97,91
SoftDrink	PreparedFood	63,964	97,887
BiscuitChocolateCandy	Pasta	71,678	97,879
BiscuitChocolateCandy	FlourandBakeryProducts	69,538	97,814
BiscuitChocolateCandy	PowderDrink	74,55	97,659
RedMeatProducts	FrozenFood	64,471	97,555
OilandMargarine	PreparedFood	63,964	97,535
TomatoandCookingSauces	Salt	52,252	97,522
Egg	Salt	52,252	97,522
BiscuitChocolateCandy	PreparedCakeandDesserts	75,901	97,478
CarbonatedSoftDrink	PreparedFood	63,964	97,447
BiscuitChocolateCandy	BreakfastAperitifs	72,748	97,446
WhiteMeatProducts	Salt	52,252	97,414
CarbonatedSoftDrink	Salt	52,252	97,414
PreparedCakeandDesserts	PreparedFood	63,964	97,359
PowderDrink	Salt	52,252	97,306
SoftDrink	FrozenFood	64,471	97,205
Egg	PreparedFood	63,964	97,183
BiscuitChocolateCandy	SoftDrink	77,477	97,166
BiscuitChocolateCandy	Bread	75,225	97,156

Bread	Salt	52,252	97,091
Bread	PreparedFood	63,964	97,007
OilandMargarine	FlourandBakeryProducts	69,538	97,004
BiscuitChocolateCandy	WhiteMeatProducts	74,324	96,97
CarbonatedSoftDrink	FrozenFood	64,471	96,943
BiscuitChocolateCandy	Egg	75,507	96,943
Pulse	Salt	52,252	96,875
RedMeatProducts	FlourandBakeryProducts	69,538	96,842
TomatoandCookingSauces	PreparedFood	63,964	96,831
Egg	FrozenFood	64,471	96,769
OilandMargarine	FrozenFood	64,471	96,769
OilandMargarine	Sugar	71,791	96,706
WhiteMeatProducts	PreparedFood	63,964	96,655
BiscuitChocolateCandy	CarbonatedSoftDrink	77,196	96,645
BiscuitChocolateCandy	OilandMargarine	76,802	96,628
WhiteMeatProducts	FrozenFood	64,471	96,594
Sugar	Salt	52,252	96,552
PowderDrink	PreparedFood	63,964	96,479
RedMeatProducts	Pulse	72,748	96,44
BiscuitChocolateCandy	RedMeatProducts	77,421	96,364
SoftDrink	Pulse	72,748	96,285
SoftDrink	FlourandBakeryProducts	69,538	96,275
OilandMargarine	Pulse	72,748	96,207
SoftDrink	TomatoandCookingSauces	73,649	96,177
Bread	FrozenFood	64,471	96,157
PreparedCakeandDesserts	FrozenFood	64,471	96,157
BreakfastAperitifs	Salt	52,252	96,121
PreparedCakeandDesserts	FlourandBakeryProducts	69,538	96,113
CarbonatedSoftDrink	FlourandBakeryProducts	69,538	96,113
RedMeatProducts	Sugar	71,791	96,078
RedMeatProducts	Pasta	71,678	96,072
RedMeatProducts	BreakfastAperitifs	72,748	96,053
SoftDrink	Sugar	71,791	96
SoftDrink	AlcoholicBeverage	60,36	95,989
Pasta	Salt	52,252	95,905
RedMeatProducts	TomatoandCookingSauces	73,649	95,872
Pulse	PreparedFood	63,964	95,863
MilkandMilkProducts	BiscuitChocolateCandy	83,727	95,831
OilandMargarine	Pasta	71,678	95,758
TomatoandCookingSauces	FrozenFood	64,471	95,721
PowderDrink	FrozenFood	64,471	95,721
CarbonatedSoftDrink	AlcoholicBeverage	60,36	95,709
SoftDrink	Pasta	71,678	95,679
SoftDrink	BreakfastAperitifs	72,748	95,666
Egg	FlourandBakeryProducts	69,538	95,628
SoftDrink	PowderDrink	74,55	95,619
CarbonatedSoftDrink	Pulse	72,748	95,588
RedMeatProducts	WhiteMeatProducts	74,324	95,53
PreparedCakeandDesserts	Sugar	71,791	95,529
OilandMargarine	BreakfastAperitifs	72,748	95,511
PreparedCakeandDesserts	Pulse	72,748	95,511
TomatoandCookingSauces	FlourandBakeryProducts	69,538	95,466
RedMeatProducts	AlcoholicBeverage	60,36	95,429
BreakfastAperitifs	PreparedFood	63,964	95,423
PreparedCakeandDesserts	TomatoandCookingSauces	73,649	95,413
RedMeatProducts	PowderDrink	74,55	95,393
Egg	Sugar	71,791	95,373

WhiteMeatProducts	FlourandBakeryProducts	69,538	95,304
PowderDrink	FlourandBakeryProducts	69,538	95,304
Bread	FlourandBakeryProducts	69,538	95,304
OilandMargarine	Egg	75,507	95,302
CarbonatedSoftDrink	Sugar	71,791	95,294
Egg	Pasta	71,678	95,287
Egg	BreakfastAperitifs	72,748	95,279
FlourandBakeryProducts	Salt	52,252	95,259
OilandMargarine	PowderDrink	74,55	95,242
SoftDrink	WhiteMeatProducts	74,324	95,227
PreparedCakeandDesserts	Pasta	71,678	95,208
OilandMargarine	TomatoandCookingSauces	73,649	95,183
CarbonatedSoftDrink	BreakfastAperitifs	72,748	95,124
Egg	Pulse	72,748	95,124
SoftDrink	PreparedCakeandDesserts	75,901	95,104
RedMeatProducts	Egg	75,507	95,078
PowderDrink	Sugar	71,791	95,059
OilandMargarine	WhiteMeatProducts	74,324	95
Sugar	PreparedFood	63,964	94,982
Pulse	FrozenFood	64,471	94,934
RedMeatProducts	PreparedCakeandDesserts	75,901	94,881
CarbonatedSoftDrink	TomatoandCookingSauces	73,649	94,878
PreparedCakeandDesserts	PowderDrink	74,55	94,864
CarbonatedSoftDrink	PowderDrink	74,55	94,789
OilandMargarine	AlcoholicBeverage	60,36	94,776
SoftDrink	Bread	75,225	94,76
RedMeatProducts	Bread	75,225	94,76
Pulse	Sugar	71,791	94,745
PowderDrink	Pasta	71,678	94,737
CarbonatedSoftDrink	Pasta	71,678	94,737
OilandMargarine	PreparedCakeandDesserts	75,901	94,733
PowderDrink	TomatoandCookingSauces	73,649	94,725
Egg	TomatoandCookingSauces	73,649	94,725
SoftDrink	Egg	75,507	94,705
WhiteMeatProducts	Sugar	71,791	94,667
WhiteMeatProducts	Pulse	72,748	94,659
PowderDrink	Pulse	72,748	94,659
CarbonatedSoftDrink	PreparedCakeandDesserts	75,901	94,659
CarbonatedSoftDrink	WhiteMeatProducts	74,324	94,621
BreakfastAperitifs	FrozenFood	64,471	94,585
CarbonatedSoftDrink	Egg	75,507	94,556
PreparedCakeandDesserts	WhiteMeatProducts	74,324	94,545
OilandMargarine	Bread	75,225	94,536
PreparedCakeandDesserts	BreakfastAperitifs	72,748	94,505
SoftDrink	OilandMargarine	76,802	94,501
RedMeatProducts	OilandMargarine	76,802	94,501
Pulse	FlourandBakeryProducts	69,538	94,494
Egg	PowderDrink	74,55	94,486
Egg	WhiteMeatProducts	74,324	94,47
CarbonatedSoftDrink	Bread	75,225	94,461
Pasta	PreparedFood	63,964	94,366
Bread	Pulse	72,748	94,35
WhiteMeatProducts	Pasta	71,678	94,344
PreparedCakeandDesserts	Bread	75,225	94,311
SoftDrink	CarbonatedSoftDrink	77,196	94,311
RedMeatProducts	CarbonatedSoftDrink	77,196	94,311
Egg	AlcoholicBeverage	60,36	94,31

Bread	BreakfastAperitifs	72,748	94,272
PreparedCakeandDesserts	AlcoholicBeverage	60,36	94,216
Bread	Sugar	71,791	94,196
TomatoandCookingSauces	Pasta	71,678	94,187
PowderDrink	BreakfastAperitifs	72,748	94,118
PreparedCakeandDesserts	Egg	75,507	94,109
WhiteMeatProducts	BreakfastAperitifs	72,748	94,04
TomatoandCookingSauces	Pulse	72,748	94,04
TomatoandCookingSauces	Sugar	71,791	94,039
Bread	TomatoandCookingSauces	73,649	94,037
CarbonatedSoftDrink	RedMeatProducts	77,421	94,036
SoftDrink	RedMeatProducts	77,421	94,036
WhiteMeatProducts	AlcoholicBeverage	60,36	94,03
Sugar	FlourandBakeryProducts	69,538	94,008
CarbonatedSoftDrink	SoftDrink	77,477	93,968
RedMeatProducts	SoftDrink	77,477	93,968
Bread	PowderDrink	74,55	93,958
BreakfastAperitifs	FlourandBakeryProducts	69,538	93,927
Sugar	FrozenFood	64,471	93,886
Bread	Pasta	71,678	93,873
PowderDrink	AlcoholicBeverage	60,36	93,843
OilandMargarine	RedMeatProducts	77,421	93,745
WhiteMeatProducts	TomatoandCookingSauces	73,649	93,731
Egg	OilandMargarine	76,802	93,695
OilandMargarine	SoftDrink	77,477	93,677
PreparedCakeandDesserts	OilandMargarine	76,802	93,622
CarbonatedSoftDrink	OilandMargarine	76,802	93,622
Egg	PreparedCakeandDesserts	75,901	93,62
TomatoandCookingSauces	PowderDrink	74,55	93,58
Sugar	Pulse	72,748	93,498
TomatoandCookingSauces	BreakfastAperitifs	72,748	93,498
Pulse	Pasta	71,678	93,48
Bread	PreparedCakeandDesserts	75,901	93,472
Bread	AlcoholicBeverage	60,36	93,47
TomatoandCookingSauces	AlcoholicBeverage	60,36	93,377
Egg	Bread	75,225	93,338
PowderDrink	WhiteMeatProducts	74,324	93,333
Bread	WhiteMeatProducts	74,324	93,333
PowderDrink	Egg	75,507	93,289
Pasta	FlourandBakeryProducts	69,538	93,279
Pasta	FrozenFood	64,471	93,275
BreakfastAperitifs	Sugar	71,791	93,176
PowderDrink	PreparedCakeandDesserts	75,901	93,175
PreparedCakeandDesserts	SoftDrink	77,477	93,169
Sugar	Pasta	71,678	93,166
OilandMargarine	CarbonatedSoftDrink	77,196	93,144
PowderDrink	Bread	75,225	93,114
PreparedCakeandDesserts	CarbonatedSoftDrink	77,196	93,071
WhiteMeatProducts	PowderDrink	74,55	93,051
Pasta	Sugar	71,791	93,02
PreparedCakeandDesserts	RedMeatProducts	77,421	93,018
WhiteMeatProducts	Egg	75,507	92,99
Bread	Egg	75,507	92,99
FlourandBakeryProducts	PreparedFood	63,964	92,958
Pulse	BreakfastAperitifs	72,748	92,957
BreakfastAperitifs	Pulse	72,748	92,957
Pulse	TomatoandCookingSauces	73,649	92,89

TomatoandCookingSauces	WhiteMeatProducts	74,324	92,879
Egg	RedMeatProducts	77,421	92,727
BreakfastAperitifs	Pasta	71,678	92,694
Pulse	WhiteMeatProducts	74,324	92,652
Bread	OilandMargarine	76,802	92,595
TomatoandCookingSauces	PreparedCakeandDesserts	75,901	92,582
WhiteMeatProducts	PreparedCakeandDesserts	75,901	92,582
Egg	CarbonatedSoftDrink	77,196	92,487
PowderDrink	OilandMargarine	76,802	92,449
Pulse	AlcoholicBeverage	60,36	92,444
TomatoandCookingSauces	Egg	75,507	92,394
Pulse	PowderDrink	74,55	92,372
BreakfastAperitifs	TomatoandCookingSauces	73,649	92,355
BreakfastAperitifs	AlcoholicBeverage	60,36	92,351
Egg	SoftDrink	77,477	92,297
WhiteMeatProducts	Bread	75,225	92,216
Pasta	Pulse	72,748	92,105
Bread	RedMeatProducts	77,421	92,073
TomatoandCookingSauces	Bread	75,225	92,066
FlourandBakeryProducts	FrozenFood	64,471	92,052
Bread	CarbonatedSoftDrink	77,196	92,05
BreakfastAperitifs	WhiteMeatProducts	74,324	92,045
PowderDrink	SoftDrink	77,477	92,006
Bread	SoftDrink	77,477	92,006
Sugar	BreakfastAperitifs	72,748	91,95
WhiteMeatProducts	OilandMargarine	76,802	91,935
PowderDrink	RedMeatProducts	77,421	91,855
BreakfastAperitifs	PowderDrink	74,55	91,843
BreakfastAperitifs	Egg	75,507	91,797
WhiteMeatProducts	RedMeatProducts	77,421	91,709
Sugar	AlcoholicBeverage	60,36	91,698
Pasta	TomatoandCookingSauces	73,649	91,667
Sugar	TomatoandCookingSauces	73,649	91,667
Pulse	Egg	75,507	91,648
Pulse	PreparedCakeandDesserts	75,901	91,543
Sugar	PowderDrink	74,55	91,541
PowderDrink	CarbonatedSoftDrink	77,196	91,539
Sugar	WhiteMeatProducts	74,324	91,439
TomatoandCookingSauces	SoftDrink	77,477	91,424
WhiteMeatProducts	SoftDrink	77,477	91,352
Pasta	BreakfastAperitifs	72,748	91,331
TomatoandCookingSauces	OilandMargarine	76,802	91,276
Pulse	Bread	75,225	91,243
TomatoandCookingSauces	RedMeatProducts	77,421	91,2
BreakfastAperitifs	Bread	75,225	91,168
Pulse	OilandMargarine	76,802	91,129
WhiteMeatProducts	CarbonatedSoftDrink	77,196	91,101
Pasta	PowderDrink	74,55	91,088
FlourandBakeryProducts	Sugar	71,791	91,059
Pasta	WhiteMeatProducts	74,324	90,985
Pasta	AlcoholicBeverage	60,36	90,951
Sugar	Egg	75,507	90,679
Pulse	RedMeatProducts	77,421	90,618
BreakfastAperitifs	PreparedCakeandDesserts	75,901	90,579
TomatoandCookingSauces	CarbonatedSoftDrink	77,196	90,518
FlourandBakeryProducts	Pasta	71,678	90,495
BreakfastAperitifs	OilandMargarine	76,802	90,469

Pasta	Egg	75,507	90,455
Pulse	SoftDrink	77,477	90,407
Sugar	OilandMargarine	76,802	90,396
Sugar	PreparedCakeandDesserts	75,901	90,356
FlourandBakeryProducts	Pulse	72,748	90,325
BreakfastAperitifs	RedMeatProducts	77,421	90,255
FlourandBakeryProducts	TomatoandCookingSauces	73,649	90,138
Pulse	CarbonatedSoftDrink	77,196	90,08
SoftDrink	BiscuitChocolateCandy	83,727	89,913
Pasta	PreparedCakeandDesserts	75,901	89,911
Sugar	Bread	75,225	89,895
BreakfastAperitifs	SoftDrink	77,477	89,826
FlourandBakeryProducts	BreakfastAperitifs	72,748	89,783
BreakfastAperitifs	CarbonatedSoftDrink	77,196	89,643
PreparedFood	Salt	52,252	89,547
BiscuitChocolateCandy	MilkandMilkProducts	89,696	89,454
Pasta	Bread	75,225	89,446
Pasta	OilandMargarine	76,802	89,37
FlourandBakeryProducts	WhiteMeatProducts	74,324	89,167
CarbonatedSoftDrink	BiscuitChocolateCandy	83,727	89,106
RedMeatProducts	BiscuitChocolateCandy	83,727	89,106
Sugar	RedMeatProducts	77,421	89,091
Sugar	SoftDrink	77,477	88,953
Pasta	RedMeatProducts	77,421	88,945
FrozenFood	Salt	52,252	88,901
FlourandBakeryProducts	PowderDrink	74,55	88,897
OilandMargarine	BiscuitChocolateCandy	83,727	88,635
Sugar	CarbonatedSoftDrink	77,196	88,621
FlourandBakeryProducts	AlcoholicBeverage	60,36	88,619
Pasta	SoftDrink	77,477	88,517
PreparedCakeandDesserts	BiscuitChocolateCandy	83,727	88,366
FlourandBakeryProducts	Bread	75,225	88,099
FlourandBakeryProducts	Egg	75,507	88,069
FlourandBakeryProducts	PreparedCakeandDesserts	75,901	88,056
Pasta	CarbonatedSoftDrink	77,196	87,965
FlourandBakeryProducts	OilandMargarine	76,802	87,83
Egg	BiscuitChocolateCandy	83,727	87,424
FrozenFood	PreparedFood	63,964	87,412
Bread	BiscuitChocolateCandy	83,727	87,29
FlourandBakeryProducts	RedMeatProducts	77,421	86,982
PowderDrink	BiscuitChocolateCandy	83,727	86,954
PreparedFood	FrozenFood	64,471	86,725
FlourandBakeryProducts	CarbonatedSoftDrink	77,196	86,579
FlourandBakeryProducts	SoftDrink	77,477	86,41
TomatoandCookingSauces	BiscuitChocolateCandy	83,727	86,147
WhiteMeatProducts	BiscuitChocolateCandy	83,727	86,079
PreparedFood	FlourandBakeryProducts	69,538	85,506
FrozenFood	FlourandBakeryProducts	69,538	85,344
Pulse	BiscuitChocolateCandy	83,727	85,071

Table A2. The top five association rules of product groups for two antecedent

Consequent	Antecedent	Support (%)	Confidence (%)
BiscuitChocolateCandy	Salt and BreakfastAperitifs	50,225	100
BiscuitChocolateCandy	Salt and Pulse	50,619	100
BiscuitChocolateCandy	Salt and Bread	50,732	100
BiscuitChocolateCandy	PreparedFood and Pulse	61,318	100
BiscuitChocolateCandy	FrozenFood and Sugar	60,529	100
Bread	Salt and PreparedFood	46,791	99,519
Bread	Salt and BreakfastAperitifs	50,225	99,215
Bread	Salt and FrozenFood	46,453	99,03
Bread	Salt and FlourandBakeryProducts	49,775	98,982
Bread	PreparedFood and FlourandBakeryProducts	59,459	98,958
BreakfastAperitifs	Salt and PreparedFood	46,791	98,797
BreakfastAperitifs	Salt and FrozenFood	46,453	98,545
BreakfastAperitifs	Salt and FlourandBakeryProducts	49,775	98,529
BreakfastAperitifs	Salt and Bread	50,732	98,224
BreakfastAperitifs	Salt and WhiteMeatProducts	50,901	98,119
CarbonatedSoftDrink	Salt and FrozenFood	46,453	99,515
CarbonatedSoftDrink	Salt and PreparedFood	46,791	99,398
CarbonatedSoftDrink	Salt and AlcoholicBeverage	41,554	99,187
CarbonatedSoftDrink	Salt and WhiteMeatProducts	50,901	99,115
CarbonatedSoftDrink	Salt and Bread	50,732	99,112
Egg	Salt and FrozenFood	46,453	99,636
Egg	Salt and PreparedFood	46,791	99,398
Egg	Salt and BreakfastAperitifs	50,225	99,327
Egg	PreparedFood and FrozenFood	55,912	99,295
Egg	Salt and WhiteMeatProducts	50,901	99,226
FlourandBakeryProducts	Salt and PreparedFood	46,791	98,195
FlourandBakeryProducts	Salt and BreakfastAperitifs	50,225	97,646
FlourandBakeryProducts	Salt and FrozenFood	46,453	97,576
FlourandBakeryProducts	Salt and Pulse	50,619	97,442
FlourandBakeryProducts	Salt and Pasta	50,113	97,303
FrozenFood	Salt and PreparedFood	46,791	91,817
FrozenFood	Salt and Pulse	50,619	91,324
FrozenFood	Salt and AlcoholicBeverage	41,554	91,192
FrozenFood	Salt and BreakfastAperitifs	50,225	91,143
FrozenFood	Salt and FlourandBakeryProducts	49,775	91,063
MilkandMilkProducts	Salt and PreparedFood	46,791	100
MilkandMilkProducts	Salt and FrozenFood	46,453	100
MilkandMilkProducts	Salt and Pasta	50,113	100
MilkandMilkProducts	Salt and BreakfastAperitifs	50,225	100
MilkandMilkProducts	Salt and Pulse	50,619	100
OilandMargarine	Salt and FrozenFood	46,453	99,879
OilandMargarine	Salt and WhiteMeatProducts	50,901	99,779

OilandMargarine	Salt and Pulse	50,619	99,778
OilandMargarine	Salt and FlourandBakeryProducts	49,775	99,774
OilandMargarine	Salt and PreparedFood	46,791	99,759
Pasta	Salt and PreparedFood	46,791	98,556
Pasta	Salt and FrozenFood	46,453	98,182
Pasta	Salt and FlourandBakeryProducts	49,775	97,964
Pasta	Salt and BreakfastAperitifs	50,225	97,758
Pasta	Salt and AlcoholicBeverage	41,554	97,696
PowderDrink	Salt and PreparedFood	46,791	99,398
PowderDrink	Salt and FrozenFood	46,453	99,394
PowderDrink	Salt and WhiteMeatProducts	50,901	99,226
PowderDrink	Salt and BreakfastAperitifs	50,225	99,215
PowderDrink	Salt and FlourandBakeryProducts	49,775	99,208
PreparedCakeandDesserts	Salt and FlourandBakeryProducts	49,775	99,661
PreparedCakeandDesserts	Salt and Pulse	50,619	99,555
PreparedCakeandDesserts	Salt and BreakfastAperitifs	50,225	99,552
PreparedCakeandDesserts	Salt and PreparedFood	46,791	99,519
PreparedCakeandDesserts	Salt and TomatoandCookingSauces	50,957	99,448
PreparedFood	Salt and FrozenFood	46,453	92,485
PreparedFood	Salt and FlourandBakeryProducts	49,775	92,308
PreparedFood	Salt and BreakfastAperitifs	50,225	92,04
PreparedFood	Salt and Pasta	50,113	92,022
PreparedFood	Salt and Pulse	50,619	91,991
Pulse	Salt and PreparedFood	46,791	99,519
Pulse	Salt and FrozenFood	46,453	99,515
Pulse	Salt and FlourandBakeryProducts	49,775	99,095
Pulse	Salt and WhiteMeatProducts	50,901	98,894
Pulse	Salt and TomatoandCookingSauces	50,957	98,785
RedMeatProducts	Salt and PreparedFood	46,791	100
RedMeatProducts	Salt and FrozenFood	46,453	100
RedMeatProducts	Salt and BreakfastAperitifs	50,225	100
RedMeatProducts	Salt and Pulse	50,619	99,889
RedMeatProducts	Salt and FlourandBakeryProducts	49,775	99,887
SoftDrink	Salt and PreparedFood	46,791	99,759
SoftDrink	PreparedFood and Pulse	61,318	99,633
SoftDrink	Salt and FrozenFood	46,453	99,515
SoftDrink	PreparedFood and WhiteMeatProducts	61,824	99,454
SoftDrink	PreparedFood and PowderDrink	61,712	99,453
Sugar	Salt and PreparedFood	46,791	98,917
Sugar	Salt and FrozenFood	46,453	98,667
Sugar	Salt and FlourandBakeryProducts	49,775	98,416
Sugar	Salt and Pulse	50,619	98,331
Sugar	Salt and Pasta	50,113	98,315
TomatoandCookingSauces	Salt and FlourandBakeryProducts	49,775	99,548

TomatoandCookingSauces	Salt and PreparedFood	46,791	99,519
TomatoandCookingSauces	Salt and Pulse	50,619	99,444
TomatoandCookingSauces	Salt and BreakfastAperitifs	50,225	99,327
TomatoandCookingSauces	Salt and FrozenFood	46,453	99,273
WhiteMeatProducts	Salt and PreparedFood	46,791	99,519
WhiteMeatProducts	Salt and FrozenFood	46,453	99,515
WhiteMeatProducts	Salt and Pulse	50,619	99,444
WhiteMeatProducts	Salt and BreakfastAperitifs	50,225	99,439
WhiteMeatProducts	Salt and FlourandBakeryProducts	49,775	99,434