ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

Application of Data Mining Technology to Support Customer Insolvency Prediction at Ethiopian Telecommunication Corporation

A Thesis Submitted to The Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Information Science.

By
Gashaw Mulatu Gessesse
June 2004

Acknowledgment

First and for most I would like to thank my bothers Abay Mulatu and Hailu mulatu for their support and encouragement during my study.

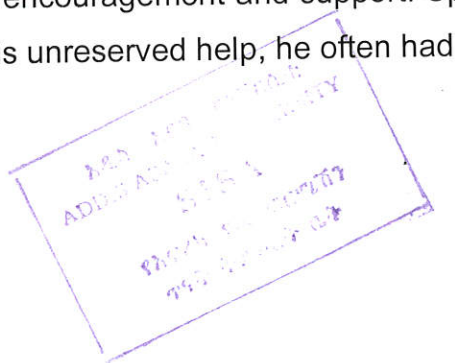
I would like to thank my advisor Dr. B. L. Desai for his constructive comments and for refining my work.

I am very grateful to the management and staff of Ethiopian Telecommunication Corporation, especially to Ato Dereje, Ato Zelalam, Ato Elias, Ato Seida, and W/o Hana, for their constant support in accessing the data and in providing relevant information to my research work.

My especial thanks goes to my beloved family for their support and encouragement and specially to my brother Biruk, who helped me in typing my paper. He was always there whenever I needed someone. My brother Teshome Mulatu deserves special thanks for his support. He was with me from the beginning to the end.

I am also grateful to my cousin Ato Abraham Cherenet for providing me relevant materials during my study. His council and encouragement was always driving me forward.

I am also grateful to my friends, staff members of SISA, and my classmates, for their encouragement and support. Special thanks to my class mate Hussen Seid for his unreserved help, he often had a solution to the problems I faced.



Dedication

I dedicate this thesis to my parents, Tsehaynesh Negede and Mulatu Gessesse, who costed everything they had to me.

List of Tables

List of Figures

Abstract

Many service-providing companies often suffer from insolvent customers who use the provided services without paying their dues. Ethiopian Telecommunication Corporation is one of these companies which is loosing considerable amount of money.

This paper reports on the findings of a research that had the objective to build a decision support system to handle customer insolvency, customers' failure to meet their payment obligation, for Ethiopian Telecommunication Corporation. The study focused on post paid mobile phone users for reason of data availability.

In the paper, the process of building a model through knowledge discovery and data mining techniques in heterogeneous as well as noisy data is described. Different statistical tools are also used for the purpose of data analysis.

The neural network backpropagation algorithm is used in the study. The particular tool used for the model building was the neural network toolbox which is incorporated in MATLAB 6.5. Different variations of the basic backpropagation algorithm were tested and the one with the best performance was selected for the model building process.

In general, a model that can classify customers, well in advance, as potentially solvent or insolvent, was built and tested. The reported findings are promising, making the proposed model a useful tool in the decision making process. And the whole research process can be a good input for further in-depth research.

Chapter one

Introduction

1.1. Background

We are living in an era where the economy is very competitive, consumer oriented, and service oriented. As a result, understanding customer behavior is important for adjusting business strategies, increasing revenues, and identifying new opportunities.

Today, in many business areas, detailed customer interaction data is abundant. We have data about purchase behavior, returns, complaints, wishes, and more. Yet, how many businesses are truly using this data effectively? The reason for this paradox is that the technology for generating, capturing, and storing data has far outpaced the human capacity to understand, analyze, and exploit it for maximum impact (Fayyad, 2003).

During the past few decades, the use of IT (Information Technology), and especially that of computer technology, has come to being from the phased automation of certain business operations, such as accounting and billing, into the present day's integrated computing environments, which offer end-to-end automation of all major business processes. Not only has the computer technology changed, but how that technology is viewed and how it is used in a business has changed. According to Bigus (1996) changes has been observed from the new hardware configuration using local and wide area networks for

distributed client/server computing to the software emphasis on object oriented programming. These changes support one overriding business requirement - process more data, faster, in ever more complex ways.

Customer focused enterprise regards every record of an interaction with a client or consider each call to customer support, each point-of-sale transaction, each catalogue order, and each visit to a company web-site as a learning opportunity. But learning requires more than simply gathering data. For learning to take place, data from many sources must first be gathered together and organized in a consistent and useful way or in short, the data has to be fed in to a data warehouse. Data warehousing allows the enterprise to remember what it has noticed about its customers. Next, the data must be analyzed, understood, and turned into actionable information. This is the point where the application of data mining is needed. In today's economy, data is the raw material that fuels business growth - if only it can be mined (Berry and Linoff, 1997).

Data Mining is the automated extraction of hidden predictive information from databases. It 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 predictions (Moxon, 1996).

Data mining is the discovery of knowledge from data, and uses a variety of tools ranging from classical statistical methods to neural networks and other new techniques originating from machine learning and artificial intelligence. Recently,

data mining has been used with substantial results in enabling and improving data base marketing and process optimization.

Data mining technology, which focuses on identifying interesting patterns and developing predictive models from data, has the greatest potential for enabling businesses to control data resources for strategic business success (Fayyad, 2003).

The first and simplest analytical step in data mining is to describe the data, summarize its statistical attributes (such as mean and standard deviation), visually review it using charts and graphs, and look for potentially meaningful links among variables (Wiley, 1997).

In the process of technological progression, more and more business data is being gathered within companies. A wealth of information is kept in company and master databases, but remains unrecognized in terms of its information content and unused in terms of company management. Special tools and modern statistics programs allow you to systematically search the master data of a company for relevant congruencies, patterns and trends. This way, a lot of information can be gathered for a more effective management and prediction of business processes of many industries (Daszezuk, 1996).

The telecommunication industry is one of the various service-giving industries. From offering local and long distance telephones services, the telecommunication industry has grown to providing many other comprehensive communication services including fax, mobile phone, e-mail, computer and web

data transmission. Changes in regulations in the industries complicate the dynamic nature of the local and cellular telephone markets. To stay in business in a changing world and very competitive market, telecommunication industries have to develop strategies that can be used for identifying market trends, detecting key characteristics and patterns for market segments, improving the quality of products and services offered, detecting fraud and insolvency early enough (Andreescu, and Zilliacus ,2002).

Senior mangers and other concerned workers of different telecommunication industries expressed that there is a need for mining the data stored in the data warehouse of telecom companies. Almost all areas of telecommunication business can benefit from data mining, but in particular the marketing and sales department can benefit much.

Andreescu and Zilliacus (2002) indicated that data used in telecommunications has many dimensions such as caller, receiver, calling time, and duration; therefore a multidimensional analysis can be performed, for identifying patterns of behavior of different groups of users, the traffic data, usage of service, etc. Usually, all these data is stored in a data warehouse that collects customer and network data, from applications like billing, marketing, sales, fraud management, performance analysis system, network switches and customer service across the company.

Telecommunication industries face considerable loss of revenue because some of the customers don't pay their dues. Therefore, detection and prevention of this phenomenon is very important for the telecommunication industry. This problem

is called customer insolvency. This problem is very difficult because sometimes the customer's behavior is the result of factors beyond his will (as social factor, etc), and is not always related to fraud. But it is still possible to conduct studies that can discriminate potentially solvent customers from potentially insolvent customers.

Examples of data used in the study are: customer profiles, usage of offered services, and financial transactions of the customers with the company. It is assumed that insolvent customers on the average behave differently from the rest of the customers, especially during a critical period preceding the due date for payment.

A number of researches have been done to show the possible applications of data mining techniques in different parts of the world. Researchers tried to prove its applicability in many organizations. In the department of information science, a number of researches have also been done so far. To mention some of these researches, Possible Application of Data Mining Technology in Supporting Loan Disbursement Activity at Dashen Bank S.C. by Askale (2001), Predictive Modeling using Data Mining Techniques in support of Insurance Risk Assessment by Tesfaye (2002), and The application of Data Mining to support Customer Relationship Management at Ethiopian Airlines by Henok (2002).

Almost all of these researches have shown the applicability of data mining to different areas (industries) in Ethiopia. It is the researcher's belief that data

mining techniques are also applicable to the Ethiopian Telecommunication Corporation, where very large data is accumulated.

This research is conducted in the form of case study at the Ethiopian Telecommunication (ETC). There is huge amount of both manual and electronic data at ETC that can be mined and used for decision support system. As to the researcher's knowledge no data mining research was conducted at ETC so far.

1.1.1. Ethiopian Telecommunication Corporation (ETC)

The introduction of telecommunication in Ethiopia dates back to 1894. During those periods, the open-wire line system was laid out connecting Addis Ababa with all the important administrative cities of the country. Most of the telecommunication network, however, was completely destroyed during the Italian Fascist aggression. It was necessary then to start the development of telecommunication facilities all over again in the country. Then the expansion of telecom services through out the nation, and training the required personnel continued. In 1996 ETC got establishment as a corporation according to Proclamation no 10/1996.

ETC was established for the following major objectives:

- To engage, in accordance with development policies and priorities of the government, in the construction, operation, maintenance and expansion of telecommunication services;

- To provide domestic and international telephone, fax and other communication services;
- To provide communication services using integrated information technology, including rebroadcast of television broadcasts;
- To repair, assemble and manufacture telecommunications equipment;
- To render training services to telecommunication personnel;
- To engage in other related activities necessary for the attainment of its purpose

In order to achieve its objectives, the corporation had undergone through series of development programs. Some of the major activities done in the history of ETC are the following:

1979 - Establishment of Satellite Communication Earth Station to facilitate international telephone, telex, telegraph and television services.

1988 - Digital exchanges go operational in Addis Ababa and other major towns for the first time.

1997 - Internet Service gets introduced - Ethio Internet established.

1999 - Mobile Telephone Service gets introduced - Ethio Mobile established

2003 - ETC introduced Pre Paid Mobile

From the different activities ETC has accomplished, which are listed above, the service provided are clear. Some of these services are fixed pone, mobile phone, Internet, and data communication. Since the focus of this research paper is on

mobile phone customers, the research would like to give a brief description of mobile service at ETC (post paid mobile phone in particular).

ETC has started post paid mobile phone service five years before. There are nearly 47,000 customers. Mobile phone services available at ETC are call diverting, call waiting, call barring, roaming, SMS (short message service), and voice mail. Users of post paid mobile phone are governmental organizations, Non governmental organizations, business organizations, international organizations, and individuals. Among these group of users, more than 70% of them are individual users.

1.2. Statement of the Problem

The major problem that made this research conducted is the insolvency problem at Ethiopian Telecommunication Corporation (ETC). Among the many problems Ethiopian Telecommunication Corporation faces, one is customer insolvency (Customers' failure to meet their payment obligation). As a result, Ethiopian Telecommunication Corporation faces considerable loss of revenue because some of the customers don't pay their dues. From the different services (Fixed phone, Mobile phone, Internet, etc) the company provides, there is high amount of uncollected money. From discussions made with officials of the company, it was learnt that the uncollected bill from mobile phone is nearly 11 million birr. This amount is seen only in five year's period, (since the service has started).

As a result of this serious problem, currently the company doesn't give post paid mobile phones to new customers except for very few individuals with special reasons. The company rather encourages its post paid mobile phone users to switch to the pre paid mobile phone. This contradicts the international trend, from what other telecom companies are doing. Telecom companies normally make every effort to make their customers use post paid mobile phones rather than prepaid to get more profit. There are two major reasons for this, one is that post paid mobile users normally don't control the duration they use, and the other reason is that they pay a certain amount of rent for a specific period of time (per month in the ETC case). But the problem of insolvent customers made ETC to do contrary to what it should have been done.

As the company continue looking for ways to maximize return from its customers, the need to reduce costs and manage business risk by evaluating customers' worthiness becomes increasingly important. Recently ETC has set one strategy that can minimize very little the problem of insolvency. The company makes follow up every three days on customers that show exaggerated phone usage particularly for international calls. This is done by looking at the call detail record of each customer within those three days period. When an exaggerated usage is observed on any number, action is taken by the company.

This problem can be solved using one data mining application area, called customer insolvency prediction. This research is conducted in the form of case study at the Ethiopian Telecommunication Corporation (ETC). It is assumed that

insolvent customers on the average behave differently from the rest of the customers, especially during a critical period preceding the due-date for payment. In this research, insolvent customers are characterized as those that refuse to pay their bill forty five days after the expiration of the deadline for payment. The whole point of the research is to confirm the hypothesis that a prediction of customer insolvency is possible by mining data of telephone service usage and customer's transaction data.

There is huge amount of both manual and electronic data at ETC that can be mined and used for decision support system with regard to customer insolvency. Developing a model that can predict customer's insolvency can help Ethiopian Telecommunication Corporation control the insolvency of current customers. That means it will enable the corporation to detect those customers early enough in case of which payment problems can be expected. This allows the mobile phone service, for example, to identify which customers should be targeted for prepaid and post paid mobile phone services.

The data mining function for this research problem is defined to be a classification problem, since the ultimate goal is to classify each customer as potentially solvent or potentially insolvent. As a suitable technique for this problem, neural network is used.

Major problems to be addressed are the following:

- Identifying patterns of behavior of different groups of users.
- Detecting as many insolvent customers as possible.

- Minimizing the number of solvent customers that would probably be wrongly classified as insolvent. This is important because maintaining loyal customers is simpler than getting new ones.

To solve the mentioned problems, a preprocessing of all customer-relevant data is performed. Specifically, cases of an insufficient credit from the past were viewed in order to create the customer profile of a customer having financial difficulties.

1.3. Objectives

1.3.1. General Objective

The general objective of the research work is to explore the potential applicability of data mining technology in developing a model that can support customer insolvency prediction for Ethiopian Telecommunication Corporation.

1.3.2. Specific Objectives

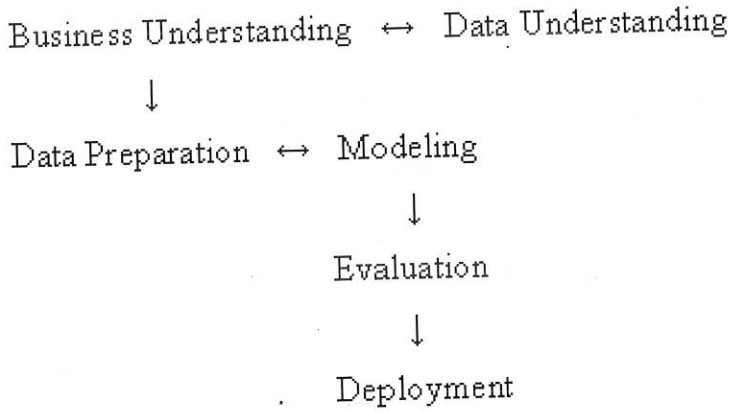
In order to achieve the above stated general objective, the study will undertake the following specific objectives:

- To identify the data - deciding what data is needed to solve the problem.
- To prepare the data using different preprocessing techniques.
- To analyze the data making it appropriate for model building.

- To build a model.
- Evaluating the accuracy of the model.
- To report the result and make recommendations.

1.4. Research Methodology

The methodology applied for this research is adopted from the techniques used by StatSoft and Two Crows. It uses the different data mining steps that need to be carried out to develop a decision support system, assuming that the business problem is well understood. The general sequence of steps for the data mining project applied in the research can be shown as follows:



The different methods to be employed for the research and how they will be applied for the study are provided below:

1.4.1. Identifying the Data Source

The data sources for this research work were customer database, billing data, and call detail record of post paid mobile phone at the Ethiopian Telecommunication Corporation. Mobile phone data is identified for the reason

that there was no sufficient data for the fixed phone. Even though the different data sources were not combined and kept in a suitably designed data warehouse, most of them were in electronic format except some reports from the billing information system.

1.4.2. Preparing Data for Analysis

Before using the raw data for analysis, it had to be converted in a format that was compatible with its end use. Hence, the collected data was cleaned into a form that was suitable for particular neural network software to be used. In spite of being cleaned, the data needed to be transformed and put in a single table to make it suitable for the mining algorithm. Among the available neural network software, MATLAB 6.5 is used for the reason that it has many algorithms that can be applied for classification problem.

At this stage pre-processing tasks like handling noisy data, unknown values, missing values, deriving new fields from the existing ones, and summarization of data was performed by taking into account the selected neural network tool.

1.4.3. Training and Building Models

The next step after clearing and transforming the collected data was training and model building. Model building, which is the core of the data mining problem, was accomplished using backpropagation algorithm on MATLAB. The training was carried out by dividing the whole data set into training, validation, and testing sets. Many models were built and those with high performance were selected.

1.5. Scope and Limitation

The scope of this research is limited to the post paid mobile phone of Ethiopian Telecommunication Corporation, where the required customer data was available. Furthermore, the study was limited in development a model that can predict current customers as potentially solvent or potentially insolvent.

The major limitations while undertaking this research was time. This has set a constraint on the amount of data collected from the call detail record (CDR). Had there been enough time, call detail records of longer period could have been obtained. The fact that ETC do not keep call detail record in back up system for more than six months made the problem worse. The absence of important attribute from the customer database of ETC was another limitation.

1.6. Thesis Organization

This research report is organized into six chapters. The first chapter briefly discusses background to the problem area and data mining technology, and states the problem, objective, research methodology, and scope and limitation of the research.

Chapters two and three review background materials necessary to understand the basic concepts and results of this research. The concepts of data mining in general are discussed in chapter two. The neural network data mining technique is reviewed in chapter three in brief.

Chapter four reviews what customer insolvency is in general and customer insolvency at Ethiopian Telecommunication Corporation in particular. In addition to this, the chapter explains how data mining technology can solve the insolvency problem.

Chapter five presents the experimentation phase of the study. This chapter discusses the data collection, data preparation, and model building process. Results of training and testing of the neural network models are also discussed in this chapter.

The final chapter provides conclusion, and recommendations given based on the findings of the undertaken research.

Chapter Two

Concepts of Data Mining

2.1. Introduction

Nowadays organizations are collecting larger and larger amounts of data. As a result databases are very huge in size. There are many reasons for this like the computerization of many businesses, scientific and governmental transaction, and advances in data collection tools. As the collected data grows in organizations, there was a need for new automated methods that can enable them to convert the collected data into useful information and knowledge. To get benefit from the collected data, there should be a way to identify relevant and useful information.

According to Witten and Frank (2000), there is a gap between the generation of data and our understanding of it. As the volume of data increases, the proportion of it that people understand decreases. There is always hidden, potentially useful information in all these data, which should be made explicit. The abundance of data and the need for powerful data analysis tools has been described as a data rich and information poor situation. The fast growing, large amount of data, collected and stored in large and many databases, has far exceeded our human ability for comprehension without powerful tools.

Kamber and Ham (2001) proposes data mining as an appropriate solution for the above problem which is the automated extraction of patterns representing knowledge stored in large databases and data warehouses. It is being used both

to increase revenues and reduce costs. The potential returns are enormous. Organizations world wide are already using data mining to locate and appeal to higher value customer, to reconfigure their product offering, to increase sales, and to minimize losses.

Data mining which was evolved from the need for the extraction of useful information and knowledge is an application specific issue, and various techniques have been developed to solve different application problems. Some examples are, mining association rules, classification, clustering and sequential patterns. In order to improve business decision, data mining tools are used, for performing data analysis on the databases or date warehouses to find data patterns.

Data mining techniques are the result of a long process of research. This evolution began when business data was first stored on computers, continued with improvements in data process, and generated technologies that allow users to navigate through their data in real time. Data mining now is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: massive data collection, powerful multiprocessor computers, and data mining algorithm (Thearling, 2002).

The purpose of data mining is to discover patterns in data so that this knowledge can be applied to problem solving. The data mining system can automatically find and show new patterns that will lead us to fresh insight. Examples of this might

be determining correlation among attributes, discriminating among subsets of the data with differing characteristics, and inferring probabilities of future events from historical data.

2.2. What Is Data Mining?

To understand the word 'data mining', it is useful to look at the literal translation of the word 'to mine', in English it means to extract. The verb usually refers to mining operation that extract from the earth hidden and precious resources. The association of this word with data suggests an in-depth search to find additional information which was not previously unnoticed in the available data.

Whitten and Frank (2000) define data mining as the extraction of implicit, previously unknown, and potentially useful information from data. The idea behind this definition is to build computer programs that shift through data warehouse, or database automatically, looking for patterns. When patterns are found in the data, prediction, association, or other generalizations can be made.

Many people treat data mining as more of a philosophy, or as a subgroup of mathematics, rather than a practical solution to business problem. The following definition by Baragoin et al (2000) shows this:

> *"Data mining is the exploration and analysis of very large data with automatically or semi-automatically procedures for*

previously unknown, interesting, and comprehensible dependencies."

The above definition omits one important aspect - the ultimate goal of data mining. In data mining the aim is to obtain results that can be measured in terms of their relevance for the owner of the database, which is a business advantage. Giudici (2003) gives us a more complete definition as follows:

"Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

From the above definition it can be seen that applying a data mining methodology means following an integrated methodological process that involves translating the business needs into a problem which has to be analyzed, retrieving the database needed to carry out the analysis, and applying a statistical technique implemented in a computer algorithm with the final aim of achieving important results useful for taking a strategic decision.

According to Giudici (2003), Data Mining, from view point of scientific research, is a relatively new discipline that has developed mainly from studies carried out in other disciplines such as computing, marketing, and statistics. Many of the methodologies used in data mining come from two branches of research, one developed in the machine learning community and the other in the statistical community, particularly in computational statistics.

As some people think, data mining is not just a set of mining algorithms, but rather a process. This process aims at solving a definite problem or making a decision. It utilizes various mathematical and computer techniques to analyze the relevant data stored in large databases. It finally finds a solution based on the discovered patterns in data and applies the solution to the predefined problem (Schumann, 2002).

According to Whitten and Frank (2000), data mining is a practical topic, and involves learning in a practical, not a theoretical, sense. People in the area are interested in techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it.

Data mining is interdisciplinary in its nature. Some of the disciplines that have contributed for data mining are machine learning, statistics, database, high performance computing, pattern recognition, neural networks, information retrieval, and image processing.

Because of the diversity of disciplines contributing to data mining, research in the area is expected to generate a large variety of data mining system. Therefore it is necessary to provide a clear classification of the data mining system. Such a classification may help potential users distinguish data mining systems and identify those that best match their needs.

One important point that should be mentioned whenever we talk about data mining is the kind of data that it is going to be mined. In principle, data mining should be applicable to any kind data. This includes relational databases, data warehouses, transactional databases, flat files, advanced databases systems, and the World Wide Web.

Another important point in data mining is the kind of patterns that can be mined. This can be specified by data mining functionalities. Some of the data mining functionalities are association analysis, classification, prediction, cluster analysis, and estimation (Kamber and Ham, 2001).

Data mining systems can be categorized according to various criteria. The first classification is according to the kind of database mined. Because database systems themselves can be classified based on different criteria, each system may require its own data mining technique. The second classification is according to the kind of knowledge mined. This kind of classification is based on data mining functionalities, such as clustering analysis, prediction and association. A comprehensive data mining system usually provides multiple and integrated data mining functionalities, the third classification is according to the kind of technique utilized. The techniques utilized in data mining systems can be described according to the degree of user interaction involved or the methods of data analysis employed. A sophisticated data mining system will often use multiple data mining techniques or work out an effective, integrated technique

that combines the merits of a few individual approaches (Kamber and Ham, 2001).

How exactly is data mining able to tell us important things that we didn't know or what is going to happen next? The technique that is used to perform these facts in data mining is called modeling. Modeling is simply the act of building a model in one situation where we know the answer and then applying it to another situation that we don't. Model building is something that people have been doing for a long time; certainly before the advent of computer or data mining technology. What happens on computer is that they are loaded up with lots of information about a variety of situation where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situation where we don't know the answer (Kamber and Ham, 2001).

2.3. Data Mining and Knowledge Discovery in Databases

Historically the activity of finding useful patterns in data has been given different names, like knowledge extraction, information harvesting, data archaeology, information discovery, pattern analysis, and data mining. The term data mining has gained popularity in the database field. A new term, knowledge discovery in databases (KDD) was created and adopted by Artificial Intelligence practitioners

by the end of the1980s to emphasize that knowledge is the end product of a data driven discovery (Fayyad, 1996)

Many people in the area agree that KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of a specific algorithm for extracting patterns from data in this context.

KDD focuses on the overall process of knowledge discovery from data, including how the data are stored and accessed, how algorithms can be scaled to massive data sets and run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported.

A number of research fields have contributed for the involvement of KDD, such as statistic, databases, machine learning pattern recognition, and expert system. The unifying goal is extracting high level knowledge from low level data in the context of large data sets. The data mining component of KDD currently relies heavily on known techniques from machine learning, pattern recognition, and statistics to find patterns from data in the data mining step of the KDD process.

But still there are some people who treat data mining as a synonym for another popularly used term, knowledge discovery in databases. Recently, as a result of the increasing attention of vendors and the media in this area, the word data

mining has been used and has come to mean, like KDD, the overall process of extracting knowledge from databases in many areas (Schumann, 2002).

2.3.1 The KDD Process

The KDD process consists of an iterative sequence of many steps. The steps are essential to ensure that useful knowledge is derived from the data. Some of the steps are outlined below.

The first step in the process is learning the application domain. This includes getting relevant prior knowledge and identifying the goal of the application or the KDD process in general.

The second step is creating a target data set. This is a step where data relevant to the analysis task are retrieved from the database. It is about focusing on a subset of attributes or records, on which discovery is to be performed.

The third step is preprocessing. This may take more than 60% of the whole effort. At this step the quality of the collected data is tested, to filter out information of no significance to a particular study and to inter-relate the heterogeneous data items in a database or a data warehouse. Basic operation of this step can be removing noise and inconsistent data and deciding on strategies for handling missing data fields (Daskalaki, 2002).

The fourth step is data reduction and transformation. This is a step where data is transformed into forms appropriate for mining by performing summary or aggregation operation. This also includes finding useful features to represent the data depending on the goal of a task.

The fifth step is choosing the data mining algorithms and selecting methods to be used for searching data patterns. This process includes deciding which models and parameters might be appropriate and matching a particular data mining methods with the overall criteria of the KDD process.

The sixth step is data mining. This is an essential process where intelligent methods are applied in order to extract data patterns. The data mining step is an essential step because it uncovers hidden pattern for evaluation. The data mining step may interact with a user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in a knowledge base (Kamber and Ham, 2001).

The seventh step is pattern evaluation and knowledge representation. This is a step where the truly interesting pattern representing knowledge base on some interesting measures are identified. This step may need returning back to the previous steps. The step may also involve visualization of the extracted pattern and models.

The final step is using the discovered knowledge. This step involves using the discovered knowledge directly, incorporating the knowledge into another system for further action, or simply documenting and reporting it to interested bodies. It also checks for potential conflicts and resolve them with previously believed knowledge.

To wind up what has been said so far, the KDD process can involve significant iteration and can contain loops between any two steps. Though most of the works in the area focus on data mining, the other steps are also important for the successful application of KDD (Fayyad and Smyth, 1996).

2.4. Data mining and other statistical tools

People have used statistical techniques for centuries to understand the natural world. These techniques included predictive algorithms which are called regression by statisticians, sampling methodologies, and experimental design. Statistics is one of the major disciplines that have contributed to data mining. It is still an important support to the field data mining (Berry and Linoff, 2000).

Data mining does not replace traditional statistical technique; it is rather an extension of statistical methods which is the result of major changes in the statistic community. The development of most statistical techniques was based on elegant theory and analytical method that worked well on the modest accounts of data being analyzed. The increased power of computers and their

lower cost, coupled with the need to analyze large data sets with millions of rows, have allowed the development of data mining technique (Two Crows, 1999).

There are many differences between statistics and data mining. Statistical analysis concerns itself with analyzing primary data that has been collected to check specific research hypothesis, while data mining can also concern itself with secondary data collected for other reasons. In addition to this, statistical data can be experimental, but in data mining, the data is typically observational.

Giudici (2003) lists three other aspects that distinguish statistical data analysis from data mining. First, data mining analyzes huge amount of data while statistical analysis limit itself to small data sets. Second, many databases do not lead to the classic forms of statistical data organizations. This created a need for appropriate analytical methods from outside the field of statistics. Third, data mining results must be of some consequences. This means that constant attention must be given to business results achieved with analysis models.

Whitten and Frank (2000) present the two fields-statistics and machine learning (as a data mining technique) by showing differences and similarities between them. With regard to their differences, statistics has been more concerned with testing hypothesis, whereas machine learning has been concerned with formulating the process of generalization as a search through possible hypothesis. And to mention their similarity, very similar schemes have been developed in parallel, in both fields. Statisticians published a book on

classification and regression trees, while almost at the same time a machine learning researcher was developing a system for inferring classification trees from examples. These two independent projects produced quite similar schemes for generating trees from examples. Another similarity between the two is that the use of nearest-neighbor method for classification. These are standard statistical techniques that have been extensively adopted by machine learning researchers.

Now the two fields have converged and applied together for specific applications. At this point, it is necessary to explain how statistical tools are used in the process of data mining.

Statistical tools are typically used to address the business problem of generating an overview of the data in a database. This is done by using techniques that summarize information about the data into statistical measures that can be interpreted without requiring every record in the database to be understood in detail. An example can be the application of statistical functions like finding the maximum and minimum, the mean, or the variance (Baragoin et al, 2001).

Some business questions addressed by statistics are:

- What is a high-level summary of the data that gives me some idea of what is contained in my database?
- Are there apparent dependencies between variables and records in my database?

- What is the probability that an event will occur?
- Which patterns in the data are significant?

There are various statistical methods and many algorithms that can be used to solve a business problem. The choice of a method depends on the problem being studied or the type of data variable. The different methods can be classified into three major classes according to the aim of the analysis. The first is descriptive method; its aim is to describe groups of data more briefly. Observations may be classified into groups not known beforehand. The second class is predictive methods. Its aim is to describe one or more of the variables in relation to all other. This is done by looking for rules of classification or prediction based on the data. These rules help us to predict or classify the future result of one or more response of target variables in relation to what happens to the input variables. The third class is local method. Its aim is to identify particular characteristics related to subset interests of the database (Giudici, 2003).

2.5. Data Mining and Data Warehouse

Nowadays, many businesses are trying to transform their data from various data sources into meaningful information that can provide their company with insights into where their business has been, is today, and is likely to be tomorrow. The process improves decision making at all levels by giving a consistent, valid, and in-depth view of a business by consolidating data from different systems into a single accessible source of information – a data warehouse.

whereas operational databases do not maintain historical data (Kamber and Ham, 2001)

Data warehouse provide a single consistent point of access to corporate or organizational data rather than access to departmental division. Data warehouse is a place where old data is published in a way that can be used to inform business decisions. The presence of data warehouse can be considered as an important preprocessing step to data mining (Whitten and Frank, 2000).

Different types of tools can be used to analyze and visualize the data from a data warehouse for different applications. The different tools can range from query and reporting to advanced analysis by data mining.

In data mining, there is a real benefit if the data to be mined is part of a data warehouse. Cleaning data for a data warehouse and for data mining are very similar. If the data has been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. In addition to this, data integration and data consolidation are also done in the construction of a data warehouse. But since constructing a large data warehouse is a huge task and time taking, sometimes it is possible to apply data mining on data that are extracted from operational or transactional databases (Two Crows, 1999).

Data warehouse are generally relational databases containing hundreds of tables described by thousands of fields. Data is brought into the system, cleaned, and

verified. Usually, there is a corresponding meta data system that is used to describe the tables, fields, reference tables, and so on. The data warehouse is often designed and then data is incrementally loaded. As the business changes, particularly by entering new markets and through merges and acquisition, data warehouse struggle to keep up with the new sources of data (Berry and Linoff, 2000).

The next important question that should be answered is "How can organizations use the information from data warehouses?" Organizations mostly use this information to support business decision making activities. Kamber and Ham (2001) put the following as important decision making activities (1) increasing customer focus, which include the analysis of customers buying patterns; (2) repositioning and managing products portfolios by comparing the performance of sale by quarter , by year, and by geographic regions, in order to fine-tune production strategies;(3) analyzing operations and looking for sources or profit; and (4) managing the customer relationship making environmental corrections, and managing the cost of corporate assets.

2.6. Data mining and On-Line Analytical processing (OLAP)

One of the most common questions from data processing professional is about the difference between data mining and OLAP (On-line Analytical Processing). Data mining and OLAP are very different tools that can complement each other.

It is important to distinguish between the capabilities of data mining from those of an On-Line Analytical Processing (OLAP) tools. OLAP uses multidimensional view of aggregate data to provide quick access to strategic information for further analysis. It enables analysts, managers, executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information. It transforms raw data so that it reflects the real dimensionality of the enterprise as understood by the user. It ranges from basic navigation and browsing to calculations, to more serious analysis, such as time series and complex modeling (Baragain et al, 2001).

OLAP is an important tool for business intelligence. It is used to explain why certain relations exist. In using OLAP, a user makes his/her own hypothesis about the possible relations between the variables and he/she looks for confirmation of his/her opinion by observing the data. Suppose a user wants to find out why some debts are not paid back; first he might say that people with a low income and lots of debts are high risk categories. So he can check his/her hypothesis, OLAP gives him a graphical representation of the empirical relation between the income, debt, and insolvency variables. Analysis of the graph can confirm his/her hypothesis. It can be seen that OLAP allows the user to extract information that is useful for business databases. Research hypothesis are suggested by a user and are not hidden from the data. OLAP can provide useful information for databases with smaller number of variables, but problems arise when there are large numbers of variables. Then it becomes very difficult and

time taking to find good hypothesis and analyze the database with OLAP tools to confirm or deny it (Giudici, 2003)

Data mining is different from OLAP because rather than verifying hypothetical pattern, it uses the data itself to uncover patterns. It is essentially an inductive process. Suppose the user who wanted to find out why some debts are not paid back were to use a data mining tool. The tool may give a result that people with low income and lots of debts are high-risk categories, it might go further and also bring out a pattern that the analyst didn't think to try, such as that age is also a determinant of risk.

According to Baragain (2000) OLAP systems have the ability to answer "who?" and "what?" question while data mining answers the question "what if?" and "why?" OLAP and data mining are complementary. OLAP can be used in the processing stages of data mining. This makes understanding the data easier, because it becomes possible to focus on the most important data, identifying special cases or looking for principal inter-relations.

The use of OLAP tools for the interactive analysis of multidimensional data of difference parts, which facilitates effective data mining, is provided by data warehouses. In addition to this, many other data mining functionalities, like classification, prediction, association, and clustering, can be integrated with OLAP operation to enhance interactive mining of knowledge at multiple levels of abstraction (Kamber and Ham, 2001).

2.7. Data Mining Technologies

There are two common categories of data mining technology: Predictive modeling and descriptive modeling. Data mining techniques differs in the approach taken to solve problems in each of these applications. There is usually a particular type of problem to be solved by each of the applications. This is to say that there is always a specific algorithm that will be used for a problem posed by people who are going to mine the data.

2.7.1. Predictive Modeling

As its name implies, predictive modeling predicts the value of a particular attribute. The goal of a predictive model is to predict the value of one column based on the value of other columns. We call these tasks, supervised modeling. In supervised modeling, there is a special attribute called the "label" that you intend to predict. By encoding the relation between the label and other attributes, the model can make predictions about new data. In addition to this, by visualizing the model itself, we can gain insight into the relationship between labels and other attributes. (Gerritsu, 1999)

The two most common predictive modeling tasks are called classification and regression. If the label is discrete (containing a fixed set of values), the task is called classification. If the label is a continuous value, the task is called regression.

Classification is the task of assigning a discrete label value to an unlabeled record. In doing so, records are divided into predefined groups. For example, a simple classification might group customer billing record into two specific classes: those who pay their bills within a month, and those who take longer than a month to pay. Classifiers can also predict the probability that the label will take on a specific value. For example, the probability that a person will pay his bill within a month can be computed.

Regression is similar to classification, except that the label is not discrete. For example, predicting salary or the price of stock is a regression, whereas predicting whether the salary is in a given range or whether a stock will go up or down is a classification task. Regression uses standard statistical techniques such as linear regression.

Competitive advantage can be obtained by applying predictive modeling in different problem areas. Risk assessment and fraud detection are major problem areas where predictive modeling can be applied in companies like insurance, bank and telecommunications (Hong & Weiss, 2002).

2.7.2. Descriptive Modeling

The goal in descriptive modeling is to discover patterns and segments of the data. These are unsupervised tasks. Unsupervised tasks provide insight to the

data as a whole by showing patterns and segments that behave similarly. The two most common descriptive modeling tasks are association and clustering.

To generate association, the task is to determine rules of implication between data attributes so that A implies B. Associations are used to find affinity groupings that discover what items are usually purchased with others. The classic affinity grouping is market basket analysis, predicting the frequency with which certain items are purchased together.

Clustering algorithms segment the data into groups of records, or clusters that have similar characteristics. Clusters help data complexity. For example, it is probably easier to design a different marketing plan for each of six targeted customer clusters than to design a specific marketing plan for each 15 million individual customers (Gerritsen, 1999).

2.8. Application of Data Mining

Data mining has become very popular because of its successful applications. Many companies are benefiting from data mining. Two critical factors for success with data mining are: large well-integrated data warehouse, and a well-defined understanding of the business process within which data mining is to be applied.

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customer, increasing revenue from existing customers, and retaining good customers. By profiling customers who have bought a particular product a company can focus attention on similar customers who have not bought that product. By profiling customers who have left, a company can act to retain customers who are at risk for leaving (Two Crown, 1999)

A credit card company uses its vast warehouse of customer transaction data to identify customer most likely to be interested in a new credit product. Such a company can also apply data mining to detect fraudulent use of their service. On such cases a decision support system can be constructed for on-line operation with temporal data like customer transaction (Thearling, 2002).

Data mining is also applicable for biomedical and DNA data analysis. Biomedical researches include the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human gene by discovering large-scale sequencing patterns. Recent research in DNA analysis has led to the discovery of new medicine and approaches for disease diagnosis, prevention, and treatment (Kamber and Ham, 2001).

Other applications of data mining can be in market analysis and management, in risk analysis, and management, for text mining, for web mining, and so on.

2.8.1. Application of Data Mining in Telecommunication

Telecommunication is one of the leading companies in applying data mining for different purposes. In the view of Kamber and Ham (2001), data mining can be used in the following ways:

Multidimensional analysis of telecommunication data: Data used in telecommunication has many dimensions, such as caller, receiver, calling time, and duration. Therefore, a multidimensional analysis can be performed for identifying patterns of behavior of different groups of users, the traffic data, usage of services, etc.

Fraudulence pattern analysis and the identification of unusual patterns: Fraud is very costly activity for the telecommunication industry; therefore companies should try to identify early the potentially fraudulent users and their typical usage patterns. Example of their actions could be attempts to gain fraudulent entry to customer accounts.

Multidimensional association and sequential pattern analysis: Based on association analysis, the telecom company can suggest to the specific customer to buy additional item, thus discovery of association and sequential patterns in multidimensional analysis is important for promotion of other its products or services.

Other two popular applications of data mining are churn prediction and customer insolvency prediction.

Churn Prediction: This is about predicting customers who are at risk of leaving a company. A mobile service provider, for example, wishing to retain its subscribers needs to be able to predict which of them may be at risk to changing services. The company should focus on those customers and make every effort to keep them from leaving. One technique of identifying such customers is to use subscribers' contractual information and call pattern changes extracted from call detail records. This allows identifying potential churners at the contract level for a specific prediction time period. This application is very important because it is usually far less expensive to retain a customer than acquire a new one (Andreesqu & Zilliacus, 2003)

Customer insolvency prediction: Customer insolvency prediction, which is the focus of this research paper, is another important application in the telecom industry. Telecommunication companies face considerable loss of revenue because some of the customers don't pay their dues. Because there is high competition in the telecommunication world, companies cannot afford the cost of insolvency. Data mining techniques can be applied for the detection of such phenomena (Daskalaki, 2002).

Data mining which is the process of building a decision support system based on large amounts of available data can be applied in telecommunication companies

to predict customer insolvency. Because telecommunication companies suffer from insolvent customers , they can get benefit from data mining by predicting customers that will refuse to pay their telephone bills in the next due date for payment, while there is still time for preventive measures.

The data mining technologies for this application can function to classify each customer as potentially solvent or potentially insolvent. Usually, researches in data mining with respect to customer insolvency investigate the hypothesis that calling habits and phone usage in general change during critical period before and right after termination of the billing period (Daskalaki, 2003).

Chapter Three

Neural Networks

There are a number of data mining techniques that can be applied for different applications. Some of these techniques are: Decision Tree, Artificial Neural Networks, and Bayesian Belief Network. This chapter briefly explains the basic concepts of neural network, which is used for this research work.

Neural network is the most widely known and the least understood of the major data mining techniques. It is an information processing system that is inspired by the way biological system, such as the brain, process information. Neural network attempts to simulate within specialized hardware and sophisticated software, the multiple layers of simple processing elements called neurons. Each neuron is linked to certain of its neighbors with varying coefficients of its connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results (Klerfors, 1998)

One important idea that should be raised here is the analogy of neural networks to the human brain. The most basic components of the neural networks are modeled after the structure of the brain. Neural networks have a strong similarity to the biological brain and therefore a great deal of terminologies is borrowed from neuroscience.

The basic element of the human brain is a specific type of cell, which are very large in number and provide us with the abilities to remember, think, and apply previous experiences to our every action. These cells are known as neurons; each of these neurons can connect with many other neurons. The power of the brain comes from the numbers of these basic components and the multiple connections between them. All natural neurons have four basic components: dendrites, soma, axon, and synapses. Basically biological neuron receives input from other sources, combines them in some way, performs generally non linear operation on the result, and then output the final result (Klerfors, 1998)

The basic unit of neural networks, the artificial neurons, simulates the four basic functions of the natural neurons. Artificial neurons are much simpler than the biological neurons. The following figures shows the basic components of an artificial neurons

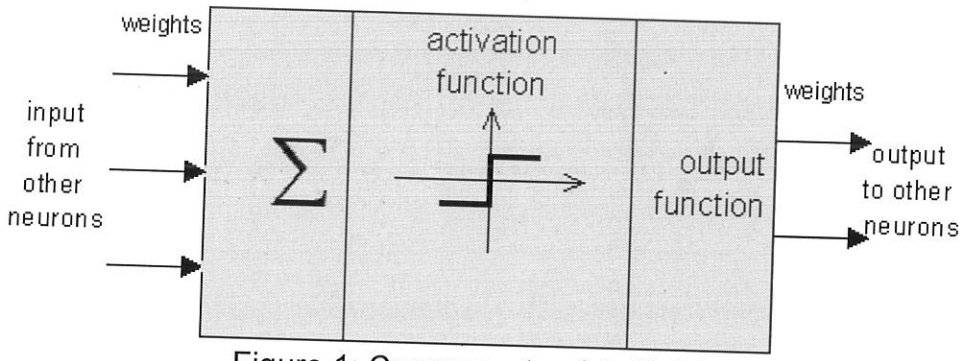


Figure 1: Components of Artificial Neuron

The various inputs to the network are represented by the mathematical system, x (n). Each of these inputs are multiplied by a connection weight, these weights are represented by w (n). In the simplest case, these products are simply summed,

fed through a transfer functions to generate a result, and then an output is provided.

3.1. Architecture of neural networks

Feed-forward networks: Feed forward artificial neural networks allow signals to travel one way only; from input to output. There is no feedback i.e. the output of any layer does not affect the same layer. Feed forward neural networks tend to be straight forward networks that associate inputs with outputs. They are extensively used in pattern recognition. They work with the back propagation algorithm. This model generalizes the network of perceptron for the architecture with hidden layers, the so called multilayered perceptron (Sima, 1998)

Feedback networks: Feedback networks can have signals traveling in both directions by introducing loops in the network. Feedback networks are very powerful and can get extremely complicated. They are dynamic; their state is changing continuously until they reach all equilibrium point. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found (Stergion and Siganos, 1996).

3.2. Network Layers

A neural network starts with an input layer, where each node corresponds to a particular variable. These input nodes are connected to a number of nodes in a hidden layer. Each input node is connected to every node in the hidden layer.

The nodes in the hidden layer may be connected to nodes in another hidden layer, or to an output layer.

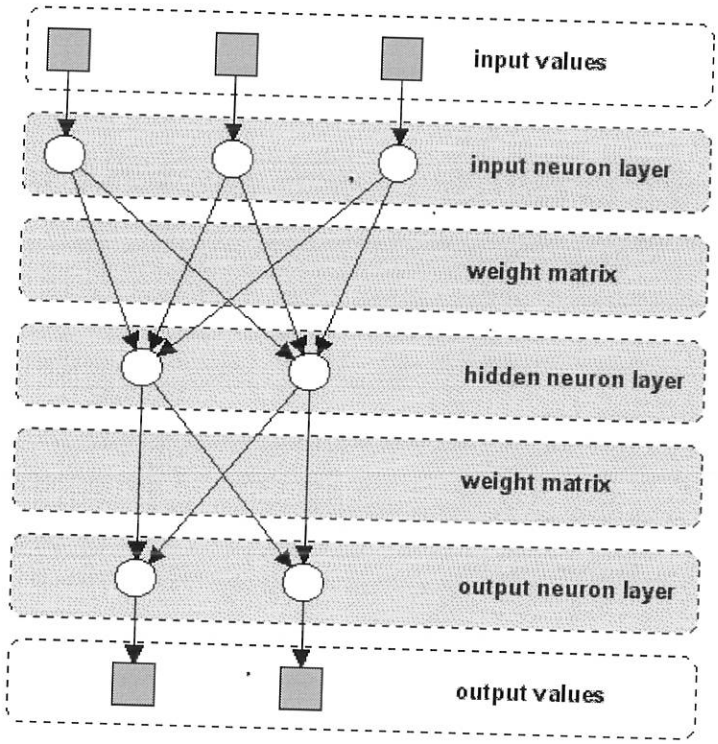


Figure 2: A neural network with one hidden layer

The output layer consists of one or more response variables. After the input layer, each node takes in a set of inputs, multiplies them by a connection weight, adds then together, applies a function (called activation function) to them, and passes the output to the node(s) in the next layer.

3.3. Training neural networks

The brain basically learns from experience. Neural networks are sometimes called machine learning algorithms, because changing of its connection weights

causes the network to learn the solution to a problem. The strength of connection between the neurons is stored as a weight-value for the specific connection. The system learns new knowledge by adjusting these connection weights (Klerfors, 1998)

Training a neural network is the process of setting the weights on the inputs of each of the units in such a way that the network best approximates the underlying function, or put in data mining terms, does the job of predicting the target variable.

The role of a training algorithm is to set the network's weight and thresholds so as to minimize predictive error by the network. The historical cases gathered are used to automatically adjust the weights and thresholds in order to minimize this error. This process is equivalent to fitting the model represented by the network to the training data available. The error of a particular configuration of the network can be determined by running all the training cases through the networks, comparing the actual output generated with the desired or targeted output. The differences are combined together by an error function to give the network error. The most common error functions are the sum squared error, where individual errors of output unit on each case squared and summed together (StatSoft, 2003)

Generally, learning in neural network can be supervised or unsupervised. A supervised neural network uses training and testing data to build a model. The

data involves historical data sets containing input variables, or data fields, which correspond to an output. The training data is what the neural network uses to learn how to predict the unknown output, and the testing data is used for validation. The aim is for the neural network to predict the output for any record given the input variables only. An important issue concerning supervised learning is the problem of error convergence, i.e. the minimization of error between the desired and computed unit values (Karahoca and Kaykayoglu, 2003).

The ultimate purpose of error correction learning is to minimize the cost function based on the error signals, such that the actual response of each output neuron in the network approaches the target response for that neuron in some statistical sense. Indeed, once a cost function is selected, error connection learning is strictly an optimization problem to which the usual tools may be brought to bear (Haykin, 1994).

3.3.1. Back Propagation Algorithm

Most software packages for building and training neural network models use some variation of the algorithm known as backpropagation algorithm. Training a backpropagation neural network has three steps:

- The network gets a training instance and, using the existing weights in the network, it calculates the out put for the instance.
- Backpropagation then calculates the error by taking the difference between the calculated result and the expected result. In order to train a

neural network to perform some task, we must adjust the weights of each unit in a way that the error between the desired output and the actual output is reduced. This process requires that the neural network compute the error derivatives of the weights. In other words it must calculate how the error changes as each weight is increased or deceased slightly (Stergion and Siganos, 1996).

- The error is used to adjust the weights. This is feeding the error back through the network.

Using the error measures to adjust the weights is the critical part of any back propagation algorithm. In classic backpropagation, each unit is assigned a specific responsibility for the error. For instance, in the output layer, one unit is responsible for the whole error. This unit then assigns a responsibility for part of the error to each of its inputs, which come from units in the hidden layers, and so on, if there is more than one hidden layer (Berry and Linoff, 2000).

3.4. Application of Neural Networks

Neural networks are good choices for most classification and prediction tasks when the results of the model are more important than understanding how the model works. Neural networks actually represent complex mathematical equations, with lots of summation, exponential functions, and many parameters.

Neural networks do not work well if there are many hundreds or thousands of input features. Large numbers of features make it more difficult for the network to find patterns and can result in long training phases that never converge to a good solution (Berry and Linoff, 2000).

Neural networks have broad applicability to real world business problems. They have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including; sales forecasting, customer research, data validation, risk management, and target marketing (Stergion and Siganos, 1996).

Chapter Four

Customer Insolvency Prediction

4.1. What is Customer Insolvency?

There has been a great interest for insolvency prediction in the business world for a long period of time. But researches in the field of data mining were restricted to very few subjects. One possible reason why insolvency prediction models have not gained greater use in the business community is because it has been difficult to calculate the results.

Customer insolvency can be defined as customer's failure to meet his/her payment obligation. Customer insolvency can be the result of insufficient credit status or bad payment moral of customers. Customer insolvency can make companies face considerable loss of revenue. The process of detecting early enough those customers, in case of which payment problems can be expected, is called customer insolvency prediction (Daskalaki, 2002).

The prediction of insolvency with statistical methods has been practical for many years. It has been applied in crediting in many areas for decades. Bankruptcy models for example, have been used in many organizations. But in the last few years, the use of data mining techniques for the prediction of insolvency has become very popular. Telecommunication companies in particular are applying the techniques, being victims of the insolvency problem, since they offer services

to their customers trusting them that they will pay their bills at the end of a billing period (Arutyunjan, 2002).

4.2. Dealing with Customer Insolvency at the Telecom Industry

Telecommunication companies face considerable loss of revenue because some customers do not pay their dues. Companies in telecommunication business take precautions against these customers; however, in most cases this refers to measures applied quite late, often with no significant effect. As a result, many unpaid bills end up in the account of uncollected debts. Thus, failure of some customers to pay their dues results in considerable loss of revenue for the company. Therefore, detection and prevention of this phenomenon is very important for the telecom industries. This can be done by building a decision support system that can handle customer insolvency. By this, customer insolvency prediction well in advance will be possible to make it useful for a company.

Telecommunication companies have made a great effort to minimize the problem of customer insolvency by developing different statistical methods. Companies have determined that statistical methods are no longer adequate to detect or prevent the problem over their complex networks. Statistical approach are often limited in scope and capacity, and are unable to serve the rapidly changing drive towards integrated data, voice, and Internet Services. In response to this need,

data mining techniques are being used providing proven decision support systems based on comprehensive and advanced techniques.

The use of data mining techniques in telecommunication companies is important because they collect high volumes data relating to different aspects of the interaction that takes place between the company and its customers. While in most cases dispersed, these data when inter-related may contain valuable information relating to the insolvency prediction problem.

By applying data mining techniques (neural network, decision tree, Bayesian network, etc), with traditional statistical analysis, decision management, and expert rules, telecommunication companies are able to predict whether a customer is solvent or insolvent (Isaac, 2003).

The following key benefits can be obtained by using the techniques for both fixed and mobile phones with respect to the insolvency problem.

- Substantially reduced financial losses and resource expense from insolvent customers.
- Ensured customer satisfaction by protecting high-value customers with reliable, secure and cost-effective services.
- Maximized efficiency of customer care and collections efforts by focusing on most critical and suspect accounts.

Examples of data that can be used for the purpose of customer insolvency prediction are: customer profiles, usage of offered services, and financial transactions of the customer with the company. It is usually assumed that insolvent customers on the average behave differently from the rest of customers, especially during a critical period preceding the due date for payment. For this reason it is important to identify behavioral patterns, which may distinguish insolvent customers from the rest, this information has to be combined with other features. In dealing with insolvency, the system to be built should enable us to detect as many insolvent customers as possible, to minimize the number of solvent customers that previously wrongly classified as insolvent, and to take actions against the identified insolvent customers (Daskalaki, 2002)

4.3. Customer Insolvency at Ethiopian Telecommunication Corporation (ETC)

The Ethiopian Telecommunication Corporation like other telecom companies worldwide suffers from insolvent customers who use the provided services without paying their dues. It is even mentioned on article 18 of the Mobile Telephone Subscription Agreement that both ETC and the subscriber acknowledge the presence of high risk of intentionally payment avoiding practices for the services subscribers used in the mobile industry. Though ETC provide many services like the Internet, fax, post and pre paid mobile phones, and fixed phone, the researcher would like to focus only on post paid phones with respect to customer insolvency, which is the purpose of this research work. It has only been five years since ETC started to offer post paid mobile phone

service. There are about 47,491 customers who are getting the post paid mobile service. There is a total of nearly birr 11 million uncollected debts in these five years period.

At this point, it is important to provide a more detailed description of the billing process at ETC, and the measures taken today against insolvency. As shown in figure 4.1, customers use their phone for one month, called the billing period. The bill is issued two weeks after the billing period and customers receive their bills approximately one week later. The due-date for payment is one week after the customer receives the bill. If a bill is not paid in this period, the company takes action on such customers. The company disconnects the phone one way, one week after payment due date for 45 days. That means the customer can only receive incoming calls and can't make out going calls for these 45 days.

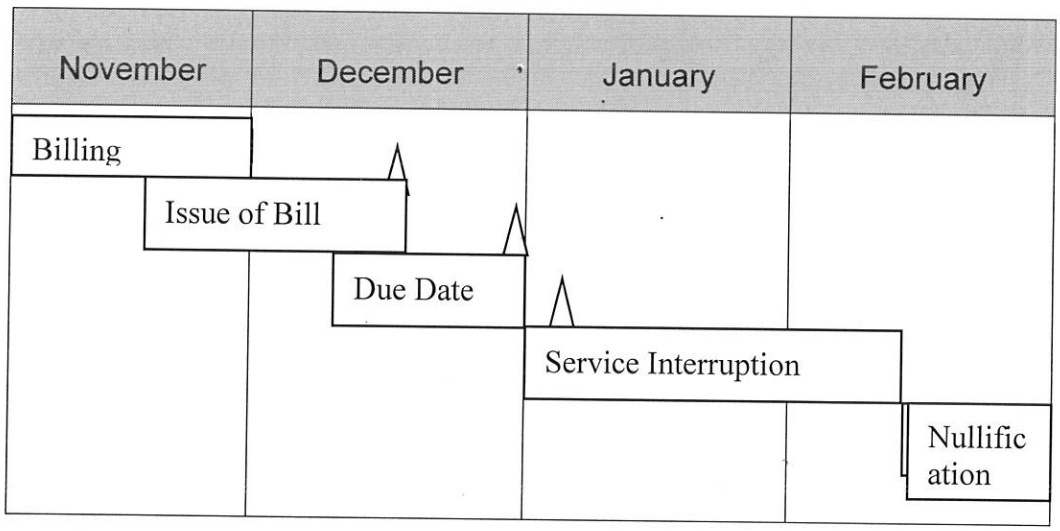


Figure 3: Time sequence of the billing process

If the customer pays, connection is reestablished and he/she will be charged a reestablishment fee. If the customer doesn't pay in the 45 days period, the company nullifies the contract and the uncollected debt will be passed to custody. Even during this period, a customer can pay his bill with extra charges and get his line back if he/she has tangible reason for not paying in the given period. In such and other similar cases a customer can pay his/her bill by installments. Normally a customer is expected to pay 50% of it at once and the rest 50% in two months time. That means the customer will be given a total of three months for installment.

Among the different approaches ETC practices to deal with insolvency problem one is collecting security deposit when a customer at first subscribes a phone. The following list shows the amounts deposited by different customer categories.

Business	408 birr
Government	220 birr
Private	134 birr
NGO/International Organization	853 birr

Source: ETC Service manual

The purpose of the security deposit is to save the loss that can result by customers who fail to pay their bills at the end of the billing period. But as it can be seen from the list that the amounts deposited are very little when compared with the amounts that potential insolvent customers refuse to pay.

Another approach ETC practices is the security-ship agreement signed by a third party and ETC. Customers, particularly if they are private customers, should bring a third party that can stand as joint guarantor for the maximum sum of to be paid in accordance with the monthly bill or a subscription charge of some amount.

One last approach ETC practice on issues related to insolvency problems is, terminating the agreement if a customer effects a late payment more than three times in any twelve month period.

Recently ETC is using one strategy to detect insolvent customers. The company makes follow up every three days on customers that show exaggerated phone usage particularly for international calls. This is done by looking at the call detail record of each customer within those three days period. When an exaggerated usage is observed on a particular telephone number, the customer will be contacted and asked about the situation. This is normally done because there is a suspicion that someone else might be using the phone. If the reasons for the exaggerated usage is convincing, the company will let the customer to use the phone for another few days. If there is still the same problem on the same number, the company disconnects the phone. If the customer claims for reestablishment of connection, he/she has to deposit some money and connection will be reestablished. Otherwise, the customer has to wait till the end of the billing period (one month), the phone being disconnected.

Chapter Five

Experimentation

This chapter details the different steps in data mining that are carried out for this particular research. The first part of the chapter explains how the target data set was created. The second part explains how the collected data is analyzed, and finally the model building and evaluation process is explained.

5.1. Creating the Target Dataset

The major task at this stage was identifying and collecting the relevant dataset for the purpose of the research work, i.e. customer insolvency prediction. Because the steps performed at this stage determine the final result of the research work, much effort was made in creating the right dataset.

As it was clearly discussed in section 4.2 of this research paper, customer behavior related data is very important for the prediction of insolvency. But customer behavior may be described by numerous characteristics, most of which are not available in the information system and telecommunication equipment of a company. For this purpose, generally two types of data were requested from those available in this context. The first group referred to static customer information (customer database), and the second to time-dependent data providing information on bills, payments and usage of telephone services, the so called Call Detail Record (CDR).

The data at Ethiopian Telecommunication Corporation were not combined and kept in a data warehouse. Hence it was necessary to collect data from different sources. The sources were databases, flat files, and manual systems.

The CDR consists of every detail of a call (call no, receiver no, call date, call time, and duration) of each telephone number. The records on the CDR were in a text format and there were nearly 15 million records on average in a one month record. The size of a one month record was about 750 MB on average.

As to the customer database, it had major fields (which are likely to be relevant to the problem of the research work) like phone number, address, category, type of security deposit, and cancellation (a field that shows if a line was interrupted for reason of not paying and connection was reestablished again).

Another source of data was identified from the billing section. In this section reports on payments, in manual format and a master list that consist of the detail of each customer usage in each month was identified.

The final source of data identified was files of customers in custody. These are customers who failed to pay their bill on the dead line set to them before the interruption of their line. Major attributes in these files are phone number, zone, month of line interrupted, and amount uncollected.

To be more specific the raw data used in this research referred to:

- Customer information from customer files,
- Usage of phone connection from switching centers,
- Billing data from the billing information system,
- Reports on payments by customer from the billing information system,
- Reports of phone disconnection due to failure of payment,
- Reports of permanent nullification of contracts.

In order to make the study representative customers from all categories were included in the dataset. These categories include governmental, NGO/International organizations, private and businesses. The numbers of records included in the original dataset are listed in table 1 below. From discussions made with domain experts it was learned that the precedence customer took post paid mobile services from ETC has some indication with regard to customer behavior. The precedence is clearly seen by the starting digits of their telephone number. Generally customers in different starting number are believed to have different characteristics and can be grouped by the first two starting digits (those who start by 20, 21, 22, 23, 24 and 25). Hence, an effort was made to make the dataset representative with respect to starting number group. Altogether 5000 customers were included in the target dataset. In order to select these 5000 records, stratified sampling was used taking category as strata. For the selected number of records again, phone starting

number was considered to make the distribution representative (with respect to phone staring number).

Category	Number of Records
Government (10)	178
Business (18)	752
Private (20)	3848
NGOs/International Organization (12)	222
Total	5000

Table 1: Number of records selected based on category

With respect to time, the data in the target dataset covered a span of 2 months. The reason for taking only a 2 month span was that, the company keeps CDR records of six recent months for reason of memory space. The total size of the collected data initially was over 4GB.

5.1.2. Description of the data collected

From the different sources described above, those that look relevant to the insolvency problem were selected. The attributes with their data types and descriptions are shown in the following lists.

Attributes collected from customer files are the following:

Attribute	Data Type	Description
Category	Text	Category of phone account
SEC_DEP	Currency	Security deposit
Address	Text	Address of customer

Attributes collected from the billing information system are the following:

Attribute	Data Type	Description
LATE_PAY	Number	Count of late pay
EXTRA_CHARGES	Number	Count of bills with extra charges
INSTALLMENTS	Number	Count of payment by installments

Attributes collected from the call detail record are the following:

Attributes	Data Type	Description
MAX_UNITS	Currency	Maximum # of units charged in any two week period during the study period.
MIN_UNITS	Currency	Minimum # of units charged in any two week period during the study period.
MAX_DUR	Number	Maximum total duration for the calls in any two-week period during the study period.
MIN_DUR	Number	Minimum total duration for the calls in any two-week period during the study period.

MAX_COUNT	Number	Maximum # of calls in any two-week period during the study period.
MIN_COUNT	Number	Minimum # of calls in any two weeks- period during the study period.
MAX-DIF	Number	Maximum # of different numbers called in any two-week period during the study period.
MIN_DIF	Number	Minimum # of different numbers called in any two-week period during the study period.

From all the collected data listed above, those that are most relevant to insolvency were selected. The selection was made after consulting domain experts. The remaining attributes were eliminated. At this stage, it was discovered that the selected data were not sufficient unless other derived attributes are incorporated which will be discussed in section 5.3.4.

5.3. Preparing Data for Analysis

This is a step where the collected data was arranged into a form that would allow the data mining tools chosen use it effectively. Steps like data cleaning, defining the data mining function, feature selection, and transformation are discussed in detail as in the following sections.

5.3.1. Data Cleaning

Before proceeding to any of the major preprocessing steps, the data from the different sources had to be arranged in a way convenient for preparation. One of these actions was splitting the Call Detail Record data using splitting software into smaller sizes since it was difficult to work on such a large size file.

According to Dasakalaki et al (2003) it is essential to test the quality of the data collected, to filter out information of no significance to the study, and to inter-relate the heterogeneous data item. This has been a tedious process in doing this research.

Examples of the performed operations during this period are included here.

Missing Values: this was a problem appeared on the data collected from all sources. Typical missing values on the call detail record were the date of call, time of call, and duration of call. One, two, or all of these three values were missing in some records of the dataset. Since it was very difficult to use other methods of dealing with missing values, it was a must to ignore such records in the study. Because a record with such missing values was irrelevant for the research problem undertaking. Missing values observed on the customer files were on the category field of customers. These values were filled manually since it was easy to get the category of a customer from documents attached on the customer's manual file.

At this stage, two attributes were eliminated since records in these attributes were not complete though they were very important for the problem of insolvency. These two attributes are LATE_PAY and EXTRA_CHARGES. The reason for the incompleteness of these attributes was that the records available cover a specific period of time.

Inaccurate Values: Such values were observed particularly on the call detail record. Some of these inaccurate values are: the receiver number being a meaningless number and call date being wrong date. As it was learned from experts of the corporation, the cause can be error generated by the digital exchange system. Because the exact cause for such inaccurate values and possible values couldn't be identified, there was no other option than ignoring such records.

Another operation performed at this stage was the elimination of inexpensive calls (charging less than 5 cents). This showed a significant decrease of the total volume of data. The elimination of these data didn't affect the result of this research given that the focus of the research was in detecting patterns of expensive calls with the ultimate goal not to be paid.

For the purpose of all the above mentioned data preparation procedures, SQL server and egrep shell of the cygwin tool were used.

5.3.3. Defining the data mining function

Predicting customer insolvency can be viewed as a classification problem, where each customer is classified in one of the two classes. Most possibly solvent or most possibly insolvent with respect to the following due data of payment and the attitude towards the amount owed in the following 45 days.

Even though there were many insolvent customers reported, it was difficult to get a significant number of them in the desired study period. As a result, the distribution of customers between the two classes was very uneven in the original dataset. Approximately 95.68% were solvent and 4.32% insolvent customers in the selected study period. Classification problem with such characteristics are difficult to solve. Therefore, a new dataset had to be created specifically for the data mining function through a stratified sampling procedure for the solvent customers.

In the new dataset, the distribution of customers between the two classes was changed to approximately 75% of solvent customers and 25% of insolvent customers. This was achieved by maintaining all cases of insolvent customers in the original dataset, while for the solvent customers a stratified sampling was performed. Group of phone account and category were used for sampling strata. The result of the above procedure was a dataset containing 864 cases in total (648 solvent and 216 insolvent).

The study period for the behavior of insolvent customer was set to be approximately a period of 2 months before the disconnection of the phone. Within the two month study period information regarding call transactions made by the customers was aggregated by two-week periods. Thus, a study period was established for each phone account called critical period during which patterns of customer behaviors were to be observed. For the accounts that were nullified due to non-payment, the critical period was defined to be the last 4 two-week periods before interruption of service. For the accounts that were not nullified, the critical period was the same 4 two-week period prior to a possible date of interruption had they not paid a certain bill.

It was based on this study period that for each account and every two-week period in the critical period several attributes were defined by counting the total number of units charged, the total duration, the number of phone calls made, and the total account of different numbers that were called during that period. The following figure shows one of these attributes, the average number of calls during their corresponding critical period.

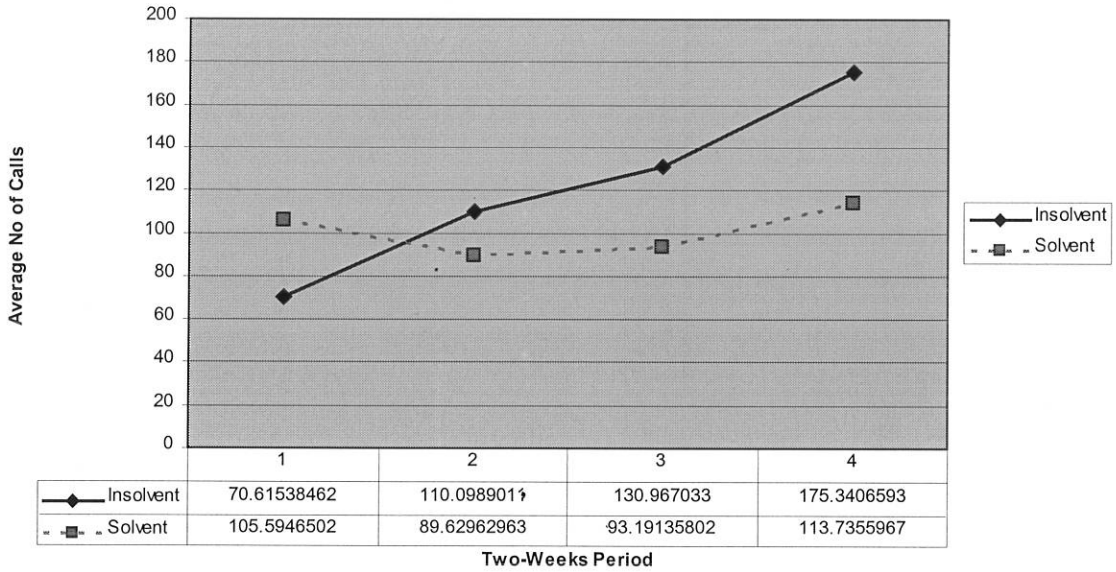


Figure 4: Average number of calls during the critical period for solvent and insolvent customers

From the above chart the difference between the solvent and insolvent customers is clearly seen. On the average the solvent customers were using their phone approximately for the same number of times during all periods, ranging from 105 to 113. On the contrary, the insolvent customers on the average were using their phone for less number of times (even less than the solvent customers) for the first few days and then their behavior changed resulting to high number of calls than solvent customers, ranging from 70 to 175.

5.3.4. Derived Attributes

The process of defining the variables for the critical period enabled the researcher creating new fields that can describe the usage of each phone account. The following were derived attributes used in the research:

Attribute	Data Type	Description
AVG_UNITS	Currency	Average # of units charged over the 4 two -week period.
STD_UNITS	Currency	Standard derivation for the units charged over the 4 two- week period.
AVG_DUR	Number	Average total duration of the calls over the 4 two- week period.
STD_DUR	Number	Standard deviation for the total duration of all calls over the 4 two -week period.
AVG_COUNT	Number	Average # of calls over the 4 two week- period.
STD_COUNT	Number	Standard deviation for the # of calls over the 4 two- week period.
AVG_DIF	Number	Average # of different # called over the 4 two-week period.
STD_DIF	Number	Standard deviation for the # of different # called during the 4 two-week period.

From discussions made with domain experts of the corporation, it was tried to evaluate some of the attributes whether they were discriminant factors for the two categories of customer (solvent and insolvent). Those that didn't provide any useful information in distinguishing solvent from insolvent were eliminated. These reduced attributes were Address and SEC_DEP.

A total of 18 independent attributes, selected as candidate discriminating factor, are listed in the following table. The dependent attribute in this experiment was taken to be the characterization of the customers as solvent or insolvent; thus the attribute solvent was assigned two values, 0 in the cases of insolvent and 1 for the cases of no such proof.

Attribute	Data Type.	Description
CATEGORY	Text	Category of phone account
INSTALLMENTS	Number	Count of payment by installments
MAX_UNITS	Currency	Maximum # of units charged in any two week period during the study period.
MIN_UNITS	Currency	Minimum units charged in any two week period during the study period.
MAX_DUR	Number	Maximum total duration for the calls in any two-week period during the study period.
MIN_DUR	Number	Minimum total duration for the calls in any two-week period during the study period.
MAX_COUNT	Number	Maximum # of calls in any two-week period during the study period.
MIN_COUNT	Number	Minimum # of calls in any two-week period during the study period.
MAX-DIF	Number	Maximum # of different numbers called in any two-week period during the study period.
MIN-DIF	Number	Minimum # of different numbers called in any two-week period during the study period.
AVG_DUR	Number	Average total duration of the calls over the 4 two-week period.
STD_DUR	Number	Standard deviation for the total duration of the calls over the 4 two- week period.

AVG_COUNT	Number	Average # of calls over the 4 two -week period.
STD_COUNT	Number	Standard deviation for the # of calls over the 4 two -week period.
AVG_DIF	Number	Average # of different numbers called over the 4 two-week period.
STD_DIF	Number	Standard deviation for the # of different numbers called over the 4 two-week
AVG_UNITS	Currency	Average # of units charged over the 4 two -week period.
STD_UNITS	Currency	Standard deviation for the units charged over the 4 two -week period.

Table 2: Selected candidate attributes

5.3.5. Feature Selection

At this phase of the data analysis process a statistical tool was used to select the most significant attributes from the candidates listed in the table of the previous section. This is done since reducing the number of input nodes to a neural network tool results a better performance as it is explained in the neural network chapter of this research paper. For this purpose attributes were cross examined for correlation and those that were strongly correlated with others were eliminated. As a result 9 attributes, out of the 18 candidate attributes, were selected as the most import to discriminate solvent from insolvent. The final 9 selected attributes are: CATEGORY, MAX_COUNT, MIN_COUNT, AVG_COUNT, STD_COUNT, MAX_DUR, MIN_DUR, AVG_DUR, and STD_DUR.

5.3.6. Data Transformation

Neural network training can be made more efficient if the network inputs and target are normalized to values between 0 and 1 or -1 and 1. Hence, the input values were normalized to a scale between 0 and 1. For this purpose, the min-max normalization method was used.

The attribute CATEGORY was changed to binary values. The existing values were changed by four combinations of binary digits (00, 01, 10, and 11). Thus, category 10 was changed by the binary numbers 00, category 12 by the binary numbers 01, category 18 by the binary numbers 10, and category 20 was changed by the binary numbers 11. This made the total number of input nodes to the neural network tool 10, since the two bits assigned for category are given as two separate nodes.

5.4 Model Building

The major task to be performed at this stage was creating and training a network that can discriminate between solvent and insolvent customer. But before discussing these major parts of the experiment, the researcher would like to explain how the data mining technique and tool used for this research work were selected.

5.4.1. Data Mining Software Selection

Even though data mining is a new field with many issues that still need to be researched in depth, there are already a great many domain specific softwares available on the market. With many data mining softwares available on the market, it is normal for some one to ask "What kind of software should I choose?" Many commercial data mining systems have little in common with respect to data mining functionality or methodology and may even work with completely different kinds of data sets. To choose data mining software that is appropriate for a specific task, it is important to have a multiple dimensional view of the software. This can include features like data types, scalability, functionality, usability, and performance (Kamber and Ham, 2001).

The data mining technique used for this research work is Neural Network. The neural network was selected because it is a good choice for classification with many advantages, which is the problem of this research work. The neural network provides a clear and simple way to search over multiple network architecture to find the best model. It also has advanced learning options and employ cross validation to govern when to stop training (Abbott et al, 1996).

Some more advantages that made neural network chosen from other software are the following:

- It is able to approximate complex non linear mapping.

- It is flexible with respect to incomplete, missing, and noisy data.
- It can be updated with fresh data, making them useful for dynamic environments.
- Its performance can be highly automated, minimizing human involvement.

It is the researcher's believe that all these advantages have something to contribute for the problem of this research work, which is a classification problem. In a classification problem, the neural network is very accurate in assigning each case to one of a number of classes or more generally to estimate the probabilities of membership of the case in each class (Carey and Collier, 1999).

5.4.1.1. MATLAB

From the many available neural network tools, MATLAB 6.5 was selected for this research work for many reasons. The neural network toolbox contained in MATLAB has many algorithms, which are variations of the basic backpropagation algorithm. It also has enormous applications, one of which is the problem of this research work – classification. The major features that made MATLAB popular and hence made chosen for this research work are the following.

Improving Generalization: One of the problems that occur during neural network training is called overfitting. The error on the training set is driven to a very small value, but when new data is presented to the network the error is

large. The network has memorized the training examples, but it has not learned to generalize to new situations. MATLAB has provided two solutions to the overfitting problem. These two solutions are regularization and early stopping with validation. While regularization involves modifying the performance function, early stopping is a technique based on dividing the data into three subsets- namely the training, the validation, and the testing set.

Pre and Post Processing: Neural network training can be made more efficient if certain preprocessing steps are performed on the network inputs and targets. Some functions included for this purpose are: Scale Minimum and Maximum, Scale Mean and Standard Deviation, Principal Component Analysis, and Post Training Analysis.

Advanced Training Options: In addition to minimizing mean squared error, which is common in most neural network tools, the MATLAB Toolbox for neural network has other options. Some of these options are: minimizing with variations of mean squared error for better generalization, training with validation to achieve appropriate early stopping, and stopping the training when the error gradient reaches a minimum.

Modular Network Representation: The modular representation in the Toolbox allows a great deal of flexibility, including the following options:

- Networks can have any number of sets of inputs, layers.
- Any input or layer can be connected to any layer with a weight.

- Each layer can have a bias or not.
- Each layer can be a network output or not.
- Weights can be partially connected
- Each layer can have a target or not

Extensible: The MATLAB Toolbox has many functions which can be applied to a broad variety of networks. It is extensible since network properties can be altered and custom properties can be added to a network object.

Custom Functions: The MATLAB Toolbox allows us to create and use many kinds of functions, giving us a great deal of control over the algorithms used to initialize, simulate, and train, our networks. Some of the functions we can create are: Simulation Functions, Initialization Functions, and Learning Functions.

5.4.1.2. Algorithm Used

The algorithm that is used for this research work is Backpropagation Algorithm. There are a number of variations of this basic algorithm which are based on other standard optimization techniques.

Properly trained back propagation networks tend to give reasonable answers when presented with input that they have never seen. Typically, a new input will lead to an output similar to the correct output for input networks used in training that are similar to the new input being presented (Demuth and Beale, 1992).

Architecture: The architecture used in this experiment is multilayer feedforward network, which is the most commonly used architecture with the backpropagation algorithm. Feedforward networks often have one or more hidden layers of sigmoid neurons followed by an output layer of linear neurons. The feedforward architecture was used to build all the models in the study.

Training function: There are a number of variations of the basic backpropagation algorithm. The MATLAB Neural Network Toolbox implements a number of these variations. From these functions three of them were tried for this research work which were suitable for the problem. These three functions are traingdm, (Gradient descent with momentum), traingdx (Gradient descent with momentum and adaptive learning rate), and trainlm (Levenberg Marquardt algorithm). Again from these three functions, trainlm was selected for it showed satisfactory results compared with the others. The MATLAB user guide also recommends users to apply trainlm as number one choice for most problems. Thus, trainlm was used for the different models built in this research work.

Performance function: For the entire model the default performance function, mean square error (mse), was used. The mean square error is the average squared error between the network outputs and the target output. During training the weights and biases of the network are interactively adjusted to minimize the network performance function (mse). The performance function is necessary in that it determines how well the neural network is doing its task.

5.4.2. Network Topology

According to Kamber and Ham (2001), deciding on network topology means mainly specifying the number of nodes in the input layer, the number of hidden layer, the number of neurons in each hidden layer, and the number of neurons in the output layer. Because there should be as many input nodes as there are attributes in the data, the numbers of input nodes for the network of all the models were ten. A single hidden-layer network with many neurons was used for the study. The number of neurons used in the hidden layer in all the models was not the same. An attempt was made to observe the performance of the models, testing for different numbers of neurons starting from small number of neuron to a large numbers. The output layer contains a single neuron, since the research deals with a two-class problem, solvent and insolvent.

The next step was to decide on the type of transfer function to be used in the hidden and output layers. The log-sigmoid transfer function was used in the output layer since the target values used in the study were between 0 and 1. In the hidden layer, hyperbolic target sigmoid (tansig) and log-sigmoid (logsig) transfer functions were tested and logsig was selected for all the models because it showed better results.

5.4.2. Data Organization for Model Building

The previous section was a preprocess activity to create a network. At this stage, it was important to organize the data into a format suitable for training, or model

goal, show, min_grad, max_fail, mem_reduc, mu, mu_inc, mu_dec, and mu_max.

The parameters epochs, goal, and time are used to determine where the training should stop. The training will stop if the number of iterations exceeds epochs, if the performance function drops below goal, or if the training time is larger than the specified time. Epochs was varied for the different models built in the study. Mem-reduc is used to control the amount of memory used by algorithm. Because memory was not a problem in this particular study mem_reduc was taken as 1, which is recommended by the tool providers. Max_fail was varied to get the best model. The max_fail values used for building the selected models are indicated in the following sections. For the other parameters, the default values were taken since they were sufficient for the problem.

The first training test was made with all the default parameters changing only the number of neurons in the hidden layer from smaller values (5 neurons) to higher values (20 neurons). Though there was a difference in accuracy in changing the number of neurons, the overall performance of the first test was not encouraging.

The second attempt made was to set some of the parameters like the epoch and max_fail to different values using the number of neurons in the hidden layer on which better performance was gained in the previous tests. Again the result was not as such encouraging; changing the activation function didn't also bring any change. Since model building is an iterative process, the researcher was forced

to go back and deal with the dataset. Two actions performed on the dataset were changing and reducing attributes. Some tests after this change showed better results and some others showed worse results.

The major problem observed in the accuracy of all the previous tests was in classifying the insolvent customers. They classify most of the insolvent customers as solvent. They had no problem in classifying the solvent customers as solvent. This actually showed that the pattern of the insolvent customers was not recognized by the model.

As a result other two actions were performed on the dataset, which actually brought a significant change on the performance of the tests carried out. These two actions were: varying the composition of inputs in the different datasets and eliminating outliers in some of the attributes. The above mentioned adjustment showed encouraging results and hence different models were built changing the different parameters and the number of neurons on the hidden layer. From which the best three models were selected and described in the following sections.

The different parameters and the number of neurons in the hidden layer used in all the three models are indicated on the following table.

Model	Epochs	Max-fail	Number of Neurons in hidden layer
Model 1	300	5	10
Model 2	500	4	15
Model 3	300	5	8

Table 3: Parameters and number of neurons used for the three best models

The training results of the three models for the parameters and number of neurons indicated in the above table, is shown in the following table. The accuracy was calculated for the test set which was reserved for this purpose.

Model	Performance	Accuracy
Model 1	0.0225	94.44%
Model 2	0.0413	94.44%
Model 3	0.0058	95.83%

Table 4: Training results for the three best models selected

The error on the training, validation, and test set are shown on the following figure. The figure shows a graph plotted during the training process for model 3, which showed the best accuracy. The graph shows a reasonable result since the

test set error and the validation set error have similar characteristics, and it doesn't appear that any significant overfitting has occurred.

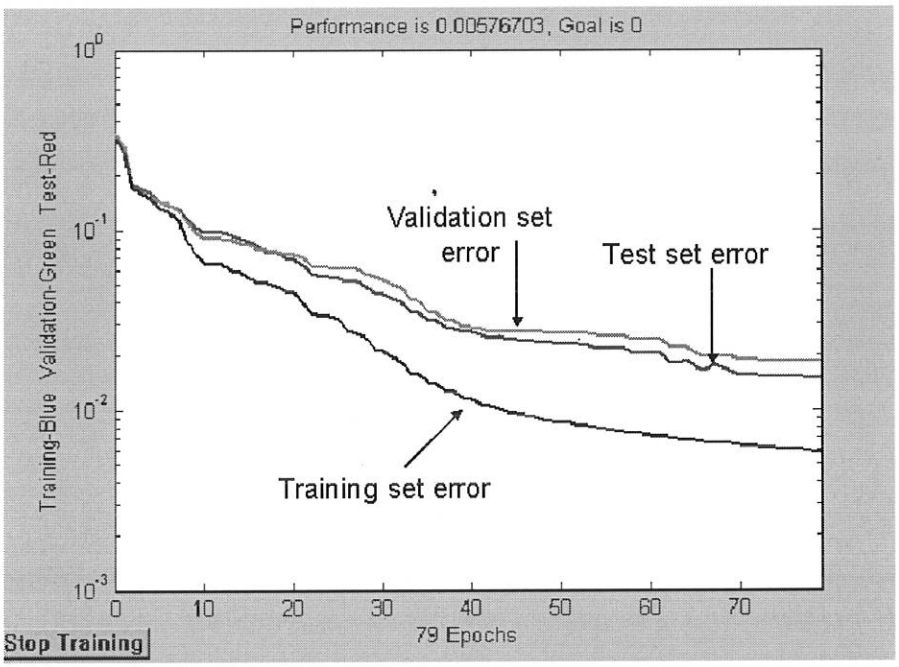


Figure 7: Training, validation, and test set error for the highest accuracy model

Even though the accuracy can be used to select the best model, it is not usually sufficient to describe every detail needed from the result of a study. Hence other methods should be used to choose which model is the best.

5.6. Evaluation and Interpretation

After building the model the next action performed was evaluating and interpreting the results of the selected models. This was an important step because accuracy by itself is not necessarily the right metric for selecting the best model. The type of errors and the costs associated with it should be known as well. One approach to deal with such cases is to use confusion matrix.

Confusion Matrix is a very useful tool for understanding the results of classification problems. Confusion matrix shows not only how well the model predicts, but also presents the details needed to see exactly where things may have gone wrong. The confusion matrix provide us an alternative to the accuracy measure, known as *sensitivity* (which shows how well the classification can recognize true positives, insolvent in this study) and *specificity* (which shows how well the classification can recognize true negatives, solvent in this study).

For the models selected in this study, the following confusion matrixes were provided using the test dataset. The sensitivity and specificity are calculated for each model as follows:

Sensitivity (Model1) = # true positives/# total positives

=44/54

=0.8148 or 81.48%

Specificity (Model1) = # true negative/# total negatives

= 160/162

= 0.9876 or 98.76%

Actual	Predicted		Total	Score
	Solvent	Insolvent		
Solvent	160	2	162	98.76%
Insolvent	10	44	54	81.48%
Total	170	46	216	94.44%

Table 5: Confusion matrix for model one

Sensitivity (Model2) = # true positives/# total positives

=47/54

=0.8704 or 87.04%

Specificity (Model2) = # true negative/# total negatives

= 157/162

= 0.9691 or 96.91%

Actual	Predicted		Total	Score
	Solvent	Insolvent		
Solvent	157	5	162	96.91%
Insolvent	7	47	54	87.04%
Total	164	52	216	94.44%

Table 6: Confusion matrix for model two

Sensitivity (Model3) = # true positives/# total positives

= 49/54

= 0.9074 or 90.74%

Specificity (Model3) = # true negative/# total negatives

= 158/162

= 0.9753 or 97.53%

Actual	Predicted		Total	Score
	Solvent	Insolvent		
Solvent	158	4	162	97.53%
Insolvent	5	49	54	90.74%
Total	163	53	216	95.83%

Table 7: Confusion matrix for model three

From the three models the best model was selected taking into account the objective of the study, i.e. maximizing the accuracy for insolvent customers. But it was also important to consider the error rate for solvent customer. As it is explained in the introduction part of this study, the model had to also minimize false alarms, i.e. the number of solvent customers that would be falsely classified as insolvent. Hence, the following table is provided to show the accuracy for insolvent customer with the error rate of solvent customers for all three models.

Model	Accuracy for Insolvent Customer	Error Rate for Solvent Customer
Model 1	81.48%	1.23%
Model 2	87.04%	3.08%
Model 3	90.74%	2.47%

Table 8: Summarized results for the three best models.

From the above table it was clearly seen that model 1 was the best taking in to account the error rate of solvent customers and model 3 was the best taking in to account the accuracy of insolvent customers. Considering both criteria, model 3 was the best since it has high accuracy for insolvent customers and relatively moderate error rate for solvent customers.

Chapter 6

Conclusion and Recommendation

6.1. Conclusion

Data mining offers great advantage to organizations where there is huge amount of data, helping them uncover patterns hidden in their data that can be used to predict the behavior of customers. The telecom industry is one of these organizations that can use data mining techniques for many applications.

This paper reports on a research project that was set to assess the applicability of data mining technology for customer insolvency problem specifically at the Ethiopian Telecommunication Corporation (ETC). The reported results are significant to ETC and other similar organization for one major reason; the study involved a real life problem. This is to say that the data used and the requirements set correspond to a real problem.

There are two limitations with this respect: the first limitation is that the scale of the experiment is not significant. The attempt to include a large percentage of customers in the study was not possible for the reason that there were only few insolvent customers reported that can be used for the specified study period. This has limited the total number of records to be included in the data set. The second limitation was the least involvement of domain experts. For reasons of

time scarcity the involvement of the domain experts was not as it needed to be for the study conducted.

In order to achieve the objective of the research, which was exploring the possibility of supporting a decision process regarding the prediction and prevention of future insolvencies from customers, a knowledge discovery in data (KDD) project was set up and executed. Creating the target data set was the first step in the KDD process, a very challenging task in conducting the research. There was no data warehouse at ETC, and hence it was necessary to integrate every relevant data from different sources, both manual and electronic. Different tools were used to extract the required data set, particularly from the Call Detail Record (CDR), which was very large in size (around 750MB for a single month record). Attributes that can discriminate the behavior of insolvent customer from that of solvent customers were also selected discussing with domain experts.

As it is believed by many people in the area, from all the phases of the KDD process, the data collection and preparation part was the most time consuming and tedious. Approximately 65% of the total time allotted for the research was consumed by this activity.

Defining the data mining function was also one phase in the process. The data mining function of the study was defined to be a classification problem, classifying customers as potentially solvent and potentially insolvent. Other tasks performed during the process were deriving attributes from the existing ones,

feature selection and transforming the data. For the purpose of feature selection correlation analysis was used. The correlation analysis reduced the final 18 candidate attribute to 9 attributes.

The experimentation part was an iterative process. The back propagation neural network algorithm was used, which is a common choice for classification problems. MATLAB 6.5 was used as a neural network tool. Different functions of the basic back propagation algorithms were tested, and one which showed better result (and also the one that is recommended by the matlab tool provider for many problems) was selected.

Using the selected function, many models were built by changing the different parameters and modifying the data set from which the best three were selected based on their accuracy. The three models were compared taking into account the accuracy of predicting insolvent customer and the error rate of solvent customer. Thus, the one with 90.74% accuracy in predicting insolvent customer and with 2.47% error rate of solvent customer was chosen as the best model.

To conclude, the result of the study can be considered as satisfactory and the proposed model can be used as one component of a decision support system that can provide advices on future insolvencies at the Ethiopian Telecommunication Corporation.

Schumann, M. (2002). *A Framework of Data Mining Application for Credit Scoring*. Available at URL: http://www.Wi2.wiso.uni-goettingen.de

Sima, J. (1998). *Introduction to Neural Networks.* Technical Report, Acadamy of Science of the Czech Republic.

StatSoft (2003). *Neural Networks*. Available at URL: http://www.StatSoftinc.com/textbook/steneunet.html

Stergion, C. and Siganos, M. (1996). *Neural Networks*. Available at URL: http://www.doc.ic.ac.uk

Tesfaye Hintsay (2002). *Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment.* Unpublished Master's Thesis. Addis Ababa University. Addis Ababa

Two Crows Corporation (1999). *Introduction to Data Mining and Knowledge Discovery*. Available at URL: http://www.twocrows.com

Wiley, J. (1997). *Introduction to Data Mining and Knowledge Discovery*. Two Crows Corporation. Available at URL: http://www.twocrows.com

Witten, H. and Frank, E. (2000).*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann Publishers.

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented as a partial degree requirement for a degree in any other university and that all sources of materials used for the thesis have been duly acknowledged.

Gashaw Mulatu Gessesse

June 2004

The thesis has been submitted for examination with my approval as university advisor

---

Dr. B. L. Desai

June 2004