ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

KNOWLEDGE DISCOVERY FOR EFFECTIVE CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOMS AUTHORITY

BELETE BIAZEN

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

KNOWLEDGE DISCOVERY FOR EFFECTIVE CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOMS AUTHORITY

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Information Science

By

BELETE BIAZEN

JUNE 2011

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF INFORMATION SCIENCE

KNOWLEDGE DISCOVERY FOR EFFECTIVE CUSTOMER SEGMENTATION: THE CASE OF ETHIOPIAN REVENUE AND CUSTOMS AUTHORITY

By

BELETE BIAZEN

Name and signature of Members of the Examining Board

Name	**Title**	**Signature**	**Date**
Ato Mahder Alemayehu	Chairperson	_____	_____
Ato Getachew Jemaneh	Advisor	_____	_____
Gashaw Kebede (PhD)	Examiner	_____	_____

DECLARATION

I declare that the thesis is my original work and has not been presented for a degree in any other university.

_____

Date

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

_____

Advisor

DEDICATION

I would like to dedicate this paper to my family, who have always been there, and supported me when I face any difficulty.

ACKNOWLEDGEMENT

First and foremost, I would like to thank the almighty GOD for giving me the ability to do this research work.

My pleasure and deepest gratitude goes to my advisor Ato Getachew Jemaneh for his patience, valuable ideas, and supportive advice. His interest and encouragement has always stimulated me to accelerate to the completion of the work.

I would like to forward my special thanks to the Ethiopian Revenue and Customs Authority employees. Many thanks go particularly to the Database Administrator of the Authority, Ato Getacher Alemu, and the customer service officers for their kind cooperation and consistent assistance in explaining the work in the problem domain.

I would also like to extend my very sincere and special thanks to Dr. Million Meshesha who provided me with the necessary assistance during the proposal writing.

I am also very thankful to my brother Endegena Biazen and my sister Zegeju Biazen for their support and encouragement in all my studies, starting from my early school age.

Lastly, my special gratitude and respects go to my friends Hilemariam Abebe, Kendie Alebachew, Shumet Taddese, Tadele Asitatikie and Tariku Adane, who supplied me with all the necessary information and valuable comments.

LIST OF ACRONYMS

AI: Artificial Intelligence

ANN: Artificial Neural Network

ARFF: Attribute Relationship File Format

ASYCUDA: Automation System Customs Data

CRISP-DM: CRoss Industry Standard Process of Data Mining

CRM: Customer Relationship Management

CSV: Comma Separated Value

DBMS: Database Management Systems

ERCA: Ethiopian Revenue and Customs Authority

IT: Information Technology

KDD: Knowledge Discovery in Databases

MOLAP: Multidimensional Online Analytical Processing

OLAP: Online Analytical Processing

ROLAP: Relational Online Analytical Processing

SAD: Single Administration Document

SOM: Self Organized Maps

SQL: Structured Query Language

WEKA: Waikato Environment for Knowledge Analysis

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

CRM is a process by which an organization maximizes customer satisfaction in an effort to increase loyalty and retain customers' business over their lifetimes. On the other hand, customer segmentation is the grouping of customers into different groups based on their common attributes and it is the main part of CRM. In order to analyze CRM data, one needs to explore the data from different angles and look at its different aspects. This should require application of different types of data mining techniques. Data mining finds and extracts knowledge hidden in corporate data warehouses.

The aim of this study is to test the applicability of clustering and classification data mining techniques to support CRM activities for ERCA using the Cios et al. (2000) KDD process model. In this study, different characteristics of the ERCA customers' data were collected from the customs ASYCUDA database. Once the customers' data were collected, the necessary data preparation steps were conducted on it and finally a dataset consisting of 46748 records was attained.

To segment customers, the K-means clustering algorithm was used. During the cluster modeling different experiments have been conducted using different cluster numbers (K=3, 4, 5, 6) and seed values. From the different experiments, the one which had better performance has been selected. Hence, the cluster model at K=5 had better performance and its output was used for the next classification modeling.

The classification modeling was built by using J48 decision tree and multilayerperceptron ANN algorithms with 10-fold cross-validation and splitting (70% training and 30% testing) techniques. Among these models, a model which was built using J48 decision tree algorithm with default 10-fold cross-validation shows better performance which is 99.95% of overall accuracy rate; hence this model was selected.

The results of this research were encouraging as very high classification accuracy has been obtained.

CHAPTER ONE

BACKGROUND

1.1 Introduction

Customer relationship management (CRM) has become one of the strategies of an organization for sustained competitive advantage. CRM in its broadest sense simply means managing all customer interaction (Trappey et al. 2009). The new millennium is in the middle of explosive change witnessing rapidly changing market conditions, volatile equity markets, reconstructed value chains and new global competitors (Kumar and Solanki 2010). Customers themselves are changing, and consider natural customer loyalty which is a thing of the past (Suresh 2002). CRM includes all measures for understanding the customers and for exploiting this knowledge to design and implement marketing activities, align production and coordinate the supply-chain (Srivastava 2002). It is an integration of technology and business to satisfy the need of the customer. CRM gains in its importance for companies that serve multiple groups of customers and exploit different interaction channels for them. CRM is a process by which a company maximizes customer satisfaction in an effort to increase loyalty and retain customers' business over their lifetimes. The primary goals of CRM are to build long term and profitable relationships with chosen customers and get closer to those customers at every point of contact (Verhoef 2003).

Segmentation can be defined as aggregating customers into groups with similar characteristics such as demographic, geographic or behavioral traits and marketing to them as a group (Parvatiyar and Sheth 2001). Consequently, each member of the segment has similar needs and wants; however, they are not completely uniform. The result was that customers often received most of what they wanted but still had to compromise on many desires (Bose 2002).

Customer segmentation is the grouping of customers into different groups based on their common attributes and it is the main part of CRM (Verhoef 2003). Segmentation requires

the collection, organization and analysis of customer data. With proper segmentations of a customer's data it is possible to identify the reliability/loyalty of customers so as to increase the revenue of the organization. CRM creates interaction of customers with the organization by using information technology (IT). Moreover, identifying customer's need/interest better and treating them accordingly can increase their life time (Verhoef 2003).

In order to analyze CRM data, one needs to explore the data from different angles and look at its different aspects. This should require application of different types of data mining techniques. There is a strong requirement for data integration before data mining. Data mining involves specialized software tools that allow users to filter through large amounts of data to uncover data content relationships and build models to predict customer behavior; data mining uses well-established statistical and machine learning techniques to build models that predict customer behavior (Suresh 2002).

Rygielski et al. (2002) define data mining as a sophisticated data search capability that uses machine learning algorithms to discover patterns and correlations in data. Data mining finds and extracts knowledge hidden in corporate data warehouses, or information that users have dropped on a website, most of which can lead to improvements in the understanding and use of the data (Rygielski et al. 2002). In simple terms, data mining is another way to find meaning in data. Data mining discovers patterns and relationships hidden in data, and is actually part of a larger process called knowledge discovery in database (KDD), which describes the steps that must be taken to ensure meaningful results (Edelstein 2002). Data mining software does not, however, eliminate the need to know the business, understand the data, or be aware of general statistical methods. Data mining helps business analysts to generate hypotheses, but it does not validate/confirm the hypotheses.

Fayyad et al. (1996) also describe data mining as the application of specific algorithms for extracting patterns from data. In addition to data mining, the additional steps in the KDD process, such as data preparation, data selection, data cleaning, are also essential to ensure that useful knowledge is derived from the data. So, to effectively segment the

customers of a certain organization it is possible to apply different kinds of data mining techniques

Classification is a data mining technique which maps the target data into the predefined groups or classes. Classification is a supervised learning technique because the classes are predefined before the examination of the target data (Deshpande and Thakare 2010). Classification is mainly used for predictive model which makes prediction about unknown data values by using the known values.

Similar to classification, clustering is also the organization of data in classes. However, unlike classification, in clustering, class labels are unknown (Han and Kamber 2006). Clustering is also referred to as unsupervised learning or segmentation and it is the partitioning or segmentation of the data into groups or clusters (Deshpande and Thakare 2010).

The present research is on the Ethiopian Revenue and Customs Authority (ERCA). ERCA has the responsibility to collect revenue for the federal government and prevent contraband. In addition to raising revenue, ERCA is responsible to facilitate the legitimate movement of people and goods across the border. Simultaneously, ERCA focuses on those people and vehicles that may involve in the act of smuggling i.e. the act of bringing into or taking out of the country goods on which customs duty and taxes are not paid and goods the importation or exportation of which are prohibited by law. ERCA conducts investigation, audits and prosecutes offenders/ criminals. In order to discharge this responsibility the ERCA works by holding the objective of using established modern revenue assessment and collection system; and provide customers with equitable, efficient and quality service, and cause taxpayers voluntarily discharge their tax obligations.

1.2 Statement of the Problem and Justification

Currently, the ERCA is using statistical analysis and assessment techniques to identify potential and low valued customers. The Authority check whether the customers discharge their responsibility or not in the revenues database and at the same time they

cross check the customs database whether the items are imported/exported with paying the required tax by using assessment and statically analysis techniques. The Authority is also using these techniques to check the credibility of the customers, i.e., whether or not they import/export goods in accordance with what they had specified in their declaration form. However, these techniques are not effective and efficient and took a considerable amount of time to treat customers according to their characteristics. For instance, international trade participants (importers, exporters) were facing some difficulties to deliver their goods to domestic and international market on time. Owing to it, importers or exporters viewed ERCA procedure with disfavor or as an obstacle that blocked the movement of international trade. These techniques are also less efficient to control taxpayers who fail to declare their actual income in order to reduce their tax bill and the federal government's revenue.

Chen et al. (2005) define CRM as the strategy integrating sales, marketing and services, which unites operating procedures and technology to better understand customers from different perspectives. ERCA has well organized customers database. Even though the organization has well-organized customers database, there is no such an integrated system or model being applied to segment customers. In addition, the ERCA has no any well-organized set of rules or procedures that are used for segmenting customers according to their culture, behavior, and characteristics to say whether the customers are potential or low valued customers. As a result, the organization is unable to collect the required revenue and control the act of smuggling.

In order to solve the problems that are described earlier, conducting a research using the data mining technology is appropriate. In the context of Ethiopia, there are some attempts made by Fekadu (2004) and Melaku (2009) in Ethiopian Telecommunications Corporation, Kumnegere (2006) in Ethiopian Shipping Lines, and Henok (2002) and Denekew (2003) in Ethiopian Airlines to solve CRM problems using the data mining technology. Moreover, different substantial research works have also been done on revenue and customs authority CRM abroad, such as Mahler and Hennessey (1996) on "Taking Internal Customer Satisfaction Seriously at the U.S. Customs Service", Doye (2010) a study on "Collaborative Border Management" and Boulding et al. (2005)

investigates on “A Customer Relationship Management Roadmap: What is Known, Potential Pitfalls, and Where to Go”. Finally, they come up with a conclusion that achieving customer satisfaction has been a keystone in the total quality management movement and in the national performance review efforts. However, to the best of the knowledge of the present researcher, there is no work done to solve CRM problems on ERCA.

Since there has been an increase in demand and supply from time to time, there is also an increase in transactions of goods in the borders of the country. So, the organization has to revisit its CRM strategy and should support it with new technology so as to satisfy its customers and reduce smuggling.

This study is attempted to address CRM related problems according to customers’ behavior, culture and characteristics. The study also focuses on grouping of customers based on the hidden information, which is found in the customers’ database of ERCA and to generate rules for identifying potential and low valued customers.

1.3 Objectives

1.3.1 General Objective

The general objective of this study is to design a model using data mining techniques for customer segmentation that helps the organization to maintain the potential CRM. This can be done through transforming customer data into meaningful categorization of customer that has been useful for designing appropriate CRM strategies for the purpose of maximizing revenue, preventing smuggling and increase customer satisfaction.

1.3.2 Specific Objectives

The specific objectives of this study are the following:

- To review the literature, to have a conceptual understanding about the study area
- To identify and preprocess the type of customers data which is found in the organizations database

- To select the data mining tool and algorithms to be used based on the objective of the study
- To prepare the data for analysis (data selection, data cleaning and organizing the training dataset)
- To build data mining model using training dataset for effective customer segmentation
- To evaluate the performance of the model

1.4 Scope of the Study

The present study focuses only on customers who perform import/export transactions through ERCA; that means the study is only restricted on the ERCA customs database. The study is also restricted on building a data mining model for segmenting the customers, interpreting the resulting segments, and developing a classification rules for each segment. The researcher uses clustering for describing the dataset and classification techniques for developing a predictive model.

1.5 Methodology

This study generally follows both quantitative and qualitative methods. The quantitative method is used to collect and analyze customers' data. On the other hand, the qualitative aspect is used to understand the business operation by making a close relationship with a domain experts and the responsible body such as the Database Administrator of the organization. In data mining or KDD process, there are different kinds of standard methodologies or models, such as Cross-Industry Standard Process for Data Mining (CRISP-DM), Anand & Buchner, Fayyad et al., Cios et al. (2000) and Cabena et al. (Kurgan and Musilek 2006). All process models consist of multiple steps executed in a sequence, which often includes loops and iterations. As Kurgan and Musilek (2006) explained the main differences between the models lie in the number and scope of their specific steps.

The following table shows the different steps of the KDD process models.

Model	Fayyad et al.	Anand & Buchner	Cios et al. (2000)	Cabena et al.	CRISP-DM
Area	Academic	Academic	Hybrid	Industry	Industry
# of steps	9	8	6	5	6
Steps	1. Developing & Understanding the Application Domain	1. Human Resource Identification	1. Understanding of the Problem Domain	1. Business Objective Determination	1. Business Understanding
	2. Creating a Target Data Set	2. Problem Specification	2. Understanding of the Data	2. Data Preparation	2. Data Understanding
	3. Data Cleaning & Preprocessing	3. Data Prospecting	3. Preparation of the Data	3. Data Mining	3. Data Preparation
	4. Data Reduction & projection	4. Domain Knowledge Elicitation	4. Data Mining	4. Analysis of Results	4. Modeling
	5. Choosing the Data Mining Task	5. Methodology Identification	5. Evaluation of the Discovered Knowledge	5. Assimilation of Knowledge	5. Evaluation
	6. Choosing the Data Mining Algorithm	6. Data Preprocessing	6. Use of the Discovered Knowledge		6. Deployment
	7. Data Mining	7. Pattern Discovery			
	8. Interpreting Mined Patterns	8. Knowledge Post-Processing			
	9. Consolidating Discovered Knowledge				

Table 1.1 Comparison of the five KDD process models Source: Kurgan and Musilek (2006)

As it is indicated in Table 1.1**,** most process models come under the categories of either industrial or academic process models. However, there are some models which combine both aspects; this kind of models are called hybrid models. An example of such kind of model is a six step KDD process model developed by Cios et al. (2000). It was developed based on the CRISP-DM model by adopting it as an academic research model. As Kurgan and Musilek (2006) explained the main difference of the Cios et al. (2000) model from other data mining model is that the Cios et al. (2000) model has the advantages of:

- providing more general, research-oriented description of the steps
- introducing a data mining step instead of the modeling step

- introducing several new explicit feedback mechanisms, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and
- modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains

So, this study employed Cios et al. (2000) KDD process model. The Cios et al. (2000) KDD process model focuses on the process of data mining projects' life cycle.

- **Understanding of the Problem Domain**

Interviews, observation and document analysis have been made to thoroughly assess the need of users and analyze the business problems. Interviews with domain experts and other officers are preferred to other methods, such as questionnaires, because interviews are used to understand the business problem very well by asking further questions or interviews is more flexible than the other techniques. Observation of the existing system of the organization is also important to understand in what manner they treat or handle their customers. Any relevant document related to customer relationship/handling would also be analyzed by the researcher to get the required information.

- **Understanding of the Data**

The data source has been the database of ERCA which contain the data of the customers engaged in import and export of different goods.

The dataset contains customer types with respect to their attributes and selection of suitable data has been made according to the organizations need and problem.

- **Preparation of the Data**

Before using the customers' data it must be converted into an appropriate form for analysis. Data preparation techniques such as analysis, editing, cleaning, integration and transformation have been preformed.

- **Data Mining**

After the data has been prepared for analysis it is used to build clustering and/or classification models using data mining techniques. This has enabled to design a better strategy for effective customer segmentation.

- **Evaluating of the Discovered Knowledge**

In order to check the output of the model's performance, counterchecking with domain experts and responsible persons (other officers) of the organization is made. Based on the confusion matrix accuracy of the system is also evaluated to see the performance of the model in segmenting customers.

- **Use of the Discovered Knowledge**

At the end of this study, the researcher has generated the final written report of the data mining engagement, including all of the deliverables, summarizing and organizing the results.

1.6 Significance of the Study

This study is important for the organization to properly segment the customers and treat them accordingly. If the organization implements proper CRM with the help of IT, the satisfaction of the customers is expected to increase, which in turn would result in an increase of revenue and reduction of smuggling. The results of this research would support the routine and strategic decision making processes of the ERCA.

1.7 Organization of the Thesis

This research paper is organized in five chapters. The first chapter deals with the background of the study which mainly introduces the problem area, states the problem, the general and specific objectives of the study, the research methodology, the scope of the study, and significance of the study. The second chapter reviews the KDD or data

mining technology. In this chapter, different data mining techniques, such as clustering and classification with their respective algorithms, are reviewed. Chapter Three discusses CRM, customer segmentation and introduces the ERCA (the organization which this research is on). Chapter Four presents the experimentation phase of the study, which mainly discusses the different stages of the experiment to build the KDD model and interprets the results of the clustering and classification experiments. The final chapter, Chapter Five, presents the conclusion of the result of the study and provides recommendation based on the investigation of the research.

CHAPTER TWO

DATA MINING

2.1 Overview

The amount of data kept in computer files and databases is growing at a phenomenal rate. At the same time, the users of these data are expecting more sophisticated information from them. A marketing manager or customer service officer is no longer satisfied with a simple listing of marketing contexts, but wants detailed information about customers past purchases as well as predictions of future ones (Dunham 2000). Simple SQL queries are not adequate to support these increased demands on information. Information retrieval (IR) is also not adequate anymore for decision-making. With huge collections of data, now there would be created new needs to help make better managerial choices. These needs are automatic summarization of data, extraction of the core information stored, and the discovery of patterns in raw data (Han and Kamber 2006). Data mining, also known as knowledge discovery in database (KDD), is able to solve these needs.

As explained by Trappey et al. (2009), data mining enables the extraction of hidden predictive information from large databases. As a result, organizations identify valuable customers, predict their future behaviors, which enable them to make positive, knowledgeable decisions. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large dataset. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data; it also includes analysis and prediction (Trappey et al. 2009).

As indicated earlier, Rygielski et al. (2002) describe data mining as a powerful data search capability that uses statistical algorithms to discover patterns and correlations in data. It is possible to apply data mining techniques on data represented in quantitative,

textual, or multimedia forms (Rygielski et al. 2002). Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns), clustering (finding and visually documenting groups of previously unknown facts), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities) (Singh and Chauhan 2009).

2.2 Data Mining Process

Data mining process is a step in KDD process which consists of methods that produce useful patterns or models from the data (Nasereddin 2009). In data mining process there are two possibilities; in some cases when the problem is known and correct data is available as well, there is an attempt to find the models or tools which will be used. On the other hand, some problems might occur because of duplicate, missing, incorrect, outlier values and there is a need to make some statistical methods. According to Nasereddin (2009), to solve these kinds of problems the data mining process passes through the following five steps:

- Defining the data mining problem
- Collecting the data mining data
- Detecting and correcting the data
- Estimating and building the model
- Model description, and validation

2.2.1 Defining the Data Mining Problem

Understanding of the projects' objectives and requirements are the first step in data mining process. Once the project have specified from a business perspective, it is possible to formulate it as a data mining problem and develop a preliminary implementation plan. Most data-based modeling studies are performed for a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement (Nasereddin 2009).

2.2.2 Collecting the Data Mining Data

This process mainly focuses on the collection of data from different sources and locations. Currently, there are two methods which are used to collect the data mining data. These are:

- Internal data: Data are usually collected from existing databases, data warehouses, and OLAP. Actual transactions recorded by individuals are the richest source of information.
- External data: In addition to data shared within a company, data items can also be collected from demographics, psychographics and web graphics.

2.2.3 Detecting and Correcting the Data

Real-world databases are usually confronted with noise, missing, and inconsistent data due to their typically huge size. Data preprocessing is commonly used as a preliminary data mining practice. It transforms the data into a format that has been easily and effectively processed by data mining algorithms. As indicated by Nasereddin (2009), there are a number of data preprocessing techniques which include:

- **Data cleaning:** This can be applied to remove noise and correct inconsistencies, outliers and missing values.
- **Data integration:** Merges data from multiple sources into a coherent data store, such as a data warehouse or a data cube.
- **Data transformations:** It is the process which improves the accuracy and efficiency of mining algorithms involving distance measurements.
- **Data reduction:** It is the reduction of the data size by aggregating, eliminating redundant features.

2.2.4 Estimating and Building the Model

According to Nasereddin (2009), the process of estimating and building the model includes the following four parts:

- select data mining task
- select data mining method
- select suitable algorithm and
- extract knowledge

Select data mining task (s): Selecting which task to use depends on the model whether it is predictive or descriptive (Two Crows Corporation 1999). Predictive models predict the values of data using known results and/or information found in large dataset. Classification, regression, time series analysis, prediction, and estimation are tasks for predictive model (Zaiane 1999). On the other hand, descriptive model identifies patterns or relationships in data and serves as a way to explore the properties of the data examined. Clustering, summarization, association rules and sequence discovery are usually viewed as descriptive (Zaiane 1999). The importance of prediction and description for particular data mining applications can be varying. That means selecting which task to use depends on the model whether it is predictive or descriptive.

Select data mining method (s): After selecting the task the next step is choosing the method of the data mining. There are a number of methods for model estimation including neural networks, decision trees, association rules, genetic algorithms, cluster detection, fuzzy logic and so on.

Select suitable algorithm: The next step is to select a suitable specific algorithm that implements the general methods.

Extracting knowledge: This is the last step in building the model which is the result (the answers for the problem solved in data mining) after making the simulation for the algorithm.

2.2.5 Model Description and Validation

In all cases, the function of the data mining models is to assist users in decision making. However, the model by itself doesn't help users to give decisions; hence, there is a need

to interpret such models because humans are unable to make good decisions on these complex models.

2.3 Data Mining as KDD Process

Data mining discovers patterns and relationships hidden in data, and is actually part of a larger process called KDD process which describes the steps that must be taken to ensure meaningful results (Rygielski et al. 2002). KDD refers to the serious extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are frequently treated as synonyms, data mining is actually part of the KDD process.

As described by Dunham (2000), the KDD process is an iterative process which consists of the following steps:

Data cleaning: Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection. The process of data cleansing is normally expensive, hence it was not possible to do with old technologies. Nowadays, faster computers allow data cleansing to be performed in an acceptable amount of time on a large amount of data (Huang 2003). In addition to KDD, data cleansing is also applied in data warehouse and total data quality management

Data integration: At this stage, multiple heterogeneous data sources are combined in a common source because the data may come from several resources. The combination of different sources of data is needed so that data is integrated. Issues of data integration include identifying similar entities, removing redundancy, detecting and removing conflicts and errors.

Data selection: At this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: Also known as data consolidation, it is a phase in which the selected data is transformed into an appropriate forms for the mining process.

Data mining: It is an essential process where intelligent techniques are applied to extract potentially useful patterns.

Pattern evaluation: Is a step in which strictly interesting patterns representing knowledge are evaluated based on given measures.

Knowledge representation: Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

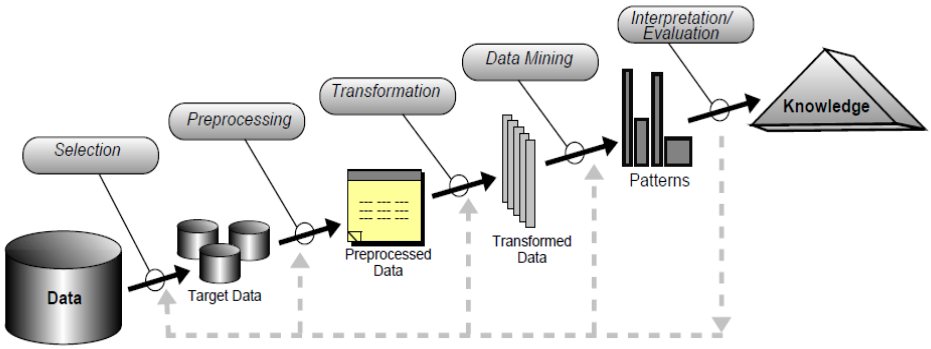

Figure 2.1 Knowledge discovery steps (Source: Fayyad et al. 1996)

2.4 Data Mining and Related Fields

2.4.1 Data Mining and Data Warehousing

Before starting any kinds of work on data mining it is better to bring all the data together because in real business environment the data is found in different departments. However, integrating data from different sources is not an easy task because different departments have been used different styles of record keeping, different conventions, different time period, different degree of data aggregation, and there have been different kinds of errors. A data warehouse is a repository/storage area of data collected from multiple heterogeneous data sources and is intended to be used as a whole under the same unified schema (Han and Kamber 2006). A data warehouse is also defined as a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process (Two Crows Corporation 1999). Data warehouse

construction can be viewed as an important preprocessing step for data mining because data warehouse involves data cleaning, data integration and data transformation. A data warehouse gives the option to analyze data from different sources under the same roof.

A data warehouse is not a requirement for data mining. However, setting up a large data warehouse that consolidates data from multiple sources resolves data integrity problems (Two Crows Corporation 1999).

As it is described by Two Crows Corporation (1999), frequently the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. There is some real benefit if data is already part of a data warehouse. For example, the problems of cleansing data for a data warehouse and for data mining are very similar. So, if the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. The data mining database may be a logical rather than a physical subset of data warehouse, provided that the data warehouse DBMS can support the additional resource demands of data mining. If it cannot, then it will be better off with a separate data mining database.

2.4.2 Data Mining and DBMS

Database management system (DBMS) provides well ground to data mining to discover knowledge from large databases. DBMSs provide a number of essential capabilities to data mining, for example as a persistent storage, a consistent data model, and a high-level query language, that allows users to request the required information. DBMS also provide the integrity of data in the database by constraint enforcement and transaction management. In addition, databases also provide a metadata description that can be used to help understand the data which is to be mined.

2.4.3 Data Mining and OLAP

There exists confusion with the difference between data mining and OLAP (On Line Analytical Processing). According to Han and Kamber (2006), OLAP is defined as the dynamic synthesis, analysis and consolidation of large volumes of multidimensional data.

OLAP is a field which is part of the decision support tools. Traditional query and report tools describe what is in a database but OLAP is more than this kind of activity; it is used to answer why certain things are true. The user of OLAP forms a hypothesis about a relationship and verifies it with a series of queries against the data.

According to Dunham (2000), OLAP tools can be classified as relational OLAP (ROLAP), multidimensional OLAP (MOLAP) or hybrid OLAP (HOLAP). During MOLAP data is modeled, viewed, and physically stored in a multidimensional database. MOLAP tools are supported by specialized DBMS and software systems capable of supporting the multidimensional data directly. With MOLAP, data is stored as n-dimensional array, thus the cube view is stored directly. Although MOLAP has extremely high storage requirements, indexes are used to speed up processing. On the other hand, with ROLAP data is stored in a relational database and ROLAP server (middleware) creates the multidimensional view for the user. Even though the ROLAP tools are less complex, it is also less efficient at the same time. Hybrid OLAP combines the best features of ROLAP and MOLAP together. Queries are stated in multidimensional terms. Data which is not updated frequently will be stored as multidimensional database, while data which is updated frequently will be stored as relational database.

Data mining is different from OLAP, because in OLAP the user should first verify hypothetical patterns and use OLAP and use the data for support where as in data mining rather than verify hypothetical patterns, it uses the data itself to discover such patterns. In addition, OLAP systems focus on providing access to multidimensional data while data mining systems deal with influence analysis of data along a single dimension (Two Crows Corporation 1999).

2.4.4 Data Mining, Artificial Intelligence and Statistics

Data mining takes advantage of advances in the fields of artificial intelligence (AI) and statistics. Both disciplines have been working on problems of pattern recognition and classification and they have made great contributions to the understanding and application of neural nets and decision trees (Two Crows Corporation 1999).

Statistics is very useful in providing a language and framework for quantifying the uncertainty, which results when one tries to conclude general patterns from a particular sample of an overall population. So, data mining does not replace these traditional statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community. The development of most statistical techniques was, until recently, based on elegant theory and analytical methods that worked quite well on the modest amounts of data being analyzed. The increased power of computers and their lower cost, together with the need to analyze enormous/ huge data sets with millions of rows have allowed the development of new techniques. Traditional statistical techniques rely on the modeler to specify the functional form and interactions.

The key point is that data mining is the application of these and other AI and statistical techniques to common business problems in a way that makes these techniques available to the skilled knowledge worker as well as the trained statistics professional (Two Crows Corporation 1999).

2.5 Data Mining Functionality

The kinds of patterns which are found in data mining tasks are specified by data mining functionalities. Generally, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks such as association analysis, clustering analysis, outlier analysis and so on, characterize the general properties of the data in the database whereas predictive mining tasks like classification and prediction perform inference on the current data in order to make predictions (Han and Kamber 2006).

Data mining functionalities can be applied to customers at all stages of the life cycle; including customer identification and customer analysis, access to new customers and upgrade existing customers value, and to maintain the valuable clients (Huang 2003).

2.5.1 Concept/Class Description: Characterization and Discrimination

There is a difference between class and customer segmentation; class is the input to find customer segment. Class, or concept description, is describing individual classes or concepts in summarized, concise, and precise terms. These descriptions can be derived via characterization and discrimination:

Characterization: Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristic rules (Han and Kamber 2006). The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

Discrimination: Data discrimination produces what are called discriminate rules and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures.

2.5.2 Association Analysis

Association analysis is the discovery of what are commonly called association rules (Witten and Frank 2005). It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets. Another threshold, confidence, which is the conditional probability that an item appears in a transaction when another item appears, is used to locate association rules. Association rules mining try to find interesting association or correlation relationship among a large set of data items. Association analysis is commonly used for market basket analysis (Bounsaythip and Rinta-Runsala 2001).

2.5.3 Classification and Prediction

Classification analysis: - Also known as supervised classification, classification analysis is the process of finding a model (or function) that describes and distinguishes data

classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown (Han and Kamber 2006). By using classification it is possible to organize data in a given classes. The classification uses given class labels to order the objects in the data collection.

Classification approaches normally use a training set where all objects are already associated with known class labels. Then the classification algorithm learns from the training set and builds a model. The model is used to classify new objects. In other words classification is a two-step process first, a classification model is built based on training data set and then the model is applied to new data for classification (Huang 2003).

Prediction: - There are two major types of predictions: numeric prediction and class label prediction. The first type of prediction that is numeric prediction predicts some unavailable data values or pending trends, and the second type of prediction (the one which is tied to classification) predicts a class label for some data.

Prediction has attracted considerable attentions by giving the potential implications of successful forecasting in a business context. Once a classification model is built based on a training set, the class label of an object can be forecasted based on the attribute values of the object and the attribute values of the classes (Two Crows Corporation 1999). Prediction is more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea in prediction is to use a large number of past values to consider probable future values.

2.5.4 Cluster Analysis

Clustering can be defined as the process of grouping a set of physical or abstract objects into classes of similar objects (Han and Kamber 2006). Clustering is also called unsupervised classification, because the classification is not dictated/ordered by given class labels. There are many clustering approaches, all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Clustering is similar to classification, but classes are not predefined and it is up to the clustering algorithm to discover acceptable classes.

Often it is necessary to modify the clustering by excluding variables that have been employed to group instances, because upon examination the user identifies them as irrelevant or not meaningful. After clusters are found that reasonably segment the database, these clusters then used to classify new data. Some of the common algorithms used to perform clustering include Kohonen feature maps and K-means.

Some peoples are confused with the difference between clustering and segmentation; however, clustering is different from segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics where as clustering is a way to segment data into groups that are not previously defined (Two Crows Corporation 1999). Clustering is useful to find natural groups of data which is called clusters. A cluster is a collection of data that are similar to one another. Clustering can be used to group customers with similar behavior and to make business decisions in industry (Huang 2003).

2.5.5 Outlier Analysis

Outliers are data elements in the database that cannot be grouped in a given class or cluster. In other words, outliers are data objects that don't fulfill the general behavior or model of the data. Outliers are also known as exceptions or surprises, they are often very important to identify. Although outliers can be considered as noise and discarded in some applications, they can produce important knowledge in other domains, so their analysis is very significant.

2.5.6 Evolution and Deviation Analysis

Evolution and deviation analysis is relevant to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which permit to characterizing, comparing, classifying or clustering of time related data. Deviation

analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

2.6 Data Mining Techniques and Algorithms

2.6.1 Clustering Technique

Clustering is a tool for data analysis, which solves classification problems. Its objective is to distribute cases (people, objects, events etc.) into groups, so that the degree of similarity can be strong between members of the same cluster and weak between members of different clusters (Hajizadeh et al. 2010). In clustering, there is no pre-classified data and no distinction between independent and dependent variables. Instead, clustering algorithms search for groups of records (the clusters composed of records similar to each other). The algorithms discover these similarities.

According to Faber (1994), clustering involves dividing a set of data points into non-overlapping groups, or clusters, of points, where points in a cluster are more similar to one another than to points in other clusters. When a dataset is clustered, every point is assigned to some cluster, and every cluster can be characterized by a single reference point, usually an average of the points in the cluster. Any particular division of all points in a dataset into clusters is called a partitioning.

Clustering algorithms are used when there have been a large complex dataset with many variables and lots of internal structures. Clustering algorithms are often used to find outliers or records that don't fit the predictive model. Many clustering algorithms have been developed including K-means (K-nearest neighbors), hierarchical, fuzzy C-means and a special type of neural network called Self Organized Maps (SOM) or kohonen net maps (Bounsaythip and Rinta-Runsala 2001).

2.6.1.1 The K-Means Algorithm

K-means clustering, which was developed by MacQueen in 1967, is the most frequently cited algorithm for database applications **(**Trappey et al. 2009**)**. Since K-means algorithm is simple it is used in various fields.

K-means is a partition clustering method that separates data into K mutually excessive groups. By iterating such partitioning, K-means minimizes the sum of distance from each data to its clusters. K-means method is very popular because of its ability to cluster huge data, and also outliers, quickly and efficiently. Even though, K-means algorithm is very sensitive in initial starting points K-means generates initial cluster randomly. When random initial starting points close to the final solution, K-means has high possibility to find out the cluster center. Otherwise, it will lead to incorrect clustering results. Because of initial starting points generated randomly, K-means does not guarantee the unique clustering results (Arai and Barakbah 2007).

Arai and Barakbah (2007) explain that the K-means method is a well known geometric clustering algorithm. Given a set of n data points, the algorithm uses a local search approach to partition the points into k clusters. A set of k initial cluster centers is chosen arbitrarily. Each point is then assigned to the center closest to it, and the centers are recomputed as centers of mass of their assigned points. This is repeated until the process stabilizes. In simple K-means clustering, there is no partitioning a single point twice, so the algorithm is guaranteed to terminate.

K-means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters especially for large or huge datasets. Since the K-means clusters are non-hierarchical they do not overlap. With a large number of variables, K-means may be computationally faster than hierarchical clustering (if K is small). K-means may produce tighter/rigid clusters than hierarchical clustering, especially if the clusters are globular.

2.6.1.2 Self Organized Maps (SOM)

SOM or kohonen feature map, is a special kind of network architecture that provides a mapping from the multi-dimensional input space to a lower order regular lattice of cells (typically two dimensional grids). SOM is important in multi-dimensional inputs, because when the set of input is multi-dimensional, traditional clustering algorithms do not offer an easy way to visualize the closeness of other clusters (Bounsaythip and Rinta-Runsala

2001). Such a mapping is used to identify clusters of elements that are similar in the original space.

Unlike other neural network approaches, the SOM network performs unsupervised training. The most common approach to neural networks requires supervised training of the networks, i.e, the network is fed with a set of training cases and the generated output is compared with the known correct output. The SOM network, on the other hand, does not require the knowledge of the corresponding outputs. The nodes in the network converge to form clusters to represent groups of entities with similar properties. The number and composition of clusters can be visually determined based on the output distribution generated by the training process.

The SOM network is typically a feed-forward neural network with no hidden layer; it has two layers of nodes, the input layer and the Kohonen layer. The input layer is fully connected to a two or one or multi-dimensional Kohonen layer. In effect multiple outputs are created and the best one is chosen (Dunham 2000).

2.6.2 Classification

Classification is one of the most common learning models in data mining techniques and it aims at building a model to predict future customer behaviors through classifying database records into a number of predefined classes based on certain criteria (Ngai et al. 2009). According to Dunham (2000), classification is viewed as a mapping from the database to the set of classes. Classes which produced by classification are predefined, none overlapping and partition the entire database. There are several classification techniques including decision trees and neural networks. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. This pattern can be used both to understand the existing data and to predict how new instances will behave (Two Crows Corporation 1999).

Issues in Classification

According to Dunham (2000), there are different issues when using classification techniques. The most common ones are the following:

Missing data: One of the most common classification issues is missing data values, which cause problems during both the training phase and the classification process itself. Missing values in the training data has to be handled and may produce an inaccurate result. There are many approaches to handle missing data such as, ignore the missing data, assume a missing value for the data (this may be determined by using some method to predict what the value could be) and assume a special value for the missing data (this means that the missing data is taken to be a specific value all of its own).

Measuring performance: The other issue in classification techniques is performance measurement. The performance of classification algorithms is usually examined by evaluating the accuracy of the classification. However, since classification is often a fuzzy problem, the correct answer may depend on the user. Classification accuracy is usually calculated by determining the percentage of tuples placed in the correct class.

2.6.2.1 Decision Tree

Decision tree algorithm is a data mining induction technique that recursively partitions a dataset of records using either depth-first greedy approach or breadth-first approach until all the data items belong to a particular class (Anyanwu and Shiva n.d). Decision trees are a way of representing a series of rules that lead to a class or value (Two Crows Corporation 1999) and it is a powerful and popular tool for classification and prediction. In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. Decision trees are also useful for exploring data to gain sight into the relationships of a large number of candidate input variable to the target variable. Since decision trees combine both data exploration and modeling, they are a powerful first step in modeling process even when building the final model using some other techniques. When a decision tree is used for classification tasks, it is more

appropriately referred to as a classification tree and when it is used for regression tasks, it is called regression tree.

The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so records falling into a particular category may be retrieved.

Hajizadeh et al. (2010) also described decision tree as it is a model which consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable. The target variable is usually categorical and the decision tree model is used either to calculate the probability that a given record belongs to each of the categories, or to classify the record by assigning it to the most likely class.

According to Anyanwu & Shiva (n.d), decision tree classification technique has tree building and tree pruning phases. Tree building is a top-down approach in which the tree is recursively partitioned until all the data items belong to the same class label. The tree growing or building phase is an iterative process which involves splitting the data into progressively smaller subsets (Bounsaythip and Rinta-Runsala 2001). Tree-building algorithms usually have several stopping rules. These rules are usually based on several factors including maximum tree depth, minimum number of elements in a node considered for splitting, or it's near equivalent, the minimum number of elements that must be in a new node (Bounsaythip and Rinta-Runsala 2001). Tree pruning, on the other hand, is a bottom-up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing over-fitting (noise or much detail in the training data set); over-fitting in decision tree algorithm results in misclassification error (Dunham 2000). There are two methods of tree pruning; the first type of pruning is pre-pruning which involves try to decide during the tree-building process when to stop developing sub trees; this technique involves look ahead during tree generation. The second type of pruning is post-pruning which involve heuristics to determine which branch to prune or using cross-validation to get a better estimate on good prunes. Sub tree

replacement and sub tree raising are examples of post-pruning methods (Witten and Frank 2005).

The disadvantage of many classification techniques like Bayesian classifier is that the classification process is difficult to understand. However, humans do easily understand and accept decision rules. So, by using decision tree method it is possible to make good decisions especially decisions that involve high costs and risks.

2.6.2.1.1 Decision Tree Algorithms

According to Bounsaythip and Rinta-Runsala (2001), the most commonly implemented decision tree algorithms which are well suited for classification techniques include Chi-squared Automatic Interaction Detection (CHAID), Classification and Regression Trees (CART), C4.5 and C5.0.

CART: - This is a technique which generates a binary decision tree. CART uses binary split based on GINI (recursive partitioning motivated by statistical prediction) techniques and in CART algorithm there exists exactly two branches from each non-terminal node. In CART algorithm pruning is performed based on measure of complexity of the tree.

C4.5 and C5.0: - The decision tree algorithm C4.5 is an improvement to ID3 and it produce tree with multiple branches per node (Dunham 2000). The number of branches is equal to the number of categories of predictor. C4.5 combines multiple decision trees into a single classifier. In C4.5 pruning is performed based on error rate at each leaf. C4.5 algorithm has also the ability to handles missing and continuous data. C5.0 is a commercial version of C4.5. Unlike C4.5 the precision algorithm used for C5.0 have not be disclosed.

CHAID: - This is a multi-way splitting algorithm using chi-square tests (detection of complex statistical relationship). The number of branches varies from two to the number of predictor categories.

2.6.2.2 Neural Network Algorithm

According to Singh and Chauhan (2009), neural networks are defined as a non-linear statistical data modeling tools that can be used to model complex relationships between inputs and outputs or to find patterns in data. The neural networks approach, like decision trees, requires that a graphical structure be built to represent the model and then that the structure be applied to the data (Dunham 2000). According to Dunham (2000), a neural network can be viewed as a directed graph with source (input), sink (output), and internal (hidden) nodes. The hidden nodes may exist over one or more hidden layers. To perform the data mining tasks, a tuple is input through the input nodes and the output node determines what the prediction is. Unlike decision tree which have only one input node (the root of the tree), a neural net has one input node for each attribute value to be examined to solve the data mining function. Unlike decision trees, after a tuple is classified, the neural network may be changed to improve future classification applications. Although the structure of the graph does not change, the labeling of the nodes and edges may change.

It is possible to collect important information from data warehouse in the process of data mining by using neural networks as a tool. Neural networks essentially consist of three pieces: the architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or trained to store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill-defined problems; in summary, to estimate sampled functions when it is difficult to know the form of the functions (Singh and Chauhan 2009).

It is possible to successfully applied neural networks in both supervised and unsupervised learning applications. In data mining tasks we can commonly use neural network methods, because neural network often produce clear and understandable models. A neural network is a computational technique which is similar to what the human brain works. It is designed to imitate or copy the ability of the human brain to process data and information and understand or realized patterns. It imitates the structure and operations of

the three dimensional web of network among brain cells (nodes or neurons, and hence the term ‘‘neural’’) (Hajizadeh et al. 2010).

The major drawback to the use of neural networks is the fact that they are difficult to explain to the end users (unlike decision trees which are easy to understand), it does not generate rules like decision trees; thus, no rules are derived for the model and unlike decision trees, neural networks usually work only with numerical data (Dunham 2000).

Neural networks except SOM can work well with noisy data and data with missing some values. Neural networks are good tool for predictions and classification. When there are lots of input features select only important features for the training phase by using other methods (e.g associations) (Bounsaythip and Rinta-Runsala 2001).

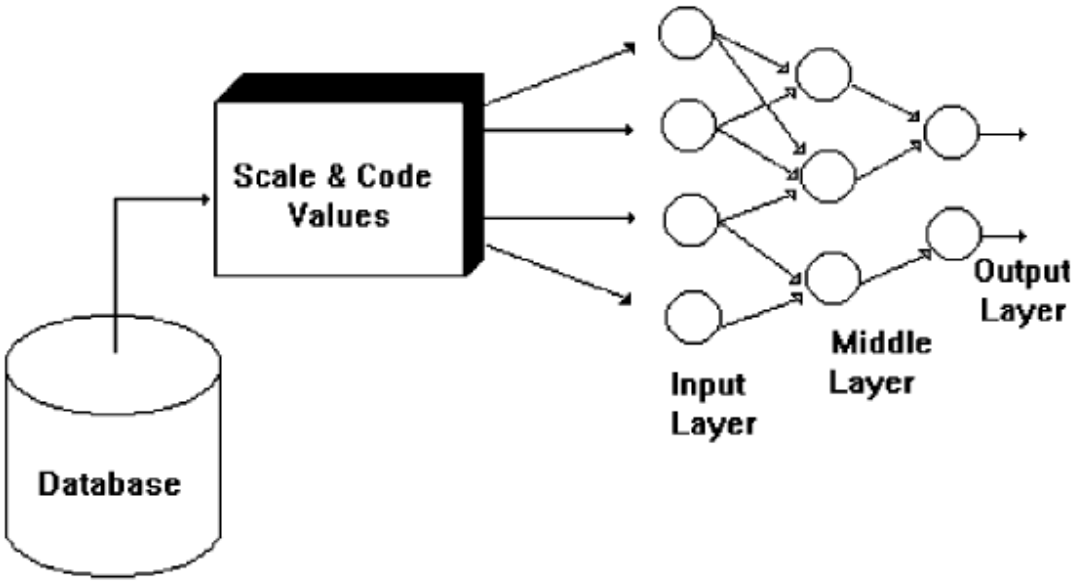

Figure 2.2 Diagram of a typical neural network (Source: Rygielski et al. (2002))

2.7 Applications of Data Mining

Data mining has so many applications in reduced costs of doing business, improved profitability, or enhanced quality of service for different fields. Areas in which such benefits have been demonstrated includes Insurance, Direct Mail Marketing,

Telecommunications, Retail, Medicine, CRM and others (Huang 2003). Data mining applications have been shown to be highly effective in addressing many important business problems. Each business is interested in predicting the behavior of its customers through the knowledge gained in data mining

Application of data mining in medical science: In healthcare data mining is applicable in many different ways. These applications can generally group as the evaluation of treatment effectiveness, management of healthcare, CRM, and detection of fraud and abuse (Koh and Tan n.d). Diagnosis of disease, health care, patient profiling and history generation etc. are the few examples of application of data mining in medical science.

Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumors/swelling. Computer-aided methods could assist medical staff and improve the accuracy of detection (Deshpande and Thakare 2010). The neural networks with back-propagation and association rule mining, used for tumor classification in mammograms.

Koh and Tan (n.d) describe the applications of data mining in healthcare as follows:

- **Treatment effectiveness:** Data mining applications can be used to assess the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis, of which courses of action verify effective (Koh and Tan n.d). For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective.
- **Healthcare management:** Identifying chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of patients and claims are an example of data mining applications that assist healthcare management.
- **CRM:** While CRM is a core approach in managing interactions between business organizations such as banks and retailers and their customers, it is also important

in a healthcare context. Customer interactions may occur through reservation, physicians' offices, inpatient settings, and ambulatory care settings.

- **Fraud and abuse:** Data mining applications that attempt to detect fraud and abuse often establish norms and then identify unusual or abnormal patterns of claims by physicians, laboratories, clinics, or others (Koh and Tan n.d).

Application of data mining in distance learning: Data mining is useful in distance learning to enhance the learning process based on the vast amount of data generated by the tutors and student's interactions with web based distance-learning environment (Deshpande and Thakare 2010). The data mining applications transfers the data into information and feedback to the e-learning environment. This solution transforms large amounts of useless data into an intelligent monitoring and recommendation system applied to the learning process.

Application of data mining in anomaly detection: Anomaly detection in the network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic in to normal traffic or abnormal traffic (Deshpande and Thakare 2010). The classification is done by if any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

Application of data mining in retail: Through the use of store-branded credit cards and point-of-sale systems, retailers can keep detailed records of every shopping transaction. This enables them to better understand their various customer segments. According to Rygielski et al. (2002), some of the data mining application for retail includes:

- **Performing basket analysis**: Basket analysis also known as affinity analysis, reveals which items customers tends to purchase together. This analysis can improve stocking, store layout strategies, and promotions.
- **Sales forecasting:** Examining time-based patterns helps retailers make stocking decisions. It used to analyze if a customer purchases an item today, when are they most likely to purchase a complementary item.

- **Database marketing**: Retailers can develop profiles of customers with certain behaviors, for example, those who purchase designer labels clothing or those who attend sales. This information can be used to focus cost-effective promotions.
- **Merchandise planning and allocation**: When retailers add new stores, they can improve merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

Application of data mining in telecommunications: The telecommunications industry was an early adopter of data mining technology and therefore many data mining applications exist. Generally there are three data mining application areas in telecommunication industries these are fraud detection, marketing/customer profiling and network fault isolation.

Telecommunications companies around the world face increasing competition which is forcing them to aggressively market special pricing programs aimed at retaining existing customers and attracting new ones. Knowledge discovery in telecommunications mainly include the following (Rygielski et al. 2002):

- **Call detail record analysis**: Telecommunications companies accumulate detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.
- **Customer loyalty:** Some customers repeatedly switch providers, or churn, to take advantage of attractive incentives by competing companies. The companies can use data mining to identify the characteristics of customers who are likely to remain loyal once and those who switch their providers' frequently, thus enabling the companies to target their expenditure on customers who will produce the most profit.

Application of data mining in banking: According to Rygielski et al. (2002), banks can utilize knowledge discovery for various applications, including:

- **Card marketing**: By identifying customer segments, card issuers and acquirers can improve profitability with more effective acquisition and retention programs, targeted product development, and customized pricing.
- **Card holder pricing and profitability**: Card issuers can take advantage of data mining technology to price their products so as to maximize profit and minimize loss of customers. Includes risk-based pricing.
- **Fraud detection*:*** Fraud is enormously costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns.
- **Predictive life-cycle management:** Data mining helps banks predict each customer's lifetime value and to service each segment appropriately (for example, offering special deals and discounts).

2.7.1 Application of Data Mining for CRM

The application of data mining tools in CRM is an emerging trend in the global economy. Analyzing and understanding customer behaviors and characteristics is the foundation of the development of a competitive CRM strategy, so as to acquire and retain potential customers and maximize customer value. Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools for making different CRM decisions (Ngai et al. 2009). Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers (Two Crows Corporation 1999).

Customer identification: Customer identification which is also known as customer acquisition is the beginning of CRM. The main target of this phase is to identify people who are most likely to become customers or most profitable to the company. Elements for customer identification include target customer analysis and customer segmentation. Target customer analysis involves seeking the profitable segments of customers through analysis of customers' underlying characteristics, whereas customer segmentation involves the subdivision of an entire customer base into smaller customer groups or

segments, consisting of customers who are relatively similar within each specific segment.

Customer attraction: This is the phase following customer identification. After identifying the segments of potential customers, organizations can direct effort and resources into attracting the target customer segments. An element of customer attraction is direct marketing. Direct marketing is a promotion process which motivates customers to place orders through various channels.

Customer retention: This is the central concern for CRM. Customer satisfaction, which refers to the comparison of customers' expectations with his or her perception of being satisfied, is the essential condition for retaining customers. As such, elements of customer retention include one-to-one marketing, loyalty programs and complaints management.

One-to-one marketing refers to personalized marketing campaigns which are supported by analyzing, detecting and predicting changes in customer behaviors. Loyalty programs involve supporting activities which aim at maintaining a long term relationship with customers. Specifically churn analysis, credit scoring, service quality or satisfaction forms are part of loyalty programs.

Customer development: This involves consistent expansion of transaction intensity, transaction value and individual customer profitability. Elements of customer development include customer lifetime value analysis, up/cross selling and market basket analysis. Customer lifetime value analysis is defined as the prediction of the total net income a company can expect from a customer.

Application of data mining in customer segmentation: All industries can take advantage of data mining to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis. That means by incorporating behavioral components into traditional segmentation techniques, organizations can target and market customers far more effectively. Behavioral attributes which used to segment customers includes product usage, buying cycle, purchase history, online activity, responsiveness to marketing materials and others.

Substantial research works in many other data mining applications has been done abroad. In the context of Ethiopia, data mining applications for CRM has been done by Henok (2002) and Denekew (2003) in Ethiopian Airlines. They performed their study by using clustering and classification techniques; for the case of clustering they use K-means algorithm, on the other hand, for classification purpose they prefer decision tree algorithm. The other research in CRM was conducted by Fekadue (2004) and Melaku (2009) in Ethiopian Telecommunications Corporation; they also conducted their research by using clustering and classification data mining techniques. Still another research was done by Kumnegere (2006) in Ethiopian Shipping Lines. She has also used clustering and classification techniques in her study, at the same time she selected K-means and decision tree algorithms. Tilahun (2009) has also conducted a research on application of data mining in the area of Banking Industry by using the same techniques.

There are also research projects done abroad on revenue and customs authority CRM, such as Mahler and Hennessey (1996), Doye (2010) and Boulding et al. (2005).

CHAPTER THREE

CUSTOMER RELATIONSHIP MANAGEMENT AND CUSTOMER SEGMENTATION

3.1 Loyalty and Customer Relationship Management

3.1.1 Overview

Customer relationship management (CRM) has become one of the leading business strategies in the new millennium. Nowadays, companies are changing their business process models and build information technology (IT) solutions that enable the acquisition of new customers, retain existing ones, and maximize the customers' lifetime.

It is difficult to find out a totally approved definition of CRM. However, there are two approaches to define CRM, i.e. from the perspective of management and IT approaches (Bose 2002, Wahab and Ali 2010). In management approach, CRM refers to an integrated approach to identifying, acquiring, and retaining customers (Bull 2003, Bose 2002). On the other hand, CRM on IT approach refers to the tools or system design to support the relationship strategy activities such as identifying, acquiring, and retaining customer (Farn and Huang 2009).

CRM is an enterprise approach to understand and influence customer behavior through meaningful communications in order to improve customer acquisition, customer retention, customer loyalty, and customer profitability (Kim et al. 2003). By using customer satisfaction as a principle of CRM, the modern marketing promotes a customer orientation (Parvatiyar and Sheth 2001). As Trappey et al. (2009) explained, business enterprises understand the value of customers, target the most profitable customers, and build high-quality relationships that increase loyalty and profits by the help of CRM; because the most crucial elements for the success of CRM are targeting the most profitable customers and evaluating customer profitability. In addition, organizations are realizing the need for in-depth and integrated customer knowledge in order to build close

cooperative and partnering relationships with their customers (Parvatiyar and Sheth 2001).

According to Parvatiyar and Sheth (2001), CRM is the creation of superior value for both the company and the customer by the comprehensive strategy and process of acquiring, retaining, and partnering with selective customers. To achieve a better efficiencies and effectiveness in delivering customer value, CRM involves the integration of marketing, sales, customer service, and the supply-chain functions of the organization.

CRM is also viewed as the strategy of transforming enterprises to become customer centric while still expanding revenue and profit is one of the main strategies in business today (Kim et al. 2003). Many business enterprises, therefore, realize the importance of CRM and the potential of these techniques to achieve and sustain a competitive advantage. To realize CRM success, business and IT executives should implement CRM processes and technologies and promote employee behavior that supports coordinated and more effective customer interactions throughout all customer channels (Kim et al. 2003). Therefore, one of the most important processes of CRM is extracting valid, previously unknown, and comprehensible information from a large database and using it for profit.

CRM is illustrated as a combination of people, processes, and technology which provide understanding of customer needs, business strategy support, and build long-term relationships with customers (Shang and Chen n.d). As a result, companies make every effort to deliver the highest value to customers through better communication, customized promotions, faster delivery, and personalized products and services (Trappey et al. 2009). To effectively address human behavioral elements appropriate business processes and organizational culture are also required for successful utilization of the integrated technology (Shang and Chen n.d).

CRM is also described as a customer-focused business strategy that dynamically integrates sales, marketing and customer care service in order to create and add value for

the company and its customers (Chalmeta 2006). As Chalmeta (2006) explained, CRM systems basically make the following things possible:

- CRM enables to have an integrated, single view of customers by using analytical tools
- Managing customer relationships in a single way, regardless of the communication channel such as telephone, website, personal visit, and so on,
- Improving the effectiveness and efficiency of the processes involved in customer relationships.

Generally, nowadays companies are becoming more and more aware of the many potential benefits provided by CRM. According to Kim et al. (2003), some of the potential benefits of CRM are the following:

- Increased customer retention and loyalty
- Higher customer profitability
- Creation of value for the customer
- Customization of products and services
- Lower process, higher quality products and services

3.1.2 Customer Loyalty

The importance of creating and maintaining customer loyalty is not new. It is a common belief among businesses that it costs more to find a new customer than to keep and grow an existing one. However, recent studies indicate that in spite of heavy investments in customer satisfaction efforts and rewards programs, loyalty remains an intangible goal in almost every industry (Mc Kinsey & Company 2001).

According to Hwang et al. (2004), customer loyalty can be defined as the index that customers would like to remain as customers of a company. That is:

Customer Loyalty = 1 – Churn Rate.

As Hwang et al. (2004) explained, churn describes the number or percentage of regular customers who discard relationship with a service provider. So, by using customer loyalty

it is possible to measure customer retention. As Farn and Huang (2009) explained, there are two fundamental methods to retain customers. One is to improve customers' satisfaction and then motivate their true attitudinal and behavioral loyalty, so that customers do not be diverted by competitive offering. The other is to increase switching barriers to discourage customers from using other suppliers, such as switching cost and changing to alternatives as disincentives. To retain customers, understanding the value of customers and the most profitable customers are also essential. Level of customer retention can be derived from churn rate. It is significant for customer cultivation and retention to consider the churn rates (Hwang et al. 2004).

3.1.3 Principles and Tasks of CRM

3.1.3.1 Principles of CRM

According to Gray and Byun (2001), personalization, loyalty and lifetime value are the three main principles of CRM implementation. These principles have the benefit of improving the organization's ability to retain and acquire customers, maximize the lifetime value of each customer and improve services without increasing cost of service.

Personalization (treat customer individually): Personalization deals about treating customers individually so that the content and services to customer should be designed based on customer preferences and behavior (Adomavicius and Tuzhilin 2001).

Acquire and retain customer loyalty through personal relationship: Once personalization is takes place, then a company needs to maintain relationships with the customer. Continuous contacts with the customer intern can create customer loyalty.

Select good customer instead of bad customer based on lifetime value: This principle mainly focuses on finding and keeping the right customers who generate the most profits for the organization. According to this principle the best customers earn the most customer care and the worst customers should be dropped.

3.1.3.2 CRM Tasks

The database marketing approach is highly company centric; however, customers were not kept loyal by the discount programs and the one-time promotions that were used in the database-marketing programs (Gray and Byun 2001). So, there is a need to follow/use the CRM approach which is customer-centric; because it is better to gain customer loyalty in customer-centric approach than in database marketing approach (Gray and Byun 2001). Rather than based on what the company wants to sell, the customer centric approach of CRM focuses on the long-term relationship with the customers by providing the customer benefits and values from the customer's point of view. So, to achieve these basic goals CRM requires the following four basic tasks (Gray and Byun 2001, Kumnegere 2006). These are:

Customer identification: It refers to the selection and or knowing of customers through marketing channels, transactions, and interactions over time for the purpose of serving or providing value to the customer (Berndt et al. 2005).

Customer differentiation: This step refers to segmenting customers into different perspective from the company's point of view; because, from the company's point of view each customer has their own lifetime value.

Customer interaction: The main purpose of this step is to analyze the customer's behavior over a long period of time in order to provide the right goods and services at the right time because customer demands are sensitive and it change over time. From a CRM perspective, the customer's long-term profitability and relationship to the company is important. Therefore, the company needs to learn about the customer continually.

Customization / Personalization: The other goal of CRM is to treat each customer uniquely so as to increase customer loyalty. Through this personalization process, the company can also increase customer. Customization is carried out by the organization in order to ensure that customer needs are met. It requires that the organization adapts its product, service or communication in such a way that have something unique for each customer. Communication can be customized to address the specific needs and profile the

customer, and organization also makes use of personalization as part of this process. Products can be customized as to the specific desires that the customer has of the organization (Berndt et al. 2005).

3.2 Customer Segmentation

3.2.1 Overview

Segmentation is the process of developing meaningful customer groups that are similar based on individual explanations characteristics and behaviors (Trappey et al. 2009). According to Bounsaythip and Rinta-Runsala (2001), segmentation also viewed as a way to have more targeted communication with the customers; and the process of segmentation describes the characteristics of the customers groups (called segments or clusters) within the data.

Greengrove (2002) explained that there are two main segmentation approaches: the first type of segmentation is the process of segmenting the customers based on understanding the needs of the end user which is called needs-based segmentation. The second type of segmentation, characteristics-based segmentation, is the process of segmenting customers based on their characteristics, attitudes or behaviors.

Customer segmentation is defined as the practice of classifying customer base into distinct groups (Farn and Huang 2009). In other words, customer segmentation is also described as the process of dividing customers into homogeneous groups on the basis of shared or common attributes (Bounsaythip and Rinta-Runsala 2001). The goal of segmentation is to know the customer better and to apply that knowledge to increase profitability, reduce operational cost, and enhance customer service. Segmentation can provide a multidimensional view of the customer for better treatment strategy.

One of the challenging tasks during segmentation is to choose and derives appropriate segmentation bases and variables in which the segmentation process is performed and resulting segments are interpreted. These variables may be demographic, geographic, psychographic or behavioral variables (Bounsaythip and Rinta-Runsala 2001).

Several segments may be formed by using customer profitability. For instance, the most profitable segment consisting of the highest-profit customers should be retained through loyalty and retention program. Another possible segment is the most unprofitable customer group who generate more costs than profit (Kim et al. 2006).

3.2.2 Applications of Customer Segmentation

According to McGuirk (2007), customer segmentation has different applications for a certain organization which is described as follows.

Making customer investment decisions: Segmentation provides a framework to help identify the optimal customer investment strategy for each unique segment. For some segments, the investment may be directed towards further developing customer relationships, while for other segments the investment is made to introduce new products and services that address unsatisfactory customer needs. Ultimately, the key factor driving customer investment decisions has been the expected return on that investment. Segmentation not only helps to determine how much to invest in a customer segment, but how to spend it.

Managing customer relationships: Segmentation also provides an excellent framework to manage the varied needs of customers. Customized customer management and development strategies can be developed for each unique segment. The development plans should include a set of objectives, goals and performance metrics that are derived from the unique opportunities and challenges present within each customer group. The segment-level plans function as a strategic roadmap, supporting business growth and attempting to maximize the potential of each customer relationship.

Tailoring marketing programs: It is true that successful data-driven marketers understand how to communicate with their customers at the right time, right place and with the right message. Distinctive customer preferences and needs represent unique opportunities and challenges that can be pursued by introducing tailored or modified programs for each segment. The make-up of each customer segment and their past and projected behaviors and needs should guide the tailored use of key marketing levers to

maximize program effectiveness. Segmentation becomes the focus, supporting program development and ongoing test and learns activities.

Guiding product development and research: A comprehensive segmentation solution has been emphasized the fact that individuals have different product needs and usage patterns. Customers in one segment may use a company's full portfolio of products quite frequently, while customers in another segment may only have a need for a single product that is used at irregular intervals (infrequently). McGuirk (2007) further explained, segments that contain less active customers often exposes opportunities to strengthen and broaden these customer relationships by introducing new or re-packaged products that meet a specific customer need. Segmentation provides the means to target research and product development activities with the goal of further stimulating customer demand.

3.2.3 Difficulties in Making Good Segmentation

As Bounsaythip and Rinta-Runsala (2001) explained, there are different factors which make the segmentation process more complex such as:

Relevance and quality of data: Only relevance and good quality of data is needed because if the company has inadequate or too much customer data, this can lead to complex and time-consuming analysis. If the organizations data are also poorly organized (different formats, different source systems) then it is also difficult to extract relevance information. Furthermore, if the company's customer data is insufficient, too much or poorly organized, the resulting segmentation can be very complicated for the organization to implement it effectively.

Intuition: Although data can be highly informative, marketers need to be continuously developing segmentation hypotheses in order to identify the right data for analysis.

Continuous process: Segmentation needs continuous development and updating as new customer data is acquired. In addition, effective segmentation strategies have been influenced the behavior of the customers affected by them; thereby necessitating revision and reclassification of customers.

Over-segmentation: Segmentation can become too small and/or insufficiently distinct to justify treatment as separate segments.

3.3 CRM in Ethiopian Revenue and Customs Authority

3.3.1 Overview

The Ethiopian Revenue and Customs Authority (ERCA) came into existence on 14 July 2008, by the merger of the Ministry of Revenue, Ethiopian Customs Authority and The Federal Inland Revenue Authority, formerly responsible to raise revenue for the Federal Government and to prevent contraband. According to Article 3 of the proclamation No .587/2008 of the *Ethiopia Federal Negarit Gazeta*, the ERCA is looked upon as "an autonomous federal agency having its own legal personality". The reasons for the merge of the previous separately administrations into a single autonomous authority are, to provide the basis for modern tax and customs administrations; to reduce or avoid unnecessary and redundant procedures which causes delay and are considered cost-inefficient; to become more effective and efficient in keeping and utilizing information, promoting law and order, resource utilization and service delivery; and to transform the efficiency of the revenue sector to a high level (ERCA official website).

According to the experts, the newly designed tax and customs procedures have brought about a positive result and it has produced beneficial effects to both the Authority and its customers or stakeholders. However, to further improve the satisfaction of customers and the beneficial effects of the Authority, there is a need to implement modern CRM in the organization.

Objective of the ERCA

The ERCA has the following objectives:

- To establish or set up modern revenue assessment and collection system; and provide customers with fair, efficient and quality service
- To cause taxpayers voluntarily discharge their tax obligations

- To enforce tax and customs laws by preventing and controlling contraband as well as tax fraud and evasion
- To collect timely and effectively tax revenues generated by the economy
- To provide the necessary support to regions with a view to harmonizing federal and regional tax administration systems.

Vision

The ERCA's vision is to see "fair and modern taxes and customs administration system that enhances proper and effective revenue collection".

Mission

"The ERCA shall promote the voluntary compliance of taxpayers, ensure integrity and develop the skill of the employees, support the modernization and harmonization of the taxes and customs administration system, contribute to economic development and social welfare through effective revenue collection".

3.3.2 Powers and Duties of the ERCA

According to the proclamation No. 587/2008 of the *Ethiopia Federal Negarit Gazeta* the ERCA has the powers and duties to establish and implement modern revenue assessment and collection system; the ERCA has also the powers and duties of provide efficient, equitable and quality service within the sector; properly enforce incentives of tax exemptions given to investors and ensure that such incentives are used for the intended purposes.

In addition, ERCA has the responsibility to collect and analyze information necessary for the control of import and export goods and the assessment and determination of taxes; compile statistical data on criminal offences relating to the sector, and disseminate the information to others respected sectors. Furthermore, the Authority provide information and appropriate support to the Federal Police in the control of illegal trafficking of goods and struggle contraband; and cause appropriate measures be taken in accordance with the

law. The powers of deciding the place where import and export goods are to be deposited and establish are also given to the ERCA.

3.3.3 Organization of the ERCA

The head office of ERCA is found in Addis Ababa. The ERCA is led by a Director General with the assistance of four Deputy Director Generals, namely Deputy Director General for Corporate Functions Sector, Operations Sector, Change Management and Support Sector, and Enforcement Sector. The appointment of both the Director General and the Deputies are made by the Prime Minister of the country. At the headquarters, the Authority has 19 business processes and 2 directorates namely Women's Affairs Directorate and Ethics Directorate. According to the proclamation No. 587/2008 of the *Ethiopia Federal Negarit Gazeta*, "business processes" is defined as "a class or some class of work flow in which a serious and succeeding group of activities are performed step by step by a case worker or a case team in a single cycle of performance to bring about a result from the external or internal inputs and information utilized in the general course of performance of the Authority". The Authority's directorates and business process are led by a Director. Among the directorates, two (Audit and Inspection Business Process and Public Relations and Image Building Business Process) are accountable to the Director General while the remaining are accountable to the four Deputy Director Generals.

Apart from the 19 business processes and the 2 directorates at the headquarters level, the Authority has 17 branch offices which can be divided as follows: 15 branch offices located in a regional state or city administrations and 2 coordination offices located outside of Ethiopia at the port of Djibouti and at the port of Burbera, Somalia. Each branch office is directed by a manager who is accountable to the Director General of the Authority.

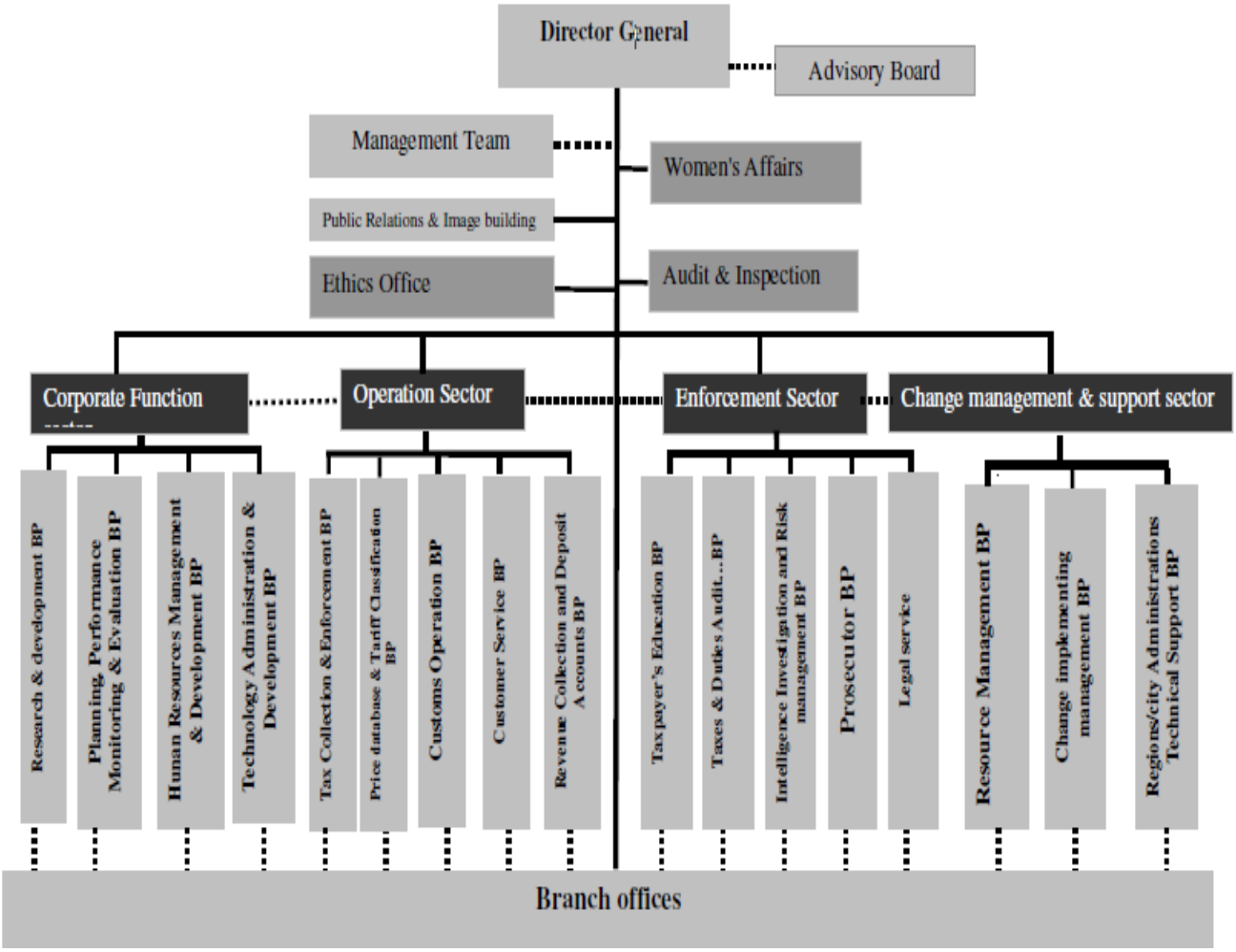

Figure 3.1 Organizational charts of the Ethiopian Revenue and Customs Authority

3.3.4 Complaints Handling in ERCA

Customer complaints are firstly a private communication between a company and a customer who communicates his/her dissatisfaction on a defined matter. Customer complaint is defined as the expression of a customer's dissatisfaction through various possible channels such as letter, email, phone call, physical claim and others (ERGEG 2010). Complaints handling and managing it accordingly, i.e dealing with customers after a service failure should be the corner stone of an organization's customer-satisfaction strategy (Johnston and Mehra 2002). Customer complaint handling is a key instrument of identifying the problem in the route of service providing and appropriate response to complaints. The most important tasks in the customer services are designing and

implementing of customer's compliant handling system and identify problems that may have been encountered and propose solutions to those problems. It is agreed that customers make complaints now and then should be considered as loyal to the organization rather than the sources of chaos (ERGEG 2010). Hence an organization should pay due attention of these parties and implement systematic compliant handling mechanism.

According to Johnston and Mehra (2002), the objectives of effective compliant handling are:

- To meet customer's satisfaction
- To build trust and confidence
- To avoid unfavorable publicity
- To provide fair, transparent and timely service
- To ease burden in the day-to-day work of legal system and
- To inform management

There are different techniques of customer complaints handling systems such as, handling complaints face-to face, handling complaints on the telephone and handling complaints in writing. When contacts in customers are difficult let customers make contact by telephone. During telephone customer complaints handling, make easy for customers to find the number; established telephone procedure to ensure all calls are dealt with punctually and train staff to be positive and friendly (ERGEG 2010).

In order to improve customer satisfaction, the ERCA tries to established customer compliant handling systems. However, to make the current strategic thinking more transparent and practical ERCA should implement modern CRM techniques.

CHAPTER FOUR

EXPERIMENTATION

4.1 Overview

This chapter is the main part of the study, which used to successfully meet the objectives of the research. To effectively achieve these research objectives, it is necessary to follow appropriate research methodology (model). This research follows all the fundamental steps of the Cios et al. (2000) KDD process model. Figure 4.1 indicates that Cios et al. (2000) model consists of a cycle that comprises six stages: understanding of the problem domain, understanding of data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge. The main reason the researcher prefers the Cios et al. (2000) model to other data mining model is that the Cios et al. (2000) model draws from both academic and industrial models and emphasizes iterative aspects; the Cios et al. (2000) model has also an advantage to identify and describe several explicit feedback loops.

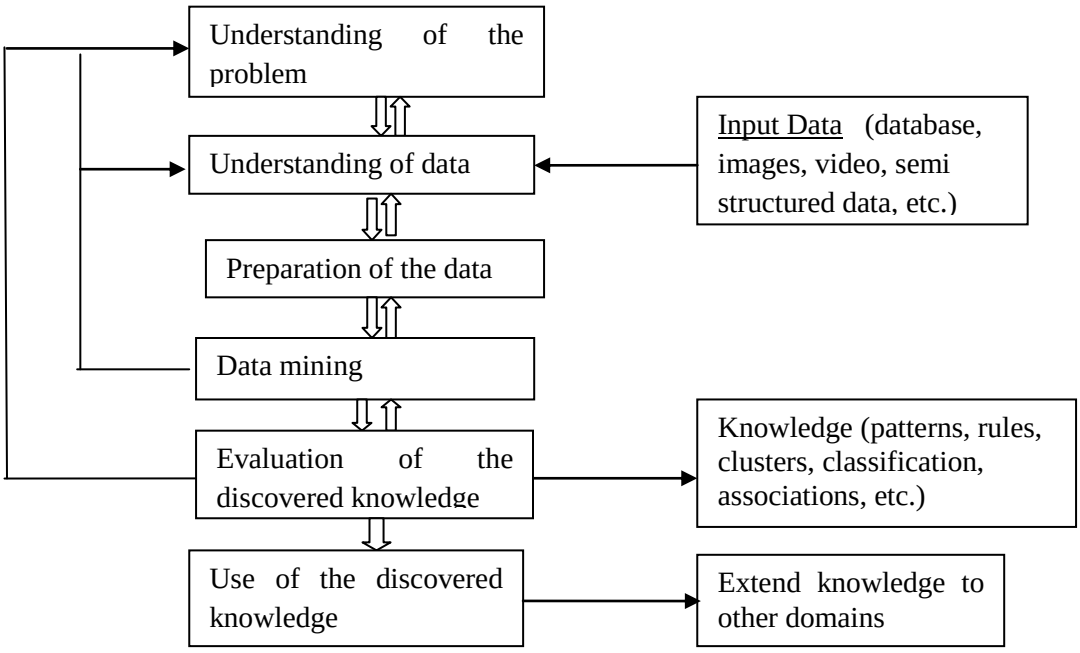

Figure 4.1 The six-step Cios et al. (2000) KDD process model

4.2 Understanding the Problem Domain

As it is explained in Section 3.2.2, which mainly deals about application of customer segmentation, there is a need to implement effective customer segmentation in the Authority, in order to increase the profitability of the Authority and to raise the satisfaction of the customers. As stated in Section 1.5, the researcher used different techniques to understand the problem domain of this investigation.

Based on the experts explanation and evaluation made, there are different kinds of customers in the Authority. Some customers are high value customers; these types of customers are those customers who generate substantial revenue for the organization by importing smaller amount of items and spending low invoice. On the other hand, there are also customers who generate lower revenue for the Authority. These types of customers are called low value customers. Low value customers import large amount of items and spend higher invoice but they generate smaller amount of revenue for the Authority; the type of goods which are imported by these low value customers were cheap. The intermediate between high and low value customers are medium value customers. Thus, the main aim of this research is to differentiate customers whether they are high, medium and low value customers based on the revenue they generate and the transaction of the data.

4.2.1 Data Mining Goals

The first data mining goal for this study was segmenting or clustering customers into different groups based on their characteristics to treat or handle them accordingly. To effectively accomplish this task, identifying important variables from the total data collected was necessary. The identified variables were in turn used to build a model by using clustering techniques. In addition, the identified variables were also used to better understand the customers. The data preparation and data analysis phases help the identification of important attributes that could serve as an input for the model building.

The other data mining goal for this research was building a classification model which used to assign new records into the identified clusters. The classification model was done

by using the most appropriate data mining classification techniques which is decision tree and neural network classification techniques and the one which performs better accuracy is selected.

4.2.2 Data mining Tool Selection

Finding and selecting the most appropriate data mining tool with better capability is one of the challenging tasks in data mining process. For this study, selection of an appropriate data mining tool was done by first listing certain criteria. Among the criteria used for the selection, the most important ones were:

- The speed and quality of the tool (performance)
- The data mining tasks that the tool is intended for (clustering, classification)
- The algorithm supported by the tool (K-means, decision tree, neural network)
- The number of records the tool can handle
- User friendliness
- Availability of the tool on the internet (freely downloaded)

By considering the above criteria, the researcher found different tools in the Internet, such as Weka Version 3.6.4, Tanagra 1.4 and Sipina research Version 3.6. After though evaluation of the tool as per the criteria the researcher selected Weka Version 3.6.4 and more or less Weka fulfilled the above criteria. Weka Version 3.6.4 is better than the other two in that it had better performance than both Tanagra and Sipina; in addition, Weka has more facility for preprocessing of the data and number of algorithms than the remaining ones.

4.3 Understanding the Data

Having defined the data mining goal, or after understanding of the problem to be addressed and selecting an appropriate tool, the next step was to analyze and understand the available customers' data. Therefore, to fulfill the data requirement, data was initially collected regarding customers' behavior from ERCA's ASYCUDA database. Owning to this, analysis of the data and its structure was done together with the Database

Administrator and other domain experts by evaluating the relationship of the data with the problem at hand and the particular data mining task were dealt with the experts in the Authority. In Cios et al. (2000) methodology during the data understanding step there are subtasks which are considered by the users of the model. These are collection of initial data, description of data and verification of data quality.

4.3.1 Collection of Initial Data

The primary step of data understanding is identifying the initial source of data that is used for the data mining process. As indicated earlier, the major and initial data source for this research was the ERCA's ASYCUDA database. The ERCA's ASYCUDA database is very huge and rich data source because every declaration of import/export items and other detail information of the customers are found in the ERCA's ASYCUDA database. For instance, the customer/company name, the item type, total revenue paid, the price of the item, total number of items imported, total amount of invoice spent to import the items etc. of every customer are found in this database. The researcher collected all the two year declaration of these customers' data which was around 110740 records. The responsibility of administering or controlling the ERCA's ASYCUDA database is given to the IT department of the Authority.

4.3.2 Description of the Data Collected

The researcher took a two-year transaction data for this study. Since every transaction of goods within the country is registered in the ASYCUDA database, the amount of data found is very huge but the researcher took only selected data. To select the data, the researcher used the Weka preprocessing facilities which is stratified remove folds by setting number of folds with two; and only around 46,748 records were taken with around 8 detail selected attributes. However, in the ERCA's ASYCUDA database there are a lot of tables and many attributes. A sample data which is found from the initial data collection is attached into this research as **Appendix 1**.

Among the different tables of the ASYCUDA database the researcher together with the domain experts, selected the customer related tables and used the following once.

SAD_GEN TABLE (SAD General Segment Table): The SAD_GEN table contains all information related to the SAD general segments. SAD means an abbreviation for single administration document.

Attribute	Data Type	Description
SAD_REG_DATE	Date	Date that the SAD was register. (registration date)
SAD_ITEM_TOTAL	Small integer	Total number of items imported
SAD_PACK_TOTAL	Char	Total number of packet
SAD_CONSIGNEE	Char	Company code
SAD_TOT_INVOICED	Decimal	Total amount of invoiced for entire declaration
SAD_TYP_DEC	Char	Type of declaration (import or export)
SAD_CTY_IDLP	Char	If the SAD is an import SAD this field indicates the country of last consignment. If the SAD is for an export declaration, then this is the field showing the country of first destination.
SAD_CUR_COD	Char	Currency code indicates that used on the invoice.
SAD_MOT_BORD	Char	Mode of transport at border.
SAD_WHS_COD	Char	If the goods are being warehoused (for re-export, later home use etc) the warehouse code is noted here.

Table 4.1 Attributes and description of the SAD general segment table

SAD_ITM TABLE: - The SAD_ITM TABLE contains all data about the declaration of items.

Attribute	Data Type	Description
SADITM_ITM_PRICE	Decimal	Declaration per item price
SADITM_GOODS_DESC	Char	Description of the goods
SADITM_PACK_KNDCOD	Char	The code used to identify the kind of package
SADITM_CTY_ORIGCOD	Char	Country of origin
SADITM_HS_COD	Char	Goods / commodity code
SADITM_GROSS_MASS	Decimal	Total weight of the imported item
SADITM_SUPP_UNITS	Decimal	Supplementary units
SADITM_PREV_DOC	Char	Previous document reference (for temporary import for instance)
SADITM_ITM_TOTAMT	Char	Total amount of duties and taxes for the item
SADITM_NUM	Integer	Declaration version (if the declaration is amend)

Table 4.2 Attributes and description of the SAD_ITEM table

UNCMPTAB (Companies Table): - This table records all companies'/customers' related data. As this table is not historized, ASYCUDA doesn't record all modifications done on a company code but only keep the last updated value

Attributes	Data type	Description
CMP_COD	Char	Company code
CMP_NAM	Char	Company name
CMP_ADR	Char	Company address
CMP_TEL	Char	Telephone number
CMP_FAX	Char	Fax number
CMP_TLX	Char	Telex number
LST_OPE	Char	Last operation
SER_STA	Integer	Server status number

Table 4.3 Attributes and description of the company table

CSH_GEN_CDI (Cash Declaration Payment In Table)

This table contains information such as declarant reference and number of declaration paid regarding cash payment in declarations. This is the master table for the multiple rows to be inserted in CSH_DEC (declaration), in CSH_TRA (other payment), in CSH_MEA (means of payment) and in CSH_OCC (if occasional exporter / importer).

Attributes	Data type	Description
RCP_YEAR	Char	Year of receipt
RCP_CUO	Char	Customs office code
RCP_SERIAL	Char	Number of the serial in which the receipt is issued
RCP_NBER	Integer	Number of the receipt within the previous serial
CDI_DEC_STM	Integer	Cash declaration statement
CDI_DECLARANT	Char	Code of SAD declarant
CDI_COMPANY	Char	Code of SAD consignee or exporter
CDI_DECL_TOT	SmallInt	Number of declaration on this receipt
CDI_AMOUNT_TOTAL	Decimal	Total amount paid on this receipt
CDI_RCPT_TIME	Char	Time when the receipt was issued
CDI_TOTAL_DECL	Decimal	Total amount paid for declarations on this receipt

Table 4.4 Attributes and description of the cash declaration payment in table

Unctytab (country table): - This table contains information's about the country where the item is imported.

Attributes	Data type	Descriptions
CTY_COD	Char	City code (ET)
CTY_DSC	Char	City description(Ethiopia)
LST_OPE	Char	City status (deleted or active)

Table 4.5 Attributes and description of the country table

4.3.3 Data Quality Verification

Generally, the researcher was satisfied with the reliability of data and completeness of the records. However, the collected data contains missing, incomplete and irrelevant data; for instance, item price and total revenue columns contained missing or incomplete values. Hence, revenue and item price column details which were missing or contain zero are either handled the missing value or discarded them from the dataset.

4.4 Preparation of the Data

The main purpose of the data preparation phase is to make the data more suitable for the next step which is the modeling (data mining in Cios et al. (2000) model). Starting from the initial data collection there were a number of transformations performed until the final dataset was found which is used for model building. In KDD process model, there are different methods for data preprocessing such as data cleaning, data construction, data integration and data formatting.

4.4.1 Data Cleaning

During the data cleaning phase, the researcher tried to handle missing values of the attribute and remove some incomplete data which does not affect the entire dataset. From the total dataset, only two attributes have missing values; these are ITEM_PRICE and TOTAL_REVENUE. The researcher replaced the missing ITEM_PRICE attribute value with the Weka mean/mode missing value replacement preprocess facility. When the total revenue is missed or became zero the researcher learns from the Database Administrator that these types of items are imported freely (without any taxation), as a result the researcher discarded theses instances from the dataset; because, these types of instances were not important for the data mining step. The researcher also discretizes one attribute called SADITM_CTY_ORIGCOD; this attribute has around 80 values which is the name of the country where the item is imported or originated. However, the researcher discretizes this value into 6 which is in their continent of origin. MS-Excel is selected for cleaning the data because the researcher is more familiar with it.

4.4.2 Data Integration and Transformation

Data integration method for retrieving important fields from different tables and data transformation method, which is attributes construction, are done in the effort to prepare the data ready for the data mining techniques to be undertaken in this research.

After selecting different fields from different tables together with the domain experts, the integration part which is bringing different attributes from different tables into a single dataset was performed by together with the Database Administrator by using SQL commands in the original ASCUDA database. The attributes were retrieved from the above tables which were stated in the description of the data collected parts.

Most of the attributes which were used for clustering and classification model building have been derived from the SAD_GEN and SAD_ITEM tables; because, these tables contains the main attributes, which describe the characteristics of the items which transfers around the border. Accordingly, many of the attributes such as registration date, total number of items, total amount of invoice, item price, descriptions of the goods, and total number of the pack, have been derived from these tables. In addition, the researcher has also derived additional attributes which help the modeling process more efficient. According to Saarenvirta (1998), data creation involves the creation of new variables by combining existing variables to form ratios, difference and so forth. REVENUE_ITEM, REVENUE_PRICE and INVOICED_REVENNUE are attributes which were derived from the other original attributes. Accordingly, the derived attributes are then defined using the existing attributes as follows:

$$\text{REVENUE\_ITEM} = \frac{\text{TOTAL\_REVENUE}}{\text{TOTAL\_ITEM}}$$

$$\text{INVOICED\_REVENUE} = \frac{\text{TOTAL\_INVOICED}}{\text{TOTAL\_REVENUE}}$$

$$\text{REVENUE\_PRICE} = \frac{\text{TOTAL\_REVENUE}}{\text{ITEM\_PRICE}}$$

4.4.3 Data Formatting

At this step the researcher changes the data into a format which was suitable for the data mining tool or algorithms. Since the preprocessing of the data was performed in MS-Excel the final dataset was also in MS-Excel format; however, the selected tool Weka 3.6.4 doesn't accept the data in Excel format. So, the researcher first tried to convert the data in comma delimited (CSV) text file in ARFF (Attribute Relation File Format) format. Comma delimited applied for a list of records where the items are separated by commas, whereas ARFF is an extension of a file format that the Weka software can read. In addition, outliers and extreme values which may mislead the K-means algorithm were also totally discarded from the dataset.

4.4.4 Attribute Selection

After the preparation of the final dataset the next step was to identify the best attributes which is used for cluster modeling. To distinguish attributes which have high information content, the researcher input all the selected attributes in the selected data mining tool which is Weka 3.6.4. The following table shows attributes with their description after the final preprocessed dataset. The dataset containing these attributes is fed into the data mining software.

Attribute	Data Type	Description
CMP_COD	Char	Company code
CMP_ADR	Char	Company address
CMP_NAM	Char	Company name
SAD_REG_DATE	Date	Date that the SAD was register. (registration date)
SAD_ITEM_TOTAL	Integer	Total number of items imported
SAD_TOT_INVOICED	Decimal	Total amount of invoiced for entire declaration
SADITM_ITM_PRICE	Decimal	Declaration per item price
SADITM_GOODS_DESC	Char	Description of the goods
SADITM_PACK_KNDCOD	Char	The code used to identify the kind of package
SADITM_CTY_ORIGCOD	Char	Country of origin
CDI_RCPT_DATE	Date	Date that the declaration was paid and the original receipt issued.
SAD_PACK_TOTAL	Char	Total number of packet
SAD_CONSIGNEE	Char	Company code
SADITM_ITM_TOTAMT	Char	Total amount of duties and taxes for the item
RCP_YEAR	Char	Year of receipt
REVENUE_ITEM	Decimal	Ratio of revenue to total number of items
REVENUE_PRICE	Decimal	Ratio of revenue to item price
INVOICE_REVENUE	Decimal	Ratio of invoice to revenue

Table 4.6 Selected attributes with their description

Once the attributes were fed in the data mining tool, the researcher evaluated the information content of the attributes using the select attribute technique with GainRationAttributeEval attribute evaluator and Ranker search method. So, the tool arranged the attributes in order of their gain ratio; hence, the researcher together with the domain experts removed those attributes which had less gain ratio.

In addition to evaluating the gain ratio of the attributes, Weka 3.6.4 explorer also enables to know the minimum, maximum, mean, standard deviation, data type, number of missing values, number of distinct values, and number of unique values of each currently selected attribute. It is possible to see these values for all other attributes by selecting

each of them one-by-one. These values in turn enable the researcher together with the domain experts to determine the threshold values of each variable for the analysis of the result.

4.5 Data Mining

There are different tasks performed during the data mining phase the major once are selection of modeling techniques and building a model.

4.5.1 Selection of Modeling Techniques

The main objective of this research being to effectively segment the customers of ERCA into various groups in which there is less similar between segments /clusters however whose members being very similar. To successful meet this research objectives there was a need of selecting appropriate modeling techniques.

Customer segmentation of this research is performed by using clustering and classification data mining techniques. These techniques are selected because segmentation problems are most of the time performed by using these techniques. Furthermore, the techniques are also well implemented in the selected data mining tool.

For the case of clustering the selected tool supports various algorithms, such as simple K-means, Cobweb, DBScan, Expectation Maximization (EM), FarthestFirst, FilteredClusterer, HierarchicalClusterer, MakeDensityBasedClusterer and OPTICS. From all these clusters, the researcher selected simple K-means algorithm; simple K-means algorithm is selected as it is better in handling discrete and numeric attributes. Moreover, K-means method is very popular because of its ability to cluster huge data, and outliers, quickly and efficiently (Kanungo et al. 2002). When using simple K-means algorithm there are different activities which are performed by the users of the algorithm, such as determining the value of K which is the number of segments/clusters. In order to decide the value of K there is a need to consult the domain experts because the number of clusters chosen should be driven by how many clusters the business can manage. The

experts suggested that the K value to be in the range of 3-6, this four cluster models were evaluated against their performance of creating dissimilar segments.

For classification, the most common classification techniques, decision tree and neural network classifier, are tested. Regarding decision tree, the selected tool holds various decision tree algorithms, such as BFTree, J48, ID3, LADTree, NBTree, RandomForest, RandomTreee and so on. From these options J48 decision tree algorithm has been selected because it handles a large amount of variables with either a continuous or discrete variable. Moreover, The J48 algorithm gives several options related to tree pruning (Han and Kamber 2006). As described in Section 2.6.2 tree pruning produces fewer, more easily interpreted results and it used as a tool to correct for potential over-fitting.

J48 decision tree algorithm employs two pruning methods. The first method of pruning is known as sub-tree replacement and the second type of pruning used in J48 is termed sub-tree raising (Han and Kamber 2006, Witten and Frank 2005).

- Sub-tree replacement: - in this method of pruning, nodes in a decision tree are replaced with a leaf. This process starts from the leaves of the fully formed tree, and works backwards toward the root.
- Sub-tree raising: - In this type of pruning, a node may be moved upwards towards the root of the tree, replacing other nodes along the way. This type of pruning is computationally complex

4.5.2 Test Design

Before passing to the actual model building, it is better to set a guide for the training and testing process. For the case of clustering the total instance of the dataset is used to train the clustering model. On the other hand, for classification model the researcher used both 10-Fold Cross-validation and percentage split. In 10-fold cross-validation, the initial data are randomly partitioned into 10 mutually exclusive subsets or “folds,” 1, 2, 3, …, 10, each approximately equal size. Training and testing is performed 10 times. In the first iteration, the first fold is reserved as a test set, and the remaining 9 folds are collectively

used to train the classifier; the classifier of the second iteration is trained on folds 1, 3, 4, …, 10 and tested on the second fold; and so on (Han and Kamber 2006). For the case of percentage split 70% was used for train the model, whereas 30% of the dataset was used for test data.

The following variables are used for model building, while the rest attributes are not used because they are less important for the model building.

- Total number of items imported (SAD_ITEM_TOTAL)
- Total amount of invoice spent to import the item (SAD_TOTAL_INVOICED)
- Total amount of revenue generated (TOTAL_REVENUE)
- The price of the item (ITEM_PRICE)
- Origin (continent) of the item where it is imported (CTY_DSC)
- Ratio of total revenue to total number of items (REVENUE_ITEM)
- Ratio of total revenue to item price (REVENUE_PRICE)
- Ratio of total invoiced to total amount of revenue generated (INVOICED_REVENUE)

There is one attribute called risk color, it is important for model building but it doesn't included during the attribute selection, because this attribute has three values (green, yellow, and red) out of the total data more than 90% has red value and only 0.9% of the total data has green value.

4.5.3 Model Building

The model building data mining phase mainly consists of the cluster modeling and classification modeling subsections.

4.5.3.1 Cluster Modeling

Once the final dataset has been selected the next step was to build the clustering models by using the selected data mining tools, Weka 3.6.4 Version. In clustering of data using simple K-means there is a need to set the number of clusters or the value of K, the number of data tuples the cluster must start with (seed size) and also there is a need to

specify the variables to be used for building the cluster. In this cluster modeling the entire population of the dataset (100%) is used to train the clustering model.

Before starting the experimentation it is important to set the threshold values of each numeric attributes of the dataset, because the threshold values are used to interpret the output of the clustered model. The threshold value has been determined with the domain experts and with the aid of the Weka's minimum, maximum, and mean values display for each attribute. So the threshold value for each attribute has been as follow.

TOTAL_ITEM Very High High Average Low Very Low

TOTAL_INVOICE Very High High Average Low Very Low

ITEM_PRICE Very High High Average Low Very Low

TOTAL_REVENUE Very High High Average Low Very Low

REVENUE_ITEM Very High High Average Low Very Low

INVOICE_REVENUE Very High High Average Low Very Low

REVENUE_PRICE Very High High Avarage Low Very Low

List of Attributes	**Threshold Values**				
	Very High	High	Average	Low	Very Low
1. TOTAL_ITEM (TIT)	>50	26-50	15-25.9	10-14.9	<10
2. TOTAL_INVOICE (TIV)	>200,000	151,000-200,000	110,000-150,999	45,000-109,999	<45,000
3. TOTAL_REVENUE (TRN)	>1,000,000	750,000-1,000,000	500,000-749,999	300,000-499,999	<300,000
4. REVENUE_ITEM (RIT)	>60,000	39,000-60,000	15,000-38,999	10,000-14,999	<10,000
5. INVOICE_REVENUE (IVR)	>5	4-5	2-3.9	1-1.9	<1
6. ITEM_PRICE (IPR)	>50,000	25,000-50,000	15,000-24,999	2,000-14,999	<2,000
7. REVENUE_PRICE (RPR)	>12,000	10,000-12,000	5,000-9,999	2,000-4,999	<2,000

Table 4.7 List of range of conditions by which a cluster result is assessed

4.5.3.1.1 Experiment 1

The first experiment is performed by adjusting the number of cluster or the value of K= 3 and using the seed size100. During this experiment all of the selected variables were input into the Weka cluster run. In clustering all the variables were set as independent variables because clustering is unsupervised data mining techniques. The output of the first experiment is summarized in the following table

Cluster distributions		
Cluster 1	Cluster 2	Cluster 3
6544(14%)	38992(83%)	1212(3%)

Table 4.8 Cluster distributions with number of iterations 4 and sum of squared errors 5199.169

During the default seed size (seed=10) the total number of instances were not fairly distributed; which means the sum of the whole cluster was 99%, but when the seed number was set to 100 the distribution of instances were significantly improved. In addition, the number of iterations and sum of squared errors were also minimized. When the seed size is 10, the number of iterations and sum of squared errors were 22 and 9409.92, respectively. However, if the seed size is set to 100 the number of iterations was 4 and sum of squared errors was 5199.169.

As describe in Table 4.8, majority of the customers were grouped in the second cluster. From the total (46748) customers 38992 (83%) of them are grouped in this cluster. The first cluster contains the second largest number of customers that is it consists of 6544 (14%) number of customers and the third cluster contains the list number of customers which is only 1212(3%) number of customers.

In an attempt to improved the data distribution of the segments, different seed values have been tried. But, the seed value at 100 gives a better data distribution in the segments. The optimal seed value has been obtained after a number of experimentations.

The following abbreviations are used in the cluster modeling interpretation

Attributes of the dataset with their description		**Abbreviated terms used in the cluster interpretation**	
Abbreviated attributes	Description	Abbreviated words	Description
TIT	Total number of items imported	VH	Very High
TIV	Total amount of invoice spent	H	High
TRN	Total amount of revenue generated	A	Average
RIT	Ratio of revenue per total number of items	L	Low
IVR	Ratio of invoiced per revenue generated	VL	Very Low
RPR	Ratio of revenue generated per item price	AS	Asia
IPR	The price of the item	EU	Europe
		AM	America
		LA	Latin America
		AF	Africa
		AU	Australia

Table 4.9 List of abbreviated words and attributes of the dataset along with their description

Segmentation (K=3, Seed=100)									
Cluster No.	**Distribution of instances (in %)**	**TIT**	**TIV**	**TRN**	**RIT**	**IVR**	**RPR**	**IPR**	**City**
1	6544 (14%)	13.33	74041.98	639766.33	69301.05	1.81	5025.53	20555.43	EU
2	38992 (83%)	12.58	105916.33	879519.78	35285.70	2.53	6414.65	32368.00	AS
3	1212 (3%)	15.83	152172.10	555970.81	33285.87	7.02	5414.98	27767.38	AM
Total	**46748 (100%)**								
Cluster No.	**Distribution of instances (in %)**								
1	6544 (14%)	L	L	A	VH	L	A	A	
2	38992 (83%)	L	L	H	A	A	A	H	
3	1212 (3%)	A	H	A	A	VH	A	H	
Total	**46748 (100%)**								

Table 4.10 Clustering result of the first experiment

Table 4.10 shows the mapping of the attributes average value of each segment with the corresponding discrete value. This mapping is used to compare the segments which are generated from the clustering algorithm.

Once the average value of each attribute in each cluster has been replaced with the corresponding discrete value, the next step is to describe each cluster according to the above table.

Cluster 1:- Cluster One of the first experiment contains customers who import small amount of total number of items, and customers who spend low amount of invoice. Customers that grouped in this cluster has generated average amount of revenue; and the ratio of total revenue per total number of items was very high. However, the ratio of total invoice with total revenue generated was low; on the other hand, the ratio of revenue generated with the item price and the price of the item which is imported by customers

that grouped in this cluster has average value. Items of this cluster have been imported from Europe countries.

Cluster 2:- The second cluster also holds those customers who import lower amount of total number of items and those customers who spend low amount of invoice. The cluster consists of high amount of revenue generated customers. The ratio of total revenue per total amount of items, the ratio of total invoice per total revenue generated, and the ratio of total revenue generated to the item price were average. The customers that grouped in this cluster have been imported high price item. This cluster customer has been imported their items from Asian countries.

Cluster 3:- The last cluster of the first experiment embraces customers that import average number of items and those spend high amount of invoice. The revenue generated by the customers of this cluster was average. The ratio of total revenue generated to total number of items and the ratio of total revenue generated to the item price were average. On the other hand, the ratio of total invoice to total revenue generated was very high. The price of the item which is grouped in this cluster was high. The cluster holds items which were imported from North America.

As stated in the business understandings of the Authority high value customers were those customers who generate high amount of revenue by importing small number of items and spending lower amount of invoice. On the other hand, low value customers were that customers who generate lower amount of revenue by importing large amount of items and spending higher amount of invoice.

By considering the above facts of the business, each cluster has been assigned in the following ranking order.

Cluster No.	**Possible Rank**
Cluster 1	2
Cluster 2	1
Cluster 3	3

Table 4.11 Cluster rank for the first experiment

As it is clearly indicated in Table 4.11**,** the second cluster is ranked first. This is due to the fact that customers of this cluster import smaller amount of items, spend lower invoice and generate high amount of revenue for the Authority. In addition, ratio of total revenue to total number of items, ratio of invoice to total revenue generated and ratio of revenue generated to item price were medium. For this segment the price of the item was high.

The first cluster is ranked second; customers in this segment generate average amount of revenue for the Authority by importing small number of items. In addition, the invoices spend by this segment customer were also low. The ratio of total revenue to total number of items was very high, whereas the ratio of total revenue to item price was medium. Again, the ratio of total invoice to total revenue generated was low and the price of the item was average.

Cluster Three is ranked last, because, although customers of this cluster generate average amount of revenue, the total number of items imported was also average. Customers of this segment also spent high amount of invoice. Moreover, the ratio of total revenue to total number of items and the ratio of total invoice to total revenue generated were average and very high, respectively.

Therefore, the second cluster customers were high value customers, because these customers generate high revenue by importing small number of items and spending lower invoice. However, customers categorize under Cluster Three were low value customers because this cluster customers spent high amount of invoice and they generated average amount of revenue from average number of items.

4.5.3.1.2 Experiment 2

The second clustering model experiment is conducted by setting the value of K=4 and seed size=1000. When the default seed size (seed=10) took the total number of instances were not properly distributed. The output of the second experiment is displayed in the following table.

Segmentation (K=4, Seed=1000)									
Cluster No.	**Distribution of instances (in %)**	**TIT**	**TIV**	**TRN**	**RIT**	**IVR**	**RPR**	**IPR**	**City**
1	6651 (14%)	13.40	73379.33	640087.27	68448.49	1.78	5031.43	20372.4	EU
2	9263 (20%)	10.35	138122.13	1800030.56	75083.10	0.08	12172.93	58714.05	AS
3	18790 (40%)	14.13	128316.08	327627.05	14652.31	5.64	3465.22	27325.13	AS
4	12044 (26%)	12.15	51504.68	1001966.74	36835.44	0.05	6495.89	19715.84	AS
Total	**46748 (100%)**								
Cluster No.	**Distribution of instances (in %)**								
1	6651 (14%)	L	L	A	VH	L	A	A	
2	9263 (20%)	L	A	VH	VH	VL	VH	VH	
3	18790 (40%)	L	A	L	L	VH	L	H	
4	12044 (26%)	L	L	VH	A	VL	A	A	
Total	**46748 (100%)**								

Table 4.12 Clustering result of the second experiment

In the second experiment four segments were formed, but these four segments didn't contain equal number of customers. The majority of the customers are grouped under Cluster 3 that is 18790 customers out of 46748. Cluster 4 consists of the second larger group (12044 out of 46748), Cluster 2, the third (9263 out of 46748) and Cluster 1 consists the least number of customers (6651 out of 46748).

Cluster 1: - This cluster consists of customers that imported low number of items, spent low amount of invoice, generated average amount of revenue, imported items which have average price, the ratio of revenue to total number of item was very high, the ratio of invoice to total number of item was low, the ratio of total revenue to item price was average and items originated from European countries.

Cluster 2: - The second cluster consists of customers that imported small number of items, spent average total invoice, generated very high revenue, had very high ratio of total revenue to total number of items, had very low ratio of total invoice to total revenue, had very high ratio of total revenue to item price, had very high item price and items originated from Asian countries.

Cluster 3: - The third segment holds customers that imported small number of items, spent average invoice, generated low revenue, had low ratio of total revenue to total number of items, had very high ratio of total invoice to total revenue, had low ratio of total revenue to item price, had high item price and that items which were imported from Asian countries.

Cluster 4: - The last cluster consists of customers that imported small number of items, spent low number of invoice, generated very high amount of revenue, had average ratio of total amount of revenue to total number of items, had very low ratio of total invoice to total amount of revenue generated, had average ratio of total revenue to item price, had average item price and items imported from Asian countries.

Rank of the second experiment is arranged in the following table.

Cluster No.	**Possible Rank**
1	3
2	2
3	4
4	1

Table 4.13 Cluster rank for the second experiment

As it is shown in Table 4.13, Cluster 4 is ranked first, as customers in this cluster generated very high amount of revenue by importing small number of items and spending low invoice. This segment consists of average ratio of total revenue to total number of item, very low ratio of invoice to total revenue, average ratio of total revenue to item price and average item price. The items for the fourth segment are imported from Asian countries. Customers in the second segment is ranked second as they generated very high revenue by importing small number of items and spending average amount of invoice. The ratio of total revenue to total number of items and the ratio of total invoice to total revenue were very high and very low, respectively. The ratio of total revenue to item price and the price of the item were very high. The second cluster also consists of items which imported from Asian countries. The first cluster is ranked third because this cluster included customers who generated average amount of revenue by importing small number of items and spending low amount of revenue. The ratio of total revenue to total number of items was very high. On the other hand, the ratio of total invoice to total revenue generated was low. The ratio of total revenue to item price and the price of the item for this cluster was average. The third segment is ranked last, because customers of this cluster generated low amount of revenue by providing average amount of invoice and bringing in small number of items. The ratio of total revenue to total number of items was low and the ratio of total invoice to total amount of revenue generated was very high. The ratio of total revenue to item price and the price of the item were low and high, respectively. Items of this cluster originated from Asian countries.

4.5.3.1.3 Experiment 3

To get the best clustering the researcher tried different experiments by varying the number of clusters (the value of K) and the seed size. So, the researcher continued with the third experiment by setting the value of K=5 and using the default seed size, which is 10. The Weka output of the third experiment is presented in the following table.

Segmentation (K=5, Seed=10)									
Cluster No.	**Distribution of instances (in %)**	**TIT**	**TIV**	**TRN**	**RIT**	**IVR**	**RPR**	**IPR**	**City**
1	17640 (38%)	11.23	132532.47	319067.19	15308.78	5.97	2985.01	28946.29	AS
2	9210 (20%)	9.52	138128.41	1797414.48	75515.58	0.08	12241.55	59053.13	AS
3	11746 (25%)	9.97	50533.09	994265.18	38099.33	0.05	6205.96	20326.46	AS
4	1486 (3%)	70.51	71479.32	694342.74	852.17	0.36	11953.49	1096.22	AS
5	6666 (14%)	13.37	73362.92	639324.67	68314.16	1.73	5027.24	20310.03	EU
Total	46748 (100%)								
Cluster No.	**Distribution of instances (in %)**								
1	17640 (38%)	L	A	L	A	VH	L	H	
2	9210 (20%)	VL	A	VH	VH	VL	VH	VH	
3	11746 (25%)	VL	L	H	A	VL	A	A	
4	1486 (3%)	VH	L	A	VL	VL	H	L	
5	6666 (14%)	L	L	A	VH	L	A	A	
Total	**46748 (100%)**								

Table 4.14 Clustering result of the third experiment

Cluster 1: - Customers that grouped under this segment generated low amount of revenue by importing small number of items and spending average amount of invoice. The ratio of total revenue to total number of items was average, the ratio of total invoice to total amount of revenue was very high, the ratio of total revenue to item price was low,

the price of the item was high and this cluster consists of items which imported from Asian countries.

Cluster 2: - The second cluster consists of customers that generated very high amount of revenue by importing very low number of items and spending average amount of invoice. The ratio of total revenue to total number of items and the ratio of total revenue to item price were very high. The ratio of total invoice to total revenue was very low; the price of the item which is grouped under this cluster was very high. This cluster holds items originated from Asian countries.

Cluster 3: - This segment holds customers that generated high amount of revenue by importing very small number of items and spending low amount of invoice. The ratio of revenue to total number of items was average; the ratio of total invoice to total amount of revenue generated was very low; the ratio of total revenue to item price and the price of items were average. This cluster also holds items which imported from Asian countries.

Cluster 4: - The fourth segment of this experiment consists of customers who generated average amount of revenue by importing very high number of items and spending low amount of invoice. The ratio of total revenue to total number of items and the ratio of invoice to revenue were very low; the ratio of revenue to item price was high; the price of item which grouped under this cluster was low and this cluster holds items which imported from Asian countries.

Cluster 5: - The last cluster of this experiment consists of customers who generated average amount of revenue by importing lower number of items and spending lower amount of invoice. The ratio of total revenue to total number of items was very high; the ratio of total invoice to total amount of revenue generated was low; the ratio of total revenue to item price and the price of the item were average. This segment holds items which were imported from European countries.

Based on the facts of the business problem of the Authority the above clusters are ranked in the following order.

Cluster No.	Possible Rank
1	5
2	1
3	2
4	4
5	3

Table 4.15 Cluster rank for the third experiment

Cluster 2 customers are ranked first because this group of customers generated very high amount of revenue by importing very small number of items and spending medium amount of invoice. In addition, both the ratio of revenue to total number of items and the ratio of revenue to item price were very high. For this segment, the ratio of total invoice to total revenue generated was also very low. So, Cluster Two customers were high value customers. The third segment customers are ranked second; this segment customers generated high amount of revenue by importing very low amount of items and spending low amount of invoice. The ratio of revenue to total number of items and the ratio of revenue to item price were average. The ratio of total invoice to total amount of revenue generated was very low. This segment consists of customers who imported medium price of item. Cluster 5 holds the third rank; this segment embraces those customers who generated average amount of revenue by importing small number of items and spending low amount of invoice. The ratio of total revenue to total number of items and the ratio of total invoice to total amount of revenue generated were very high and low, respectively. The ratio of total revenue to item price and the price of the item were average. Cluster 4 holds the fourth rank; this cluster consists of customers that generated average amount of revenue by importing very large number of items and spending low amount of invoice. For this cluster, the ratio of revenue to total number of items and the ratio of invoice to total amount of revenue generated were very low but the ratio of total revenue to item price is high that means the types of items which is imported by this group of customers were low price items.

Cluster 1 is ranked last because the customers of this segment generated low amount of revenue. Even thought the customers of this cluster imported low amount of items they spent average amount of invoice. So, according to the business problem the first segment

customers are low-value customers, whereas Cluster 2 customers are considered high-value customers.

4.5.3.1.4 Experiment 4

The fourth experiment is continued by adjusting the number of clusters or value of K=6 and using the default seed size which is 10. The Weka output of the fourth experiment looks like as follows.

segmentation (K=6, Seed=10)									
Cluster No.	**Distribution of instances (in %)**	**TIT**	**TIV**	**TRN**	**RIT**	**IVR**	**RPR**	**IPR**	**City**
1	13854 (30%)	12.10	157070.73	243884.97	13300.12	7.59	2276.08	33073.11	AS
2	6662 (14%)	8.84	104125.23	1908117.51	82442.08	0.06	13668.66	72714.90	AS
3	10740 (23%)	10.84	47558.55	729263.28	30509.01	0.07	5426.38	18969.74	AS
4	1120 (2%)	78.36	75492.02	795137.70	929.30	0.36	12229.10	994.73	AS
5	6610 (14%)	13.45	73674.61	638412.01	68930.34	1.79	5036.62	20336.02	EU
6	7762 (17%)	9.94	109094.76	1303954.71	45333.99	0.08	7950.09	19114.95	AS
Total	**46748 (100%)**								
Cluster No.	**Distribution of instances (in %)**								
1	13854 (30%)	L	H	VL	L	VH	L	H	
2	6662 (14%)	VL	L	VH	VH	VL	VH	VH	
3	10740 (23%)	L	L	A	A	VL	A	A	
4	1120 (2%)	VH	L	H	VL	VL	VH	VL	
5	6610 (14%)	L	L	A	VH	L	A	A	
6	7762 (17%)	VL	L	VH	H	VL	A	A	
Total	**46748 (100%)**								

Table 4.16 Clustering result of the fourth experiment

Cluster 1: - The first cluster of the last experiment consists of small number of items, high amount of invoice, very low amount of revenue, low amount of ratio of revenue to

total number of items, very high amount of ratio of invoice to revenue, low amount of ratio of revenue to item price, high amount of item price and items which originated from Asian countries.

Cluster 2: - The second segment of this experiment embraces very low number of items, low amount of invoice, very high amount of revenue, very high amount of ratio of revenue to total number of items, very low amount of ratio of invoice to revenue, very high amount of ratio of revenue to item price, very high item price and items which imported from Asian countries.

Cluster 3: - This cluster consists of small number of items, low amount of invoice, average amount of revenue, average amount of ratio of revenue to total number of items, very low amount of ratio of invoice to revenue, average amount of ratio of revenue to item price, average amount of item price and items which originated from Asian countries.

Cluster 4: - The fourth segment holds very high number of items, low amount of invoice, high amount of revenue, very low amount of ratio of revenue to total number of items, very low amount of ratio of invoice to revenue, very high amount of ratio of revenue to item price, very low item price and items which originated from Asian countries.

Cluster 5: - The fifth cluster consists of low number of items, low amount of invoice, average amount of revenue, very high amount of ratio of revenue to total number of items, low amount of ratio of invoice to revenue, medium amount of ratio of revenue to item price, average amount of item price and items originated from European countries.

Cluster 6: - The last cluster of this experiment consists of very low number of items, low amount of invoice, very high amount of revenue, high ratio of revenue to total number of items, very low ratio of invoice to revenue, average ratio of revenue to item price, average item price and items which imported from Asian countries.

Based on the above description and the business problems of the Authority the above clusters are ranked in the following order.

Cluster No.	Possible Rank
1	6
2	1
3	3
4	5
5	4
6	2

Table 4.17 Cluster rank for the forth experiment

The second cluster generated very high amount of revenue, very high amount of revenue per total number of items and very high amount of revenue per item price. Since the price of items was very high Cluster Two customers are imported very low number of items by spending low amount of invoice; therefore, this cluster is ranked first. The sixth cluster held the second rank, because this cluster also generated very high amount of revenue, high amount of revenue per total number of items and average amount of revenue per item price. This cluster customers imported very low amount of items by spending low amount of invoices. The price of item which is grouped under this segment was average. So, Cluster Two customers are high value customers. Cluster 6, 3, 5 and 4 customers held the next rank this cluster generally consist of medium-level value customers. The first segment held the last rank, because this cluster generated very low amount of revenue by spending high amount of invoice and importing small number of items. The ratio of revenue per total number of items was also low.

4.5.3.1.5 Choosing the Best Clustering Model

To get the best segmentation of customers, four cluster modeling experiments were conducted by varying the number of clusters or the value of K (3, 4, 5 and 6) and using different seed size. Though various experiments with different seed size are tried, only the seed size which resulted in fair distribution of segments is reported.

From the four experiments, the best clustering are selected by considering the output of clusters, which contains less inter-cluster similarity and high intra-cluster similarity. Moreover, within cluster sum of squared errors and the domain experts' judgment were also taking into consideration. From all of the above experiments, Experiment 3 (K=5)

and Experiment 4 (k=6) have relatively small number of within cluster sum of squared errors. Moreover, the clusters have good distribution of instances. From these two clusters, the domain experts chase the third experiment (K=5); this is because the experts justified that the fourth experiment did not create dissimilar clusters; the experts further clarified that, in the fourth experiment Cluster 2 and Cluster 6 as well as Cluster 3 and Cluster 5 customers were similar customers. So, according to the experts, the third experiment (K=5) is the best cluster model from the other models experimented in this research.

4.5.3.2 Classification Modeling

As described in Section 2.6.2, classification is a learning model in data mining techniques which aims at building a model to predict future customer behaviors through classifying database records into a number of predefined classes based on certain criteria. In this study, most common classification techniques, such as decision tree and neural network classification techniques were tested; for decision tree J48 algorithm and for neural network MultilayerPerceptron algorithm were investigated.

4.5.3.2.1 Decision Tree Model Building

The output of the selected clustering model was fed to the J48 decision tree algorithm. Here the cluster index is used as the dependant variable, whereas the remaining all attributes which are selected for the cluster model building, are fed as independent variables. The J48 decision tree provided a descriptive classification model of the clusters, thus enabling exploration and detection of the characteristic of each cluster.

During the generation of a classification model, both options, the 10-fold cross-validation and percentage split (with 70% train and 30% test), were investigated.

When using the J48 decision tree algorithm with 10-fold cross-validation default parameter value, the output of the tree consisted of 91 nodes and 54 leaves. The output of the confusion matrix for this learning algorithm looks as follows.

Actual	Predicted					Total	Accuracy Rate
	Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		
Cluster 1	17633	0	1	0	6	17640	99.96%
Cluster 2	0	9209	1	0	0	9210	99.98%
Cluster 3	0	0	11742	0	4	11746	99.96%
Cluster 4	0	2	0	1484	0	1486	99.96%
Cluster 5	3	0	3	0	6660	6666	99.90%
Total	17636	9211	11747	1484	6670	46748	99.95%

Table 4.18 Output from J48 decision tree algorithm with 10-fold cross-validation default parameter value

The output of the confusion matrix shows that, in this experiment from the total 46748 amounts of data, 46728 (99.95%) of the records were correctly classified, while the remaining 20 (0.05%) of the records were incorrectly classified.

Table 4.18 clearly shows that, from 17640 records, 17633 (99.96%) of the records were correctly classified as Cluster One (low-value customers), while the remaining 1 record was incorrectly classified as Cluster Three, and the other 6 were incorrectly classified as Cluster Five. From 9210 records, 9209 (99.98%) of the records were correctly classified as Cluster Two (high-value customers), whereas the remaining 1 record was incorrectly classified as Cluster Three. Out of 11746 records, 11742 (99.96%) records were correctly classified as Cluster Three, while the remaining 4 records were incorrectly classified as Cluster Five. From 1486 records, 1484 (99.96%) records were correctly classified as Cluster Four, whereas the remaining 2 records were misclassified as Cluster Two (high-value customer). Out of 6666 records, 6660 (99.90%) records were correctly classified as Cluster Five; while the remaining 3 records were misclassified as Cluster One (low-value customers) and the other 3 records were also incorrectly classified as Cluster Three.

Although the above experiment had good accuracy, the researcher tried to find the best classification with small tree size; because small tree size decision tree classifications are easier to generate rules. So, the researcher tried various experiments by changing the default value of the J48 decision tree 10-fold cross-validation parameter values. The

default values of 10-fold cross-validation are the minimum number of instances per leaf (minNumObj) with 2 and the confidence factor used for pruning (confidenceFactor) with 0.25. The different values of minNumObj with 5, 10, 15, 20, 25 were tested and the output of the confusion matrix for minNumObj=25 displayed in the following table.

Actual	**Predicted**					**Total**	**Accuracy Rate**
	Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		
Cluster 1	17583	0	29	20	8	17640	99.67%
Cluster 2	0	9209	1	0	0	9210	99.98%
Cluster 3	0	6	11727	0	13	11746	99.83%
Cluster 4	5	2	0	1479	0	1486	99.52%
Cluster 5	12	0	16	14	6624	6666	99.36%
Total	17600	9217	11773	1513	6645	46748	99.73%

Table 4.19 Output from J48 decision tree algorithm with minNumObj=25 and confidenceFactor=0.25

Although increasing the value of the minimum number of instance per leaf (minNumObj) as 25 reduces the accuracy rate, the tree was pruned significantly. At this experiment the tree is generated by 34 leaves and 55 nodes. The overall accuracy rate of the algorithm was 99.73%, which means, from the total dataset (46748 records), the algorithm correctly classified 46622 number of records, whereas the remaining 126 number of records were misclassified. The accuracy was minimized when compare with the pervious experiment, which had 99.95% of accuracy.

Moreover, out of 17640 numbers of low value customers in Cluster One, only 17583 (99.67%) of them were correctly classified, while the remaining 57 (0.33%) customers were incorrectly classified. Compared with the first experiment, which had (99.96%) of accuracy, the second algorithm's accuracy rate was reduced by 0.29%. From 9210 numbers of high value Cluster Two customers, 9209 (99.98%) of them were correctly classified; only one customer was misclassified and its accuracy rate was similar to the pervious experiment. Out of 11746 number of medium value Cluster Three customers, 11727 (99.83%) of them were correctly classified, whereas the remaining 19 (0.17%)

customers were incorrectly classified, so its accuracy was reduced by 0.13% from the previous experiment. From 1486 average value customers of Cluster Four, 1479 (99.52%) of them were correctly classified, while the rest 7 (0.48%) were misclassified; when it was compared with the previous experiment, the accuracy rate was reduced by (0.44%). Out of 6666 average value customers of Cluster Five, 6624 (99.36%) of them were correctly classified, while the remaining 42 (0.64%) customers were misclassified; hence, the accuracy rate was reduced by 0.54% from the previous experiment.

The experimentation of the J48 decision tree model building is continued by using percentage split with 70% of the records (dataset) as training and the remaining 30% of the record as testing. From the total 30% (14024) testing data, the algorithm correctly classified 13975 (99.65%) records, whereas the remaining 49 (0.35%) records were incorrectly classified. The following table shows the result of the percentage split experimentation.

Actual	**Predicted**					**Total**	**Accuracy Rate**
	Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		
Cluster 1	5264	0	10	7	0	5281	99.67%
Cluster 2	0	2748	0	0	0	2748	100%
Cluster 3	0	6	3559	0	8	3573	99.60%
Cluster 4	2	0	0	437	0	439	99.54%
Cluster 5	6	0	4	6	1967	1983	99.19%
Total	5272	2754	3573	450	1975	14024	99.65%

Table 4.20 Summary of the confusion matrix with default parameters value and 70 % for training and 30 % for testing dataset

Compared with the previous experiments, both 10-fold cross-validation experiments, this experiment had the least overall accuracy rate. Even in the individual cluster classification, except for the second cluster, the 10-fold cross-validation had better classification accuracy than the percentage split experimentation.

Hence, from all of the above experimentations, the first model, 10-fold cross-validation with the default parameter values, was selected because this model had better accuracy both in the overall and individual cluster classification.

4.5.3.2.2 ANN Classification Model

The other common classification technique is the Artificial Neural Network (ANN) classification model. According to Han and Kamber (2006), ANN classification model learn very fast when the attributes' values fall in the range [-1, 1]. So, to put the attribute value in the range of [-1, 1], the researcher used the Weka normalizing preprocessing facilities. Consequently, all numeric attributes were normalized; that is their value fell in between the range of [-1, 1]. However, CTY_DSC (the origin of items) is a nominal attribute but as it is stated in Section 2.6.2.2, ANN classification model usually works only with numerical data. So, there are six distinct values in this attribute and each of them is assigned a numeric value from 1-6. After mapping the nominal value into the numeric value, the Weka preprocessing facility normalizes all values to fall in the range of [-1, 1]. Hence, the same attributes that were used to build the decision tree models, were also used in the neural net modeling.

Like the J48 decision tree algorithm in this experiment also the researcher used both the 10-fold cross-validation and percentage split tests. During the experimentation various runs are performed by using the default values and changing the hidden layers, learning rate and momentum parameter values. The default hidden layers, learning rate and momentum parameters are shown in the following table.

Parameter	Description	Default value
hiddenLayers	This defines the hidden layers of the neural network. This is a list of positive whole numbers.	'a' = (attributes + classes) / 2 a=(8 (number of attributes +5 (number of clusters))/2=6.5 (7)
learningRate	The amount the weights are updated	0.3
momentum	Momentum applied to the weights during updating	0.2
10-folds cross-validation		

Table 4.21 Parameters and their default values of the neural network classifier

From the different runs of 10-fold cross-validation better accuracy rate are found by using 0.5 learning rate, 8 hidden layers, and 0.4 momentum. At this specific run the output of the confusion matrix looks as follows.

Actual	Predicted					Total	Accuracy Rate
	Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		
Cluster 1	16744	1	49	0	0	16794	99.70%
Cluster 2	1	8834	8	0	38	8881	99.47%
Cluster 3	13	12	12305	0	0	12330	99.79%
Cluster 4	0	0	0	1168	0	1168	100%
Cluster 5	0	13	0	0	7562	7575	99.83%
Total	16758	8860	12362	1168	7600	46748	99.71%

Table 4.22 10-fold cross-validation output from MultilayerPerceptron ANN algorithm with hiddenLayers=8, learningRate=0.5 and momentum=0.4

As it is shown in Table 4.22, the multilayerperceptron ANN algorithm had good overall and individual cluster classification accuracy rate. From the total dataset (46748) records, 46613 (99.71%) records were correctly classified by this algorithm; only 135 (0.29%) records were misclassified.

For the percentage split run also better performance accuracy rate was found when hiddenLayers=8, learningRate=0.5 and momentum=0.4. For this split run like in the above J48 decision tree experiment, 70% of the records were used for training and 30% of the records were used for testing. The output of the confusion matrix for this split run is presented in the following table.

Actual	**Predicted**					**Total**	**Accuracy Rate**
	Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5		
Cluster 1	5048	0	44	0	0	5092	99.04%
Cluster 2	0	2659	0	0	9	2668	99.66%
Cluster 3	0	2	3641	0	0	3643	99.95%
Cluster 4	0	0	0	345	0	345	100%
Cluster 5	0	0	0	0	2276	2276	100%
Total	5048	2661	3685	345	2285	14024	99.60%

Table 4.23 Split output from MultilayerPerceptron ANN algorithm with hiddenLayers=8 learningRate=0.6 and momentum=0.4

The output of the split run shows that from the total 14024 testing dataset, 13969 (99.60%) of the records were correctly classified, while the remaining 55 (0.40%) records were misclassified. For individual cluster level classification the split run better correctly classified the medium-value (Cluster 4 and 5) customers and high-value (Cluster 2) customers. However, this run was still less efficient on classifying low-value (Cluster 1) customers. Moreover, the split run had less overall accuracy rate than 10-fold cross-validation.

From the above two MultilayerPerceptron ANN classification models, the first model built using 10-fold cross-validation with hiddenLayers=8 learningRate=0.5 and momentum=0.4 was selected; because this model had better accuracy rate than the second split model.

4.5.3.3 Comparison of Decision Tree and Neural Network Models

So far two classification models were tested; the next step was comparing the above two classification models, which were the J48 decision tree model and MultilayerPerceptron ANN classification model. The purpose of the comparison was to choose the best from these two algorithms, which was appropriate for the problem domain of this research, CRM.

From the above two classification models the best algorithm was selected based on the following three parameters.

- The overall classification accuracy rate
- The model accuracy in classifying high value customers
- The model accuracy in classifying low value customers

So, based on the above criteria the algorithm, which had high overall accuracy rate and high accuracy in correctly classifying high value and low value customers in their clusters were selected. Consequently, the comparison of the decision tree and ANN models are described as follows.

10-fold classification model	Overall accuracy (46748 records)		High-value customers (Cluster 2) accuracy		Low-value customers (Cluster 1) accuracy	
	Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified	Correctly classified	Incorrectly classified
Decision tree model	46728 99.95%	20 0.05%	9209 99.98%	1 0.02%	17633 99.96%	7 0.04%
Neural network model	46613 99.71%	135 0.29%	8834 99.47%	47 0.53%	16744 99.70%	48 0.30%

Table 4.24 Summary of the accuracy level of the decision tree and neural net classification models

Table 4.24 shows that the decision tree model had better overall accuracy rate than the neural net model. From the total (46748) dataset, the decision tree model correctly

classified 46728 (99.95%) records. Only 20 (0.05%) records were misclassified, while the neural net algorithm correctly classified 46613 (99.71%) records and the remaining 135 (0.33%) records were misclassified. Moreover, decision tree had also better accuracy in classifying high value customers; the model had an accuracy of 99.98% for classifying high value customer only one customer was classified as other cluster. However, the neural network model had relatively less accurate in classifying high-value customers; this model had an accuracy of 99.47% for classifying high-value customer as Cluster Two. Furthermore, the decision tree model had better classification accuracy for classifying low-value customers; it had an accuracy of 99.96%, whereas the neural network model had only 99.70% accuracy.

From the above comparison of the two classification models, it is possible to conclude that the decision tree classification model is the best classifier for CRM applications, because the result shows that from the three parameters the decision tree model had better accuracy rate than the neural network classification model. Moreover, the decision tree classification model had also advantages in generation rules easily. Sample decision tree rules are attached in **Appendix 3**

Finally, the model developed with the J48 decision tree algorithm using the 10-fold cross validation with the default parameter values is tested with 17012 testing dataset to evaluate the predicting performance of the model. This developed model scored a prediction accuracy of 99.87 on the given testing dataset.

4.6 Evaluation of the Discovered Knowledge

Unfortunately, the real life data, which was collected from the business organization, consisted of missing value and outliers. Moreover, the collected data were also stored in different tables and in the format to which the selected data mining tool was not accepted. Accordingly, the researcher had taken considerable time for the data preparation task. During the preprocessing phase the researcher handled those records, which contained missing values and detected and removed outliers. Furthermore, data integration and transformation tasks were carried out during the data preparation phase. Finally, the

preprocessed data were converted in the format (ARFF) to which the selected tool was accepted.

Once the data was preprocessed, the next step was to build a model, which could effectively segment the customers of ERCA. The model-building phase was divided into two. The first phase embraced building different clustering model using the K-means algorithm. The cluster model, which best differentiated high, medium and low value customers, was selected. The next phase was to build a classification model with the J48 decision tree and multilayerperceptron neural net algorithm using the cluster index as the dependent variable. The classification task was carried out using the 10-fold cross-validation and percentage split (70% for training and 30% for testing) test options. Among the various models of J48 decision tree and multilayerperceptron neural net algorithms, the best one was selected with the criteria of overall classification accuracy rate, accuracy in classifying high value customers and accuracy in classifying low value customers. In all of these criterions the decision tree scored better classification accuracy. As a result, the decision tree with 10-fold cross-validation, which scores an overall accuracy of 99.95 %, was selected as the best classification model. Then the J48 decision tree model was tested with 17012 separate testing dataset and scored an accuracy of 99.87%.

Different rules are generated from the decision tree developed by the J48 algorithm. Some of the sample rules are the following:

Rule # 1: If TOTAL_REVENUE <= 652333.11 AND
CTY_DSC = AS AND
TOTAL_ITM <= 37 AND
REVENUE_ITEM <= 162139.61, then a customer is classified into **cluster 0**

Rule # 2: If 630615<TOTAL_REVENUE < 652333.11 AND
CTY_DSC = AS AND
TOTAL_ITM <= 37 AND
REVENUE_ITEM >162139.61, then a customer is classified into **cluster 2**

Rule # 3: If TOTAL_REVENUE <= 652333.11 AND
CTY_DSC = AS AND

37<TOTAL_ITM <= 43 AND

TOTAL_INVOICED > 87408.66, then a customer is classified into **cluster** 3

Rule # 4: If TOTAL_REVENUE<=478053.49

CTY_DSC = AM AND

TOTAL_ITM <=42, then a customer is classified into **cluster 0**

Rule # 5: If 652333.11<TOTAL_REVENUE <= 1393587.35 AND

TOTAL_ITM <= 39 AND

CTY_DSC = EU, then a customer is classified into **cluster 4**

Rule # 6: If TOTAL_REVENUE > 1393587.35 AND

CTY_DSC = AS AND

TOTAL_ITM > 75, then a customer is classified into **cluster 3**

Rule # 7: If TOTAL_REVENUE >1393587.35 AND

CTY_DSC = AM AND

TOTAL_ITM < 9, then a customer is classified into **cluster 1**

Rule # 8: If TOTAL_REVENUE <= 498887.78 AND

TOTAL_ITM > 9, then a customer is classified into **cluster 4**

The above-generated rules are consistent with the business rules of the company, which helps to identify potential and low valued customers. For instance, Rule # 7 shows those customers who generated higher revenue and import small number of items. And this is one of the characteristics of potential customers. On the other hand, Rule # 1 shows those customers who generated small amount of revenue and import large amount of items, which is one of the characteristics of low value customers. In addition to this the generated rules showed that the items origin, item price, and revenue per number of item attributes are important for identifying potential and low value customers.

Those items imported from North America, Europe, and Latin America, item price and revenue per number of item is high showed potential customers. On the other hand, those items imported from Asia, item price low, and revenue per number of item low showed low valued customers.

4.7 Use of the Discovered Knowledge

The above-generated rules in this research are encouraging. These discovered knowledge/rules could be used for identifying high, medium, and low value customers and treating them accordingly. For the company to use these generated rules effectively and efficiently the company should have to first design a knowledge base system, which can provide advice for the domain experts. The company can also incorporate the newly generated rules with those current rules used for identifying potential and low value customers in order to improve the decision making process.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The main objective of this research was to effectively segment the customers of ERCA based on their revenue generated, invoice spent, total number of items imported, the price of items imported, and the items originated behaviors. This, in turn, could enable the Authority to identify high, medium, and low value customers. The identification of these different groups of customers would enable the ERCA to create good CRM with high value customers.

This investigation is carried out using Cios et al. (2000) hybrid data mining process model. The Cios et al. (2000) data mining model consists of understanding of the problem domain, understanding of the data, preparation of the data, data mining, evaluation of the discovered knowledge, and use of the discovered knowledge.

The collection and preparation of data, to make the data suitable for the data mining task, were the major tasks which took considerable time in this research. Next, K-means clustering algorithm was applied to segment the customer's data into meaningful groups. As a result, four different clustering models were built by changing the value of K (3, 4, 5 and 6) and the seed size value. Among the four models, a model which best segment high, medium and low value customers were selected together with the domain experts. Consequently, from the four experiments, the third (K=5) experiment was selected.

The classification models were built with J48 decision tree and multilayerperceptron neural net algorithms. From these two classification models the best classification model was selected by comparing the overall accuracy, accuracy in classifying high value customers and accuracy in classifying low value customers. The model which was developed with J48 decision tree algorithm had 99.95% overall accuracy rate, whereas the multilayerperceptron neural net algorithm had 99.71% of overall accuracy. For classifying high value customers, the decision tree algorithm had 99.98% of accuracy,

while the neural net algorithm had 99.47% of accuracy. Moreover, in classifying low value customers the decision tree algorithm had 99.96% of accuracy, while neural net algorithm had 99.70% of accuracy. Since, the decision tree model had scored better performance in all these evaluation parameters, it is the researcher's belief that decision tree classification model has an appropriate technique for this research on CRM.

In general, the results from this study were encouraging. It was possible to segment customers' data using data mining techniques that made business sense. To this effect, related literature on data mining techniques, CRM and customer segmentation was reviewed.

5.2 Recommendations

The researcher makes the following recommendations based on the findings of this study.

Build a Customer Data Warehouse

This study has taken considerable time for the data preparation phase. This time can be shortened if the data warehouse is established beforehand. So, the researcher strongly recommends that the ERCA to develop data warehouse which contains customer behavior/profiles. The data warehouse could be used not only for data mining but also for other statistical analysis (OLAP) and report generation.

Model Performance Improvements

The model building process in this investigation was carried out in two sub-phases. For clustering, the researcher used the simple K-means algorithm, whereas for classification J48 decision tree and multilayerperceptron neural net learner algorithms were tested. Though the results found are encouraging, refinement to both the segmentation and multilayerperceptron neural net classification were needed. Based on this, the researcher recommends the following for further data mining research:

- Inclusion of additional customer attributes: Though the attribute selection was done together with the domain experts, only limited numbers of all the possible

member of attributes were used. However, the researcher believes that better results were found through the use of as much attributes as possible.

- Integration of customs and revenue database: The present investigation only focuses on the customs ASYCUDA databases. This database only contains revenue which is generated when customers import an item. However, further investigations by integrating the customs and revenue databases will bring about better customer segmentation in ERCA.
- Comparison of the K-means clustering algorithm with other clustering algorithms: The simple K-means algorithm which was used in this investigation has good performance to segment high, medium, and low value customers. However, further investigation by using other clustering algorithms such as HierarchicalClusterer, DBScan, FarthestFirst, FillteredClusterer and SOM, would have either confirmed or adjusted the result of this investigation.
- Additional trial by changing different parameter value of multilayerperceptron neural net algorithm: As Melaku (2009) recommended that by increasing the number of records and adjusting the default parameter values of multilayerperceptron neural net algorithm, it is possible to increase the accuracy of the classification model. As a result, in this research, by increasing the dataset size from 10,090 to 46,748 and adjusting the default momentum parameter value from 0.2 to 0.4, the overall accuracy rate increases from 98.62% to 99.71%. However, in this investigation only momentum, learning rate and hidden nodes of multilayerperceptron neural network algorithm's parameters were tried at different values to improve the classification accuracy of the resulting model. So, the researcher further recommends testing the remaining parameter values, such as value of number of iteration (training time), validation set size, and validation threshold.
- Reducing within cluster sum of squared error: During the cluster experimentation, the cluster sum of squared error was relatively large, which was around 3101. So, it is the researcher's believe that by further adjusting some parameter values, the sum of squared error may be reduced.

Enhancing CRM and Data Mining Understanding

To increase the satisfaction of customers so as to generate higher revenue, the ERCA should expose its employees to the importance of CRM. Moreover, data mining techniques for customer segmentation should enable the Authority to identify potential customers. Therefore, to use this capability, awareness on the advantages of CRM among employees at all levels should be created.

REFERENCES

Adomavicius, Gediminas and Alexander Tuzhilin. 2001. *Using data mining methods to build customer profiles.* New York: IEEE 1: 74-82

Anyanwu, N. Matthew and Sajjan G. Shiva. n.d. *Comparative analysis of serial decision tree classification algorithms.* International Journal of Computer Science and Security, (IJCSS) 3: 230-240

Apte, Chidanand, Bing Liu, Edwin Pednault and Padhraic Smyth. 2002. *Business applications of data mining.* Communications of the ACM 8:49-53

Arai, Kohei and Ali Ridho Barakbah. 2007. *Hierarchical K-means: an algorithm for centroids initialization for K-means.* Saga University: 1:25-31

Arthur, David and Sergei Vassilvitskii. 2006. *How slow is the K-means method*. http://www.cs.duke.edu/courses/spring07/cps296.2/papers/kMeans-socg.pdf (access date February 17, 2011)

Berndt, Adele, Frikkie Herbst, and Lindie Roux. 2005. *Implementing a customer relationship management program in an emerging market.* Journal of Global Business and Technology 1: 81-89

Bose, Ranjit. 2002. *Customer relationship management key components for IT success*. Industrial Management and Data System 2:89-97

Boulding, William, Richard Staelin, Michael Ehret, & Wesley J. Johnston. 2005. *A customer relationship management roadmap: what is known, potential pitfalls, and where to go*. Journal of Marketing 1:155–166

Bounsaythip, Catherine and Esa Rinta-Runsala. 2001. *Overview of data mining for customer behavior modeling.* VTT Information Technology 18: 1-53

Bull, Christopher. 2003. *Strategic issues in customer relationship management (CRM) implementation.* Business Process Management Journal 9: 592-602

Chalmeta, Ricardo. 2006. *Methodology for customer relationship management*. The Journal of Systems and Software 79:1015–1024

Chen, Ruey-Shun, Ruey-Chyi Wu and J. Y. Chen. 2005. *Data mining application in customer relationship management of credit card business*. IEEE Annual International Computer Software and Applications Conference 2:730-3157

Cross, Glendon and Wayne Thompson. 2008. *Understanding your customer: segmentation techniques for gaining customer insight and predicting risk in the telecom industry*. SAS Global forum data mining and predictive model http://www2.sas.com/proceedings/forum2008/154-2008.pdf (access date February 10, 2011)

Denekew, Abera. 2003. *The application of data mining to support customer relationship management at Ethiopian Airlines.* Unpublished Master's Thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

Deshpande, S. P. and V. M. Thakare. 2010. *Data mining system and application: a review*. International Journal of Distributed and Parallel systems (IJDPS) 1:32-44

Doye, Tom. 2010. *Collaborative border management*. World Customs Journal 4:333-340

Dunham, H. Margaret. 2000. *Data mining techniques and algorithms partial draft of forthcoming book from prentice hall.* http://www.cba.ua.edu/~mhardin/dunham.pdf (access date January 5, 2011)

Edelstein, H. 2002. *Data mining: exploiting the hidden trends in your data*. Technology in Society 4:483-502

European Regulators Group for Electricity and Gas (ERGEG). 2010. *GGP on customer complaint handling, reporting and classification.* Ref: E10-CEM-33-05

Faber, Vance. 1994. *Clustering and the continuous K-means algorithm*. Los Alamos Science 22: 138-144

Farn, Cheng-Kiang and Li Ting Huang. 2009. *A study on industrial customer's loyalty to application service providers: the case of logistics information services*. International Journal of Computers 3: 151-160

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *From data mining to knowledge discovery in databases.* American Association for Artificial Intelligence 37-54

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *Knowledge discovery and data mining: towards a unifying framework.* *http://www-aig.jpl.nasa.gov/kdd96* (access date March 2, 2011)

Federal Negarit Gazeta of the Federal Democratic Republic of Ethiopia. 2008. Proclamation No. 587/2008, Proclamation Page 4123

Fekadu, Mekonnen. 2004. *Application of data mining to support customer relationship management at Ethiopian Telecommunications Corporation*. Unpublished Master's Thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

Gray, Paul and Jongbok Byun. 2001. *Customer relationship management*. http://www.wcfia.harvard.edu/us-japan/research/pdf/06-13.ueno.pdf (access date February 15, 2011)

Gray, Paul. 2001. *Customer relationship management*. http://www.crito.uci.edu/papers/2001/crm.pdf (access date February 16, 2011)

Greengrove, Kathryn. 2002. *Needs-based segmentation: principles and practice*. USA: International Journal of Market Research 44: 405-421

Hajizadeh, Ehsan, Hamed Davari Ardakani and Jamal Shahrabi. 2010. *Application of data mining techniques in stock markets.* Journal of Economics and International Finance 7:109-118

Han, Jiawei and Micheline Kamber. 2006. *Data mining: concepts and techniques*. 2nd ed.USA: Morgan Kaufmann

Henock, Woubishet. 2002. *Application of data mining techniques to support customer relationship management at Ethiopian Airlines.* Unpublished Master's Thesis, Deparment of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

http://www. Ethiopian Revenues and Customs Authority (access date January 7, 2011)

Huang, Yingping. 2003. *Infrastructure, data cleansing and mining for support of scientific simulation.* Department of Computer Science and Engineering, University of Notre Dame, Indiana

Hwang, Hyunseok, Taesoo Jung and Euiho Suh. 2004. *An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry*. Expert Systems with Applications 26: 181–188

Johnston, Robert and Sandy Mehra. 2002. *Best-practice complaint management*. Academy of Management Executive 4:145-154

Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. *An efficient k-means clustering algorithm: analysis and implementation.* IEEE Transactions on Pattern Analysis and Machine Intelligence 24:881-892

Kim, Jonghyeok, Euiho Suh, and Hyunseok Hwang. 2003. *A model for evaluating the effectiveness of CRM using the balanced scorecard.* Journal of Interactive Marketing 2: 5-19

Kim, Su-Yeon, Tae-Soo Jung, Eui-Ho Suh, and Hyun-Seok Hwang. 2006. *Customer segmentation and strategy development based on customer lifetime value: a case study*. Expert Systems with Applications 31: 101–107

King, F. Stephen and Thomas F. Burgess. 2008. *Understanding success and failure in customer relationship management.* Industrial Marketing Management 37: 421–431

Koh, Chye Hian and Gerald Tan. n.d. *Data mining applications in healthcare*. Journal of Healthcare Information Management 2:64-72

Kumar, Ela and Arun Solanki .2010. *A combined mining approach and application in tax administration.* International Journal of Engineering and Technology 2:38-44

Kumneger, Fekrie. 2006. *Application of data mining techniques to support customer relationship management for Ethiopian Shipping Lines (ESL)*. Unpublished Master's Thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

Kurgan, A. Lukasz, and Petr Musilek. 2006. *A survey of knowledge discovery and data mining process models.* United Kingdom: Cambridge University Press 21: 1-24

Mahler, J. Juliannem and Thomas Hennessey. 1996. *Taking internal customer satisfaction seriously at the U.S. customs service*. Jstor's M.E Sharpe 4:487-497

McGuirk, Mike. 2007. *Customer segmentation and predictive modeling.* http://www.iknowtion.com/downloads/Segmentation.pdf (access date March 2, 2011)

McKinsey and Company. 2001. *The new era of customer loyalty management.* htttp://www.marketing.mckinsey.com (access date February 4, 2011)

Melaku, Girma. 2009. *Applicability of data mining techniques to customer relationship management (CRM): the case of Ethiopian Telecommunications Corporation's (ETC) code division multiple access (CDMA) telephone service*. Unpublished Master's Thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

Nasereddin, H. O. Hebah. 2009. *Stream data mining*. International Journal of Web Applications 4:183-190

Ngai, E.W.T. 2005. *Customer relationship management research: an academic literature review and classification.* Marketing Intelligence and Planning 6: 582-605

Ngai, E.W.T., Li Xiu , D.C.K. Chau. 2009. *Application of data mining techniques in customer relationship management: a literature review and classification*. Expert Systems with Applications 36:2592-2602

Parvatiyar, Atul and Jagdish N. Sheth. 2001. *Customer relationship management: emerging practice, process, and discipline.* Journal of Economic and Social Research 2:1-34

Rygielski, A. Chris, Jyun-Cheng Wang B, David C. Yen. 2002. *Data mining techniques for customer relationship management.* Technology in Society 24:483–502

Saarenvirta, G.. 1998. *Mining customer data*. http://www.db2mag.com/db_area/archives/1998/q3/98fsaar.html (access date January 6, 2011)

Shang, S. C. Shari and Chih-Hsiang Chen. n.d. *Human processes in customer relationship management.* 11th Pacific-Asia Conference on Information Systems http://www.pacis-net.org/file/2007/1259.pdf (access date January 6, 2011)

Singh, Yashpal and Alok Singh Chauhan. 2009. *Neural networks in data mining*. India: Journal of Theoretical and Applied Information Technology 37-42

Srivastava, Jaideep. 2002. *Data mining for customer relationship management (CRM)*. Advances in Knowledge Discovery and Data Mining 2336:14-27

Suresh, Hemamalini. 2002. *Customer relationship management: an opportunity for competitive advantage*. India: PSG Institute of Management http://www.realmarket.com/required/psginst1.pdf (access date February 6, 2011)

Tilahun, Muluneh. 2009. *Possible application of data mining techniques to target potential visa card users in direct marketing: the case of Dashen Bank s.c.* Unpublished Master's Thesis, Department of Information Science, Faculty of Informatics, Addis Ababa University, Addis Ababa

Trappey, V .Charles, Amy J.C. Trappey, Ai-Che Chang, and Ashley Y.L. Huang. 2009. *The analysis of customer service choices and promotion preferences using hierarchical clustering*. China: Journal of the Chinese Institute of Industrial Engineers 5:367-376

Two Crows Corporation. 1999. *Introduction to data mining and knowledge discovery*. 3rd ed. ISBN: 1-892095-02-5, Potomac, MD 20854 (U.S.A.)

Verhoef, C. Peter. 2003. *Understanding the effect of customer relationship management efforts on customer retention and customer share development*. Jstor the Journal of Marketing 4: 30-45

Wahab, Samsudin and Juhary Ali. 2010. *The evolution of relationship marketing (RM) towards customer relationship management (CRM): a step towards company sustainability*. Information Management and Business Review 1: 88-96

WeiWang and Shidong Fan. 2010. *Application of data mining technique in customer segmentation of shipping enterprises*. China: IEEE 1:4

Witten, H. Ian and Eibe Frank. 2005. *Data mining practical machine learning tools and techniques.* 2nd ed. USA: Morgan Kaufmann

Zaiane, R. Osmar .1999. *Principles of knowledge discovery in databases*. University of Alberta http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/slides/ch0s.pdf (access date February 5, 2011)

APPENDICES

Appendix 1 Partial View of the Initial Collected Sample Data

SAD_CONSIGNEE Nominal	CMP_NAM Nominal	SAD_REG_DATE Nominal	SAD_ITM_TOTAL Numeric	SAD_PACK_TOTAL Nominal	SAD_TOT_INVOICED Nominal	SADITM_ITM_PRICE Nominal	SADITM_HS_COD Numeric	CTY_DSC Nominal	CDI_AMOUNT_TOTAL Nominal
0000148762DR02	AHMMED YOUSUF AWOL	12-30-2009	1.0	17.0	57,218.86	57,218.86	7210.0	China	8,577,287.28
0001600437ET02	A.M.Y. GENERAL TRADING ...	07-08-2009	21.0	81.0	12,460.80	224.4	8409.0	China	2,092,902.53
0000148668DR02	ABDUSELAM MOHAMMED A...	01-05-2009	1.0	3,400.00	54,672.00	54,672.00	1511.0	Netherla...	5,635,364.62
0000148668DR02	ABDUSELAM MOHAMMED A...	11-06-2009	1.0	5,000.00	61,875.00	61,875.00	1006.0	Thailand	5,635,364.62
0000145906DR02	ADEM MOHAMMED ABDULAHI	02-14-2009	1.0	11,990.00	297,352.00	297,352.00	1511.0	Malaysia	266,151.85
0001583210OR02	SHEMSHEDIN MOHAMED H...	11-04-2009	1.0	2,000.00	30,960.00	30,960.00	1511.0	United A...	562,159.17
0000143390DR02	ABDULJEBAR ABDUREHMA...	12-21-2009	2.0	2.0	32,400.00	16,200.00	8704.0	Japan	1,533,369.66
0002298177DR03	AWALEY GENERAL TRADIN...	07-15-2009	4.0	2,300.00	39,653.00	7,200.00	1104.0	Malta	10,698,674.41
0000760571HRRI	HARAR BREWERY SHARE C...	07-25-2009	1.0	107.0	6,741.00	6,741.00	3505.0	United K...	20,665,101.42
0000037161ET03	JUGEL PRIVATE LIMITED C...	06-20-2009	16.0	100.0	109,000.00	202.4	8714.0	India	10,569,689.14
0004332627AA02	EYASU YIRADU KIDANE	01-22-2009	19.0	1,256.00	19,540.07	2,310.64	9403.0	China	291,121.08
0004332627AA02	EYASU YIRADU KIDANE	01-28-2009	11.0	292.0	19,253.08	180.0	9401.0	China	291,121.08
0004603267DR02	AMIR AHMED ALI	09-17-2009	7.0	13.0	23,338.48	696.17	8443.0	Italy	506,903.47
0000322072DR02	JEMAL ABUBEKER MUMED	04-27-2009	42.0	505.0	25,410.25	189.4	3306.0	United A...	739,378.02
0000322072DR02	JEMAL ABUBEKER MUMED	10-27-2009	78.0	568.0	23,967.26	85.2	3306.0	Saudi Ar...	739,378.02
0000722883DR02	SEBLEWENGEL ABEBE ADM...	06-01-2009	31.0	352.0	58,964.62	50.0	5802.0	China	48,329.76
0003234229DR02	ZINAT ABDURASHID AHMED	02-06-2009	1.0	6,240.00	151,632.00	151,632.00	1511.0	United A...	541,442.05
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	28.0	8516.0	China	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	77.7	8415.0	China	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	510.0	8518.0	China	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	255.0	8518.0	China	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	536.0	8518.0	India	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	300.0	8527.0	China	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	06-08-2009	92.0	1,088.00	38,877.92	360.0	8510.0	United S...	725,161.84
0003980843DR02	BIRZAF GEBREMEDHIn TEK...	11-06-2009	96.0	1,479.00	48,453.74	960.0	8471.0	Taiwan, ...	1,450,323.68
0003835108ETFI	HABESHA STEEL MILLS PRI...	06-26-2009	53.0	175.0	47,658.96	11.5	3506.0	India	35,698.14
0003835108ETFI	HABESHA STEEL MILLS PRI...	06-26-2009	53.0	175.0	47,658.96	67.2	4016.0	India	35,698.14
0000148734DRRI	WONDWOSEN WERKU MOLLA	05-22-2009	138.0	15,503.00	189,101.45	100.0	8215.0	Portugal	317,619.75
0000329779DRRI	BERHANE GEBREHIWOT AT...	08-19-2009	77.0	1,378.00	214,813.87	484.02	9403.0	China	56,624.03
0000329779DRRI	BERHANE GEBREHIWOT AT...	08-19-2009	77.0	1,378.00	214,813.87	220.0	8536.0	China	56,624.03
0000007947ET02	AFATCO TRADING PRIVAT...	09-26-2009	29.0	1,337.00	31,716.59	717.24	6809.0	China	364,324.86
0000743299DR03	Abdiweli Jebreal Musa	10-24-2009	33.0	33.0	134,436.24	293.2	8421.0	India	645,788.32
0000743299DR03	Abdiweli Jebreal Musa	10-24-2009	33.0	33.0	134,436.24	248.41	4016.0	India	645,788.32
0000743299DR03	Abdiweli Jebreal Musa	10-24-2009	33.0	33.0	134,436.24	851.15	7320.0	India	645,788.32

Appendix 2 Sample of the Decision Tree Generated With 10-Fold Cross-Validation Technique

J48 pruned tree

TOTAL_REVENUE <= 652333.11

| CTY_DSC = AS

| | TOTAL_ITM <= 43

| | | TOTAL_ITM <= 37

| | | | REVENUE_ITEM <= 162139.61: cluster0 (16044.0)

| | | | REVENUE_ITEM > 162139.61

| | | | | TOTAL_REVENUE <= 630615.86: cluster0 (432.0)

| | | | | TOTAL_REVENUE > 630615.86: cluster2 (21.0)

| | | TOTAL_ITM > 37

| | | | TOT_INVOICED <= 87408.66: cluster0 (257.0)

| | | | TOT_INVOICED > 87408.66: cluster3 (71.0)

| | TOTAL_ITM > 43

| | | TOTAL_REVENUE <= 33962.11: cluster0 (26.0)

| | | TOTAL_REVENUE > 33962.11: cluster3 (628.0)

| CTY_DSC = EU: cluster4 (4210.0)

| CTY_DSC = AM

| | TOTAL_REVENUE <= 478053.49

| | | TOTAL_ITM <= 42: cluster0 (353.0)

| | | TOTAL_ITM > 42

| | | | TOTAL_ITM <= 49: cluster0 (8.0)

| | | | TOTAL_ITM > 49: cluster3 (3.0)

| | TOTAL_REVENUE > 478053.49: cluster4 (61.0)

| CTY_DSC = AF

| | TOTAL_REVENUE <= 448994.77: cluster0 (348.0)

| | TOTAL_REVENUE > 448994.77

| | | TOTAL_REVENUE <= 498887.78

| | | | TOTAL_ITM <= 9: cluster0 (7.0/1.0)

| | | | TOTAL_ITM > 9: cluster4 (8.0/1.0)

| | | TOTAL_REVENUE > 498887.78: cluster4 (22.0)

| CTY_DSC = LA

| | TOTAL_REVENUE <= 490154.39: cluster0 (50.0)

| | TOTAL_REVENUE > 490154.39: cluster4 (10.0)

| CTY_DSC = AU

| | TOTAL_REVENUE <= 435824.59: cluster0 (39.0)

| | TOTAL_REVENUE > 435824.59: cluster4 (2.0)

TOTAL_REVENUE > 652333.11

| TOTAL_REVENUE <= 1393587.35

| | TOTAL_ITM <= 39

| | | CTY_DSC = AS

| | | | REVENUE_ITEM <= 1327510.21

| | | | | TOTAL_REVENUE <= 665252.37

| | | | | | TOTAL_ITM <= 14

| | | | | | | TOTAL_REVENUE <= 653509.24

| | | | | | | | TOTAL_ITM <= 5: cluster2 (7.0)

| | | | | | | | TOTAL_ITM > 5: cluster0 (6.0)

| | | | | | | TOTAL_REVENUE > 653509.24: cluster2 (129.0)

| | | | | | TOTAL_ITM > 14

| | | | | | | TOT_INVOICED <= 27476.95

| | | | | | | | TOTAL_ITM <= 21: cluster0 (4.0)

| | | | | | | | TOTAL_ITM > 21: cluster2 (21.0)

| | | | | | | TOT_INVOICED > 27476.95: cluster0 (66.0)

| | | | | TOTAL_REVENUE > 665252.37: cluster2 (11304.0/1.0)

| | | | REVENUE_ITEM > 1327510.21

| | | | | TOTAL_REVENUE <= 1337529.49: cluster2 (2.0)

Appendix 3 Output of the K-means Cluster Modeling with Different K and Seed Values

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 5199.169161839557
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute             Full Data         0            1            2
                        (46748)    (6544)      (38992)      (1212)
=============================================================================
TOTAL_ITM               12.7659   13.3325      12.5755      15.8317
TOT_INVOICED        102653.6474 74041.9775 105916.3269 152172.1036
ITM_PRICE             30595.149  20555.431   32368.0044    27767.376
CTY_DSC                      AS         EU           AS           AM
TOTAL_REVENUE       837569.5832 639766.3282 879519.7837 555970.8062
REVENUE_ITEM          39995.4777 69301.0492  35285.7009  33285.8713
INVOICED_REVENUE          2.5481     1.8051       2.534      7.0145
REVENUE_PRICE           6194.2725  5025.5301   6414.6447   5414.9808
```

Output of cluster run with K=3 and seed=100

```
kMeans
======

Number of iterations: 19
Within cluster sum of squared errors: 3326.4367631392697
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute             Full Data         0            1            2            3
                        (46748)    (6651)       (9263)      (18790)      (12044)
=============================================================================
TOTAL_ITM               12.7659   13.4004      10.3455       14.131      12.1473
TOT_INVOICED        102653.6474 73379.3309 138122.1337 128316.0764    51504.681
ITM_PRICE             30595.149  20372.429   58714.0539   27325.129   19715.8375
CTY_DSC                      AS         EU           AS           AS           AS
TOTAL_REVENUE       837569.5832 640087.2653 1800030.5388 327627.0446 1001966.7408
REVENUE_ITEM          39995.4777 68448.4872  75083.0995  14652.3088  36835.4426
INVOICED_REVENUE          2.5481     1.7769      0.0819      5.6359      0.0536
REVENUE_PRICE           6194.2725  5031.4321  12172.9275   3465.2206   6495.8878
```

Output of cluster run with K=4 and seed=1000

```
kMeans
======

Number of iterations: 24
Within cluster sum of squared errors: 3101.601334182573
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute             Full Data         0            1            2            3            4
                        (46748)   (17640)      (9210)      (11746)      (1486)       (6666)
=============================================================================
TOTAL_ITM               12.7659   11.234       9.5163       9.9662      70.5074      13.3708
TOT_INVOICED        102653.6474 132532.474 138128.4144  50533.0909 71479.3151 73362.9194
ITM_PRICE             30595.149 28946.2881  59053.1257  20326.4601   1096.2238 20310.0252
CTY_DSC                      AS         AS           AS           AS           AS           EU
TOTAL_REVENUE       837569.5832 319067.1864 1797414.4812 994265.1809 694342.7351 639324.6652
REVENUE_ITEM          39995.4777  15308.7763   75515.58  38099.3343   852.1654  68314.1596
INVOICED_REVENUE          2.5481      5.9743      0.082      0.0529      0.3601      1.7733
REVENUE_PRICE           6194.2725  2985.0073  12241.5459   6205.9612  11953.4863   5027.2402
```

Output of cluster run with K=5 and seed=10