ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

The Role of Data Mining in the Risk Assessment of Customs: (With special reference to The Ethiopian Customs Authority)

A thesis submitted to the Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Masters of Science in Information Science.

By

Girma Belew

July 2004

ADDIS ABABA UNIVERSITY
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE

The Role of Data Mining in the Risk Assessment of Customs: (With special reference to The Ethiopian Customs Authority)

By

Girma Belew

July 2004

Name and Signature of Members of the Examining Board

to Getachew Jemaneh,
hairman of the Examining Board _____

r. B. L. Desai,
dvisor _____

r. Osei Nana Adjei
xternal Examiner _____

ACKNOWLEDGMENT

First and foremost my thankfulness is to the Almighty God for guiding me throughout my journey and who made it happen. Whatever I say I cannot thank him enough. Next to that I would like to forward my gratitude to my lovely mother who unreservedly devoted her whole life to me.

I am also very much indebted to my elder sister Degenesh who stood beside me at such a critical time in my life. I would also like to thank Hirut, Mimi, Enan, Yared and Yeshi for all your love, affection and support through out my endeavor.

I would also like to thank my advisor Dr. L. B. Desai for his support during my work and Ato Mesfin Getachew for his constructive comments on my work. I am also grateful to all the staff of SISA for their contribution during my stay at the school.

My special thanks goes to the management and staff of the Ethiopian Customs Authority for all their support and for giving me access to their data and specially to Ato Fethanegest T/Haimanot DGM of operations, Ato Getchew Arega, and Ato Eshetu Geref for their understanding and constant support.

The friends I met at SISA were all sources of inspiration since we met at the beginning of graduate school. I will definitely miss the times we had together.

My special thanks also goes to Ato Abebaw Alemayehu who had always been unreservedly helpful in all the times I required his assistance. I am grateful to my friends Fisesha Tade, Hailesh, Belete, Mesfin, Sami, Mimi, and all others who are not mentioned here in name but assisted me in many ways.

Last yet very important. My wholehearted gratefulness is forwarded to my sweet wife Bebe for all your love, support, understanding, and your faith in me. You are responsible for this in more ways than you know. To have you is a blessing. ***I love you with all my heart.***

TABLE OF CONTENT

LIST OF TABLES

LIST OF ABBREVIATIONS

ASYCUDA	Automated System for Customs Data
ICT	Information Communication Technology
SISA	School of Information Studies for Africa
UNCTAD	United Nations Conference for Trade and Development

Abstract

Customs Organizations are responsible for two opposing yet equally important responsibilities. These are the provision of efficient services to traders for the smooth flow of shipments and the protection of the country from any kind of risk threats that is associated with international trade.

Reform and modernization of customs services through automation and setting transparent working procedures has resulted in the provision of efficient services. However addressing the issue of implementing a proper risk management strategy remains a challenge.

The Ethiopian Customs Authority at present is handling the control of customs risks using subjective methods. The subjective method of handling risks solely depends on experts' judgment of selecting shipments for physical examination. The subjective method is essential since the knowledge and experience of customs experts and their observation of the behaviors of intervening agents like traders and clearing agents is very important. However depending only on subjective analysis for strategic risk management has its shortcomings.

This study was aimed at supporting the current selective physical examination system of incoming shipments in the Ethiopian Customs Authority with objective methods using data mining. The study was conducted through the analysis of customs fraud cases seized in the past. For this study one of the data mining techniques known as decision tree was employed. The dataset used in the study consisted of 10364 records out of which 170 cases were fraud cases.

The distribution of the two classes was highly imbalanced. To deal with the class imbalance problem the over-sampling approach was used. Five experiments were conducted by varying the rate of over-sampling. After over-sampling the five dataset had a ratio of 90:10, 80:20, 70:30, 60:40 and 50:50 all non-fraud to fraud.

Using an independent dataset that also contain 2616 non-fraud and 39 fraud cases all the models generated by the five datasets were validated. The dataset with a proportion of 70:30 has shown the best result in terms of correctly classifying the fraud cases. The model correctly classified 22 fraud cases out of the 39 cases.

Combining the subjective method with the subjective methods can improve the efficiency of risk assessment and selective physical examination of shipments in the Ethiopian customs authority. Models developed by automatic analysis can also be used across all the different customs offices in the country consistently.

CHAPTER ONE

INTRODUCTION

1.1 Background

In the current age of international trade, i.e., both import and export, has become a massively huge universal day to day activity. The rate of growth in the value of international trade in the past decade was around 6% per annum [15]. Every country in the world participates in this activity. However, the magnitude of their involvement may vary based on the level of technological development, products, resources, etc. Nations have been engaged in international trade in the past but the magnitude and complexity in today's world is quiet different from what has been the case a decade or two before. In those times the trade tie between nations was restricted by many barriers [15]. In today's globalization[1] era the marketplace has changed entirely. Globalization has resulted in an extraordinary increase in the international flow of people, goods, capital and information.

From the economic perspective, globalization is viewed as a process that can lead to fully internationally integrated markets, with free movement of goods, services, labor and capital. Thus a single market in inputs and outputs and full national treatment for foreign investors so that, economically speaking, there are no foreigners [37]. It can be stated as "the borderless world".

[1] Globalization is a term used to describe the extraordinary expansion and opening out of the international market place. The concept also incorporates the increasingly complex interrelationship of individuals and countries lives and futures, including politics, economic issues, cultures etc [37].

The development of ICT and e-commerce which enabled traders to have a global perspective and the chance to easily assess and evaluate various opportunities, to communicate and interact with a wide range of actual and potential customers, suppliers etc. has also increased the sources as well as magnitude and complexity of international trade.

Governments being the prime movers in facilitating all aspects of the economic activity to foster the development of their nation play a vital role in providing a fertile environment for the smooth operation of international trade [14].

The above changes in turn have radically changed the role of governments. The role of governments, especially in the developing countries, which was characterized as control based and isolated is now required to be substituted by liberalization involving, a freeing up of markets and reduction in the role of government in terms of ownership and control with a process going to a free movement of goods, services, labor and capital.

While facilitating and creating a fertile environment to encourage international trade and attract new investment at one end, governments also has the responsibility to guarantee the legality of trade. Moreover to finance their operation, governments have to collect various taxes.

In every country the government body in charge of collection of taxes and control of international trade operations is the respective customs administration. Customs is often the first contact foreign businesses and travelers have with the government. A satisfactory encounter with customs can encourage continued commercial activities [37].

Based on the country's level of development, the volume of transaction and other relevant factors, the organizational structure and size of customs offices may vary. Fundamentally, all customs organizations discharge some how similar duties. The basic duties include:

- Facilitation of legitimate international trade, i.e., both import and export
- Collection of taxes
- Control and monitor illegal or unlawful transactions.

Customs administrations are in the forefront of the liberalization and reform process. Traditionally customs administrations, especially in developing countries, are characterized by inefficiency, corruption and highly bureaucratic procedures and outmoded work habits [13].

The inefficiency of customs especially with regard to clearance and processing time and the associated transaction costs is a burden that traders are forced to bear. This is against the interests of traders. As a result customs procedures are perceived by traders as cumbersome and an obstacle to the smooth movement of trade.

Customs while facilitating trade and investment are equally responsible for protecting the society. Now more than ever, one of the key goals of customs is to reconcile the facilitation and control of trade, whilst protecting the society [37].

They are responsible for ensuring that all import and export activities comply with the country's laws and regulations. Thus they are faced with two opposing and equally competing responsibilities.

imbalance, and disadvantage to the law abiding trader. Hence the efficient performance of customs offices in tackling this problem is of high importance.

Customs have only a limited amount of resources at their disposal to discharge these duties. Moreover while the volume of trade and investment growth being very rapid; customs are not in a position to secure matching resource to keep pace with this growth. Hence they have to rely more on improving their competence and in using their resources more wisely. This requires a mechanism that will assist their control activity to focus on the more risk prone areas.

Customs offices perform the above mentioned duties by examining documents as well as physical examination of import and export shipments. The physical examination is usually conducted at custom points. Attempting a total examination of shipments severely obstruct the smooth flow of trade due to the massive volume of transaction passing through their gates. As a consequence, in most cases, customs perform only selective Physical examination of cargo.

Selective examination should focus on the risk prone shipments. Focusing on risk prone shipments is done by assessing the risks associated with each shipment. Risk assessment is a method by which risks associated with shipments are evaluated. There are many variables[2] that are important in conducting risk assessment. These include traders' clearing agents, country of origin, the commodity, the associated taxes and a good many other variables.

[2] In this study the terms variables and attributes have the same meaning and are used interchangeably.

Modern customs administrations nowadays employ the use of well developed methods of risk management. In such cases the use of an integrated data warehouse from external as well as internal sources is used.

Attaining the proper balance between the competing aspects of the two equally important considerations, (i.e., legitimate trade facilitation on one hand and proper enforcement to control fraud on the other) is a challenge to every customs administration and the Ethiopian Customs Authority is no exception.

Risk Management can be defined as a technique for the systematic identification and implementation of all the measures necessary to limit the likelihood of risks occurring. Its implementation is done by collecting data, analyzing the various kinds of risks, assessing the magnitude and potential consequences of the various risks and prescribing proper action to monitor and control risk.

1.2 Statement of the problem

Developing nations, like Ethiopia, are in a struggle to have their share in the ever expanding world economy. This requires the ability to respond effectively to the challenges of the global market place.

The challenge of the 21st century requires more competence in several fronts. One of these fronts is the customs administration. The effectiveness and efficiency of a country's customs organization is integral in meeting these objectives.

However, the Ethiopian Customs Authority at present has no objective methods or established procedures to apply in performing selective examination. All incoming cargos to Ethiopia are subject to physical examination. Upon entry to the country all

shipments will first go to the customs warehouse of a particular custom office and until the examination process is over they will stay at the warehouse. Until very recently, in this system, shipments may stay at the warehouse for unspecified duration. This had an adverse effect on the traders' business.

It also discouraged many investors from investing in businesses and thereby contribute to the economic development of the country. Due to globalization and other similar factors, today there are many attractive and competing investment locations all over the world that try to win foreign trade and investment. In such situations, an inefficient customs organization can cost a country capital inflows that will go to more efficient competitors. This demands countries to ***lower trade barriers***, ***release cargo more quickly*** and ***reduce physical inspection requirements*** [14].

Assessing the inefficiency of its Customs Authority the Government of Ethiopia launched Customs modernization program. The program began by re-establishing the authority with new organizational structure. Facilitated customs clearance process and transparent working procedures are now put in place. The maximum time limit for any shipment to stay in the customs warehouse before release is now limited to a maximum of two days. This improvement of the customs process has a marked effect on delivering quality services and saving the delays and costs for traders.

This in turn has demanded the control activity to find improved methods to address the risk factor. Earlier the customs control used it own pace without any time constraint. Now it has to keep pace with the new requirement of releasing cargo within a maximum of two days.

To be more efficient in addressing the physical examination within the time limit required the adoption of new methods. In the re-established structure a risk management unit is formed. This unit at present has implemented a method to classify shipments for various levels of physical examination.

In this new system, prior to the arrival of shipments, documents will be analyzed by experts. Based on the analysis, the shipments will be classified to one of the three groups, i.e., low, medium or high. For each group there is an associated level of physical examination. The low and medium groups will undergo sample examination with different sample sizes and the high risk group will undergo total examination.

Experts in the unit use their own judgment in classifying the shipments for selective examination. The base for the experts' classification includes traders, clearing agents, commodity types, country of origin etc. Among the aforementioned variables high emphasis is given to traders' and clearing agents with fraud history. Commodity types and country of origin also receive due consideration.

This subjective method of classification is important because the experts utilize their explicit as well as tacit knowledge in classifying shipments for physical examination. However, relying only experts' judgment is not sufficient. This is mainly due to the following reasons:

- Lack of any objective measure
- There are many variables that can be used as fraud detection parameters. As the number of variables increase it becomes difficult to understand or comprehend hidden relationships.

- Experts' judgment may be influenced by shipment volumes of a particular period and the workforce available for physical examination[3] – the human element.
- In the absence of experts' the classification task may suffer.

In general subjective method alone can not provide sufficient results. From this it is evident that there is an urgent need for adopting a consistent risk management model that guarantees a better selection and examination of shipments with the probability of fraud.

The Ethiopian Customs Authority is one of the beneficiaries of ASYCUDA[4] software. The ASYCUDA software has been designed mainly for trade facilitation. It houses a large amount of transactional data. The data, apart from being used as an information source for customs administrations, is serving as a source of reliable and timely external trade and fiscal statistics for government.

At present a latest version of this software known as ASYCUDA++ is available. Among the new features supported by this system is what is known as the selectivity module[5]. The Ethiopian Customs Authority is in the process of transforming its system to ASYCUDA++. Since the software provides facilities for selective examination the Ethiopian Customs Authority can enhance its classification procedure based on this facility.

---

[3] During discussion with the experts it was found out that the volume of shipment at a particular time has an influence.
[4] ASYCUDA stands for Automated SYstem for CUstoms DAta. The software is developed as a general customs database be UNCTAD. Further discussion on the software is done in chapter three.
[5] The selectivity module accepts a set of rules and shipments that match the rules will be subject to selective examination automatically by the system.

Therefore it is the aim of this study to explore the possibility of supporting the present system of selective examination of shipments in the Ethiopian Customs Authority using data mining techniques.

The study mainly aims to analyze the essential characteristics of seized customs fraud cases using datamining techniques. Analyzing such cases will help to find fraud patterns in seized commodities. The experience of other countries indicates that the analysis of seized fraud operations is an important element of objective risk evaluation that may reveal representative patterns of operations with propensity to fraud [24]. Such analysis should be realized, preferably, by automatic means, based on statistic or artificial intelligence models, due to the volume of information to be processed [24].

The output of this study will be complementary to the subjective risk assessment. Patterns that can be identified through analyzing fraud cases can be converted into rules. By analyzing the rules those rules with sound meaning or which seem to agree with the experts' opinions can be used as an input for the selectivity module of the ASYCUDA++ database.

Though the particular risk may vary from one country to the other, all countries are facing the problem of customs fraud and they are attempting to properly address the problem. An attempt is made to see the experience of other countries. One of the proposed models referred to was the Brazilian Customs risk management model. This model combines subjective and objective methods to determine the selection of an import operation to be checked. Objective risk is determined based on attributes that

characterize the merchandise being imported and the operation itself, which describe risk profiles or fraud parameters [24].

The output of the objective model is further supplemented by subjective analysis which employs risk indicatives like traders and clearing agents.

1.3 Objective of the Study

The general and specific objectives of the proposed study are the following.

1.3.1 General Objective

The general objective of this research is to explore the possibility of supporting the selective examination of shipments in the Ethiopian Customs Authority using data mining techniques. This is done by studying the essential characteristics of fraudulent transactions and identifying hidden patterns for the generation of rules that will support and improve the current selective examination system.

1.3.2 Specific Objectives

In addition, the specific objectives of the study are:

- To identify and analyze the type of data residing in the ASYCUDA database with the aim to use for customs risk assessment purpose.
- To assess the current classification of risk groups and the basis of this classification.
- To review literature on data mining technology in general and its application of data mining with reference to risk assessment in the customs sector.
- To explore and identify relevant attributes and patterns that can be used to enhance selective examination.

- To select the appropriate data mining tool to be used to develop the models.
- To collect clean and prepare the data in a format acceptable by the selected data mining tool.
- To develop models and evaluate their performance.
- Report the findings and forward recommendations.

1.3 Methods

The following methods were employed in conducting this research.

1.4.1 Literature Review

An intensive review of literature was made in the course of this study. Books, journals, magazines, manuals, WWW and other pertinent materials on data mining, customs and international trade have been consulted. In addition the following data collection methods were adopted.

1.4.2 Data Collection

1.4.2.1 Primary data collection

The primary source of information with regard to this study was the Ethiopian Customs Authority. Hence the researcher conducted an in depth discussion and interview with various experts from Enforcement, Operations and Automation departments of the Ethiopian Customs Authority.

1.4.2.2 Secondary data collection

The researcher had also gone through available documents related to the subject, and the user manual of the ASYCUDA software that are maintained by the organization.

The data for analysis was collected from two departments of the Ethiopian Customs Authority. These are the Automation and Data Processing department, the Enforcement and Legal Department. The Automation and Data Processing department is responsible for the management of the ASYCUDA database. The records of fraudulent transactions was collected from the Enforcement and Legal department

1.4.3 Data Preparation and preprocessing

The data collected was subject for data preparation. The first task in the data preparation activity was to tag[6] the fraudulent transactions on the transactional data. The other tasks included selecting the task relevant data and relevant variables for the modeling. The variable selection was done in consultation with the domain experts and considering the experience of similar studies [24]. Accounting for missing values, cleaning and checking for consistency was also performed.

1.4.4 Training and building models

The tool used for this study is knowledgeSTUDIO. The free availability of this software is the major factor for consideration. The KnowledgeSTUDIO software supports two types of decision tree algorithms, i.e., KnowledgeSEEKER, and

[6] The word tag or tagging is used to describe the process of classifying the dataset as fraudulent or not based on the fraud records.

HeatSEEKER. According to the developers KnowledgeSEEKER is a powerful, flexible algorithm that is especially good for exploration purposes and manual tree building [2].

KnowledgeSEEKER can handle a large amount of variables with either a continuous or discrete dependent variable [2]. On the other hand HeatSEEKER is a fast algorithm that is especially good for automatically generating a tree. When inserting a HeatSEEKER tree, one can have a number of options that control how the tree is automatically grown. HeatSEEKER can easily handle a very large number of records but performs better with fewer variables [2]. For this study mainly KnowledgeSEEKER is used.

Two different kinds of experiments were conducted. The first experiment was conducted using the dataset consisting of 10634 records. Out of these records only 170 cases belong to the fraud class. The remaining 10464 records were of the non-fraud class. The model built was not able to provide the desired result. This was due to the problem of class imbalance.

After recognizing the inherent problem of class imbalance for classification tasks, an attempt was made to overcome the class imbalance problem using over-sampling technique. By employing over-sampling five different datasets were prepared and different trees were generated and Comparison of their performance is also presented.

1.6 Thesis organization

The remainder of this thesis is organized as follows. Chapter two is devoted to the discussion of data mining. The history of data mining, the knowledge discovery process, the different applications and the reasons behind the popularity of data mining are discussed in this chapter. The final section of this chapter will discuss the tool employed in this research, i.e., decision tree.

The first part of chapter three, i.e., section 3.1 discusses risk management in general and risk management with in the context of customs. In 3.2 the ASYCUDA software will be elaborated followed by 3.3 the Ethiopian Customs Authority.

Chapter 4 focuses on the design of the system. In this chapter the declaration and clearance process is discussed. The structure and organization of a data warehouse and profile reports from the view point of risk analysis is also presented.

Chapter five presents a detailed discussion on the experiment conducted and the results obtained. In chapter 6 conclusions and recommendations for future study are presented.

CHAPTER TWO

DATA MINING

2.1 Introduction

We are in an age often referred to as the "information age". In this age information is reasonably quiet easy to capture. The storage cost is also fairly inexpensive. The digital revolution has resulted in the growth of collections of data both in size, and complexity.

Advancement in scientific data collection using remote sensors or from space satellites, the widespread introduction of bar codes for almost all commercial products and the computerization of many businesses and government transactions have generated a tremendous amount of data [17]. Stimulated by progress in computer technology and electronic data acquisition, recent decades have seen the growth of huge data bases in fields ranging from supermarket sales and banking through astronomy, particle physics, chemistry and medicine to official and government statistics [19]. The technological advances have resulted in our ability to meaningfully analyze and understand the data we gather lagging far behind our ability to capture and store these data [16]. A reasonable question that naturally arises as a result is what to do with such volume of data [16].

Data in itself provides no judgment or interpretation and therefore provides no basis for action. Putting data into context is what turns it into information. Connecting pieces of available information leads to the knowledge that can be used to support decisions [27]. The basic assertion is that it is certain that there is much valuable

information in them, information that has not been tapped, and data mining is regarded as providing a set of tools by which that information may be extracted [19]. This indicates that data mining is an inductive way of learning.

Of course many kinds of conventional reports and analysis are generated from these collections and many decisions are being made. Statisticians are basically concerned with the primary data analysis. In this case the data is collected with a particular question in mind [43]. The classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products [17]. How ever, large collections of data, however well structured, conceal implicit patterns of information that cannot be readily detected by conventional analysis techniques [27]. With a billion data points even a scatter plot may be useless [43].

The explosive growth of many business, government and scientific databases far outpaced our ability to interpret and digest this data, creating a need for a new generation of tools and techniques for automated and intelligent data base analysis [17]. These tools and techniques are the subject of data mining or knowledge discovery in databases. These techniques essentially seek to build a better understanding of data, and in building characterizations of data that can be used as a basis for further analysis [25], and extract value from volume [33]. In this context we can consider data mining as the process of secondary data analysis of large data bases aimed at finding unsuspected relationships which are of interest or value [19].

Data Mining is defined as the systematic extraction of non trivial, previously unknown and potentially useful information from large data bases. When we say previously unknown we mean that the information that will be generated as a result of the data mining task is not just a simple extraction of simple summaries that can easily be derived in any other traditional methods like observation or statistical summarization. It goes beyond such things and tries to find relationships between the various attributes that are hidden and could not be understood otherwise.

Now days the value of collecting data that indicate business or scientific activities to achieve competitive advantage is very well recognized. Powerful systems for collecting data and managing it in large databases are becoming common especially in large and mid-range companies. Now the major hurdle for companies is how to turn this data into meaningful information. In other words the challenge lies in the difficulty of extracting knowledge about the system you study from the collected data.

2.2 History of data mining

Data Mining is a relatively new discipline. It has only existed for about a decade [5]. Going back 50 years its origins can be traced to the early developments in artificial intelligence. At that time, developments in the areas of pattern recognition, rule based reasoning etc were providing the fundamental building blocks on which today's data mining is based. Although they were not given the name data mining at the time, many of the techniques that we use today have been in use, primarily for scientific applications [5].

Following a clear articulation of the business problem KDD process iteratively proceeds from raw data collections to some form of discovery of new knowledge. The process consists of the following steps [6].

Data discovery: after a clear definition of the business problem this step helps to get to know the data. It helps to identify what it covers and what it does not cover. This even helps to decide what other data sources to include in accomplishing the task.

Data Cleaning: is a phase in which noise data and irrelevant data are removed from the collection. Records in most of the cases contain erroneous data as a result of mistake in data entry, omission of values, inconsistencies etc. hence the KDD process cannot succeed without a serious effort to 'clean' the data [6].

Data cleaning needs to be addressed very cautiously. As [6] addresses it data cleaning is a double edged sword. This is because while in the process of data cleaning what seems to be an anomaly or an outlier may actually be a crucial indicator of an occurrence.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

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

Data transformation: also known as data consolidation is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: this is the most essential step. It is the crucial step in which various techniques are applied to extract patterns that are potentially useful.

Pattern evaluation: during this step interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: upon the successful completion of the knowledge discovery this is the phase where the discovered knowledge is visually represented to the user. This step uses visualization techniques to help users understand and interpret the data mining results.

Although the above mentioned exhaustive description is helpful to conceptualize the whole process, it is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Again Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. During the knowledge discovery process the iterative process may go from one step to the next and for some reason may come back to earlier steps. Again after the presentation of the discovered knowledge to the user, in order to enhance the evaluation measures or to further refine the mining output, new data can be selected or further transformed, or new data sources can be integrated. This iterative and dynamic approach can be used in order to get different, more appropriate results.

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since

mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

2.5 Tasks Solved by Data Mining

Based on the data mining task employed there are various kinds of patterns that can be discovered. Basically there are two types of data mining tasks, i.e., descriptive data mining (to describe the general properties of the existing data), and predictive data mining (that attempt to predict based on inference on available data).

The functionalities of data mining extend to a large variety of knowledge. Some of the functionalities and their description are discussed here under [17].

Classification: Classification analysis is the organization of data in predetermined classes. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches make use of 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 map or label new objects into one of the discrete classes. For example, after starting a credit policy, a manager could analyze the customers' behaviors with regard to their credit performance and label them accordingly as "safe", "risky" and "highly risky". In this case all the customers will be labeled in one of the three classes.

In this case the aim of the data mining task is to develop a model that will classify future credit applicants into one of the three classes. The model developed could then be used to aid the decision of either to accept or reject the credit requests.

Prediction: Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions. One can either try to predict some unavailable data values or trends, or predict a class label for some data. The latter one is very much similar to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of continuous numerical values.

Characterization: this is a descriptive task. It summarizes the general features of objects in a target class, and produces what is called characteristic rules [17]. For the purpose of characterization first the task relevant data to a user-specified class is retrieved. Next the retrieved data runs through a summarization module to extract the essence of the data at different levels of abstractions or granularity. For example, one may want to characterize the customers who regularly make transactions more than a given number of times in a year. Using concept hierarchies on the attributes describing the target class, the attribute oriented induction method can be used, for example, to carry out data summarization.

Discrimination: Data discrimination produces what are called discriminant 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 [17]. For example, one may

want to compare the general characteristics of the customers who frequently visited and made purchase from a given store with those whose visit and purchase is minimal. 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 between the two classes [17].

Association analysis: Association analysis is the discovery of what are commonly called association rules. They can predict any attribute, not just the class and this gives them the freedom to predict combinations of attributes too [42]. 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 pinpoint association rules. Association analysis is also known as market basket analysis. Using this technique is useful for processing transactional data in order to find groups of products that are sold together. One also searches for directed association rules identifying the best product to be offered with a current selection of purchased products or to design a well segmented and well targeted marketing campaign.

Clustering: This is the organization of data in to classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches. But the main principle behind all is the principle of maximizing the similarity between objects in a same class (intra-class similarity).

while minimizing the similarity between objects of different classes (inter-class similarity).

Outlier analysis: Outliers are data elements that cannot be grouped in a given class or cluster. There are cases where it is important to identify and intensively analyze the rare cases that are other wise can be rejected simply as outliers. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable. A good example may be an in depth study of an outstanding athlete or a person whose IQ is exceptionally high. Outlier events are also known as exceptions or surprises.

Evolution and deviation analysis: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent 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. In this kind of knowledge discovery it is common that users may not have a clear idea of the kind of patterns they can discover or need to discover from the data at hand. It is therefore important to have a versatile and inclusive data mining system that allows the discovery of different kinds of knowledge and at different levels of abstraction.

2.6 Reasons for the growing popularity of Data Mining

As mentioned in the introductory part of this section the main reasons that gave rise to the growing in popularity of the field of data mining are the growth in data volume, limitations of human analysis and the relatively low cost of machine learning. These factors are discussed as follows:

a. ***An escalating growth in data volume*** The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various business, scientific, and governmental organizations around the world is daunting. According to information from GTE research center, only scientific organizations store each day about 1 TB (terabyte) of new information. The academic world is also one of leading magpie of new data. Hence it is beyond the grasp of traditional analysis techniques to deal with such an overwhelming volume of data.

b. ***Limitations of Human Analysis*** There are two major problems that appear when human analysts process data. These are the inadequacy of the human brain to search for complex multifactor dependencies in data, and the lack of objectiveness in such an analysis. A human expert is always a hostage of previous experiences and bias. Sometimes this helps, sometimes this hurts, but it is a fact and almost impossible to get rid of it.

c. ***Low Cost of Machine Learning*** The rapid advancement in electronic computers has resulted in a rapid reduction of costs associated with an increase in

performance. Employing the use of computers or automated systems has a much lower cost than hiring an army of highly trained professional statisticians. While data mining does not eliminate or substitute human participation in solving the task completely, especially in decision making, it significantly simplifies the job and improves performance.

All of these have resulted in the growth of databases into larger and larger sizes. Today, databases can hold terabytes of data. Within these masses of data is buried information that is of strategic importance. As the saying goes, when there are so many trees, therefore, how can one draw a meaningful conclusion about the forest? This is the main motive behind data mining.

2.7 Decision Trees

2.7.1 Overview

Decision trees are one of the main techniques frequently used in data mining. Their use however is not limited to data mining. They are successfully employed in the field of artificial intelligence especially in expert systems. Using decision trees, one can make prediction or classify objects into a predetermined class. This is usually done by analyzing the relationship between the class variable and one or more predictor variables.

The available techniques in decision trees have much in common with the techniques used in the more traditional methods of discriminant analysis, cluster analysis, non parametric statistics and non linear estimation [10].

- Decision tree learning methods are robust to errors - both errors in classifications of the training examples and errors in the attribute values that describe these examples.

- The training data may contain missing attribute values.
  - Decision tree methods can be used even when some training examples have unknown values (e.g., humidity is known for only a fraction of the examples).

2.7.3 Decision Tree Representation

A decision tree is a predictive model that can be viewed as a tree. Specifically, each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification [23]

A decision tree can be used for data exploration in one or more of the following ways [10]

- To reduce a volume of data by transforming it into a more compact form which it preserves the essential characteristics and provides an accurate summary.
- Uncovering a mapping from independent to dependent variables that is useful for predicting the value of the dependent variable in the future. [10].

In general, decision trees represent a disjunction of conjunctions of constraints on the attribute-values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests and the tree itself to a disjunction of these conjunctions.

More specifically, decision trees classify instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute.

An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated at the node on this branch and so on until a leaf node is reached.

2.7.4 Advantages of Decision trees

The most appreciated practical advantage of decision trees is their cognitive nature [43]. In other words, they allow a human expert to easily understand the solution of a problem. Further more, the additional advantage of graphical representation of the tree representation makes it possible even for non-experts to read and interpret abstract notions and complex logistics. The practical value of cognitive simplicity should not be underestimated - it is well known that excellent modeling methods fail miserably in practice if excessively complex or abstract [10]. Perhaps for this reason, trees have a long tradition in supporting management decision.

In the context of knowledge modeling, decision trees have been from the very beginning used as a technique of inductive learning [43]. A decision tree is induced from a set of examples, whether to be used in a search process directly, or translated

into a set of decision rules for the developed knowledge-based system. In practice, the trees induced from examples often stabilize as further examples arrive, in which case they are said to have good generalization properties.

2.7.5 Limitations of Decision Trees

Decision trees are not at all with out any limitations. Perhaps the weakest point of decision trees in modeling is their sensitivity to changes in data, hence also to noise [43]. In contrast to the rather stable problems of learning of the classical kind, such as a feed-forward artificial neural network, the problem of learning decision trees is inherently ill-posed. Again when compared to the neural network they lack the ability to visualize the non linear relationship the may exist in the independent variables.

the context of customs. The last two parts of this chapter will present the ASYCUDA software and the Ethiopian Customs Authority.

3.2 Risk Management

3.2.1 The Nature of Risk

Before going into the discussion of risk management we will first focus our attention on the concept of risk. Risk is not associated only to institutions. Every individual experiences risk as a part of daily life. Risk is characterized by the fact that it is partially unknown and increases with the lack of knowledge [38]. Hence in order to understand and minimize risk knowledge is a key factor.

Another characteristic of risk is that it changes with time. However with the help of knowledge it is manageable. Managing risks is a natural reaction to minimize the level of potential harm that comes with not knowing what the future might hold [38]. There are many definitions proposed to describe and characterize risk. Some of the definitions are context free while others are contextual. The contextual definitions are some how based on the context free definition and modified to accommodate to the particular subject domain.

A simple context free definition of risk is the possibility of loss, injury, disadvantage or destruction [38]. Risk is defined as "...the potential of occurrence of an event or as the variability around the expectation of occurrence of an event" [32]. The first definition emphasizes the undesirability of risk while the second one illuminates the surprising element of risk. Both are inherent characteristics of risk.

adopting strategies to avoid, mitigate or otherwise accommodate the issues identified or any surprises.

Risk management is the sum of all activities within a system that intend to assess and acceptably accommodate the possibility of failures in elements of the operation. In other words assessing the elements of risks that are inherent with in a system and their consequences helps to handle them and ultimately to minimize them. Risk management should also constitute a proactive approach. When ever any risk element surfaces one has to learn from it and avoid its re-occurrence.

3.2.3 Risk Management in Customs

There is always the risk of customs circumvention. In some cases the deception may be for the sole purpose of tax evasion. Duties and taxes collected by customs are very important in providing sources of funds for the government budget. Due to this reason customs services are required to be highly productive. The impact of tax evasion is not limited to affecting the state revenue. It also negatively affects the law abiding traders because they will not be in an equal position to compete.

However the deception may go beyond the economic aspect. In such cases the threat is far damaging to the security of a nation. Hence customs administrations are endowed with a double edged responsibility. On the one hand they are required to improve their performance and provide efficient service while maintaining the control aspect as well.

Balancing these seemingly conflicting responsibilities is quiet a challenge. Most of the customs administrations in the quest to become more efficient and discharge their

responsibilities has undergone through a reform process. This reform is mainly geared towards the provision of efficient services, facilitation of international trade through transparency, cutting the bureaucratic red tape, and forming partnership with law abiding customers. The use of information technology has also played a pivotal role.

Similarly in order to efficiently discharge their duty with regard to the control aspect customs administrations utilize risk management. This involves identifying and realizing the various aspects of the problem and the associated risk, and then adopting strategies to avoid, mitigate or otherwise accommodate the issues identified according to some priorities suitable for the program.

Risk management is a favorite term in customs sector. Within the context of customs risk is the degree of exposure to the chance of non-compliance [38]. The integral components of a customs risk management program constitute risk assessment, profiling, and selectivity [32]. In a well established risk management system customs controls are handled at the pre-clearance, clearance and post clearance stage.

The general concept of risk assessment, profiling and selectivity is to select high risk consignments, travelers or importers for closer scrutiny, while facilitating the clearance of those that are low risk [32]. This obviously requires a good understanding of what constitutes a high risk. In practice a good portion of traders are considered as compliant and law abiding practitioners. Customs has accepted the fact that many importers have an act of complying with import laws and do not present a risk that justifies a significant allocation of resources [38]. Hence there is no point at all to use the limited human capital to perform limited inspections and reviews on all the imports [38]. The fact is that some imports present much more significant risk

than others, therefore it is clearly more effective to perform extensive through reviews and inspections on a smaller percentage of imports [38]. This logic leads to performing a risk analysis and risk assessment to determine who and what merits attention [38]. Profiling and proactive measures play a significant role.

Customs risks vary from one country to the other. The concerns of the customs risks of one country may be the smuggling of arms. Another country may have a severe problem of drugs. Yet a third country may highly emphasize the protection of the local industry by prohibiting the import of similar items. Hence customs administrations must first determine what constitutes a high risk [32]. Risk assessment is the tool to understand, as to what they are and the likelihood of their occurrence [32] and find out risks and their magnitude with in the context of the particular custom administration.

The second important component of risk management is profiling. Profiling is another concept much entertained with in the context of customs risk management. Profiling is a continuous method of registering and documenting with the aim of identifying the methods associated with fraudulent activities. The most effective source of information for determining profiles are standardized seizure reports that detail standard information related to past customs enforcement activity [32]. Findings of compliance as well as violations must be recorded and analyzed with in the risk management process [38].

As a basic requirement for a well handled profiling is the use of standardized seizure reports. Standardized seizure reports must contain standard fields of information from which profiles can be developed and risk indicators can be identified [32]. The key to

risk management is data [38]. As a result the importance of risk profiling is recognized and being implemented in many customs administrations. Some institutions explicitly state in their manuals that a risk profile should be prepared in each customs office where declarations are presented [13].

This is very important because in fraud activities there is most likely some form of modus operandi. Usually when ever a fraud tactic is adopted by illegal traders unless it is discovered they are likely to use it again and again. Even when the tactic is exposed in one custom point it will continue to be used at other custom entry points. For instance most often high duty cosmetics are brought under the disguise of medicines. In such cases of profiling, the mapping of the items concealed helps to exchange information from one custom point to the others so that the practice can be controlled properly.

Risk assessment and evaluation can be done on retrospective and prospective methods. In the retrospective method, infractions committed in the past are used to determine behavior standards for each infractor identify its modus operandi and find regular characteristics in these occurrences – risk profiles – which would help in determining future occurrences alike [24]. Since retrospective method is mainly based on the historical pattern of the fraud operations it is not able to forecast new circumstances nor detect new types of fraud.

Forecasting new kinds of methods and prevention of the occurrence of risk is done by the prospective method of risk evaluation. The prospective analysis tries to determine the way infractors adapt themselves to the new customs controls and to the changes on geopolitical and economical conditions [24]. It demands registering and analyzing

these critical events, which could affect the behavior of importers, like the alteration of the import tariff of a specific good or the concession of a new benefit [24].

Combining retrospective and prospective methods is of paramount importance to effectively implement an all round customs risk management. However the scope of this study is focused on the use of the retrospective method.

The main reason behind a consistent profile format is standardization. Standardization of information is very crucial for analysis purposes. Unless there is standardization essential elements may be omitted that will make the risk assessment less efficient. Profiling supports proactive intelligence work. For instance the risk associated with raising the duty on a particular type of good can be predicted based on past profiles [32]. Without tracking and reporting the risk management wheel goes flat, the process stalls and we never make forward progress [38].

As we have seen profiling is very crucial to collect the relevant data about operations and to understand the nature of frauds for risk assessment. However, there are also other very important sources of information. These are the external sources. With out the input of the data from these sources the risk assessment will suffer heavily.

The external sources are institutions like the revenue and tax authorities, ministry of Trade and similar authorities. The information that comes from these agencies mainly constitutes important information about traders and clearing agents. Customs risk is not a natural phenomenon. It is a man made event. It is the traders and other parties that orchestrate its happening. As a result what better way is there more than understanding the main players?

It is customary to assume that if a person intends to go out of the legal path and commit infringements in his other business activities or known to be a tax evader he will likely do it when ever there seem to be opportunities. Hence the basic reason behind the collection and analysis of data from external sources is to assess their track record in their business operations and their compliance behavior with regard to tax and the like. As a result modern customs organizations develop an integrated data warehouse from wide range of sources and employ it intensively in their risk assessment.

Customs control is a multi step process. Many custom administrations adopt the strategy of performing customs control in three stages. These are at the pre-clearance, clearance and post-clearance stages.

Pre-clearance, as the name implies, is a stage where prior to the arrival of the shipment, the traders present their documents. These documents are known as declaration documents or alternately as manifests. The main source of information and thus customs control at this stage is the declaration documents or passenger cargo manifests. These documents are reviewed in order to identify those transactions that match a profile and have other risk indicators [32].

The customs administrations must have a procedure in place in order to receive the documents in advance of the arrival of the shipment [24]. Additionally an importers' history database, i.e., the history of compliance or non-compliance, a data base of the values for various types of goods and other similar information is vital [24, 32]. The output of this stage is an evaluation as to the degree of risk associated to the shipment.

The second stage, i.e., the clearance stage control is conducted when the shipments arrive at the custom point. The analysis of the pre-clearance stage is the basic input for this stage. Based on the evaluation result the shipments will be subject to further control. Physical examination is the main tool of control at this stage. Review of other additional documents like accompanying documents explaining the specific nature of the commodities etc are also used. The evaluation of the pre-clearance stage helps to focus the search and to identify those that match a profile [32]. Profiling is the major output of this stage.

Post-clearance customs control is primarily post entry audit. During this stage of control customs can visit the traders' stores, review records and conduct financial audits. In developed societies customs organizations use post entry audit for the purposes of trade facilitation. When such system is put in place the traders will be relieved form physical examinations.

The over all purpose of choosing a method is to focus the limited amount of resource to effectively control and enforce the control activity while protecting the legitimate traders from the hassles of the examination process.

Modern customs now days are also adopting much advanced risk management methods. In such systems the traders as well as shipments will be categorized into various risk groups. The most common method widely used is to classify into three risk groups. These groups are known as high, medium and low. The high risk groups, as the name implies, constitute traders as well as consignments that present the highest level of risk. In such cases the shipments will be subject to an in depth or

exhaustive physical examination. The medium risk group is also subject to physical examination but at a less rigorous level. The low level risk groups are subject to only a small portion of physical examination or none at all.

Based on this method many customs organizations now a day have provisions for the automatic clearance of some shipments. There are many technical issues associated with this kind of provision. Post-entry or post-clearance control is the main tool for such customs institutions. However the technique as well as this kind of classification is beyond the scope of this research. Mention is done mainly to give an insight to the more advanced risk management methodologies.

To conclude this discussion, the purpose of risk management is to reduce the gap between the things that we know and that we don't know. A risk management program increases the knowledge about some subject and with this knowledge reduces the uncertainty in the procedures, and change behaviors and actions [38]. When risk management is handled in a systematic manner it can be successful.

3.3 The ASYCUDA database

ASYCUDA is an acronym for Automated SYstem for CUstoms DAta. It is a program funded by UNCTAD and has been in operation since 1985. The aim of this program is to provide Customs administrations with a modern, flexible and reliable tool capable of delivering a first class Customs service as well as accurate and timely trade statistics to governments [3].

The United Nations Conference for Trade and Development known as UNCTAD realizing the importance of the use of the technology in enhancing international trade

and the many similarities of the transactions and processes in customs operations worldwide developed a homogeneous platform for automation of Customs data [3].

As stated in [3] "the target was to provide Customs administrations ... with a modern, flexible and reliable tool capable of delivering a first class Customs service as well as accurate and timely trade statistics to governments". It has assisted many customs administrations to get rid of outdated procedures and practices and to incorporate international practices and standards and thereby increase customs revenues and improve clearance times [13].

Today, ASYCUDA has become a worldwide network of institutions, companies and people working together for development [3]. It is implemented in more than 80 countries. These include all regions of the world and of all economy. The ASYCUDA system has also proven its ability to bring substantial financial benefits of scale. This understandably is the economies of scale due to the application of similar software for many countries.

The benefits of the ASYCUDA system include [3]:

1. Assisting countries in international trade facilitation and bringing about a better management of Government finances through an institutional strengthening of Customs and to provide a reliable source of information on external trade.
2. To speed up the Customs clearance process through computerization and the simplification of procedures. It facilitates trade activity by implementing streamlined and simplified customs clearance procedures thus shortening the time of the clearance process.

country's foreign trade by introducing computer based procedures and is working towards providing more efficient services with regard to customs clearance.

In the past the authority has been accused of corruption and malpractices. It was also notorious for being inefficient. Although there is still a very large room for improvement and enhanced customer service, it has scored a recognizable achievement in the past few years.

Apart from the automation of the customs procedures the main reasons behind this achievement is stated as a continuous reform process geared towards more efficient service, the employment of qualified manpower, the implementation of clear and transparent rules and procedures. Due consideration is also given for preserving discipline and integrity of employees. As a result at present there is a time limit set for customs clearance. This ranges between two to five days.

The department which is responsible for the automation of the customs system is the Automation and Data Processing Department[10]. The Addis Ababa Customs office and the Addis Ababa airport customs offices are networked with the department. Hence the transactions of these two customs offices are entered into the database in real time basis. The other offices record their transactions in their system and periodically send their transactions to the department for updating on monthly basis.

The department is moving towards upgrading the system to the next version of the software known as ASYCUDA++ [3].

[10] Refer to the organizational structure in fig. 3

The present examination has another deficiency. As mentioned in the fist part of this section the examination is done on samples. How ever there is no standard procedure to determine either the sample size or the sample selection. Both are done on the subjective judgment of the person in charge. This subjectivity and lack of any standard measurement limits the system to properly define the level of examination conducted.

The lack of risk profiling is the other major problem that is not properly addressed in the Ethiopian Customs Authority. As will be explained in the following section the present risk profiling system of the Ethiopian Customs Authority is far from being of much use. At present fraud records are kept manually. There is no standard profile format. While trying to associate the fraud records with the transactional data in the ASYCUDA database the researcher has encountered many problems including repetitive declaration numbers and other inconsistencies.

Te researcher came to realize that on the side of the Ethiopian Customs Authority there is a move towards a more organized system of risk management and this includes a risk profiling system. However developing a proper risk profiling system requires adequate time and experience. There should also be a standard format that will include all the necessary variables needed for analysis. This can only be achieved through a systematic and meticulous approach.

CHAPTER FOUR

DESIGN OF SYSTEM

4.1 The declaration and clearance Process

At present the Ethiopian Customs Authority provides two alternate customs clearance procedures. These procedures will be discussed in the following sections.

Transit: under this scheme the trader presents all pertinent documents and self assessed information of how much tax he should pay. In such a case the Ethiopian Customs Authority will accept the self assessment of the trader and collects the tax payment. The detailed procedure is as follows

- ***a.*** The trader will present his import documents, detailed specification of the commodity and self-assessed tax amount to the customs office through his customs or transit agent
- ***b.*** The front office will receive the document and forward it to the customs officer
- ***c.*** The customs officer checks the documents and either accepts or rejects the document. (rejection will usually happen if documents are not complete or the correct tax rate is not applied)
    - ***i.*** If rejected the document will be returned to the trader
    - ***ii.*** When accepted the customs officer will register the transaction into the ASYCUDA database. And sends the document to the assessors.

d. The assessor will verify the assessment presented by the trader and if accepted will send it to the cashier and another copy to the customs agent.

e. The cashier collects the payment and sends the document to the customs officer

f. The customs officer registers the duties and taxes collected. Then sends the document to the examination section.

g. The customs agent initiates transit using the cash receipt obtained from the cashier

h. Upon the arrival of the commodity the shipment will be inspected and if any deviation is found measures will be taken.

In the transit process all taxes and other payments will be paid before the arrival of shipments. The maximum time limit for this process is two days.

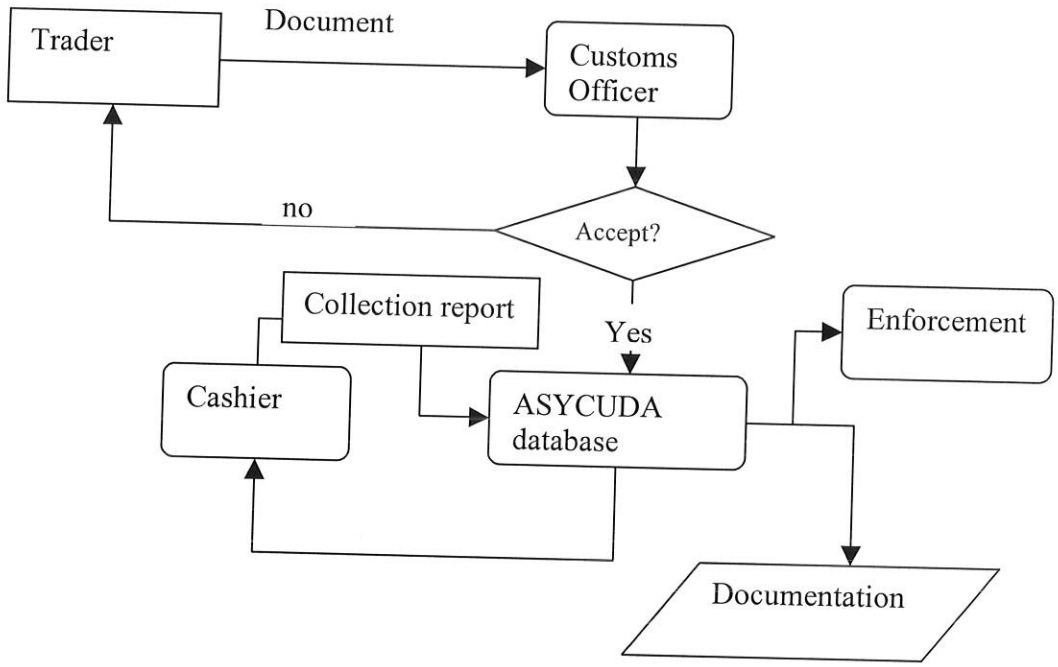

Figure 4: The Transit Import Process

In the second process the trader will not specify as to the details of the commodity and no self assessment will be submitted. For such kinds of transactions the document will be registered and a deposit amount that is approximately similar to the value of the declared item will be retained by the customs. In this instance assessment is done by the customs officers after the commodity arrives at the customs warehouse. The time limit for these kinds of transactions is five days.

4.2 The ASYCUDA selectivity module

Selecting the examination procedure for the goods is assisted by the selectivity module of the ASYCUDA++ database. Based on information in the rules files the ASYCUDA++ system will allocate a channel status of either green or red. In the case

of red the cargo will be released without any examination. In the case of red the cargo will be released only after the result of the physical examination is reported back to the system by authorized person.

The rules file is built using sound and meaningful rules generated through both the subjective and objective methods.

Modern customs are implementing the highly challenging task of implementing a data warehouse to acquire and utilize data from heterogeneous sources. The proposed project aims to get all the data available in different parts of the organization to work together and be stored in a central repository for decision-making purposes at all levels.

Data warehousing is currently being used as a strategic tool across thousands of organization to aid them in their decision making process. The data warehouse serves as a consolidated source of information, which facilitates easier, faster, timely and integrated decision making within the organization.

Profiling: This is another source of information. A standard profile format that consists of relevant fields will be maintained by the risk management unit. A record of compliance is also important to identify those traders with a clean track record. When companies preserve a clean track record for a long period of time they should receive good treatment by the customs system

The profile report, which is the output of the physical examination process, helps the system to update itself. Both compliant and non-compliant cases should be recorded by the physical examination system. When a sufficient number of new fraud cases are collected the new models with these records will be conducted. This is very important to update the system in a continuous manner and to cope up with new fraud patterns.

CHAPTER FIVE

EXPERIMENTATION

This chapter describes the experimentation and evaluation of the study. The chapter also presents discussion on the acquisition of data as well as the techniques and procedures used during the preprocessing.

5.1 Data Collection

Data mining principally depends on data. A precondition to any data mining task is the data itself [1]. This requires understanding of the data. The data employed for this experiment was obtained from two departments of the Ethiopian Customs Authority. The major component of the data was obtained from the Customs Automation department[11] while fraud records were acquired from the Risk Management unit[12] of the Enforcement department.

The Customs Automation department provides various statistical extractions for a large number of external users like government offices, businesses, research and similar institutions. One of the statistical extractions is a table known as Eurotrace. It is a tabular format to extract data from the ASYCUDA database for the purpose of providing it to external users like Ethiopian Statistical Authority, Ministry of Trade etc. Eurotrace is a concatenation of three different tables namely Declaration based,

[11] This is the department responsible for management of the ASYCUDA database

[12] This unit is formed as a result of the re-establishment program of the Ethiopian customs Authority. It has been in operation for only five months.

Item Based and Tax based data for each transactional item. The three tables can either be extracted separately or in concatenated form based on users' request.

For the initial familiarization of the data and exploration of the content a transaction record of six months that contains all the variables was obtained. This data was provided to the researcher in Excel format. With a size of 100 megabytes the data consists of 45,522 records with 140 attributes. Familiarization and understanding of this data has taken quiet a good part of the research time.

An in-depth exploration of the data and frequent consultation with the ASYCUDA staff has revealed that a good part of the variables were redundant. This was primarily due to the reason that the three tables were concatenated and the extraction was done without any filtering for the variables that are part of all the three tables. For instance the unique record identifier was a concatenation of nine variables. The variables are D23OFF, D05Y, D04M, D03D, D60SR, D21CPC, D61DNO, D62INO, and DTYPE.0 All of these variables were included in the dataset. Hence removal of this repetition reduced the dimension by 18 columns.

The importance of generating a single variable that serves as a key identifier was recognized at this level. Some or all of these variables may not be relevant for the analysis. On the other hand a key field that is used as record identifier is necessary. Since employing the use of nine variables as a key was pointless generating a key identifier helped to evaluate the variables independent of their use as key identifier.

The dataset was also explored to look for missing values, inconsistencies and other interpretable observations. The data collected had a large number of variables with

values missing. The following table summarizes the variables and the associated number of missing values.

NO.	Attribute	Percentage of total
1	D50LOC	100%
2	D90TAX6	100%
3	D90BASE6	100%
4	D90RATE6	100%
5	D90RATAMT6	100%
6	D27ACCNT	100%
7	D29PREC	100%
8	D45FINDEST	100%
9	D34REF	100%
10	D48AGREEM	100%
11	D86RATAMT2	77%
12	D89TAX5	72%
13	D89BASE5	72%
14	D89RATE5	72%
15	D89RATAMT5	72%
16	D88TAX4	47%
17	D88BASE4	47%
18	D88RATE4	47%
19	D88RATAMT4	47%
20	D42MANLINE	5%
21	D68INFOG	1%

Table 1 Variables with missing Values[13]

In the above table the first 10 attributes does not contain any value[14]. The attributes listed from number 11 to 19 are all tax related variables and most of them again contain a large number of missing values. While the values indicate the applied tax

[13] Refer to Appendix II for a complete list of the Variables and their description

[14] Since the ASYCUDA database is developed as a general system some variables may not be applicable for a particular country. Again the applicability of some variables may change over time. For instance tax related variables may not be applicable at a particular time.

for the particular item the absence of a value indicates the tax is not applicable for the particular item. In other words this means those attributes denoting taxes and that indicate missing values are equivalent to zero. Hence the missing values are filled with zero values.

The remaining two values have fewer numbers of missing values. Both attributes are nominal attributes. D42MANLINE represents a bill of lading number and D68INFOG is a column to write any kind of general information by customs officers while entering the manifest information. For nominal variables filling the missing value with the modal class is appropriate. However these two variables were removed due to their irrelevance for the analysis task.

During the study it was also found out that 33 variables have only one type of value[15]. Since an attribute with only one type of value have no contribution in the classification process all the 33 variables were removed from the dataset.

The aforementioned and other preprocessing tasks were performed as a data familiarization step. This has helped to understand the data residing in the ASYUDA database and to focus on the right attributes while obtaining the data during the second extraction.

The other source of data was the Risk Management Unit. Since the Enforcement and Legal department discharges the control aspects of the customs functions it is supposed to maintain records of the control operations. At present there is no standard profile format. The department maintains fraud records in a manner that is far from being systematic. The reports are recorded in a text like format. Apart from

[15] See appendix III for details

computing some summaries like the total number of fraud cases in a given period of time, the total value of items etc no other statistical analysis is possible.

Moreover a recent record was not available. The only 'compiled' record that could be found was a report on the Addis Ababa Customs branch for the Ethiopian fiscal year[16] 1993 containing 420 fraud cases.

This report was claimed to be a highly 'classified' document. The declaration numbers (also known as the manifestation numbers) of the fraudulent documents was the only information available to the researcher. This is the data used for tagging[17] fraudulent transaction.

Finally, based on the tagging data the transactional data of the 1993 Ethiopian fiscal year was obtained from the ASYCUDA database. During the second extraction the redundant variables and other non-relevant[18] variables were excluded. Moreover since the fraudulent transactions were exclusively that of the Addis Ababa Customs branch the data extracted was accordingly that of the Addis Ababa Branch.

The data obtained during this time was extracted from two different files. In the ASYCUDA database each file contains a single year transactional data. The year begins on January 1st and ends on December 31. On the other hand the Ethiopian fiscal year[19] begins on July 7 and ends on July 6. As a result the first six months data

---

[16] The Ethiopian Fiscal year begins on July 7 and ends on July 6. The aforementioned period was from July 7, 2000 to July 6, 2001

[17] The word tagging is used to describe the process of classifying the dataset as fraudulent or not based on the fraud records. This is the only use of the data acquired from the risk management unit.

[18] Warehouse no. location, particular shelf location in the warehouse and similar variables that are maintained for facilitating the internal handling of shipments are non relevant and excluded from the dataset.

[19] All government institutions and a good number of other institutions also use the Ethiopian fiscal year

from the year 2000 file containing 24,748 records and the other half from the 2001 file containing 35,982 records was obtained.

5.2 Data preprocessing

The data that will be used for modeling purpose had to be carefully prepared. This is the stage where most of the sanitization has been done. As is well known real world data is subject to noise and inconsistency. Missing data is also common. Such problems arise during acquisition, entry or updating. Another cause for such problems is the size of the data [19]. Maintaining data quality is a very challenging task and the challenge grows exponentially as the size of the data grows.

In this study, maintaining the data quality begun by ***preliminary transformation*** of the data. During this process the first task was to join the two separate records and create one table. This was done by copying the records of the second period, i.e., year 2001, and adding it to the records of the first six months period of the year 2000. The total of the two records became 60730. Following the creation of a unified dataset a sequential number that serves as a key field was generated.

The tagging data obtained from the risk management unit consisted of the declaration document numbers. Practically a single declaration document can contain a number of different items. The number of items in a single declaration ranges from 1 up to 96.

During the tagging process this created a problem. The dataset obtained from the ASYCUDA database was at item level. This means when a single declaration document containing, for example, 10 different items is tagged to the dataset there will be 10 entries. When tagging the records, based on the fraudulent transaction

Statistics	
Total records	2657
Correctly predicted	2615
Percentage	98.42%

Figure 7: The Accuracy of the First Experiment

The accuracy of this experiment, i.e., 98.42% was impressive. This is a classical decision tree model shortcoming when used with imbalanced[24] data sets. The above classifier simply assigned all the cases to the majority class, i.e., F-Found=0. When one class is rare, decision tree algorithms often prune the tree down to a single node that classifies all instances as members of the common class leading to poor accuracy on the instances of minority class [35].

Measuring the adequacy of a model solely on accuracy may not be sufficient. Fraud detection data being imbalanced is the norm. Usually there are many more legitimate than fraudulent examples. This means that by predicting all instances to be legal, a very high success rate is achieved without detecting any fraud [3]. Such a model would be literally accurate but practically worthless [12].

To overcome the shortcoming of the accuracy measure confusion matrix [34] can be used. For a two class problem this can be done using a 2 by 2 matrix. In the confusion matrix it is customary to use the columns as the predicted class and the rows as the actual class.

[24] Imbalanced datasets or class imbalance occurs when the different classes in the dependent variable do not have equal number of cases.

	Predict negative	Predict positive
Actual negative	TN	FP
Actual positive	FN	TP

Table 2 Confusion matrix

In the confusion matrix TN is the number of negative examples correctly classified (True Negative), FP is the number of negative examples erroneously predicted as positive (False Positive), FN is the number of positive examples incorrectly predicted as negative (False Negative), and TP[25] is the number of positive examples correctly classified (True Positive). The ideal situation here is to get both TP and TN correctly. However in the absence of attaining the ideal situation it may be more interesting to accurately classify one of the classes. In other words the misclassification cost associated with the different classes may not be equal. In a classification problem with imbalanced datasets the rare event is usually the more interesting class. Paradoxically a classifier induced from an imbalanced data set has, typically, a low error rate for the majority class and an unacceptable error rate for the minority class [35].

The confusion matrix of the first experiment as shown below has the highest error for the minority class while correctly classifying the majority class.

Confusion Matrix - F_Found

		Predicted	
		0	1
Actual	0	2615	0
	1	42	0

Figure 8: Confusion matrix of the first experiment

[25] In our case we consider positives as fraudulent transactions

The lack of satisfactory performance of this experiment was mainly the result of class imbalance problem. Before addressing the experiment with regard to this claim it may be more appropriate to point out some facts about the essential characteristics of class imbalance.

Datasets are said to be balanced if there are approximately as many positive examples of the concept as there are negative ones [22]. The above explanation implicitly assumes the problem as a two class problem. However in general the class imbalance problem occurs when, in a classification problem, there are many more instances of some classes than others.

The problem of class imbalance is that while addressing a classification problem standard classifiers tend to be overwhelmed by the large classes and ignore the small ones [35]. This is because those classifiers attempt global quantities such as the error rate, not taking the data distribution into consideration [22]. As a result examples from the majority class are well classified whereas examples from the minority class tend to be misclassified [22].

This problem is prevalent in many applications, including: fraud, intrusion detection, risk management, text classification, and medical diagnosis [35], helicopter gearbox fault monitoring, detection of oil spills, detection of fraudulent telephone calls etc [22]. In certain domains the class imbalance is inherent to the problem. For example, within a given setting, there are typically very few cases of fraud as compared to the large number of honest use of the offered facilities [35].

The problem of class imbalance has been an active research area. Addressing the problem of class imbalance grew as more and more researchers realized that their

data sets were imbalanced and that this imbalance caused suboptimal classification performance [22]. This increase in interest gave rise to two workshops held in 2000 and 2003 at the AAAI and ICML conferences, respectively [35]. At present there are quiet a number of issues that are not sufficiently addressed.

Out of the many learning methods affected by class imbalance decision trees are found on the top of the list [22]. This is mainly because decision tree learning is highly influenced by the majority class in the tree building process. Again during the pruning process majority classes have higher chances of influencing newly generated nodes.

All the above discussion supports the assertion that the reason for the failure of the first experiment is the problem of class imbalance. First of all the in the dataset the proportion of fraud cases to non-fraud cases is 98.5 to 1.5. Thus the degree of class imbalance can be considered very high[26].

Another important point to be noted is the accuracy measure of the test result is similar to the ratio of the two classes. The decision tree by simply identifying all the cases to the majority cases has achieved this result. If the problem was the accurate classification of the majority class this tree would solve it. However in most of the cases the minority classes are the most interesting. Learning from data sets that contain very few instances of the minority (or interesting) class usually produces biased classifiers that have a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class [34].

[26] In a study conducted on class imbalance problem the ratio of 32 : 1 was considered as a case of highest class imbalance

An attempt was made to analyze the nature of the error with regard to the fraud cases. Out of the 131 fraud cases the tree built by dataset 1 misclassified 51 cases. The tree built on dataset 2 misclassified 19 cases and the third tree misclassified only 5 cases and correctly classified the remaining 126 fraud cases. This tree has attained more than 95% accuracy on the fraud cases.

The attempt to overcome the class imbalance problem through over-sampling provided some result. Further attempt to analyze these results based on the rules and in consultation with the domain experts was earlier planned. This kind of analysis will help to assess the soundness of the rules. However due to the time limitation this analysis was not done and is left for future work.

One of the problems of duplication is that a relatively higher duplication of the minor class will create a bias so that the classifier will learn more of the duplicated instances. When the classification task requires accuracy of classifying both classes duplication poses some problem. Over-sampling would shift the error distribution from false positive to false negative. This will not be of much help in cases where it is essential to preserve a low false negative error rate while learning the false positive error rate. This is not a problem for this study. While the misclassification cost for fraudulent classes is being harmful, the misclassification of the other class doesn't pose any problem. Rather this misclassification can be considered as a benefit of the classifier because at the end of the day the classifier is wanted to point out transactions that are similar to those transactions identified as fraudulent. The physical examination process will ultimately indicate whether the transactions are fraudulent or not.

CHAPTER SIX

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

The Globalization process, accelerated by information technology and other enabling elements has brought an extraordinary increase in the international flow of people, goods, capital and information. Every country in the world is involved in this flow.

The role of governments has changed radically. To accommodate this change many governments, especially in the developing countries are engaged in a reform process towards liberalization and provision of efficient services.

Of the many governments customs are found in the forefront of this scenario. Customs Organizations play a key role. Customs procedures have a powerful impact on the economic competitiveness, and the modernization of its procedures must be a target for any country.

Customs administrations are in the forefront of the liberalization and reform process. Traditionally customs administrations, especially in developing countries, are characterized by inefficiency, corruption and highly bureaucratic procedures and outmoded work habits. An inefficient customs administration jeopardizes the prosperity of the nation by driving away business and investment. On the contrary a modern efficient Customs operation can contribute to development through efficient services to facilitate trade.

Customs Organizations, in this new environment, must also be more effective in the protection of the economy and the welfare of the society. In order to strike a balance

6.2 Recommendation

While approaching this problem many sub problems have been encountered. Many of them require a full research by their own merit.

1. An integrated data warehouse is a basic requirement for any modern customs organization. This data warehouse will collect important information from a wide range of external sources. The implementation of such a data warehouse will strengthen the risk assessment capability there by enabling it to address the problem proactively.
2. The enrichment of the internal data collection should be based on standard profiling formats. Such profile formats should be designed in such a way that important variables that are required for a proper risk assessment should be included.
3. Applying and fully implementing the above recommendations require a reasonable amount of time. In the short run the selectivity of physical examination should be more systematic. The small experiment conducted in this study has shown promising results. By analyzing the rules generated and even employing those rules for selecting shipments for physical examination will help to enhance the selection process to be more focused.

References

1. Aldana, A. (2000). *Data Mining Industry: Emerging Trends and New Opportunities.* Masters thesis at the University of MIT. Available at URL: http://scanner-group.mit.edu/PDFS/Aldanaw.pdf
2. Angoss Software Corporation. (2001). KnowledgeSTUDIO Data Mining User Guide. Available at URL: http://www.angoss.com
3. ASYCUDA WORLD. Unpublished UNCTAD document
4. Baker, S.W. (2002) Customs Risk Management Program Responds to Security Needs. Available at URL: http://www.swbakerlaw.com/0503.htm
5. Baragoin, C., Andersen, C.A., Bayerl, S., Bent, G., Lee, J., & Schommer, C. (2001). *Mining Your Own Business in Telecom Using DB2 Intelligent Miner for Data.* Available at URL http://www.ibm.com/redbooks
6. Brachman, R.J., and Anand, T. (1996). *The Process of Knowledge Discovery in Databases: A Human-Centered Approach.* In Advances in Knowledge Discovery and Data Mining, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, AAAI Press/The MIT Press, Menlo Park, CA. pp. 37-57.
7. Berry, M. J. A. and Linoff, G. (1997). Data Mining Techniques: for Marketing, Sales, and Customer support. New York; John Willy & sons, Inc.
8. Chen, H. *Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms.* Available at URL: http://ai.bpa.arizona.edu/papers/mlir93/mlir93.html

9. Chen, H. P., Hsu, R., Orwig, L., Hoopes, and Nunamaker, J.F. (1994). *Automatic concept classification of text from electronic meetings. Communications of the ACM*, 37(10): 56-73.

10. Classification trees. Available at URL: http://www.statsoft.com.textbook/stclatre.htm

11. Costa, E. (2000). *Classification problems: an old question with new solutions* Available at URL: http://www.dcs.napier.ac.uk/coil.summerschool/docs/problem_costa.pdf

12. CRISP-DM. (2000). CRISP-DM 1.0 step-by-step data-mining guide. Available at URL:// www.crisp-dm.org

13. Customs challenges for developing countries. Available at URL: http://www.wto.org/english/tratop_e/cusval_e/seminar_nov02_e/kunda_e.pdf

14. Customs related technical assistance for trade capacity building. A resource guide. Available at URL: http://www.nathanninc.com/NATHAN/files/ccPageContextdocfilename149116705546Customs(final).pdf

15. Enabling Environment for LDCs. Paper 2. Available at URL: http://www.intracen.org/bsrt/enablenviron.pdf

16. Fayyad, U; Piatesky-Shapiro, G and Smyth, P. (1996). *From data mining to knowledge Discovery in Data Bases.* AI magazine. Available at URL: http://www.Kdnuggets.com

9. Chen, H. P., Hsu, R., Orwig, L., Hoopes, and Nunamaker, J.F. (1994). *Automatic concept classification of text from electronic meetings. Communications of the ACM*, 37(10): 56-73.

10. Classification trees. Available at URL: http://www.statsoft.com.textbook/stclatre.htm

11. Costa, E. (2000). *Classification problems: an old question with new solutions* Available at URL: http://www.dcs.napier.ac.uk/coil.summerschool/docs/problem_costa.pdf

12. CRISP-DM. (2000). CRISP-DM 1.0 step-by-step data-mining guide. Available at URL:// www.crisp-dm.org

13. Customs challenges for developing countries. Available at URL: http://www.wto.org/english/tratop_e/cusval_e/seminar_nov02_e/kunda_e.pdf

14. Customs related technical assistance for trade capacity building. A resource guide. Available at URL: http://www.nathanninc.com/NATHAN/files/ccPageContextdocfilename149116705546Customs(final).pdf

15. Enabling Environment for LDCs. Paper 2. Available at URL: http://www.intracen.org/bsrt/enablenviron.pdf

16. Fayyad, U; Piatesky-Shapiro, G and Smyth, P. (1996). *From data mining to knowledge Discovery in Data Bases.* AI magazine. Available at URL: http://www.Kdnuggets.com

Appendices

Appendix I: List of Customs Offices under the Ethiopian Customs Authority

1. Addis Ababa Airport Customs
2. Addis Ababa Lagare Customs Office
3. Addis Ababa Parcel Post customs
4. Alamata Customs
5. Bahirdar Customs
6. Combolcha
7. Dire Dawa Airport Customs
8. DireDawa Lagare Customs
9. Dewele
10. Djibouti (Coordination office)
11. Gondar
12. Humera
13. Jijiga
14. Mekele
15. Metema
16. Moyale
17. Nazareth
18. Rama
19. Zala Anbessa

Appendix II: Variables and their description in the ASYCUDA database

filed name	description	type	width	dec
D23OF_A	office	character	4	
D05Y_A	registration year	character	2	
D04M_A	registration month	character	2	
D03D_A	registration day	character	2	
D60SR_A	serie number	character	1	
D21CPC_A	custom code CRC	character	2	
D61DNO_A	registration number	character	5	
D62INO_A	item number	character	2	
DTYPE_A	type (R REG, A ASS. S-TO)	character	2	
D22CPC	custom code CPC	character	2	
D14COD6	tarrif code (in 6 digit)	character	6	
D14COD4	tarif code (in 4 digit)	character	4	
D44CY	country origin	character	3	
D51MASNET	net mass	Numeric	11	2
D52MASGRO	gross mass	Numeric	11	2
D53QUANT	quantity	Numeric	11	2
D76FOBNC	FOB national in birr	Numeric	11	2
D54CU	FOB in declared currency	character	3	
D71CUVAL	customs value	Numeric	11	2
D73INSCO	insurance	Numeric	11	2
D74FRTCO	freight	Numeric	11	2
D75OTHCO	other cost	Numeric	11	2
D37LICENCE	licence number	character	11	
D38NOPACK	number of packages	Numeric	5	
DMARKNO	mark and no of packages	character	35	
DATTACHDOC	attached dcuments	character	34	
DADJUST	adjustment coefficient	character	6	
D78ASTDE	total assessd/ rembursed	character	11	
D79TOTAX	tax total	Numeric	2	
DQTYTAX	quantity of taxes	Numeric	11	2
D48AGREEM	agreement	character	17	
D50LOC	localization	character	4	
D39NATPACK	nature and package code	character	11	
D80LOSS	tax loss	character	8	
HS_CODE	Commodity code	Numeric	18	
FOURDIG	commodity code1	character	4	
CHAPT	chapter	character	2	
CPC	CPC	character	4	
D23OF_B	office	character	4	
D05Y_B	registration year	character	2	
D04M_B	registration month	character	2	
D03D_B	registration day	character	2	
D60SR_B	serie number	character	1	
D21CPC_B	custom code CRC	character	2	

Appendix II
Cont'd

filed name	description	type	width	dec
D61DNO_B	registration number	character	5	
D62INO_B	item number	character	2	
DTYPE_B	type (R REG, A ASS. S-TO)	character	2	
D85TAX1	duty tax code	character	2	
D85BASE1	duty tax base	character	2	
D85RATE1	duty tax rate	character	6	
D85BASECAL	duty base calculated	Numeric	11	
D85DTVAL01	duty tax value	Numeric	11	
D85RATAMT1	duty rate amount	character	6	
D85TAXAMT1	duty tax amount	Numeric	11	
D86TAX2	excise tax code	character	2	
D86BASE2	excise tax base	character	2	
D86RATE2	excise tax rate	character	6	
D86BASECAL	excise base calculated	Numeric	11	
D86DTVAL02	excise tax value	Numeric	11	
D86RATAMT2	excise rate amount	character	6	
D86TAXAMT2	excise tax amount	Numeric	11	
D87TAX3	sales tax code	character	2	
D87BASE3	sales tax base	character	2	
D87RATE3	saes tax rate	character	6	
D87BASECAL	sales base calculated	Numeric	11	
D87DTVAL03	sales tax value	Numeric	11	
D87RATAMT3	sales rate amount	character	6	
D87TAXAMT3	sales tax amount	Numeric	11	
D88TAX4	sur tax code	character	2	
D88BASE4	sur tax base	character	2	
D88RATE4	sur tax rate	character	6	
D88BASECAL	sur base calculated	Numeric	11	
D88DTVAL04	sur tax value	Numeric	11	
D88RATAMT4	sur rate amount	character	6	
D88TAXAMT4	sur tax amount	Numeric	11	
D89TAX5	duty tax code	character	2	
D89BASE5	duty tax base	character	2	
D89RATE5	duty tax rate	character	6	
D89BASECAL	duty base calculated	Numeric	11	
D89DTVAL05	duty tax value	Numeric	11	
D89RATAMT5	duty rate amount	character	6	
D89TAXAMT5	duty tax amount	Numeric	11	
D90TAX6	duty tax code	character	2	
D90BASE6	duty tax base	character	2	
D90RATE6	duty tax rate	character	6	
D90BASECAL	duty base calculated	Numeric	11	
D90DTVAL06	duty tax value	Numeric	11	
D90RATAMT6	duty rate amount	character	6	
D90TAXAMT6	duty tax amount	Numeric	11	
D80LOSSTAX	loss of revenue	Numeric	11	
DUTY	duty	Numeric	10	
MONTH	month	character	2	

Appendix II
Continued

filed name	description	type	width	dec
D23OF_C	office	character	4	
D05Y_C	registration year	character	2	
D04M_C	registration month	character	2	
D03D_C	registration day	character	2	
D60SR_C	serie number	character	1	
D21CPC_C	customs number	character	2	
D61DNO_C	item number	character	5	
DTYPE_C	type (R REG, A ASS. S-TO)	character	2	
D62ITEMQTY	Number of items in one declaration	Numeric	2	
D33FRONT	Frontiers office	character	2	
D34REF	reference number	character	4	
D25IMPEX	importer exporter	character	5	
D24DCLAR	declarant	character	17	
D27ACCNT	account holder	character	17	
DEXAM	examining officer	character	17	
D43CY	country const destination	character	5	
D41NAT	nationality of transport	character	3	
D32BANK	bank and branch number	character	3	
D29PREC	suceedng declaration number	character	18	
D301WHDUR	duration in ware house	character	10	
D45FINDEST	final destination	character	3	
D35MANIF	manifest number	character	3	
D67ASSDATE	assessment date	character	17	
D66ASSNO	serie and assessment number	character	6	
DSTATUS	status	character	13	
D57POSTCHI	access key post entry (child)	character	2	
D57POSTPAR	access key post entry (par)	character	15	
D57POSTINI	ACC. K of initial par. Of post	character	15	
DSTATREIMB	status reimbursed assess	character	2	
D78TOTASS	amount reimbursed assess	Numeric	11	
D301WHCODE	ware house code	character	5	
DNOGLOBTAX	quan. Of global tax code	Numeric	2	
D26CONS	consignee	character	18	
D46TPACD	terms of payment	character	3	
D47TODCD	terms of delivery	character	3	
D42MANLINE	B/ lading (manifest item) no	character	17	
D31SHED	shed number	character	6	
D40MT	mode of transport	character	4	
DNEWNO	new reference number	character	8	
DTOTMASS	total gross mass	Numeric	11	
DTOTFOB	total FOB	Numeric	11	
DTOTFRET	total freight	Numeric	11	
DTOTINS	total insurance	Numeric	11	
DTOTCOST	total other costs	Numeric	11	
D68INFOG	general information	character	38	
MONTH	month not in asycuda	character	2	

Appendix III: Attributes with one type of Value

1. D23OFF
2. D60SR
3. D21CPC
4. D62INO
5. DTYPE
6. D22CPC
7. D37LICENCE
8. D78ASTDE
9. D80LOSS
10. CPC
11. D85TAX1
12. D85BASE1
13. D86TAX2
14. D86BASE2
15. D87TAX3
16. D87BASE3
17. D88TAX4
18. D88BASE4
19. D89TAX5
20. D89BASE5
21. D90BASECAL
22. D90TDVAL06
23. D90TAXAMT6
24. D80LOSS
25. DUTY
26. D62ITEMQTY
27. D33FRONT
28. D301WHDUR
29. DSTATUS
30. D57POSTCHI
31. DSTREIMB
32. D301WHCODE
33. DNEWNO

Declaration

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

_____

Girma Belew

July 2004

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

_____

Dr. B. L. Desai
July 2004