

*Addis Ababa  
University  
(Since 1950)*



**ADDIS ABABA UNIVERSITY  
COLLEGE OF NATURAL SCIENCE  
SCHOOL OF INFORMATION SCIENCE**

**An Integration of Prediction Model with Knowledge Base System for Motor  
Insurance Fraud Detection: The Case of Awash Insurance Company S.C**

**By  
ABDI CHALI**

A Thesis Submitted to the College of Natural Sciences of Addis  
Ababa University in Partial Fulfillment of the Requirements for the Degree of  
Master of Science in Information Science

February 2016  
Addis Ababa, Ethiopia

**ADDIS ABABA UNIVERSITY**  
**COLLEGE OF NATURAL SCIENCE**  
**SCHOOL OF INFORMATION SCIENCE**

**An Integration of Prediction Model with Knowledge Base System for Motor  
Insurance Fraud Detection: The Case of Awash Insurance Company S.C**

**By**  
**ABDI CHALI**

February 2016  
Addis Ababa, Ethiopia

**ADDIS ABABA UNIVERSITY**  
**COLLEGE OF NATURAL SCIENCE**  
**SCHOOL OF INFORMATION SCIENCE**

**An Integration of Prediction Model with Knowledge Base System for Motor  
Insurance Fraud Detection: The Case of Awash Insurance Company S.C**

**By**  
**ABDI CHALI**

February 2016

Name and signature of Members of the Examining Board

---

Chair person, Examining Board

---

Signature

---

Advisor

---

Signature

---

Examiner

---

Signature

---

Examiner

---

Signature

## Declaration

This thesis is my original work and has not been submitted as a partial requirement for a degree in any university.

---

ABDI CHALI KEJELA

February 2016

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

DEREJE TEFERI (PhD)

---

## DEDICATION

I would like to dedicate this thesis work to my family and to all my friends.

## ACKNOWLEDGMENT

First and foremost my special thanks goes to the almighty God for his kindness, blessings and forgiveness of my everyday sins, the courage and endurance to successfully complete this research work.

Next to this I would like to express my sincerest gratitude and heartfelt thanks to my advisor, Dr. Dereje Teferi, for his keen insight, guidance, and unreserved advising. I am really grateful for his constructive comments and critical readings of the study. I am thankful to Dr. Million Meshesha and Dr. Tibebe Beshah for their support in proposal work.

My special thanks go to my family for their moral support and encouragement during my study and for their best coffee I drink.

I would like to thank Awash Insurance Company S.C (AIC) management information system (MIS) department head Mr. Dejene Tsegaye with their all staff and Claim Department Head Mr. Jibat Alemneh and staff Ms. Horenus Abebe. Although, Mr. Tamiru Fufa and Mr. Fekeru Olani(for his support during the data-preprocessing phase) for providing me the necessary data and assistance with their unreserved help throughout the study time.

I thank all my friends Biniyam(boss) A. , Bazezew(Baze) B. , Robel Ch. , Geme B. , Sura D. , Amenti C. , Oliyad E. , Alula M. , Dawit Hassen(the best), Dandi Merga, my cousin Urji T., Tamerat H. , Zemdkun (zeme) D. , Yilekal D. , Rev Jijo menase, Sagni Beyene and all my classmates for your Information and Experience Sharing.

## ACRONYMS

|                  |                                                 |
|------------------|-------------------------------------------------|
| <b>ARFF:</b>     | Attribute Relation File Format                  |
| <b>AIC:</b>      | Awash insurance company S.C                     |
| <b>ANN:</b>      | Artificial Neural Network                       |
| <b>CSV:</b>      | Comma Separated Values                          |
| <b>DM:</b>       | Data Mining                                     |
| <b>KB:</b>       | Knowledge base                                  |
| <b>KBS:</b>      | Knowledge based System                          |
| <b>CRISP-DM:</b> | Cross Industry Standard Process for Data Mining |
| <b>KDD:</b>      | Knowledge Discovery in Databases                |
| <b>WEKA:</b>     | Waikato Environment for Knowledge Analysis      |
| <b>PROLOG:</b>   | Programming in Logic                            |
| <b>NN:</b>       | Neural Network                                  |

## Table of Contents

|                                                           |      |
|-----------------------------------------------------------|------|
| DEDICATION .....                                          | I    |
| ACKNOWLEDGMENT .....                                      | II   |
| ACRONYMS .....                                            | II   |
| LIST of TABLES .....                                      | VI   |
| LIST of FIGURES.....                                      | VII  |
| ABSTRACT .....                                            | VIII |
| CHAPTER ONE.....                                          | 1    |
| INTRODUCTION .....                                        | 1    |
| 1.1    Background.....                                    | 1    |
| 1.2    Statement of the Problem .....                     | 3    |
| 1.3    Objectives of the Study.....                       | 5    |
| 1.3.1    General objective .....                          | 5    |
| 1.3.2    Specific objectives .....                        | 5    |
| 1.4    Research Methodology .....                         | 5    |
| 1.4.1    Understanding problem domain .....               | 7    |
| 1.4.2    Understanding the data .....                     | 7    |
| 1.4.3    Preparation of the data.....                     | 7    |
| 1.4.4    Data mining .....                                | 7    |
| 1.4.5    Use of the knowledge .....                       | 8    |
| 1.4.6    Evaluation of the discovered knowledge.....      | 8    |
| 1.4.7    Research Tool selection.....                     | 8    |
| 1.5    Scope and Limitations of the Study.....            | 9    |
| 1.6    Significance of the Study.....                     | 10   |
| 1.7    Organization of the study .....                    | 10   |
| CHAPTER TWO.....                                          | 12   |
| Literature Review .....                                   | 12   |
| 2.1    Types of fraud affecting insurance .....           | 13   |
| 2.2    Fraud Detection Using Data Mining Techniques ..... | 14   |
| 2.3    Data Mining and Knowledge Discovery .....          | 16   |
| 2.4    Data mining tasks .....                            | 18   |
| 2.4.1    Classification .....                             | 19   |
| 2.4.1.1    Decision tree .....                            | 20   |
| 2.4.1.2    Neural networks.....                           | 21   |

|                                                                      |                                                                             |    |
|----------------------------------------------------------------------|-----------------------------------------------------------------------------|----|
| 2.4.1.3                                                              | Rule based classification .....                                             | 22 |
| 2.4.2                                                                | Attribute selection Measures .....                                          | 24 |
| 2.4.2.1                                                              | Information Gain .....                                                      | 24 |
| 2.4.2.2                                                              | Gain Ratio.....                                                             | 25 |
| 2.4.2.3                                                              | Gini Index.....                                                             | 25 |
| 2.5                                                                  | Knowledge based system.....                                                 | 26 |
| 2.5.1                                                                | Types of Knowledge.....                                                     | 27 |
| 2.5.2                                                                | Architecture of Knowledge Based System .....                                | 28 |
| 2.6                                                                  | AI Programming Languages.....                                               | 32 |
| 2.7                                                                  | Related works .....                                                         | 34 |
| CHAPTER THREE .....                                                  |                                                                             | 36 |
| Knowledge acquisition using data mining.....                         |                                                                             | 36 |
| 3.1                                                                  | Claims handling processes.....                                              | 38 |
| 3.2                                                                  | K-Means Clustering.....                                                     | 39 |
| 3.3                                                                  | Rule based algorithm .....                                                  | 42 |
| 3.4                                                                  | Data selection and preparation .....                                        | 43 |
| 3.5                                                                  | Data reduction and Processing .....                                         | 45 |
| 3.6                                                                  | Experimentation .....                                                       | 45 |
| 3.6.1                                                                | Experimental set up .....                                                   | 45 |
| 3.7                                                                  | Creating Predictive Model.....                                              | 52 |
| CHAPTER FOUR .....                                                   |                                                                             | 58 |
| Integration of data mining Results with Knowledge Based system ..... |                                                                             | 58 |
| 4.1                                                                  | System design and Architecture .....                                        | 58 |
| 4.2                                                                  | Structure of PART Rule and PROLOG .....                                     | 60 |
| 4.3                                                                  | High level conceptual design of integration process .....                   | 62 |
| 4.4                                                                  | Implementation of Discovered Rules to Knowledge Base System Integrator..... | 69 |
| CHAPTER FIVE .....                                                   |                                                                             | 72 |
| IMPLEMENTATION AND EXPERIMENTATION .....                             |                                                                             | 72 |
| 5.1                                                                  | Suspicious Fraud diagnosis .....                                            | 73 |
| 5.2                                                                  | Explanation facility .....                                                  | 75 |
| 5.3                                                                  | Recommendation for Fraud detection .....                                    | 75 |
| 5.3.1                                                                | General Information Provider.....                                           | 75 |
| 5.3.2                                                                | Technical Support and Further Investigation .....                           | 76 |
| 5.4                                                                  | Testing and Evaluation .....                                                | 77 |

|                                                                              |                                  |     |
|------------------------------------------------------------------------------|----------------------------------|-----|
| 5.4.1                                                                        | User Acceptance Testing .....    | 78  |
| 5.4.2                                                                        | System Performance Testing ..... | 79  |
| CHAPTER SIX .....                                                            |                                  | 82  |
| CONCLUSION AND RECOMMENDATION .....                                          |                                  | 82  |
| 6.1                                                                          | Conclusion.....                  | 82  |
| 6.2                                                                          | Recommendations .....            | 83  |
| Reference .....                                                              |                                  | 85  |
| Appendix I: Rules generated from PART data mining algorithm .....            |                                  | 90  |
| Appendix II: Initial list of original attributes with their description..... |                                  | 93  |
| Appendix III : Sample values of the final selected attribute .....           |                                  | 94  |
| Appendix IV: User Acceptance Test .....                                      |                                  | 95  |
| Appendix V: Integrator java code .....                                       |                                  | 96  |
| Appendix VI: Prolog code for knowledge representation .....                  |                                  | 103 |

## LIST of TABLES

|            |                                                                              |    |
|------------|------------------------------------------------------------------------------|----|
| Table 3-1  | Types of Insurance Services AIC Provides.....                                | 37 |
| Table 3-2  | Default parameters and values for algorithms .....                           | 46 |
| Table 3-3  | List thresholds used to cluster result of claim report date gap.....         | 46 |
| Table 3-4  | List thresholds used to cluster result of claim Regain amount.....           | 47 |
| Table 3-5  | The Final List of Attributes used in the Study.....                          | 48 |
| Table 3-6  | Training of the first experiment by the default parameter values.....        | 49 |
| Table 3-7  | Cluster summary of the first experiment.....                                 | 49 |
| Table 3-8  | Training of the second experiment.....                                       | 50 |
| Table 3-9  | Cluster summary of the 2nd experimentation.....                              | 50 |
| Table 3-10 | Training of the third cluster experiment.....                                | 51 |
| Table 3-11 | Cluster Summary of the third.....                                            | 52 |
| Table 3-12 | performance of J48 classifier.....                                           | 52 |
| Table 3-13 | performance of REPTree classifier.....                                       | 53 |
| Table 3-14 | performance of PART classifier.....                                          | 53 |
| Table 3-15 | performance of JRip classifier.....                                          | 54 |
| Table 3-16 | performance of classifiers.....                                              | 54 |
| Table 3-17 | Confusion Matrix for PART classifier.....                                    | 55 |
| Table 3-18 | Precision, Recall and F- Measure of Classifiers with respect to classes..... | 55 |
| Table 3-19 | Sample rule set Generated Using PART algorithm.....                          | 57 |
| Table 4-1  | Sample PART rules for Fraud and Non-Fraud.....                               | 61 |
| Table 4-2  | Tokens in PART and PROLOG rule.....                                          | 62 |
| Table 5-1  | User acceptance evaluation.....                                              | 79 |
| Table 5-2  | Confusion matrix for evaluation of the system KBS .....                      | 80 |
| Table 5-3  | Performance evaluation based on TP, Precision, Recall and F-measure.....     | 81 |

## LIST of FIGURES

|             |                                                                     |    |
|-------------|---------------------------------------------------------------------|----|
| Figure 2-1  | Different types of fraud affecting insurance companies.....         | 14 |
| Figure 2-2  | The Process of Knowledge Discovery in Databases.....                | 17 |
| Figure 2-3  | Architecture of Knowledge Based System.....                         | 28 |
| Figure 3-1  | Diagrammatic representation of K-Means.....                         | 41 |
| Figure 3-2  | Attribute evaluation using Gain Ratio.....                          | 44 |
| Figure 3-3  | Time taken to build model by classifiers.....                       | 56 |
| Figure 3-4  | TP rate of classifiers.....                                         | 56 |
| Figure 4-1: | General Framework of integration.....                               | 59 |
| Figure 4-2  | Conceptual Design of the Integration Process.....                   | 63 |
| Figure 4-3  | work flow diagram for mapping PART into PROLOG.....                 | 64 |
| Figure 4-4  | Algorithm for rule tokenization.....                                | 66 |
| Figure 4-5  | Algorithm for rule parser.....                                      | 67 |
| Figure 4-6  | Algorithm for rule reversing.....                                   | 68 |
| Figure 5-1  | Sample rules from the rule base.....                                | 73 |
| Figure 5-2  | Integrator application screen shot for Fraud detection system ..... | 73 |
| Figure 5-3  | Question and answer in fraud detection.....                         | 74 |
| Figure 5-4  | Prolog screen shot for Further Investigation description.....       | 75 |
| Figure 5-5  | Prolog code for menu for Fraud claim.....                           | 76 |
| Figure 5-6  | Prolog screen shot for Technical Support description.....           | 77 |

## ABSTRACT

The insurance industry has been historically a growing industry. It plays an important role in insuring the economic well being of a country. Fraud risk poses a very big challenge for the insurance sector. Business leaders are aware of the need to address this risk, but the lack of a comprehensive and integrated approach to fraud risk management continues to be a concern. Insurance fraud is very costly and has become a concern of every insurance company in recent years. Fraudulent claims account for a significant portion of all claims received by insurers, and cost billions of dollars annually. Nowadays, efforts have been made to develop models to identify potentially fraudulent claims for special investigations using data mining and knowledge base system.

Data Mining is extracting information from huge set of data. In other word, data mining is inducing hidden knowledge from dataset. Knowledge based system has come across with a variety of approaches based on the knowledge representation method; Rule-based reasoning is one of the simplest and popular approach used in knowledge-based system and SWI-Prolog software is used to build a knowledge base.

In this study, an attempt is made to integrate data mining with knowledge based system for fraud detection. The research has tried to apply first the clustering algorithm followed by classification techniques for developing the predictive model. K-Means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud and then applied PART classifier on Awash Insurance dataset. The integrator application basically created using Eclipse IDE with JDK 7 and then links the model created by PART classifier to knowledge based system. The integrator understands both the PART classifier and PROLOG and converts rules from PART to PROLOG understandable format.

The experiments have been conducted following the six-step Cios et al. (2000) process model. For the experiment, the collected insurance dataset is preprocessed to remove outliers, fill in missing values and select attributes. The preprocessing phase and building the advising Knowledge base took the highest portion of the study time.

Finally, the performance of the system is evaluated by preparing test cases. The cases are given to domain experts for system performance testing. For user acceptance test users evaluated the

system. Generally, the system has scored 81.2 % accuracy with an average Precision and Recall of 81 % and 81.2%, which is a promising result. Further researches should be done to increase the merits of integrating Data mining induced knowledge with knowledge base system.

**Keywords:** - Insurance Fraud detection, data mining, rule based fraud detection, integration with Knowledge Base System, PART classifier

# CHAPTER ONE

## INTRODUCTION

### 1.1 Background

Every organization that exchanges money with customers, service providers or vendors risks exposure to fraud and abuse. Insurance companies around the world lose more and more money through fraudulent claims each year. They need to recoup this lost money so they can continue providing superior services for their customers [1].

Insurance is a contract (policy) in which an individual or entity receives financial protection or reimbursement against losses from an insurance company. The company pools clients' risks to make payments more affordable for the insured. Now days, customer choosing insurance policy for three reason, risk factor, investment and tax. Today, human life is getting some unexpected movement in their life such as accident and health problem [2]. The accident will make a huge loss for human being. So every one of our people wants to save their own life and they need a back up to their life. The insurance company helps people who met accident and suffering from health problem.

But, these days insurance sector facing more problems due to two reason internal and external reasons. For example, Customer claims frauds in insurance sector and lack of knowledge of internal auditor [2]. Customer entering all the particulars through paper based documents. Then the document scanned and converted as pdf format. After that, the data entered into data base manually by the internal auditor. Due to carelessness and lack of knowledge of internal auditor some misrepresentation will occur. This thing will make unexpected loss for insurance sector and also for customer. Every customer should surrender all the legal documents at the time of joining in any insurance sector and also provide right document to collect amount from insurance sector. It is not happening at the time of insurance policy processing. Rather, Fraud is nothing but wrongful or criminal trick planned to result in financial or personal gains [1].

Data Mining (DM) combines techniques from machine learning, pattern recognition, statistics, database theory, and visualization to extract concepts, concept interrelations, and interesting

patterns automatically from large corporate databases [3]. Its primary goal is to extract knowledge from data to support the decision-making, planning and problem solving process. Two primary functions of DM are prediction and description [4]. Prediction involves finding unknown values/relationships/patterns from known values; and description provides interpretation of a large database. Classification is useful for prediction, whereas clustering, pattern discovery and deviation detection are for description of patterns in the data.

DM can be applied in detecting fraudulent claims in the insurance industry. Modern success stories for data-mining fraud detection include applications in credit-card fraud and telecommunications fraud, which are much easier to model than insurance fraud [5]. Woodfield [5] further noted that, credit card and telecommunications fraud is easily identified within a relatively short period of time because a customer contests a bill or reports a stolen credit card. However, a large gap remains between the results a data mining system can provide and taking actions based on them [3]. This implies that there is a gap in extracting knowledge employing data mining and using it for action.

A knowledge base system is a computer system which represents knowledge about a specific domain and can be used to apply this knowledge to solve the problems from the problem domain [4].

The application of a data mining approach in knowledge base development involves a set of techniques for searching through data sets, looking for hidden correlations and trends which are inaccessible using conventional data analysis techniques [5].

This study is aimed at designing a system that identifies fraud claims by using an automatically constructed knowledge base, based on knowledge acquired from data mining models and providing advice for claim analyst.

## 1.2 Statement of the Problem

Insurance companies lose millions of dollars each year through fraudulent claims, largely because they do not have a way to easily determine which claims are legitimate and which may be fraudulent. To ensure that adjusters target claims which have the greatest likelihood of adjustment, many insurance companies have incorporated data mining into their investigating and auditing processes [6].

Companies in the Insurance industry collected enormous amounts of data about their clients. This is invaluable information about customers' behavior, activities, and preferences. To extract information from this large amount of raw data, Insurance firms lost time and efforts. Knowledge extracting process resulted in developing new products and services to meet customers' needs. It unearthed information on customers, markets, and competitors. But some essential dependencies and patterns cannot be discovered from terabytes of data by using human experts.

Insurance underwriting has to do with the selection of subjects for insurance in such a manner that general company objectives are met. The profitability of underwriting depends on accurate fraud detection. Hence one of the main goals of the underwriting management is to protect the company against high claim cost due to fraudulent claims. An increasing claim cost can lead to very serious problem for the company and, more often than not, adequate supervision at an early stage could help avert these problems [7].

Fraudulent claims are a serious financial burden on insurers and result in higher overall insurance costs [1]. According to IBM Corporation, Insurance companies lose millions of dollars each year through fraudulent claims, largely because they do not have a way to easily determine which claims are legitimate and which may be fraudulent. Insurance fraud affects not only the financial health of the insurers, but also of innocent people seeking effective insurance coverage. As indicated by IBM Corporation, fraud can range from exaggerated losses to deliberately causing an accident for the payout.

As the researcher understood from the discussion conducted with experts of Awash insurance company S.C, there are obstacles in processing claims. For example, for processing a single claim takes much time, which delays the response time for the policyholders. Due to the huge number

of claims submitted each day, it would be far too expensive for the company to have employees check each claim for symptoms of fraud. Generally, this costs the company a lot both financially and time wise. So the manual processing of claims is not effective and efficient in identifying fraudulent cases.

There are also studies conducted to investigate the application of DM in the insurance business. Tesfaye [8] used the DM technique for insurance risk assessment in Nyala insurance company. Tariku [9] attempts Mining Insurance Data for Fraud Detection in the area of motor insurance at Africa Insurance Share Company. He tries to do the applicability of the data mining technology in developing models that can detect and predict fraud suspicious in insurance claims.

However, there is a gap between the results a data mining system can provide and taking action based on them. Both researchers [8] and [9] developed predictive model for fraud detection and classify the data set as fraud or Non-Fraud.

There are a number of researches done on mining insurance data excluding integrating with advising knowledge base system. Automated advising system can help the user to apply the DM technology with knowledge base system to fraud detection. It updates the knowledge base by learning from claims occurred.

As can be seen from the current practice of the company, there is no automated predictive model or advising knowledge base system, which will help the company to show those claims which have a high degree of exposition to be fraudulent.

Therefore, the aim of this research is to identify patterns which are fraudulent and represent these patterns in knowledge base system and make the data mining findings easily accessible for the user.

To this end, an attempt is made to explore and answer the following questions:

- ❖ What is the pattern that characterizes whether a given claim is fraudulent or not?
- ❖ Which data mining technique is more appropriate to identify and determine factors for fraud detection?

## 1.3 Objectives of the Study

### 1.3.1 General objective

The general objective of this study to apply the DM techniques for constructing predicting model and integrate automatically with the knowledge Base System, in order to increase effectiveness and efficiency in Motor insurance fraud detection.

### 1.3.2 Specific objectives

For the realization of the general objective stated above, the following specific objectives are formulated.

- ❖ To understand the business environment so as to identify fraudulent transaction and their characteristics. l
- ❖ To collect the required insurance data from Awash Insurance Company S.C.
- ❖ To prepare proper data for analysis and model building, by cleaning, attribute selection and transforming the data into a format suitable for the selected DM algorithms.
- ❖ To construct an optimal classification model that will help to predict and detect fraudulent insurance claims.
- ❖ To develop a knowledge base system in which induced knowledge base is represented.
- ❖ To build an integrator aiming at representing data mining findings in a knowledge base system.
- ❖ To evaluate the performance of the prototype for its usability.

## 1.4 Research Methodology

The integration of Data Mining model with knowledge Base System can be discussed and procedures that are performed to conduct this research can be explained as follows [10]. In addition, research works which attempted to integrate data mining with knowledge base system are also referred.

For the purpose of conducting this Experimental research Hybrid process model (Cios et al.) is selected. The development of academic and industrial models has led to the development of hybrid models, i.e. models that combine aspects of both. One such model is a six-step KDP model

developed by Cios et al. It was developed based on the CRISP-DM model by adopting it to academic research [11]. It provides more general, research-oriented description of the steps, introduce several new explicit feedback mechanism, (the CRISP-DM model has only three major feedback sources, while the hybrid model has more detailed feedback mechanisms) and modification of the last step, since in the hybrid model, the knowledge discovered for a particular domain may be applied in other domains.

A description of the six steps follows

**Understanding of the problem domain:** - this initial step involves working closely with domain experts to define the problem and determine the project goals, identifying key people, and learning about current solutions to the problem.

**Understanding of the data:** - This step includes collecting sample data and deciding which data, including format and size, will be needed.

**Preparation of the data:** - this step concerns deciding which data will be used as input for DM methods in the subsequent step. It involves sampling, running correlation and significance tests, and data cleaning, which includes checking the completeness of data records, removing or correcting for noise and missing values, etc.

**Data mining:** -here the data miner uses various DM algorithms to derive knowledge from preprocessed data. This incorporate testing different data mining algorithm and selecting best algorithm based on their True positive rate percentage.

**Evaluation of the discovered knowledge:** - evaluation includes understanding the results, checking whether the discovered knowledge is novel and interesting, interpretation of the results by domain experts, and checking the impact of the discovered knowledge.

**Use of the discovered knowledge:** - this final step consists of planning where and how to use the discovered knowledge. The application area in the current domain may be extended to other domains. Here, we can represent data mining generated rules into knowledge base. Finally, the discovered knowledge is deployed by designing integrator application.

### **1.4.1 Understanding problem domain**

To clearly identify, understand, and analyze the business problems, the primary (observation and informal interview), and secondary (database analysis, literature and document review) data collection methods is employed. After understanding the business domain, informal interview was employed to define features selection with the domain experts while observation and database analyze are conducted to understand some complex business processes.

### **1.4.2 Understanding the data**

The data to be used in this research was collected, regarding motor insurance claim, from Awash Insurance Company S.C database. Careful analysis of the data and its structure is done together with domain experts by evaluating the relationships of the data with the problem at hand and the particular DM tasks to be performed.

### **1.4.3 Preparation of the data**

Today's real-world databases are highly susceptible to noisy, missing values and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources [4]. It usually consumes about half of the entire research effort. Sample size of 14576 instances is taken for this study. Data cleaning (or data cleansing) routines is applied to fill in missing values (with ReplaceMissingValues in the weka software for data preprocessing), smooth out noise (by removing error/outliers from the record), then the smoothed data is transformed into common format and data reduction is applied to select important attribute done by Gain ratio attribute evaluation technique. The cleaned data further processed by data mining algorithms.

### **1.4.4 Data mining**

Discovering regularities for detecting fraudulent claims within the insurance dataset in particular, is certainly the purpose of this study. The unsupervised clustering technique and the supervised classification technique are adopted. The clustering technique is selected because of the reason that the dataset, which obtained from Awash Insurance Company S.C may doesn't have a feature indicating whether a claim was fraud or not. So, clustering technique is used to group the

insurance claims that are susceptible to fraud, into a similar cluster. For doing this we have employed simple K-Means clustering algorithm since we have two classes only.

The clustered dataset is then used as an input for the classification algorithm for classifying instances of the dataset into similar class labels. For this purpose, the **PART rule** based algorithm method is used to create model for detecting fraudulent insurance claims.

#### 1.4.5 Use of the knowledge

To use the knowledge discovered from the data mining we need a knowledge representation tool. Data mining findings represented as a rule which is conditional statement that links given conditions to action or outcomes. These if-then rule statements are used to formulate the conditional statements that constitute knowledge base. Rule based representation is highly expressive, it is easy to interpret and easy to generate [7]. To design integration application for the advising knowledge base system, Java programming language has been used. Java is chosen because it is easy to integrate with other systems like weka tool and familiarity of the researcher with the language.

#### 1.4.6 Evaluation of the discovered knowledge

The classifier algorithm and the advising knowledge base system have been evaluated together with the domain experts regarding its accuracy and usability. Knowledge base which hold the facts generated by Data mining technology is tested whether classifying accordingly by the domain experts and easy to use for the user of the system.

#### 1.4.7 Research Tool selection

Selection of appropriate tools and techniques depends on the main task of the Fraud detection process. Moreover, the accessibility and familiarity of these tools for the researchers can be also another factor for selection. Depending on these factors, Weka, SWI-Prolog, Microsoft- Excel and Java Programming language were selected because the researcher found them to be very effective tools for conducting this research.

To this end, while conducting this research, WEKA (Waikato Environment for Knowledge Analysis) version 3.7.0 (for Mac OS) DM software is chosen. Weka is chosen because of its widespread application in different DM researches and familiarity of the researcher with the software. Moreover, to design the integration application for the advising knowledge base system, Java programming language has been used. Java is chosen because it is easy to integrate with other systems like WEKA tool and the familiarity of the researcher with the language. Furthermore, SWI-Prolog software is used to build a knowledge base.

## **1.5 Scope and Limitations of the Study**

The main aim of this research is integrating Data mining model with knowledge base system for fraud prediction and detection in the insurance sector. This research focuses only to the claim department of Awash Insurance Company S.C. Specifically the mined data for knowledge discovery is obtained from all branches, which are networked to each other with the main office. Awash Insurance Company S.C offers General Insurance (Non- Life) and Life Insurance, data related to only Motor insurance is considered. Motor insurance and Fire & Lighting are one of the most effective working areas of Awash Insurance Company S.C. First, the research is designed to conduct on Fire & Lighting insurance however due to unavailability of data in the area the research direction is changed into Motor insurance. Motor insurance is selected because there is enough amounts of data to conduct this research and it is the major area which is prone to fraud as domain experts explained.

As a limitation there is no record related to different stakeholders such as traffic police report, age, sex, health status, driving experience, and marital status of driver and the place where and when the accident has occurred, which are important for fraud detection. In addition, the PART classifier algorithm has generated 54 rules and 32 of them are Fraud types and the rest 22 are Non-Fraud, from the result shown the size of class labels are unbalanced for model building process. To keep the originality of the model building, the researcher take the class labels as it is.

The knowledge acquisition is done through data mining technique from the dataset collected from the company database. Data mining results, PART rules are mapped or integrated to knowledge base by using integrator application. Here, the integrator application automatically builds

knowledge base after rules are mined from the dataset. Whenever update is found in the size (number) of dataset the integrator application rebuild rules by overwriting previous rules.

This advising Knowledge base system targets mainly beginner claim analysts in claim department.

## **1.6 Significance of the Study**

The results of this research can help Awash Insurance Company S.C and other related companies:

- ❖ To improve the decision making process of claim analysts especially in the claim and underwriting departments.
- ❖ To identify those fraudulent areas and take timely measure accordingly.
- ❖ To discover procedures for investigating the fraudulent exposure of claims.
- ❖ To save time and money which is wasted for processing and surveying different claims reported to the organization every day.
- ❖ To integrate data mining models and knowledge based system in other fields of studies especially in areas where there is shortage of domain experts to acquire knowledge but there exist plenty of data such that knowledge can be acquired automatically from the data.

## **1.7 Organization of the study**

This thesis is organized into six chapters. The first chapter briefly discusses background to the problem area and DM technology and Knowledge base system, and states the problem, objective of the study, research methodology, scope and limitation, and significance of the results of the research. It provides highlight about insurance fraud, Data mining and Knowledge base system.

The second chapter deals about literature review. About data mining and Knowledge base system methods/techniques and algorithms, the different methodologies, the DM process, the different tasks of DM, and its application in the insurance sector. The literatures focused on employing data mining model for construction of knowledge based system are discussed.

The third chapter provides discussions about acquisition process. The focus here is on automatic knowledge acquisition techniques through data mining and about the different DM methods and algorithms that are used in this research the results of WEKA classifier algorithms are analyzed, interpreted, evaluated and compared with each other.

The fourth chapter provides introduction about AIC and discussions about the different DM technique with knowledge base system. This includes the business understanding, data understanding, and data preprocessing phases.

The fifth chapter provides a detailed discussion about the experimentation part of this study. This includes all about fraud detection system and advising knowledge base system. Evaluation of the discovered knowledge is also discussed at this section.

The last chapter is devoted to concluding remarks and recommendations forwarded based on the research findings of the present study.

## CHAPTER TWO

### Literature Review

With the growing use of computers, there is a great amount of the data being generated by such systems. We have grown gradually to the fact that there are tremendous volumes of data filling our computers, networks and lives. Government agencies, scientific institutions and businesses have all dedicated enormous resources to collecting and storing data. In reality, only a small amount of these data will ever be used because, in many cases, the volumes are simply too large to manage, or the data structures themselves are too complicated to be analyzed effectively.

Banks and Insurance companies lose billions of dollars every year to fraud. Traditional methods of fraud detection play an important role in minimizing these losses. However increasingly sophisticated fraudsters have developed a variety of ways to elude discovery, both by working together, and by leveraging various other means of constructing false identities. While no fraud prevention measures can ever be perfect, significant opportunity for improvement can be achieved by looking beyond the individual data points, to the connections that link them. Oftentimes these connections go unnoticed until it is too late— something that is unfortunate, as these connections oftentimes hold the best clues. Understanding the connections between data, and deriving meaning from these links, doesn't necessarily mean gathering new data. Significant insights can be drawn from one's existing data [12].

Ernst & Young [13], global leader in assurance, tax, transaction and advisory services states that: there are different types of insurance frauds, which occur in all areas of insurance. Insurance crimes also range in severity, from marginally exaggerated claims to ones that deliberately cause accidents or damage. Insurance companies have five key areas of risk exposure. These are related to: – Claims or surrender, Premiums, Applications, Employee-related fraud, Vendor- related third party fraud. Nearly 27% of the respondents find that the insurance companies face maximum fraud risk exposure in the area of claims or surrender.

The need to understand large, complex information- rich data sets is common to virtually all fields of business, science, and engineering. In the business world, corporate and customer data

are becoming recognized as a strategic asset. The ability to extract useful knowledge hidden in these data and to act on that knowledge is becoming increasingly important in today's competitive world. The entire process of applying a computer-based methodology, including new techniques, for discovering knowledge from data is called data mining. On other hand, Knowledge based systems are a branch of artificial intelligence, which is a computer program that attempts to replicate the reasoning processes of a human expert and it can make decisions and recommendations and perform tasks based on user input. Artificial Intelligence is all about how to make the system think, or act like human. The expert's knowledge is available when the human expert might not be and so that the knowledge can be available at all times and in many places, as necessary. Expert systems derive their input for decision making from prompts at the user interface, or from data files stored on the computer [14].

The primary intent of expert system technology is to realize the integration of human expertise into computer processes. This integration not only helps to preserve the human expertise but also allows humans to be freed from performing the more routine activities that might be associated with interactions with a computer-based system. There exist a number of different definitions for a knowledge based system, but in this study the definition discussed by [15] reviewed and presented as follow because it seems more detail and create the overall understanding of KBS definition.

## **2.1 Types of fraud affecting insurance**

According to Ernst & Young [13] company survey, nearly 31% of the respondents indicate that insurance companies are most affected by mis-selling due to premeditated fabrication and/or fraudulent misrepresentation of material information, as compared to the different types of frauds that affect insurance companies adversely. Insurance continues to be miss-sold with senior citizens being the softest targets as they do not understand new products. As Figure 2-1 illustrates Collusion between parties (29%) and fake documentation (24%) are the other critical types of fraud that affect insurance companies. The former mainly occurs in collusion with patients, who are undergoing a particular treatment and are made to sign for more expensive procedures by hospitals. In some cases, hospitals claim money for patients who have not actually been admitted to these hospitals. Such patients are paid a small amount by such hospitals, which claim large amounts from insurance companies. When we come to our country Motor insurance and Fire and

lighting are the common ones according to the Addis Ababa Fire Brigade, 357 fire perils resulting in the death of 7 people, 32 heavy & light injuries and a material loss worthing Birr 32,842,489 were registered in nine months starting from July 2007 up to March 2008.

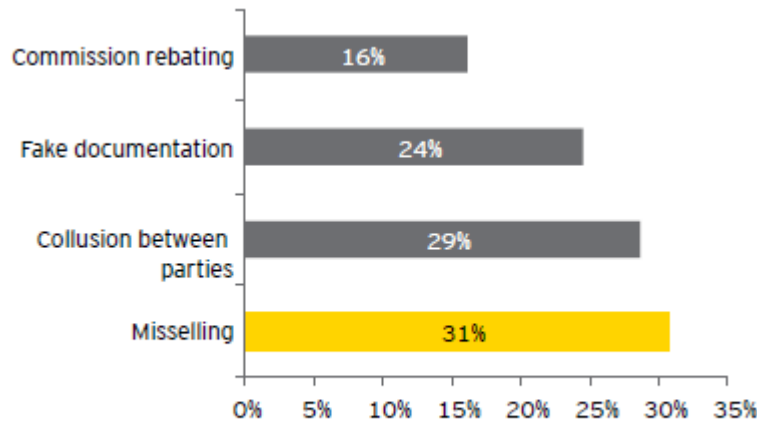


Figure 2-1 Different types of fraud affecting insurance companies according [13].

## 2.2 Fraud Detection Using Data Mining Techniques

Different scholars provided different definitions about DM. According to Berry and Linoff [16] and Han and Kamber [4], DM is the process of extracting or “mining” knowledge from large amounts of data in order to discover meaningful patterns and rules. Witten and Frank [17], have also noted that DM is valuable to discover implicit, potentially useful information from huge data stored in databases via building computer programs that sift through databases automatically or semi-automatically, seeking meaningful patterns.

DM is an interdisciplinary approach involving tools and models from statistics, artificial intelligence, pattern recognition, data visualization, optimization, information retrieval, high end computing, and others [3] and [4]. DM methodology often can improve upon traditional statistical approaches for solving business solutions by finding additional, important variables, by identifying interaction among terms and detecting nonlinear relationships [5]. Models that predict relationships and behaviors more accurately lead to greater profits and reduced costs.

Data mining is an iterative process within which progress is defined by discovery, through either automatic or manual methods. Data mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an “interesting” outcome [18]. Data mining is the search for new, valuable, and nontrivial information in large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers.

Nowadays, auditors are using data mining tools and techniques to examine the entire population of transactions in order to select samples for test controls and identify fraud. A large insurance provider needs to accurately determine which claims are fraudulently filed so it can concentrate on preventing revenue loss. Over the years, this organization collected audit results on insurance claims. The organization seldom used historical records to identify probable fraudulent claims in the future. Their previous methods often missed opportunities to collect money and adjusters spent too much time reviewing legitimate claims. Data mining now enables companies to predict which insurance claims are likely to be fraudulent. This gives adjusters the power to determine what returns they should target, thereby, recouping millions of dollars otherwise lost and saving adjusters many hours of valuable time [1].

In a competitive environment, fraud can become a business critical problem if it is very prevalent and if the prevention procedures are not fail-safe. Fraud detection, being part of the overall fraud control, automates and helps reduce the manual parts of a screening/checking process. This area has become one of the most established industry/government data mining applications. For many companies where they have interactions with up to millions of external parties, it is cost-prohibitive to manually check the majority of the external parties’ identities and activities. So the riskiest ones determined through data mining output such as suspicion scores, rules, and visual anomalies will be investigated [19].

Once you design models that predict fraudulent activity, you need to strategically deploy your results to the people who can use that information to detect fraud and recoup money. Strategic deployment means integrating designed models into your company’s daily operations [1]. Strategic deployment empowers you to put timely, consistent information into the right hands.

Everyone in your organization is on the same page and can act more quickly to recoup the most money for your organization.

## 2.3 Data Mining and Knowledge Discovery

The accessibility and abundance of data today makes Knowledge Discovery and Data Mining a matter of considerable importance and necessity. Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets. Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) -which are the essence of useful knowledge. The model is used for understanding phenomena from the data, analysis and prediction [20].

Data mining (DM) is the subfield of Machine Learning that enables finding interesting Knowledge (patterns, models and relationships) in every large database. It is the most essential part of the knowledge- discovery process, which combines databases, statistics, artificial intelligence and machine learning techniques [7].

The knowledge discovery process (Figure 2-2) is iterative and interactive, consisting of nine steps. Note that the process is iterative at each step, meaning that moving back to adjust previous steps may be required [20]. Following is a brief description of the nine-step KDD process.

1. Developing an understanding of the application domain: It is first step to understand the organizational domain. It opens the chance for understanding what should be done with the many decisions (about selection, transformation, algorithms, representation, etc.).
  - ❖ It is necessary to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior knowledge).
2. Selecting and creating a data set on which discovery will be performed. Having defined the goals, the data that will be used for the knowledge discovery should be determined. This process is very important because the Data Mining learns and discovers from the available data.

3. Preprocessing and cleansing: In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers.
4. Data transformation: Here the data is transformed into a format appropriate for further Data Mining process. Methods here include dimension reduction (such as feature selection and extraction, and record sampling), and attribute transformation (such as discretization of numerical attributes and functional transformation). This step is often crucial for the success of the entire KDD project, but it is usually very project-specific

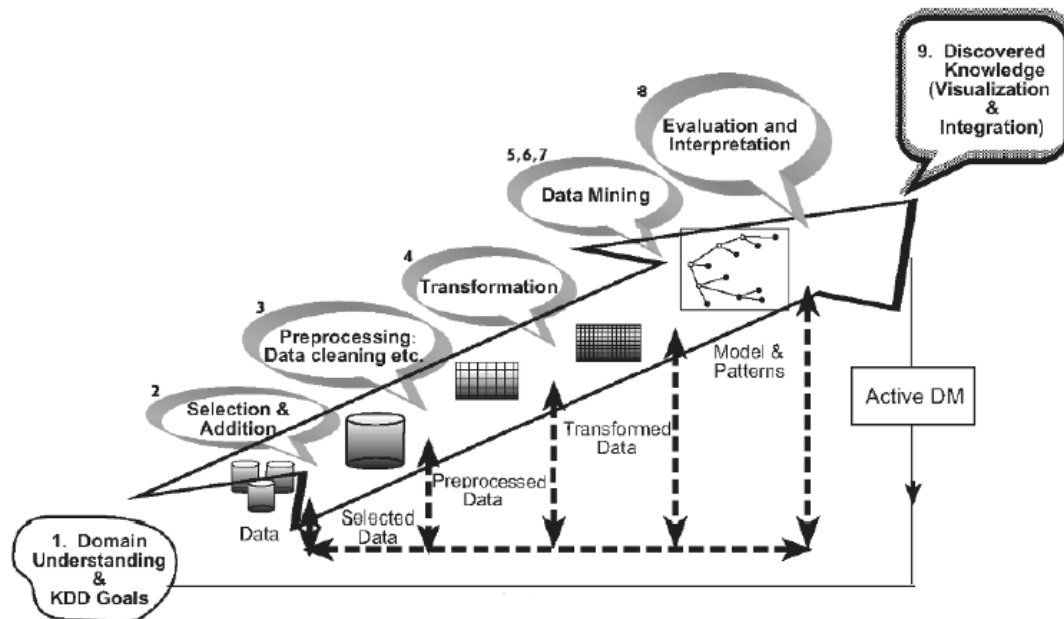


Figure 2-2. The Process of Knowledge Discovery in Databases[20]

5. Choosing the appropriate Data Mining task: We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. We discuss about this later in this chapter.

6. Choosing the Data Mining algorithm: Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers).
7. Employing the Data Mining algorithm: Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.
8. Evaluation: In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step. Here we consider the preprocessing steps with respect to their effect on the Data Mining algorithm results (for example, adding features in Step 4 and repeating from there).
9. Using the discovered knowledge. We are now ready to incorporate the knowledge into another system for further action. The knowledge becomes active in the sense that we may make changes to the system and measure the effects.

Data mining is part of an iterative process called knowledge discovery in database which can assist to extract this knowledge automatically [2]. Data mining has attracted a great deal of attention in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [4]. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, intrusion detection, customer retention and in medical diagnosis.

## 2.4 Data mining tasks

There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining. Most data mining techniques are based on inductive learning, where a model is constructed explicitly or implicitly by generalizing from a sufficient

number of training examples [20]. The underlying assumption of the inductive approach is that the trained model is applicable to future cases.

Maimon and Rokach [19] stated that, Discovery methods are those that automatically identify patterns in the data. The discovery method branch consists of prediction methods versus description methods. Descriptive methods are oriented to data interpretation, which focuses on understanding (by visualization for example) the way the underlying data relates to its parts. Prediction-oriented methods aim to automatically build a behavioral model, which obtains new and unseen samples and is able to predict values of one or more variables related to the sample.

It also develops patterns, which form the discovered knowledge in a way which is understandable and easy to operate upon. Some prediction-oriented methods can also help provide understanding of the data. In this study, predictive models are used for acquiring knowledge.

#### **2.4.1 Classification**

Classification builds up and utilizes a model to predict the categorical labels of unknown objects to distinguish between objects of different classes. Classification is the process of identifying a set of common features and proposing models that describe and distinguish data classes or concepts. Classification is the process of finding a model, which describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [4]. Classification problems aim to identify the characteristics that indicate the group to which each instance belongs. Classification can be used both to understand the existing data and to predict how new instances will behave.

PHUA [21] stated that, most of the fraud detection studies using supervised algorithms since 2001 have abandoned measurements such as true positive rate (correctly detected fraud divided by actual fraud) and accuracy at a chosen threshold (number of instances predicted correctly, divided by the total number of instances). In fraud detection, misclassification costs (false positive and false negative error costs) are unequal, uncertain, can differ from example to example, and can change over time. In fraud detection, a false negative error is usually more costly than a false positive error.

Common classification techniques are neural networks, K-nearest neighbor, the naïve Bayes technique, decision trees, support vector machines and rule based learning. Such classification tasks are used in the detection of credit card, health care and insurance fraud, and corporate fraud. In this report, decision tree and rule based learning are discussed.

#### 2.4.1.1 Decision tree

Decision tree is a predictive model which can be used to represent classifiers. In the opinion of many researchers decision trees gained popularity mainly due to their simplicity and transparency. Decision tree are self-explained. There is no need to be an expert in data mining in order to follow a certain decision tree [20]. Decision trees are trees that classify instances by sorting them based on feature values [22]. They are a way of representing a series of rules that lead to a class or value. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values [21]. In DM, a decision tree is a predictive model, which can be used to represent both classifiers and regression models [23].

To classify an item, we start at the root, and apply the predicate at the root to the item. If the predicate is true, go to the left child, and if it is false, go to the right child. Then repeat the same process at the node visited, until a leaf is reached. That leaf is classified the item as liked or not.

The initial state of a decision tree is the root node that is assigned all the examples from the training set. If it is the case that all examples belong to the same class then no further decisions need to partition the examples and the solution is complete. If examples at this node belong to two or more classes then a test is made at the node that will result in a split. The process is recursively repeated for each of the new intermediate nodes until a completely discriminating tree is obtained. A decision tree at this stage is potentially an over-fitted solution i.e. it may have components that are too specific to noise and outliers that may be present in the training data [24].

ID3, C4.5 and CART (Classification and Regression Tree), in their respective order of invention and usage, are algorithms used in decision tree construction.

The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent rules [23]. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved [9]. Decision tree has its own properties. The following are some of them:

- Learns with positive and negative examples
- Noise tolerant
- General-to-specific search (reverse for pruning)
- Follows Divide-and-Conquer strategy, which has weaknesses of fracturing and diminishing training data.
- Learns discriminating rules

## **J48 classification algorithm**

C4.5 is an evolution of ID3, presented by the same author [25]. It uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. C4.5 can handle numeric attributes. J48 is an implementation of Quilan algorithm (C4.5). J48 classifier build a decision tree for the given data set, whose nodes represent discrimination rules acting on selective features by recursive partitioning of data using depth- first strategy [7].

The algorithm used each attribute of the data to make decision by splitting the data into smaller subjects. All the possible tests are considered during decision making based on information gain value of each attribute.

### **2.4.1.2 Neural networks**

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, they imitate the way the human brain learns and use rules inferred from data patterns to construct hidden layers of logic for analysis [26]. Neural networks constitute the most widely used technique in DM. Neural network is a massively parallel-distributed processor that has a natural tendency for storing experiential knowledge and making it available for use.

A neural network is first and foremost a graph, with patterns represented in terms of numerical values attached to the nodes of the graph and transformations between patterns achieved via simple message-passing algorithms [27]. Generally, a neural network can be described as a directed graph in which each node performs a transfer function of the form.

$$y_i = (f(\sum_{j=1}^n W_{ij} X_j - Q_i))$$

Where  $y_i$  is the output of the node  $i$ ,  $x_j$  is the  $j$ th input to the node, and  $W_{ij}$  is the connection weight between nodes  $i$  and  $j$ .  $Q_i$  is the threshold (or bias) of the node. Certain of the nodes in the graph are generally distinguished as being input nodes or output nodes, and the graph as a whole can be viewed as a representation of a multivariate function linking inputs to outputs. Numerical values (weights) are attached to the links of the graph, which parameterize the input/output function and allowing it to be adjusted via a learning algorithm.

Learning in ANN's can roughly be divided into supervised, unsupervised, and reinforcement learning [28]. Supervised learning or Associative learning is based on direct comparison between the actual output of an ANN and the desired correct output, also known as the target output. Reinforcement learning is a special case of supervised learning where the exact desired output is unknown. It is based only on the information of whether or not the actual output is correct. Unsupervised learning or Self-organization is solely based on the correlations among input data. No information on "correct output" is available for learning.

### 2.4.1.3 Rule based classification

A rule is represented by the IF-THEN form, where the IF part is called the condition and the THEN part is called the action [29]. The basic unit and format of knowledge in rule-based reasoning is the rule. The IF-THEN rules are quite natural for humans and are easily understood by both programmers and domain experts. However, accurate description of the domain expert's knowledge in simple rules is often difficult.

**Rule : ( condition ) => X**

The rule-based reasoning inference engine is constructed on the concept that IF the information supplied by the user satisfies the conditions of a rule, THEN the actions of the rule are executed [29]. In this research PART Rule based algorithm is used.

Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover [49]. PART, OneR and JRIP are algorithms which are rule based classifiers.

- **PART**

PART is a separate-and-conquer rule learner. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and Repeated Incremental Pruning to Produce Error Reduction (RIPPER) rule learning [49].

- **JRip**

JRIP implements a propositional rule learner. JRip proposed a Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules--based learner (RIPPER) that can be used to classify elements with propositional rules. The RIPPER algorithm is a direct method used to extract the rules directly from the data [49].

JRip (Weka's implementation of the RIPPER rule learner) is a fast algorithm for learning "IF-THEN" rules. Like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret.

- **OneR**

OneR or "One Rule" it is a simple algorithm which builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. The algorithm is based on ranking all the attributes based on the error rate. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class [49]. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random. The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate.

## 2.4.2 Attribute selection Measures

In most of the cases, the discrete splitting functions are univariate. Univariate means that an internal node is split according to the value of a single attribute [20]. Consequently, the inducer searches for the best attribute upon which to split. There are various univariate criteria. These criteria can be characterized in different ways, such as:

- ❖ According to the origin of the measure: information theory, dependence, and distance.
- ❖ According to the measure structure: impurity based criteria, normalized impurity based criteria and Binary criteria.

This section describes three popular attribute selection measures – information gain, gain ratio, and gini index.

### 2.4.2.1 Information Gain

Information gain is the change in information entropy, which is equal to the total entropy for an attribute if for each of the attribute values a unique classification can be made for the result attribute. The attribute with the highest information gain is selected as the splitting attribute. This attribute minimizes the information needed to classify the instances in the resulting partitions and reflects the least impurity in these partitions. Entropy is a measure of unpredictability of information content. High entropy means the attribute is from a uniform distribution where as low entropy means the attribute is from a varied distribution [7].

Information gain can be calculated as:

$$\text{InformationGain}(a_i, S) = \text{Entropy}(y, S) - \sum_{v_{i,j} \in \text{dom}(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot \text{Entropy}(y, \sigma_{a_i=v_{i,j}} S)$$

where:

$$\text{Entropy}(y, S) = \sum_{c_j \in \text{dom}(y)} - \frac{|\sigma_{y=c_j} S|}{|S|} \cdot \log_2 \frac{|\sigma_{y=c_j} S|}{|S|}$$

### 2.4.2.2 Gain Ratio

The gain ratio “normalizes” the information gain as follows [25].

$$GainRatio(a_i, S) = \frac{InformationGain(a_i, S)}{Entropy(a_i, S)}$$

Note that this ratio is not defined when the denominator is zero. Also the ratio may tend to favor attributes for which the denominator is very small. Consequently, it is suggested in two stages. First the information gain is calculated for all attributes. As a consequence, taking into consideration only attributes. As a consequence, taking into consideration only attributes that have performed at least as good as the average information gain, the attribute that has obtained the best ratio gain is selected. It means that attribute with the maximum gain ratio is selected as the splitting attribute.

### 2.4.2.3 Gini Index

It measures the divergences between the probability distributions of the target attribute's values. The Gini Index is used in CART.

$$Gini(y, S) = 1 - \sum_{c_j \in dom(y)} \left( \frac{|\sigma_{y=c_j} S|}{|S|} \right)^2$$

Consequently the evaluation criterion for selecting the attribute  $a_i$  is defined as:

$$GiniGain(a_i, S) = Gini(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{|\sigma_{a_i=v_{i,j}} S|}{|S|} \cdot Gini(y, \sigma_{a_i=v_{i,j}} S)$$

To summarize, the three measures for the attribute selection are used commonly. Information gain is biased towards multi valued attributes. Whereas Gain ratio tends to prefer unbalanced splits in which one partition is much smaller than the others. Gini index is biased to multi valued attributes and has difficulty when the number of classes is large [7].

## 2.5 Knowledge based system

Knowledge based systems are a branch of artificial intelligence, which is a computer program that attempts to replicate the reasoning processes of a human and it can make decisions and recommendations and perform tasks based on user input [30]. It consists of a repository of expert knowledge with utilities designed to facilitate the knowledge retrieval in response to specific queries, along with learning and justification or to transfer expertise from one domain of knowledge to another. In particular, KBS focus on using knowledge based techniques to support human decision making, learning, and action [20].

Knowledge based system seek to embed the knowledge of a human expert (e.g. a highly skilled physician or lawyer) in a 'computerized consulting service' that -- because such systems do not get bored, or tired, or old -- preserve and disseminate the knowledge so that it can be useful to others. An expert system provides advice derived from its knowledge base, using a reasoning process embedded in its inference engine, the 'thinking' part of the system [14].

An expert/knowledge based system is a computer program that is designed to mimic the decision making ability of the decision makers/experts in a particular narrow domain of expertise. Berhanu Aebissa [30], in his research states that: to fully understand and appreciate the meaning and nature of this definition, four major component pieces are highlighted in details as follow.

**Knowledge based system** is a computer program: A computer program is a piece of software, written by a “programmer” as a solution to some particular problem or client need. Because knowledge based systems are software products they inherit all of the problems associated with any piece of computer software.

**Knowledge based system is designed to mimic the decision making ability:** The specific task of a knowledge based system is to be an alternative source of decision making ability for organizations to use; instead of relying on the expertise of just one or a handful of people qualified to make a particular decision. KBS attempts to capture the reasoning of a particular person for a specific problem. Usually this system is designed and developed to capture the scarce, but critical decision making that occurs in many organizations. It is often feared to be replacements for decision makers, however, in many organizations; these systems are used to free up the decision maker to address more complex and important issues facing the organization.

**Knowledge based system uses decision makers (experts):** Webster's dictionary [31] defines an expert as "One with the special skill or mastery of a particular subject." The focal point in the development of KBS is to acquire and represent the knowledge and experience of a person(s) who have been identified as possessing the special skill or mastery.

**Knowledge based system is created to solve problems in a particular narrow domain of expertise:** Some of the most successful development efforts of KBS have been in domains that are well scoped and have clear boundaries.

An expert system provides advice derived from its knowledge base, using a reasoning process embedded in its inference engine, the 'thinking' part of the system. The primary goal of knowledge based systems is to make expertise available to decision makers who need answers quickly. Expertise is often unavailable at the right place and the right time. Computers loaded with in- depth knowledge of specific subjects can bring the knowledge available to the decision maker.

### 2.5.1 Types of Knowledge

Frequent attempts have been made to give a systematic description of knowledge. Some attempts have been based on cognitive theories, whereas others have been formulated to serve as a basis for instructional design theories. Still another approach is to characterize knowledge from an epistemological point of view. Epistemological approaches are task dependent [32]. According to Ton de Jong [32], there are four types of knowledge: situational knowledge, conceptual knowledge, procedural knowledge, and strategic knowledge.

*Situational knowledge:* it is knowledge about situations as they typically appear in a particular domain. Knowledge of problem situations enables the solver to sift relevant features out of the problem statement (selection perception) and, if necessary, to supplement information in the statement.

*Conceptual knowledge:* it is static knowledge about facts, concepts, and principles that apply within a certain domain. Conceptual knowledge functions as additional information that problem solvers add to the problem and that they use to perform the solution. Also it is known as by the term *declarative knowledge*. We use the term *declarative* as a quality characteristic, being the opposite of *compiled knowledge*.

*Procedural knowledge:* contains actions or manipulations that are valid within a domain. Procedural knowledge helps the problem solver make transitions from one problem state to another. It can have a specific, domain-bound (strong) character, or it can be more general (weak).

*Strategic knowledge:* helps students organize their problem solving process by directing which stages they should go through to reach a solution. A strategy can be seen as a general plan of action in which the sequence of solution activities is laid down.

Elements of knowledge belonging to the first three types are specific, applicable to certain of problems in a domain, whereas the last type, strategic knowledge is applicable to a wider variety of types of problems within a domain.

### 2.5.2 Architecture of Knowledge Based System

Knowledge Based System is a computer program that simulates the judgment and behavior of a human that has expert knowledge and experience in a particular field. It contains a knowledge base containing accumulated experience and a set of rules. Knowledge Based System provides high quality experience, domain specific knowledge; apply heuristics, forward or backward reasoning, uncertainty and explanation capability.

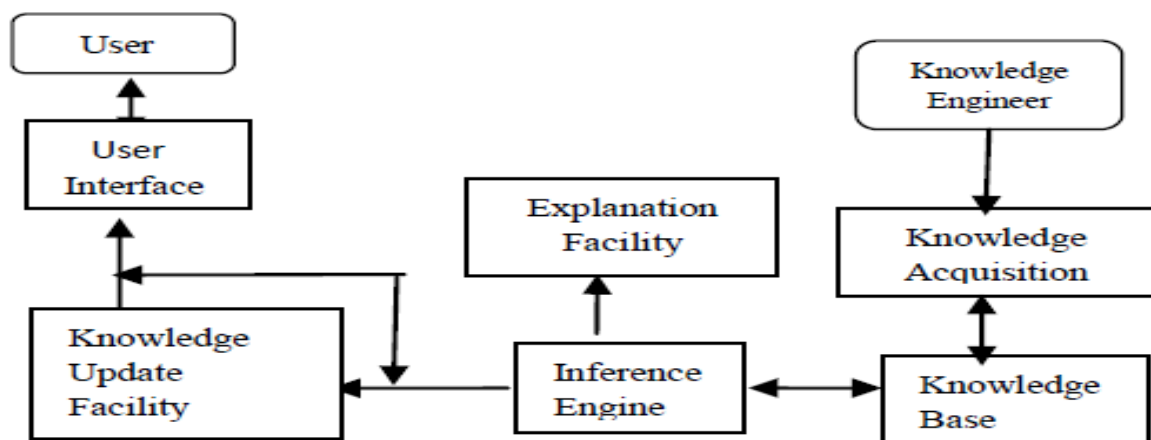


Figure 2-3: Architecture of Knowledge Based System (adopted from [33])

According to Tripathi [33], the core components of expert systems (Knowledge Based System) are the knowledge base and the reasoning engine. It is discussed as follows:

**Knowledge Base:** The knowledge base consist rules (as a rule base) and facts( as a database) which are necessary for understanding, formulating and for solving problems. It is a warehouse of the domain specific knowledge captured from the human expert via the knowledge acquisition module. To represent the knowledge production rules, frames, logic, semantic net etc. is used.

The knowledge available in the previous steps must be represented in an efficient manner. Knowledge engineer must have clear understanding of the representation scheme so that problem can be easily handled. Knowledge can be declarative or procedural. Declarative knowledge is a sentence depicting facts while procedural knowledge is a group of sentences, which can be carried in a series of steps by building relationships. Knowledge must be encoded either as a part of expert systems or should be easily available to them. An Inference engine detects these knowledge representation schemes which are in the form of the following: rules, semantic nets, frames and ontology. At times, there could be multiple knowledge representation schemes that describe the given problem.

Rule based reasoning as described in previous section depicts how an expert gives explanation about the conclusion with the help of rules. But it cannot describe why these rules apply on the problem. This could be because of insufficient knowledge about a particular subject domain in the knowledge base system. Semantic net is a graphical representation of nodes (objects) labeled and followed by links (relationships) to describe knowledge in a pictorials manner. Knowledge engineers understand these networks easily as they are presented in the hierarchical manner of using expert's knowledge. Frames are also hierarchal in nature. They also give a simplified method of storing knowledge in form of a table. The knowledge used in one frame can inherit properties used in another frame. Ontology refers to representations of specific standard symbols which are used in knowledge, expert systems and knowledge base systems for the particular purpose [34].

**Inference Engine:** Inference Engine is a brain of expert system. It uses the control structure (rule interpreter) and provides methodology for reasoning. It acts as an interpreter which analyzes and processes the rules. It is used to perform the task of matching antecedents from the responses

given by the users and firing rules. The major task of inference engine is to trace its way through a forest of rules to arrive at a conclusion.

The purpose of the inference engine is to seek information and relationships from the knowledge base and to provide answers, predictions, and suggestions in the way a human expert would. The inference engine must find the right facts, interpretations, and rules and assembles them correctly [30].

There are two broad kinds of inference engines used in rule-based systems: forward chaining and backward chaining systems

**Forward chaining** starts with the facts and works forward to the conclusions. In a forward chaining system, the initial facts are processed first, and keep using the rules to draw new conclusions given those facts.

**Backward chaining** is the process of starting with conclusions and working backward to the supporting facts. In a backward chaining system, the hypothesis (or solution/goal) we are trying to reach is processed first, and keep looking for rules that would allow concluding that hypothesis. As the processing progresses, new sub-goals are also set for validation. Forward chaining systems are primarily data driven, while backward chaining systems are goal-driven [35].

**Knowledge engineers:** Knowledge Engineering plays a vital role in the development of various technologies. These technologies are expert systems, neural network, artificial intelligence, intelligent systems, data mining, decision support systems, knowledge based systems etc. It design, build, and debug the knowledge base in consultation with domain experts. In consequence, knowledge engineering is no longer seen as a direct transfer process of human knowledge in a production rule implementation. Instead, a model of the required functionality of the KBS and the knowledge that is required to achieve the functionality is built to bridge the large gap between informal requirements and human expertise on the one hand and an implemented system on the other hand. Meanwhile, the modeling paradigm has become quite dominant in the knowledge engineering area and a number of principles are shared by most approaches [34].

**Knowledge Acquisition:** is the accumulation, transfer and transformation of problem-solving expertise from experts and/or documented knowledge sources to a computer program for constructing or expanding the knowledge base. It is a subsystem which helps experts to build

knowledge bases. For knowledge acquisition, techniques used are protocol analysis, interviews, and observation.

There exist several methods to extract human knowledge; some commonly used approaches of knowledge acquisition include interviews, questionnaires, record reviews and observation to acquire factual and explicit knowledge [36]. Knowledge acquisition is comprised of two tasks: knowledge elicitation and knowledge representation. In knowledge elicitation, domain knowledge is obtained through various means including interviews with experts and book and journal references. In knowledge representation, the elicited knowledge is converted to a form for efficient computer manipulation [37].

Knowledge acquisition is the process in which knowledge engineers acquire knowledge from human experts, group of experts, documentation, manuals, which helps in building complete, accurate and well organized knowledge based systems. Knowledge is acquired in the form of interviews (unstructured, structured, event recall, vast thinking) that can introduce some important topics, observations, filling questionnaires from patients/relatives, Expert Consultants/Practitioners etc. The role of recalling focuses on completion of knowledge obtained by cross checking expert's knowledge obtained in the form of reasoning how and why. Group of experts giving their respective opinion can create internal conflicts. They might use different reasoning steps. Therefore, there must be a judge who is specialized in that problem domain with more experiences gathered to conclude the steps to be followed in a specific decision making process [38].

***Explanation Facility:*** It is a subsystem that explains the system's actions. The explanation can range from how the final or intermediate solutions were arrived at to justifying the need for additional data. Here user would like to ask the basic questions why and how and serves as a tutor in sharing the system's knowledge with the user. It includes the design and implementation of explanation capability. For example, the ability to answer questions such as why a specific piece of information is needed by the computer or how a certain conclusion was derived by the computer [7].

Explanation subsystem helps in justification of Expert System conclusions by tracing conclusions to their sources and showing how was a certain conclusion reached. One of the key characteristics of an expert system is the explanation facility. With this capability, an expert system can explain how it arrives at its conclusions. The user can ask questions dealing with the what, how, and why aspects of a problem. The expert system will then provide the user with a trace of the consultation process, pointing out the key reasoning paths followed during the consultation [30].

**User Interface:** It is a means of communication with the user. It provides facilities such as menus, graphical interface etc. to make the dialog user friendly. Responsibility of user interface is to convert the rules from its internal representation (which user may not understand) to the user understandable form.

The user interface is the part of the program that allows the user to communicate with the system. The user interface will ask questions, present menu driven choices, and communicate to the user the answer or solution once it has been found [39].

To build the expert system is known as Knowledge Engineering. Personnel involved in expert system development are domain expert, user, knowledge engineer and system maintenance personnel. Domain expert has special knowledge, judgment, experience and methods to give advice and solve problems. It provides knowledge about task performance. Knowledge engineer is involved in the development of the inference engine, structure of the knowledge base and user interface.

## 2.6 AI Programming Languages

Programming Languages in Artificial Intelligence (AI) are the major tool for exploring and building computer programs that can be used to simulate intelligent processes such as learning, reasoning and understanding symbolic information in context [40]. The most common ones are Prolog and Lisp.

**Prolog** is a relational language. That is, a Prolog call may be regarded as expressing a relation between its arguments. This can be contrasted with functional languages where a call may be regarded as a mathematical function. Both functional and relational languages however are considered declarative, that is they break away from the idea of a programming language as a set

of instructions to manipulate the memory of a computer (such languages are referred to as imperative), and instead concentrate on providing a precise notation in which computation is a form of reasoning. Prolog originates from attempts to provide a computer-executable form of mathematical logic notation (its name comes from Programming in Logic), thus it and similar languages are often referred to as the logic languages [41].

According to Matthew [41], Prolog Procedures for a k-ary relation (i.e. one which takes k arguments), named proc takes the form:

*Proc(X1, ... Xk) :- call1, ..., calln.*

Where X1,... Xk are Prolog variables, and call1,..., calln Prolog procedure calls. Note that a variable name may be any alphanumeric string starting with a Capital letter, while a procedure name may be any alphanumeric string starting with a small letter. The underscore character '\_' is treated as an additional capital letter in names. Each of call1,..., call n may contain variables from X1, ... Xk among its arguments, or new variables, or constants.

**Lisp** it is Artificial Intelligence language developed in 1958 by J. McCarthy at MIT. It is the first functional programming language: It was invented to support symbolic computation using linked lists as the central data structure (Lisp stands for List processor). Lisp lists may be used to model every other data structure. However, Prolog allows the use of named structures. A structure is written simply with a structure name (which may be any Prolog string) and a list of its arguments enclosed within brackets [41].

## Machine Learning

Machine learning is all about learning rules from data. Machine learning is all about making computers learn from experience. These techniques contain concept learning. Weka is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It is designed so that users can quickly try out existing machine learning methods on new datasets in very flexible ways. It provides extensive support for the whole process of experimental Data Mining, including preparing the input data, evaluating learning schemes statistically, and visualizing both the input data and the result of learning [20].

## 2.7 Related works

Local researches are conducted to assess the application of DM in the different sectors like Airlines, Banking, Insurance, HealthCare, and Customs. Tariku [9] tried to develop models that can detect and predict fraud in insurance claims with a particular emphasis on Africa Insurance Company. He tried to apply clustering algorithm followed by classification techniques for developing the predictive model. K-means clustering algorithm is employed to find the natural grouping of the different insurance claims as fraud and non-fraud. In his research he used a total of 17547 datasets for the purpose of conducting this study. Henock [42] and Denekeew [43] for example, conducted a research on the application of DM for customer relationship management in the airlines industry as a case study on Ethiopian Airlines. Both [42] and [43] used clustering and classification techniques with k-Means and decision tree algorithms. In addition, Kumneger [44] has also tried to study the application of DM techniques to support customer relationship management for the Ethiopian Shipping Lines. Kumneger [44] has applied clustering and classification techniques with k-Means and decision tree algorithms.

Shegaw [45], also conducted a research on the application of DM in predicting child mortality in Ethiopia as a case study in the Butajira Rural Health Project. The methodology which was used by the researcher had three basic steps. These were collecting of data, data preparation and model building and testing. BrainMaker and See5 softwares were employed by the researcher so as to build models using neural net work and decision tree techniques respectively. The researcher found that both the techniques yield comparable results for misclassification rates. However, unlike the neural network models, the results obtained from decision tree models provided simple and understandable rules that can be used by any health care professionals to identify cases for which the rule is applicable. In fact, the accuracy obtained from decision tree models also outperforms neural networks. Shegaw [45], also found that best models using neural network technique by modifying default parameters of the program. For example, Tilahun [46] has tried to assess the possible application of DM techniques to target potential VISA card users in direct marketing at Dashen Bank. Melkamu [47] also conducted a research to assess the applicability of DM techniques to CRM (Customer Relationship Management) as a case study on Ethiopian Telecommunications Corporation (ETC).

Many researchers applied different DM techniques and algorithms for detecting fraudulent claims in the insurance industry. Rejesus et al [48], conducted a research to study the application of DM technology in detecting crop insurance fraud for the US Department of Agriculture. Applied the outlier analysis DM technique in helping to detect anomalous behaviour in an agricultural datasets. Guo [3] also applied k-Means clustering, decision tree algorithm, and logistic regression for modeling insurance risk in the property/casualty insurance. In addition to this, Koh and Gervais [50] applied DM techniques and algorithms for detecting motor insurance fraud. Koh and Gervais [50] used the clustering and expectations approaches. Further, DM techniques such as outliers clustering and decision trees were used to generate the clustering results and expected/predicted values in this research. Woodfield [5] also conducted a research for Predicting Workers' Compensation Insurance Fraud Using SAS Enterprise Miner and SAS Text Miner by applying the clustering technique.

There are a number of researches conducted in this country towards the application of DM in the insurance Industry. To the knowledge of the researcher there is no attempt conducted to integrate data mining findings with Knowledge base system using local data. Tariku [9], tried to develop model that can detect and predict fraud in insurance claims with a particular emphasis on Africa Insurance Company. Mesfin [51], had tried to study the possible applications of predictive DM techniques in the renewal process of personal accident policies taking Ethiopian Insurance Corporation as a case. The other research, which is undertaken by Tesfaye [8], has tried to develop a predictive model using DM techniques in support of insurance risk assessment as a case study on Nyala insurance share company. He applied classification technique and neural network algorithm for developing the model, which determines the risk exposure of motor insurance policies.

On the other hand, Abdulkarim [7] has tried to integrate data mining induced knowledge with knowledge based system using KDDcup'99 dataset available at ACM Knowledge discovery [online]. He tried rule based intrusion detection and advising knowledge based system to utilize hidden knowledge extracted by employing induction algorithm of data mining, specifically JRip from sampled KDDcup'99 intrusion data set. He employed the classification technique, rule based learning (production rules) and decision tree algorithms to detect network intrusion detection.

## CHAPTER THREE

### Knowledge acquisition using data mining

Awash insurance company S.C. (AIC) is one of the first few pioneer private insurance companies in Ethiopia launched following the liberalities of the financial sector in 1994. The company has subscribed capital of 80,000,000 birr [52]. AIC has 33 branches all over the country, 21 branches located in the capital, Addis Ababa, and 12 branches situated in major town across the country. Ever since its establishment starting from very low levels the company has been registering remarkable sustained increase and improvement in sales and production of life and non life institutional, financial, human resource, materials and technological capacities. The management's friendliness and responsiveness to the researcher's enquiries to get access to and collect the required data for this study encouraged the researcher to conduct this research at AIC.

The management of the company constitutes Marketing, Finance, Underwriting, Claims Handling, Reinsurance and Human Resource Development Managers that forms the management team of AIC, which is led by the Board of Directors. From the whole departments of AIC the claims department is responsible for handling insured's claims. Today, AIC provides insurance covering Life, Property and Liability Risks. Currently, the company carries out transactions of all types of insurance policies. Table 3-1 shows all services Awash insurance company S.C. (AIC) provides.

| Type of Insurance                  | Description                                                                                                                                                                                                                     |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Motor                              | Insurance purchased for cars, trucks, and other vehicles                                                                                                                                                                        |
| Fire & lighting, and Allied perils | Insurance against loss due to fire                                                                                                                                                                                              |
| Marine cargo                       | Insurance covers for the loss or damage of ships                                                                                                                                                                                |
| Workmen's Compensation             | Provides wage replacement and medical benefits for employees who are injured in the course of employment, in exchange for mandatory relinquishment of the employee's right to sue his /her employer for the tort of negligence. |
| Personal & Group                   | Provides protection against death or dismemberment arising solely and directly from an accident                                                                                                                                 |

|                                                   |                                                                                                                                                           |
|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
| Accident                                          |                                                                                                                                                           |
| Money in safe and transit                         | Provides cover for money when carried by the insured or employees                                                                                         |
| Fidelity Guarantee                                | Covers crime and theft by a firms employees                                                                                                               |
| Engineering( also called the munich Re, policies) | Coverage against losses associated with an engineering firm                                                                                               |
| Burglary and House Breaking                       | Cover against loss or damage by burglary and/or by housebreaking                                                                                          |
| Public Liability                                  | Coverage against harm that an insured have caused to a third party                                                                                        |
| Plate Glass                                       | Cover for accidental loss or damage to the plate glass                                                                                                    |
| Bonds                                             | Guarantees scheduled payments of interest and principal on a bond other security in the event of a payment default by the issuer of the bond or security. |

Table 3-1 Types of Insurance Services AIC Provides

Motor Policy is a contract concluded between the insurer and the insured to cover the insured's legal liability and/or property up to an agreed limit as stipulated on the contract on payment of a sum of money called a premium to the insurer for an agreed period of time [52].

According to AIC Underwriting Department policies are classified according to the indemnity, which they provide and the indemnities provided by the company as follows:

*Comprehensive cover:* a comprehensive cover is the widest form of cover available, although it cannot protect against every conceivable risk. It not only covers for third party claims after an incident, but also covers for damage caused to the insurer's vehicle. It also covers all damage/loss to the vehicle and its accessories caused by any accidental, external and physical means as a result of Collision, Overturning, Fire, Self-Ignition, Lightning, Explosion, Burglary, and House Breaking, Theft and Malicious acts.

Generally, the main comprehensive covers can be summarized as [9]:

- ❖ Insured's legal liability for death or bodily injury.
- ❖ Insured's legal liability for damage caused to the property of third party, arising out of the use of the insured's vehicles.
- ❖ Indemnity for damage to the insured's vehicle.

*Third party:* covers injuries caused to other people in an accident, but not damages caused to their property or vehicle and it is mandatory in the Ethiopian law enforcement

*Third party fire and theft:* third party can be extended to cover insured's vehicle against the risk of fire and theft in addition to the third party.

In the insurance practice claims that are reported quickly with private policy type and fast accident report date are rated as suspicious of non-fraudulent claims. Apart from this, those claims with a slow accident report date, Third Party policy type, and with a very high regain amount of the insured are also rated as suspicious of fraudulent claims.

### 3.1 Claims handling processes

The claim department of AIC undertakes processes, such as, notification of claim, completion of claim form, recording, reporting, survey/investigation, and claims payment approval. Immediate notification of accident enables insurers to investigate circumstances while they are fresh and before witness memory about the incidence is faded and its mandatory to report if incident happen in 24hours. Underwriting and accounts verification helps to get the information about Name of insured, Period of insurance, Type of coverage, vehicles Plate number, Chassis Number, Engine Number, and Year of make, Excess, Sum insured, Outstanding premium, Bank or other interest (if any), and Previous claims records if there exist any.

Claim related to motor insurance needs further survey and investigation. This helps the insurer to compare garage repair estimates with the estimates given by the surveyor and reach into a compromise or in cases of total loss claims, it helps the insurer to determine the market value of the object (vehicles) immediately before the loss and the salvage value of the object.

There is always a line of authority on the decision made on claims approval. The line of authority in AIC takes the payment up to Birr 30,000.00 is undertaken by the Claim officer while payment above Birr 30,000.00 is undertaken by the Managing Director (CEO).

For the purpose of data collection, all units were chosen from the branches of AIC. All these service units are connected with each other at the head office of the company, which is found at twins' tower around Beherawi. So, all the necessary data about each service unit is collected from the company database of the main office.

Fraud, or criminal deception, will always be a costly problem for many profit organizations. Data mining can minimize some of these losses by making use of the massive collections of customer data, particularly in insurance, credit card, and telecommunications industries [3].

Data mining is to find out information with special meaning from a great number of data by some technology as the procedure to discover knowledge from the database. So, the data mining is to adopt the mining procedure to discover unknown knowledge and rules from plentiful data, [3]. The result of the DM is base for the knowledge base system and we can use the findings of DM to design predictive model.

In this research, unsupervised approaches (K-Means clustering) have been used to segment the insurance data into clusters for supervised approaches (PART rule based algorithms is used). Because the dataset contains unlabelled data the clustering algorithm is applied to segment into different clusters depending on the fraudulence nature of the claims.

### **3.2 K-Means Clustering**

Clustering is an important data mining task employed in dataset exploration and in other settings where one wishes to partition sets into related groups. Among the algorithms typically used for clustering, k-means is arguably one of the most widely used and effective clustering methods. Successful use of k-means, however, requires a carefully chosen similarity measure that must be constructed to fit the task at hand [53].

Clustering is also known as data segmentation or partitioning and is regarded as a variant of unsupervised classification. It is an unsupervised data mining technique, deals with the problem

of dividing a given set of entities into meaningful subsets. Claims with similar characteristics are grouped together and small population clusters are flagged for further investigation [2].

### **K-Means algorithm**

Kiri Wagsta et al [54], defined K-means clustering as a method commonly used to automatically partition a data set into  $k$  groups. It proceeds by selecting  $k$  initial cluster centers and then iteratively refining them as follows:

- ❖ Each instance  $d_i$  is assigned to its closest cluster center.
- ❖ Each cluster center  $C_j$  is updated to be the mean of its constituent instances.

The algorithm converges when there is no further change in assignment of instances to clusters.

K-means algorithm runs in the following steps:

- ❖ Initial cluster seeds are chosen (at random).
  - These represent the “temporary” means of the clusters.
- ❖ The squared Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster.
- ❖ For each cluster, the new centroid is computed –and each seed value is now replaced by the respective cluster centroid.
- ❖ The squared Euclidean distance from an object to each cluster is computed, and the object is assigned to the cluster with the smallest squared Euclidean distance.
- ❖ The cluster centroids are recalculated based on the new membership assignment.
- ❖ Steps 4 and 5 are repeated until no object moves clusters.

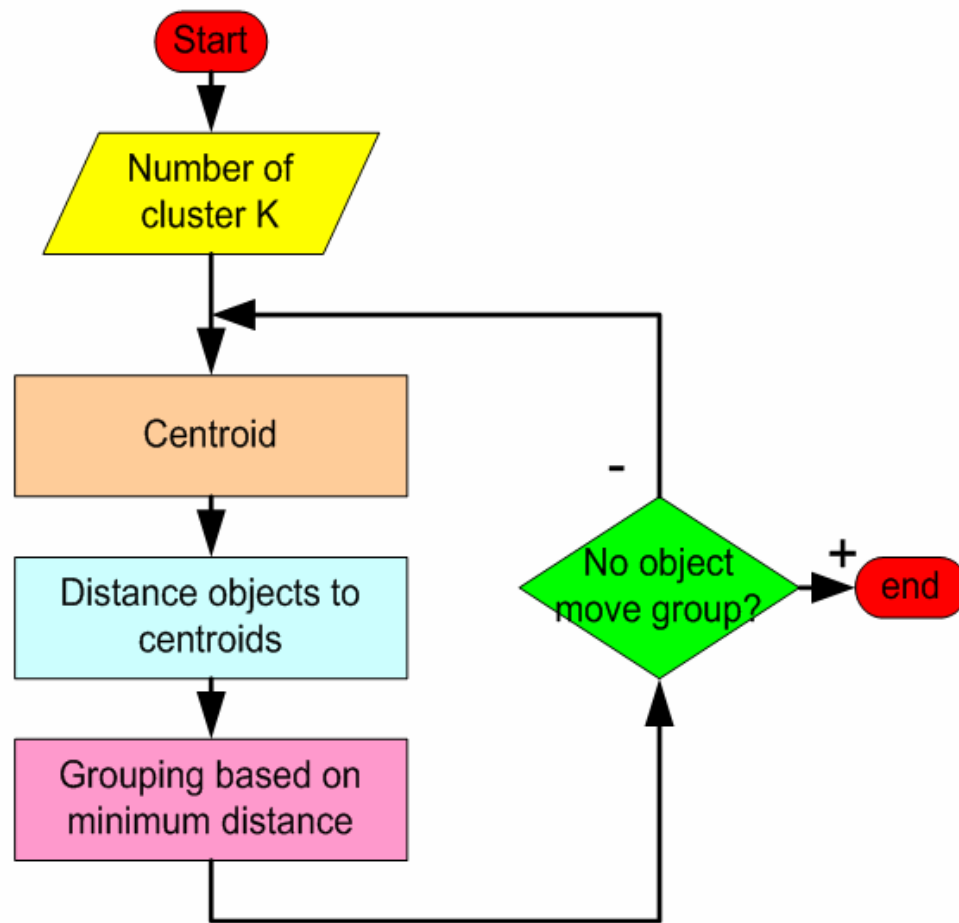


Figure 3-1: Diagrammatic representation of K-Means by Kardi Teknomo [55].

After we use clustering technique, the grouped data should have to be interpreted. There are a number of techniques to interpret the data. We use the classification technique PART Rule based algorithm to drive rules explaining how to assign new records to the correct cluster.

The most common clustering techniques are the K-nearest neighbor, the Naïve Bayes technique and K means. K means is selected because of the following reasons:

- ❖ K-means is a simple procedure for extracting clusters from data.
- ❖ K-Means algorithm is implemented in the WEKA data-mining tool, which has been applied in this research. Besides, the researcher found it easy to interpret the clustering results obtained from the K-Means algorithm.

The algorithm converges when there is no further change in assignment of instances to clusters.

K-means algorithm runs in the following steps:

- ❖ Initial cluster seeds are chosen (at random).
  - These represent the “temporary” means of the clusters.
- ❖ The squared Euclidean distance from each object to each cluster is computed, and each object is assigned to the closest cluster.
- ❖ For each cluster, the new centroid is computed –and each seed value is now replaced by the respective cluster centroid.
- ❖ The squared Euclidean distance from an object to each cluster is computed, and the object is assigned to the cluster with the smallest squared Euclidean distance.
- ❖ The cluster centroids are recalculated based on the new membership assignment.
- ❖ Steps 4 and 5 are repeated until no object moves clusters.

### 3.3 Rule based algorithm

To design Knowledge based system it needs to embed the rules we get from the classification algorithm (Fraud and Non-fraud). We use rule based classification algorithm to represent the knowledge. Rule based algorithm uses IF – THEN form to represent a knowledge.

The basic form of a rule is the following:

```
if <conditions>
then <conclusion>
```

Where <conditions> represents the conditions of a rule, whereas <conclusion> represents its conclusion. The conditions of a rule are connected between each other with logical connectives such as, AND, OR, NOT, etc, this forming a logical function. When sufficient conditions of a rule are satisfied, the conclusion is derived and the rule is said to fire (or trigger). Rules represent general knowledge regarding a domain, Jim Prentzas et al [56].

To build a rule-based classifier, we need to extract a set of rules that show the relationships between the attributes of a dataset and the class label. Each classification rule is a form  $R: (\text{Condition}) \Rightarrow y$ . Here the (Condition) is called the rule antecedent, which is a conjunction of the attribute test condition.  $y$  is called the rule consequent and it is the class label. A rule set can consist of multiple rules  $RS = \{R1, R2, \dots, Rn\}$ . The Accuracy of a rule is the fraction of instances that satisfy both the antecedent and consequent of a rule, normalized by those satisfying the antecedent [57].

We can represent these rules in prolog (Programming Logic) programming language. A rule consists of a head (a predicate) and a body (a sequence of predicates separated by commas). Head and body are separated by the sign ‘:-’ and, like every Prolog expression, a rule has to be terminated by a full stop.

### **3.4 Data selection and preparation**

Data from year 2013 to 2015 from all branches of Awash insurance company which is a total of 33 branches connected to each other to central server is used in this study. Some attributes instances did not contain their branch abbreviation in their policy number. Policy Number have its own format having first Insurance company name, second Branch name and third the vehicle use like commercial or private. Policy Number identifies the user uniquely and the year of the insured. By consulting the domain experts and by referring further sources Branch name is assigned on each policy number. E.g. AIC/ADD/CMTP/066818/14.

A total of 17547 instances with 13 attribute are collected for this study from Awash insurance company database. The company computerized its activities in 2013. According to the information researcher get from the company claim manger, the company relied on pioneer customers stayed with company from the beginning. But when the company migrate the existing data they exclude to move customer profile (which include the customer’s years stayed) into the computerized system. Due to this the researcher cannot include data related to customer’s year stayed. Its five years since the company start to computerize its activities and three years after fully operational. Awash insurance company has been in existing for the last 20 years after its establishments.

From the data set some information are incomplete and insufficient. For example claim closed reason attribute is essential for the fraud detection but only 152 of 14578 are filled. This implies that there is a problem in filling information in the company.

In the preprocessing, weka tool is used to replace missing values .Replaces all missing values for nominal and numeric attributes in a dataset with the modes and means from the training data.

```
weka.filters.unsupervised.attribute.ReplaceMissingValues
```

Weka provides both attribute evaluator and search method. In the attribute evaluation there are number of attribute evaluation methods. GainRatioAttributeEval and InfoGainAttributeEval attribute evaluator are selected because the researcher is familiarized with attribute evaluator and Rank search method.

GainRatioAttributeEval: - evaluates the worth of an attribute by measuring the gain ratio  
with respect to the class.

InfoGainAttributeEval: - evaluates the worth of an attribute by measuring the information  
gain with respect to the class.

```
=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 Pur_Veh):
  Gain Ratio feature evaluator

Ranked attributes:
0.47237  10 POLICY_TYPE
0.03348   1 VEH_TYPE
0.03252   8 Branch Unit
0.01617   3 Peril_Name
0.0067    4 Cause_of_Loss
0.00436   7 Claim_Clsd_Rsn
0.00391   6 Claim_Status
0.00347   5 CLM_REG_AMT
0.00304   2 CLAIM_REP_DI_G
0         9 Policy_Status

Selected attributes: 10,1,8,3,4,7,6,5,2,9 : 10
```

Figure 3-2: Attribute evaluation using GainRatioAttributeEval

GainRatioAttributeEval is selected due to its suitability with the data at hand. The result of the ranking leads us to the potential attributes to detect fraud. According to the GainRatioAttributeEval evaluator the following attributes get high gain ratio and are selected for the research.

### **3.5 Data reduction and Processing**

Data reduction is one of the tasks in data mining that needs to be done before the actual mining task takes place. Although, 17547 instances are collected, after preprocessing the redundant instance and substitute missing values, the researcher re-sampled the dataset into 14576 instances. After preparing the required sample dataset, the data is converted into comma delimited Excel file (CSV format), then to the *arff* file suitable for mining using WEKA 3.7.

### **3.6 Experimentation**

Having finished the data preprocessing and preparation task in format suitable for WEKA data miner, the next task is undertaking the experiment involving the selected classifier algorithms.

#### **3.6.1 Experimental set up**

The sampled data set contains 14576 instances with 13 attributes and 10 of them are taken for this research based on the attribute evaluation technique. Default value of parameters in WEKA Data Mining tool for the classifier algorithms is taken into consideration, since it allows achieving better accuracy compared to modifying the default parameters values.

| Parameters           | Description                                                                                                                 | Default values |       |       |         |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------|----------------|-------|-------|---------|
|                      |                                                                                                                             | PART           | JRip  | J48   | REPTree |
| binarySplit          | Tells about whether to use binary splits on nominal attributes when building the partial trees.                             | -              | -     | False | -       |
| minNumObj            | The minimum number of instances per rule                                                                                    | 2              |       | 2     | -       |
| numFolds             | Determines the amount of data used for reduced-error pruning. One fold is used for pruning, the rest for growing the rules. | 3              | 3     | 3     | 3       |
| seed                 | The seed used for randomizing the data when reduced-error pruning is used.                                                  | 1              | -     | 1     | 1       |
| checkErrorRate       | Whether check for error rate $\geq 1/2$ is included in stopping criterion                                                   | -              | True  | -     | -       |
| debug                | Whether debug information is output to the console.                                                                         | False          | False | False | False   |
| useLaplace           | Whether counts at leaves are smoothed based on Laplace.                                                                     | -              | -     | False | -       |
| minNum               | The minimum total weight of the instances in a leaf                                                                         | -              | -     | -     | 2.0     |
| binarySplits         | Whether to use binary splits on nominal attributes when building the partial trees.                                         | False          | -     | -     | -       |
| optimizations        | The number of optimization runs.                                                                                            | -              | -     | -     | -       |
| usePruning           | Whether pruning is performed.                                                                                               | True           | -     | -     | -       |
| reducedError Pruning | Whether reduced-error pruning is used instead of C.4.5 pruning.                                                             | False          | -     | False | -       |

Table 3-2 Default parameters and values for algorithms

According to the insurance policy the claim notification time is 48 hours although they tolerate claim to be filled for ten days according to the information from the under writer. The customer can report the report from anywhere by telephone.

Threshold for claim report date gap

| No | Claim_Re_Gap |          |
|----|--------------|----------|
| 1. | 0- 5         | Fast     |
| 2. | 6-7          | Moderate |
| 3. | > 8          | Slow     |

Table 3-3 Thresholds used to cluster result of claim report date gap

But if the claim dates exceed 10 days without notification to the insurance company the insurance will never accept the claim and it will be rejected.

Claim regain amount: - Is the difference between the estimated loss by the company surveyor and total expense paid for the loss. The claim officers categorize this depending on the amount of percent exceeded

| No | Cl_regain_amt |           |
|----|---------------|-----------|
| 1  | >10%          | Low       |
| 2  | 10% - 15 %    | Moderate  |
| 3. | <= 30%        | High      |
| 4  | >30%          | Very High |

Table 3-4 Thresholds used to cluster result of claim Regain amount.

The percentage difference is computed in excel from the *Estimated loss To Total Claims Expenses Paid* using the following formula:

$$(Estimated\_loss - Total\ Claims\ Expenses\ Paid) / Estimated\_loss$$

Datapreparator 1.7 is used to process the data. It is a tool lets you import information from *aff*, *csv* or plain *text file* format. This software is chosen to handle outliers and missing values. To handle outliers the program uses Z score.

Consequently, the researcher together with the domain experts removed those attributes, which have less significance to the research. The Table 3-4 below shows the final list of attributes that have been used in this study

| No | Attribute Name | Data Type | Description                                                                                               | Remark  |
|----|----------------|-----------|-----------------------------------------------------------------------------------------------------------|---------|
| 1  | VEH_TYPE       | Varchar   | The length of accident report date since the occurrence of accident                                       |         |
| 2  | CLAIM_REP_DT_G | Number    | The length of accident report date since the occurrence of accident                                       | Derived |
| 3  | Peril_Name     | String    | A specific risk or cause of loss covered by an insurance policy.                                          |         |
| 4  | Cause_of_Loss  | String    | The cause of accident that has occurred                                                                   |         |
| 5  | CLM_REG_AMT    | Number    | The amount of recovered money from the previous claim by the insured                                      | Derived |
| 6  | Claim_Status   | String    | Status of the claim                                                                                       |         |
| 7  | Branch Unit    | Varchar   | Branch service unit. The data type this attribute was Number initially.                                   |         |
| 8  | Policy_Status  | Varchar   | Refers to the status of the policy                                                                        |         |
| 9  | POLICY_TYPE    | Varchar   | Type of policy the insured buy from the insurance                                                         |         |
| 10 | Pur_Veh        | String    | The type of vehicle use. It is to mean that whether the vehicle is used for private or commercial purpose |         |

Table 3-5 The Final List of Attributes used in the Study

The experiments are conducted by changing the default parameters of the simple K-Means algorithm. These experiments are presented and discussed in this section. From the three different experiments conducted in this study, the best is one is selected for developing the final cluster model of this study.

## Experimentation I

The first experimentation is done for K=2, with the default seed value and default distance function. All of the final selected attributes and 14576 records are used as an input for the experiment. As can be seen from this table, the distribution of the dataset for each cluster is presented.

| Clustering Result of the 1st Experiment |                    |            |                      |            |
|-----------------------------------------|--------------------|------------|----------------------|------------|
| K                                       | Distance Function  | Seed Value | Cluster Distribution |            |
|                                         |                    |            | C1                   | C2         |
| 2                                       | Euclidean Distance | 10         | 10484 (72%)          | 4092 (28%) |

Table 3-6 Training of the first experiment by the default parameter values

The first experiment is conducted with default values of the simple K-Means algorithm (K = 2, Seed = 10, and Euclidean distance function).

The table below shows results of the first experiment. The description for each segment of the cluster has been done as shown in the below Table 3-6. The ranking is determined based on the fraudulence nature of the insurance claims.

| Cluster # | Description                                                                                                                                                                                                                                                                                                        | Rank |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| 1         | Public_Transport_Buses_over_45_seats Vehicle Type, Fast report date, Motor own Damage Peril name, Overturning cause of loss, very high Regain amount, pending Claim Status, Settled claim closed Reason, Finfinne_Main Branch, renewed policy status, Commercial policy type, Public Transport Purpose of vehicle. | 1    |
| 2         | Pickup_OVR_2350cc Vehicle Type, Slow report date, Motor own Damage Peril name, Road condition cause of loss, very high Regain amount, pending Claim Status, settled claim closed Reason, Gotera Branch, renewed policy status, Commercial policy type, Own Goods Purpose of vehicle.                               | 2    |

Table 3-7 Cluster summary of the first experiment for K=2, seed=10, Euclidean distance and rank of clusters

According to the domain experts, fraudulent suspicious claims are those with late accident report date, in which the cause of loss is unknown, and a claim for a very high-recovered amount of money.

## Experimentation II

The second cluster experiment is done for K=2, changed default seed value (10) to 100 and with the default distance function (Euclidean distance). Similar to the first experiment, all of the final selected attributes and 14576 records are used as an input for conducting the experiment.

| Clustering Result of the 2 <sup>nd</sup> Experiment |                    |            |                      |            |
|-----------------------------------------------------|--------------------|------------|----------------------|------------|
| K                                                   | Distance Function  | Seed Value | Cluster Distribution |            |
|                                                     |                    |            | C1                   | C2         |
| 2                                                   | Euclidean Distance | 100        | 8897 (61%)           | 5679 (39%) |

Table 3-8 Training of the second experiment by changed Seed value =100 and other default parameter Values.

The description of the resulting cluster model is then interpreted and briefly explained as follows.

| Cluster # | Description                                                                                                                                                                                                                                                                                   | Rank |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| 1         | Private_Cars_FM Vehicle Type, Slow report date, Motor own Damage Peril name, Others cause of loss, very high Regain amount, Settled Claim Status, settled claim closed Reason, Teklehaimanot Branch, renewed policy status, THIRD_PARTY policy type, Own_service Purpose of vehicle.          | 1    |
| 2         | Pickup_OVR_2350cc Vehicle Type, Moderate report date, Motor own Damage Peril name, Windshield cause of loss, very high Regain amount, Notification Claim Status, settled claim closed Reason, GofaMazoria Branch, renewed policy status, PRIVATE policy type, Private_Use Purpose of vehicle. | 2    |

Table 3-9 Cluster summary of the 2nd experimentation for K=2, seed=100, Euclidean distance function and rank of clusters

As can be seen from the above Table 3-8, the first cluster is ranked 1st because of the reason that the claims that are assigned in this cluster have slow accident report date, clearly unknown cause of loss, and high-recovered amount of money. In addition, the type of accident that has occurred is collision with third party. The second cluster is ranked second because it has moderate accident report date, and although very high recovered amount of money. The domain expert explained that most of the time Pickup\_OVR\_2350cc are company owned vehicles and rare chance for fraud. Cluster 1 is considered as suspicious of fraudulent claims while cluster 2 is considered as non-fraudulent claims. Compared to the first run experimentation this one seems to create dissimilar clusters and essential for further processing and distinctive clustering results.

### Experimentation III

The result of the third cluster experiment, with K=2, seed =1000, and Manhattan distance function.

| Clustering Result of the 3 <sup>rd</sup> Experiment |                    |            |                      |            |
|-----------------------------------------------------|--------------------|------------|----------------------|------------|
| K                                                   | Distance Function  | Seed Value | Cluster Distribution |            |
|                                                     |                    |            | C1                   | C2         |
| 2                                                   | Manhattan Distance | 1000       | 10130 (69%)          | 4446 (31%) |

Table 3-10 Training of the third cluster experiment with K=2, Seed=1000 and Manhattan distance function

The description of the resulting cluster model is then interpreted and briefly explained as follows.

| Cluster # | Description                                                                                                                                                                                                                                                                                          | Rank |
|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| 1         | OG_TP___TRUCKS Vehicle Type, Slow report date, Motor_TP_Compulsory_Property_Damage Peril name, Others cause of loss, very high Regain amount, Settled Claim Status, settled claim closed Reason, Wallo_Sefer Branch, renewed policy status, THIRD_PARTY policy type, Own service Purpose of vehicle. | 1    |

|   |                                                                                                                                                                                                                                                                             |   |
|---|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|
| 2 | CAR_HIRE and TOUR Vehicle Type, Slow report date, Motor own Damage Peril name, Others cause of loss, very high Regain amount, Settled Claim Status, settled claim closed Reason, Shagar Branch, renewed policy status, PRIVATE policy type, Private_Use Purpose of vehicle. | 2 |
|---|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---|

Table 3-11 Cluster Summary of the third Experiment for K=2, seed=1000, Manhattan Distance Function and Rank of Clusters

Although, the third Experiment resulted with different segmentation from that of the first and the second cluster, Discussion with the domain experts revealed that the third experiment has low segmentation value, because the major clustering attributes are found in both clusters.

### 3.7 Creating Predictive Model

At this stage, four predictive models involving J48, REPTree, PART and JRip classifier algorithms are constructed. J48 and REPTree are tree based classifiers in WEKA whereas PART and JRip are rule based classifiers. All the four classifiers are capable of generating rules.

#### ❖ Experiment 1 – J48 classifier

This experiment is undertaken using J48 classifier involving its default value of parameters. J48 is an open source Java implementation of the C4.5 algorithm in the WEKA data mining tool. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan [58]. 10-Fold cross-validation is selected as test option. The tree size is 2007 and number of leaves is 1958. The algorithm correctly classified 13901 which are 95.36 % and only 675 which are 4.63 % instances are classified incorrectly taking 0.26 seconds to build the model.

| Classifier |           | Class   |           |        |           |
|------------|-----------|---------|-----------|--------|-----------|
|            |           | TP Rate | Precision | Recall | F-Measure |
| J48        | Non-Fraud | 0.921   | 0.912     | 0.921  | 0.917     |
|            | Fraud     | 0.966   | 0.970     | 0.966  | 0.968     |

Table 3-12 performance of J48 classifier

### ❖ Experiment 2 – REPTree classifier

The second experiment is based on REPTree algorithm. The default parameters with respective values and 10-Fold Cross-Validation test mode is used. According to this algorithm the number of Correctly Classified Instances are 14205 (97.45 %) and the number of Incorrectly Classified Instances are 371 which is 2.54% while taking 0.53 seconds to build the model. Although, the size of tree generated is 4174.

| Classifier | Class     |         |           |        |           |
|------------|-----------|---------|-----------|--------|-----------|
| REPTree    |           | TP Rate | Precision | Recall | F-Measure |
|            | Non-Fraud | 0.947   | 0.960     | 0.947  | 0.954     |
|            | Fraud     | 0.985   | 0.980     | 0.985  | 0.982     |

Table 3-13 performance of REPTree classifier

### ❖ Experiment 3 – PART classifier

In this experiment PART rule induction algorithm is employed. It generated 54 rules by involving all the attributes of the dataset and 10- fold cross-validation test option. In this experiment 14374 instances from 14576 which is 98.61% are Correctly Classified and Incorrectly Classified Instances are 202 which is 1.38% taking 0.28 seconds to build the model.

| Classifier | Class     |         |           |        |           |
|------------|-----------|---------|-----------|--------|-----------|
| PART       |           | TP Rate | Precision | Recall | F-Measure |
|            | Non-Fraud | 0.968   | 0.982     | 0.968  | 0.975     |
|            | Fraud     | 0.993   | 0.988     | 0.993  | 0.990     |

Table 3-14 performance of PART classifier

### ❖ Experiment 4 – JRip classifier

In this experiment JRip algorithm is employed. This classifier implements propositional IF-THEN rules from the data set with default values of the parameter and 10 – fold cross-Validation test model is employed. JRip Correctly Classified Instances are 13852 which is 95.03 % from 14576 instances and incorrectly Classified Instances are 723 which is 4.96 % taking 17.48 seconds to build the model. In addition, it generates 50 rules.

| Classifier | Class     |         |           |        |           |
|------------|-----------|---------|-----------|--------|-----------|
|            |           | TP Rate | Precision | Recall | F-Measure |
| JRip       | Non-Fraud | 0.941   | 0.877     | 0.941  | 0.908     |
|            | Fraud     | 0.950   | 0.977     | 0.950  | 0.963     |

Table 3-15 performance of JRip classifier

As we can see from the above experiments, two decision tree and two decision rule induction algorithms are developed. All the selected algorithms allow generating rules from the data set. The results of the algorithms are evaluated based on prediction accuracy in classifying the instances of the data set into Fraud or Non-Fraud.

As indicated in the table Table 3-15, the classifiers performed almost the same. There is a slight difference among the classifiers in terms of classifying the dataset correctly.

| Classifier  | Correctly classified instances |                | Incorrectly classified instances |                | Time take to build the model( in the second) |
|-------------|--------------------------------|----------------|----------------------------------|----------------|----------------------------------------------|
|             | No                             | Percentage (%) | No                               | Percentage (%) |                                              |
| J48         | 13901                          | 95.36 %        | 675                              | 4.63           | 0.26                                         |
| JRip        | 13852                          | 95.03          | 723                              | 4.96           | 17.48                                        |
| REPTree     | 14205                          | 97.45          | 371                              | 2.54           | 0.53                                         |
| <b>PART</b> | <b>14374</b>                   | <b>98.61</b>   | <b>202</b>                       | <b>1.38</b>    | <b>0.28</b>                                  |

Table 3-16 performance of classifiers

As we can see from the result of the experiments in the above tables, there is a slight difference between the results. PART has selected the best accuracy by classifying 14374 instances out of 14576 correctly. Results of J48, JRip and REPTree show nearly equal number of incorrectly classified instances. The highest incorrect classification is scored by JRip.

| Actual    | Predicted |       |
|-----------|-----------|-------|
|           | Non-Fraud | Fraud |
| Non-Fraud | 3893      | 130   |
| Fraud     | 72        | 10481 |

Table 3-17 Confusion Matrix for PART classifier

Apart from prediction accuracy, classifiers are also evaluated to measure how they correctly classified each class to their correct class or incorrectly classified to another class. Hence, to evaluate the performance of the classifiers employed in this study True Positive rate, Precision, Recall and F- measure are used.

| Classifier |         |           |        |           | Class     |
|------------|---------|-----------|--------|-----------|-----------|
| J48        | TP Rate | Precision | Recall | F-Measure |           |
|            | 0.921   | 0.912     | 0.921  | 0.917     | Non-Fraud |
|            | 0.966   | 0.970     | 0.966  | 0.968     | Fraud     |
| REPTree    | 0.947   | 0.960     | 0.947  | 0.954     | Non-Fraud |
|            | 0.985   | 0.980     | 0.985  | 0.982     | Fraud     |
| PART       | 0.968   | 0.982     | 0.968  | 0.975     | Non-Fraud |
|            | 0.993   | 0.988     | 0.993  | 0.990     | Fraud     |
| JRip       | 0.941   | 0.877     | 0.941  | 0.908     | Non-Fraud |
|            | 0.950   | 0.977     | 0.950  | 0.963     | Fraud     |

Table 3-18 Precision, Recall and F- Measure of Classifiers with respect to classes

With regard to FP (False positive) rate, J48, REPTree and JRip have registered almost similar values for both clusters. PART has registered the least FP rate (2.5%) for both clusters as compared to the other three algorithms.

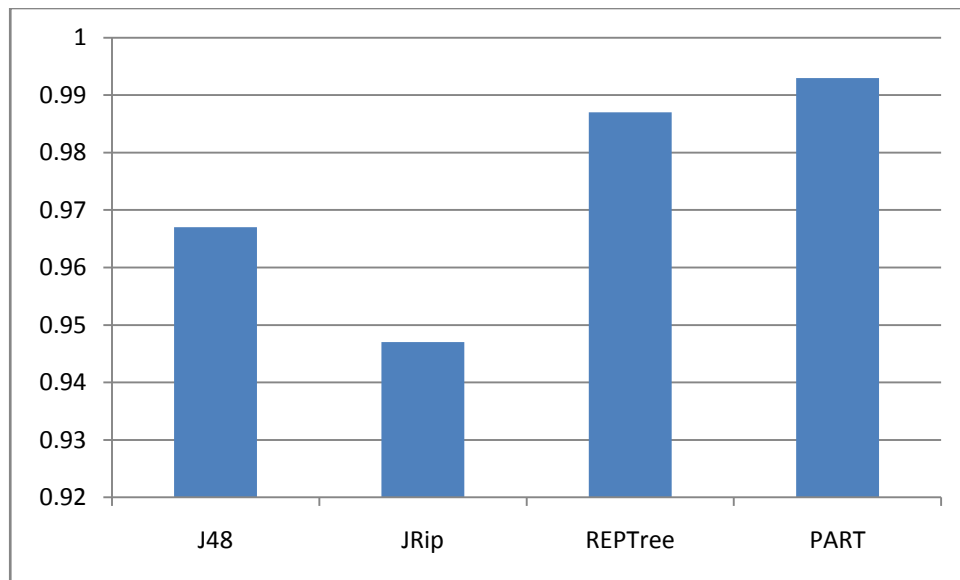


Figure 3-3: Time taken to build model by classifiers

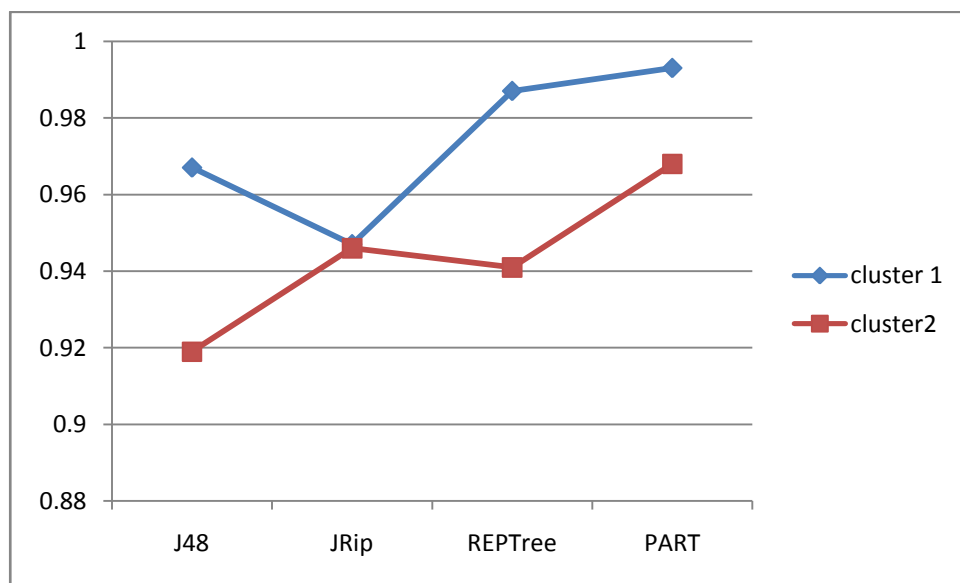


Figure 3-4: TP rate of classifiers

The rule acquired from the classifier algorithms is used for constructing knowledge base. So as to develop effective knowledge base system, acquiring relevant rules is paramount. Hence from the four algorithms the researcher selected the classifier which best performed on classifying the data set.

PART has best performance among the four classifiers. Its prediction accuracy and TP rate for Non-Fraud is 96.8% and for Fraud 99.3% which is a promising performance in predicting Fraud claims from the rest. PART has registered the least FP rate (2.5%) for both clusters as compared

to the other three algorithms. This shows the model developed using PART is acceptable for constructing the rule base of the knowledge base system. However, as shown in figure 3-3 PART classifier takes much time than others.

PART classifier algorithm has generated 54 rules as shown in Appendix I. 32 of them are Fraud types and the rest 22 are Non-Fraud. One can generalize that if certain claim satisfies one of the 32 rules it is Fraud otherwise it is Non-Fraud claim type.

In consultation with domain experts, the rules are evaluated to make sure whether they identify fraud suspicious claims. Based on the feedback from the domain experts, the rules have the capability of identifying Fraud suspicious but the idea raised is further investigation should be done to include different attributes especially the Traffic police report and garage report in the incident.

| Rule# | Rule                                                                                                                                              | Class     |
|-------|---------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 1     | Cause_of_Loss = Others AND POLICY_TYPE = THIRD_PARTY: Fraud (2687.0)                                                                              | Fraud     |
| 3     | Cause_of_Loss = Others AND POLICY_TYPE = COMMERCIAL: Fraud (1420.0/1.0)                                                                           | Fraud     |
| 4     | POLICY_TYPE = THIRD_PARTY AND Claim_Clsd_Rsn = settled AND VEH_TYPE = Private_Cars_FM: Fraud (489.0)                                              | Fraud     |
| 7     | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Peril_Name = Motor_TP_Passengers: Non-Fraud (199.0)                                  | Non-Fraud |
| 8     | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Peril_Name = Motor_Own_Damage AND Branch_Unit = Awash_Head_Office: Non-Fraud (157.0) | Non-Fraud |
| 9     | POLICY_TYPE = PRIVATE AND VEH_TYPE = Pickup_OVR_2350cc: Non-Fraud (181.0)                                                                         | Non-Fraud |
| 10    | POLICY_TYPE = PRIVATE AND VEH_TYPE = General_Cartage_Tr: Non-Fraud (83.0/2.0)                                                                     | Non-Fraud |

Table 3-19 Sample rule set Generated Using PART algorithm.

## CHAPTER FOUR

### Integration of data mining Results with Knowledge Based system

Findings of data mining are the base for the development of knowledge base system through the integrator application. Facts we get from Data mining can be represented as a rule in knowledge base. After knowledge acquisition is done using PART rule induction algorithm, which performs on data collected from awash insurance company S.C the facts extracted is represented in Knowledge base system. The main challenge here is how to use hidden knowledge extracted from data mining for knowledge based system.

#### 4.1 System design and Architecture

The general framework shows the data mining tasks used for generating knowledge from collection of dataset. Following the approval of rules, the generated knowledge (rule set) is represented in knowledge base. The framework discussed in below Figure 4-1.

Dataset: - the data set is created by collecting data from awash insurance company database. The data which has been used from 2013 to 2015 from all branches of Awash insurance company S.C. Since the research area is on Claim incidents, the data extracted is about 17,547 related to claim cases from the company databases.

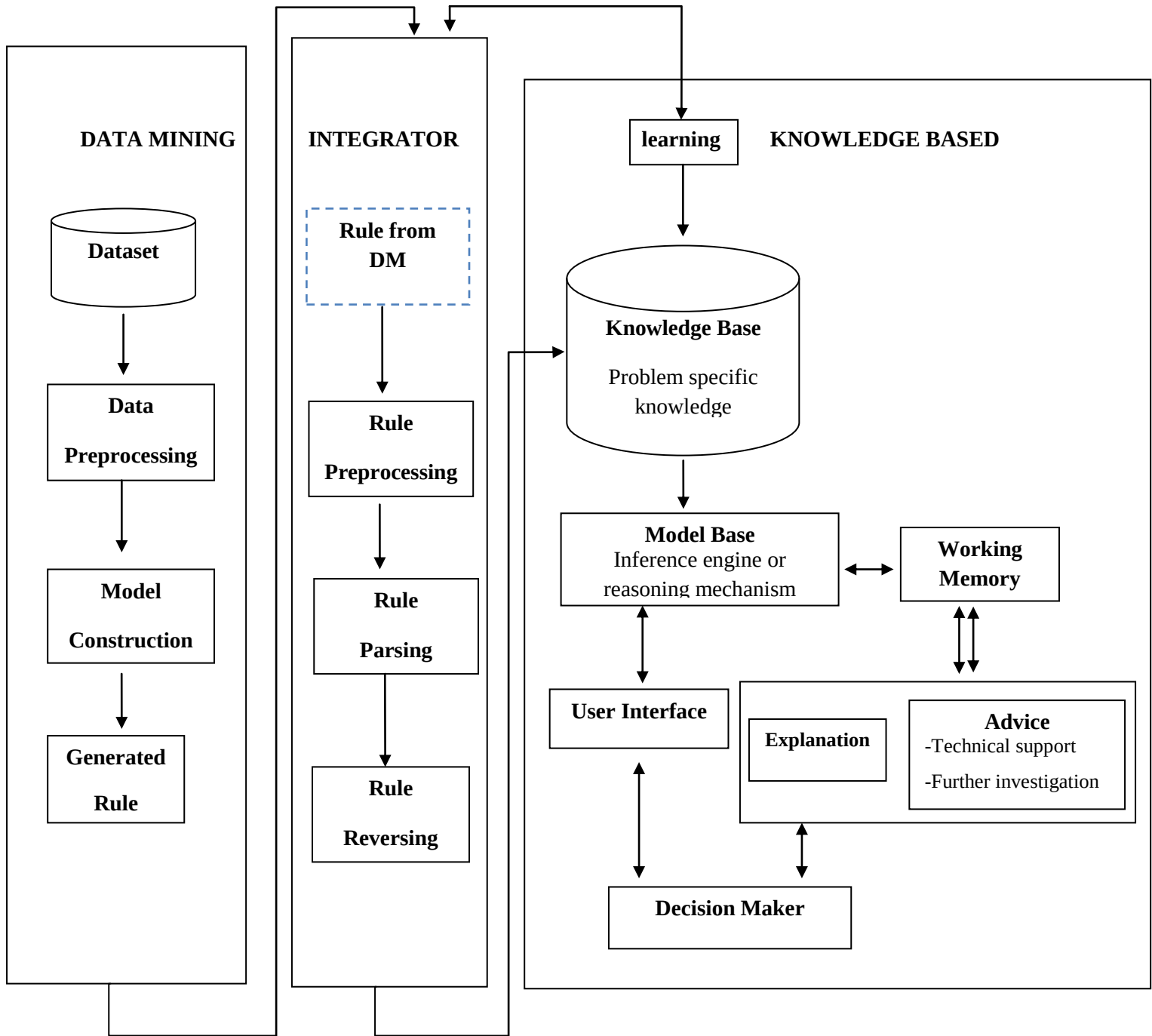


Figure 4-1: General Framework of integration of data mining model with knowledge based system

**Data preprocessing:-** According to Han [4], Today's real world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple heterogeneous sources. Low quality data will lead to low- quality mining results. The data set from the Awash Insurance Company S.C, which I used for this research, is highly prone to errors containing, missing and redundant values. From the data 17.11% is redundant data, which means from total of 17,547 records only 14,577 is valid for the research.

**Rule induction:** -after clustering the dataset into Fraud and Non-Fraud class, the next step is creating predictive model that classify instances in to labeled classes. In this study, algorithms such as J48, PART, JRIP and REPTree which are capable of generating rules are employed.

**Knowledge validation:** - Induction algorithms generate knowledge in the form of rules from the dataset that should be validated to make sure that they are reflection of the data set.

Therefore, before using the generated rules as part of the knowledge base, the rules are evaluated in consultation with domain experts in the area of claim analysis.

**Knowledge Base:** -is used as a repository of knowledge in understandable forms. This rules generated by PART rule induction algorithms after mapped by the integrator to PROLOG.

**User Interface:** - it is where the user and system interact with one another. It may have the Natural Language Processing facility and Customized presentation of information according to user profile. Command line interface (CLI) can be used as user interface.

## 4.2 Structure of PART Rule and PROLOG

From the four algorithms used PART is selected as the best algorithm. PART algorithm generates rules in the form of (condition)...then... (conclusion) format. The algorithms generated 54 rules from the dataset, The format of some of the rules is indicated in the table below. The condition part contains attribute, comparison operator and value. Two or more conditions are joined by 'AND. After the conditions the symbol ":" meaning implies follows. The conclusion part of the rule has the format "Fraud" it means the claim is fraud suspicious.

| No | Rule                                                                                                                   | Conclusion             |
|----|------------------------------------------------------------------------------------------------------------------------|------------------------|
| 1  | POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Self_Ignition AND Claim_Clsd_Rsn = settled:                              | Fraud (171.0)          |
| 2  | POLICY_TYPE = PRIVATE AND VEH_TYPE = Land_Cruisers AND Claim_Clsd_Rsn = settled AND Branch_Unit = Hayahult_Mazoria_Br: | Non-Fraud (138.0/17.0) |
| 3  | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Peril_Name = Motor_TP_Passengers:                         | Non-Fraud (199.0)      |

Table 4-1: Sample PART rules for Fraud and Non-Fraud.

As shown in the Table 4-1 the rules are in IF.... Then format. For example:

```
POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Self_Ignition AND
Claim_Clsd_Rsn = settled: Fraud(171.0)
```

The rule can be interpreted as follows

```
IF POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Self_Ignition AND
Claim_Clsd_Rsn = settled THEN Fraud(171.0)
```

The attributes Policy\_Type and Cause\_of\_Loss, here the policy type have the value Third\_Party, Private OR Commercial and Cause of Loss attribute it may have the value Self Ignition, break failure, overturning, road condition, windshield and others. The claim closed reason attribute have different attributes like settled, closed, withdrawal, duplicate claim and so...on.

For example, for certain claim accident conclusion to be Fraud in the rule, policy type should be third party; the cause of loss may be road condition, self ignition or other and claim report date gap should be slow or moderate. On the other hand conclusion can be Non-Fraud, in which the policy type should be private; the cause of loss may be road condition, self ignition or other and claim report date gap should be fast.

However, PROLOG does not work in IF.... THEN format, rather it works in reverse order. PROLOG starts with a goal and then goes to the facts which can proof the goal as true. Therefore, the above rule has to be formatted as:

```
Fraud:    -POLICY_TYPE    =    THIRD_PARTY,    Cause_of_Loss    =    Self_Ignition,
Claim_Clsd_Rsn = settled.
```

As shown above, the conclusion comes first with predicate ‘class’ and followed by ‘:-’ replacing ‘:’ in PART algorithm and then antecedents joined by ‘,’ replacing ‘AND’ in PART rule. Finally, PROLOG rules terminate by period ‘.’, Whereas PART rules terminate with new line.

| Token                              | PART tokens                                                                                                               | PROLOG equivalent | Tokenization option    |
|------------------------------------|---------------------------------------------------------------------------------------------------------------------------|-------------------|------------------------|
| Comparison Operator                | =                                                                                                                         | Same              | Keep them              |
| Logical operator                   | AND                                                                                                                       | ,                 | Replace by ‘,’         |
|                                    | :                                                                                                                         | :-                | Replace by ‘:-’        |
| Miscellaneous words and characters | PART decision list,<br><br>‘-----’(        16<br>subtraction signs)<br><br>: Fraud (254.0/41.0),<br>‘Number of Rules: 54’ |                   | Replace by empty space |

Table 4-2 Tokens in PART and PROLOG rule

### 4.3 High level conceptual design of integration process

Reversing the order of rules is needed to represent PART algorithm generated (IF...THEN format) rules in to PROLOG (THEN...IF format). Integrator application is needed to convert PART rules to following PROLOG syntax.

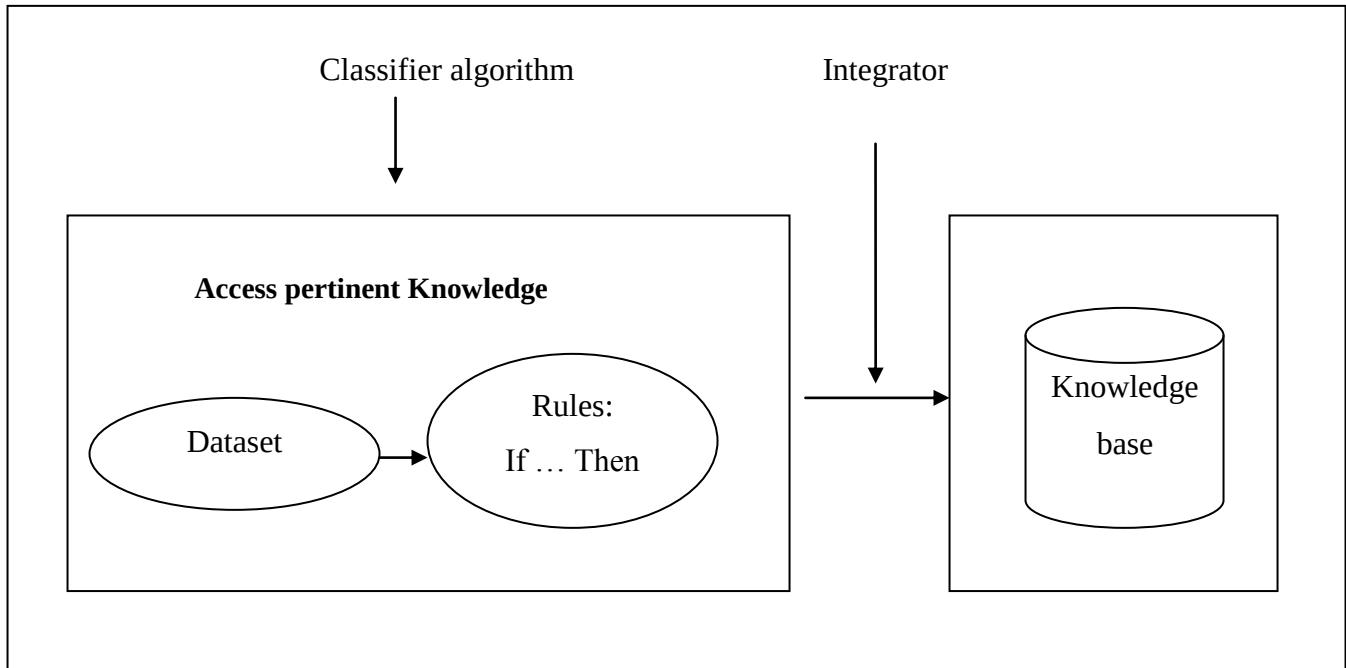


Figure 4-2 Conceptual Design of the Integration Process

As we can see from the above conceptual design there is acquisition of knowledge from the dataset using data mining techniques. Data mining technique represent knowledge in the form of rule generated through PART algorithm. Following, the rule can be converted to a format understandable by PROLOG using the integrator to form a knowledge base.

Tasks undertaken in the process of integrating PART generated rules with knowledge based system.

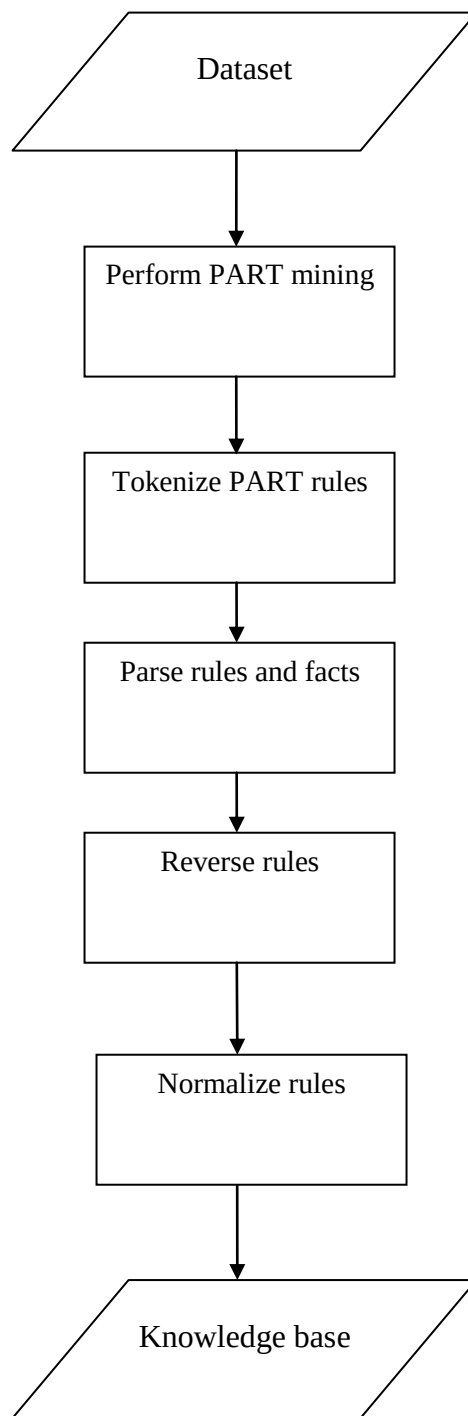


Figure 4-3 work flow diagram for mapping PART into PROLOG (adopted from [7] ).

Abdulkerim [7] illustrates, how to map any Data mining algorithm into prolog understandable format. The researcher adopted these 7 steps from his work to map PART into PROLOG. Each step can be discussed as follows:

- ❖ Dataset: - we get the dataset from Awash Insurance company S.C database. These data are collected from claim related areas of the company. The data extracted from centralized server of all branches of the insurance company.
- ❖ Perform PART mining: -this is done using weka 3.7 version software and applying PART algorithm on the dataset. In addition, rules extracted involving the PART classifier.
- ❖ Tokenize rules: -According to Verma [59], the main use of tokenization is identification of meaningful keywords. It focus on removing characters which are undesirable and replacing some tokens with other tokens. Some special characters and symbols can be replaced with other characters. For example, character ':' in PART is replaced by ':-' in equivalent character which means IF in PROLOG. As well as conjunction operator 'AND' is replaced with ',' in PROLOG. The tokens '*PART decision list*', '-----' (16 subtraction signs) which are at the beginning of rules has no relevance for the desired mapping of PROLOG rules.

In PART rule set, tokens '*Fraud (254.0/41.0)*', '*Number of Rules: 54*' and white space around the beginning and end of PART rules they have no usage then they are removed from the rules.

- ❖ Parse rules and facts: - So we require a parser which processes the tokenization of the rules. This may be trivial but parsing can analyze the components of PART rules. A PART composed of: (condition) and (conclusion). The condition part composed of an attribute, comparison operator and value. *Attribute* defines a property of an object or element and hold value for a given instance. *Comparison operator* compare attribute with value which can be number or string. For example, '*POLICY\_TYPE = THIRD\_PARTY*' here '*policy\_Type*' is attribute that describe about entity and '*THIRD\_PARTY*' its value of an attribute.

```

function tokenizer(line)
    loop i=0 till length_of_line-1
    if length of line>0
    if token is in '-----' then
        remove token
    if token is in 'PART decision list ' then
        remove token
    if token is in ['(', '\.', ')', '/',] then
        remove token
    if token is in 'Non-Fraud' then
        replace it by Non_Fraud
    if token is in 'Number of Rules :54', '[0-9]' then
        remove token
    else if token is='AND' then
        replace it by comma(,)
    else if token is ':' then
        replace it by ':-'
    else if token has the format '(digit/digit)' at the end then
        remove it
    end if
end if
end loop

```

Figure 4-4 Algorithm for rule tokenization

- ❖ **Reverse Rule:** - at this stage the algorithm reverse one position into the opposite position. One can manually reverse the right hand side of the rule to left hand side to come up with desired output. But doing this manually is very tedious and error prone especially when the number of rules are large. Moreover, the reversing process has to be performed periodically as data is added to the dataset.

The aim for reversing the PART generated rules is, PROLOG understands the reversed rule format. That means PROLOG starts with conclusion and goes for conditions that make the conclusion.

```
function rule_parser(line)
    read rule                // input rule
    R_Component = Split rule by ':' // rule component
    Condition_Part = left of ':'
    Conclusion_Part = right of ':'
    Conditions [] = condition split by ','
    Loop:
        i=0 up to number_of (,)
        ANTECEDENT[i] = Conditions [i]
        ATTRIBUTE [i] = left_of_Comparision_operator
        Value [i] = right_of_comparison_operator
    End loop
```

Figure 4-5 Algorithm for rule parser

PART algorithm rule is reversed from the format IF ... THEN to the format THEN... IF PROLOG understandable format. For example:

Rule:  $attribute_1 = value_1$  AND  $attribute_2 = value_2$  AND  $attribute_z = value_z$  THEN (conclusion)

Reversed rule format: (conclusion):-  $attribute_1 = value_1$ ,  $attribute_2 = value_2$ ,  $attribute_z = value_z$ .

Finally, period (.) is placed at the end of all reversed rules to tell PROLOG end of statement.

- ❖ Normalize rules: -its process of transforming tokens into a single form. The rational for normalization is that PROLOG understands a token which starts in upper case as a variable. In PART rule both in conclusion and condition part tokens starts with upper case

letter. So, normalization is necessary to convert the upper case letter to lower case letter as shown in the example bellow:

conclusion:-attribute<sub>1</sub>= value<sub>1</sub>, attribute<sub>2</sub> = value<sub>2</sub>, attribute<sub>z</sub> = value<sub>z</sub>.

```
function rule_ reverse(line)
loop:
iterate till end of line
if token_length not equal to zero then
    concatenate (reversed_rule, current_token)
end loop

loop:
i = token_length-1 until i greater than token_length-1
    if i less than zero
        append the current token with reversed token and space
        continue
    end if
end loop

loop:
else if
    i=0 until i is less than token length- 2
        append the current token with reversed token and space

end loop
else
    if the token in the stuck not equal to space
        append newline to reversed tokens

end if
```

Figure 4-6 Algorithm for rule reversing

## **Investigation**

First one has, to identify suspicious claims that have a higher possibility of being fraudulent. This can be done by computerized statistical analysis and if any using fraud detecting Knowledge base system or by referrals from claims adjusters or insurance agents. The public can provide tips to insurance companies, law enforcements and other organizations regarding suspected, observed or admitted insurance fraud perpetrated by other individuals. Regardless of the source, the next step is to refer these claims to investigations for analysis.

Insurance investigator examines matters pertaining to insurance claims that are suspicious or otherwise in doubt for some reason. Insurance company has their own in – house investigation teams and also if the claim is fraud suspicious they use more investigators again and again. Although such investigations are often conducted to combat fraud.

### **4.4 Implementation of Discovered Rules to Knowledge Base System Integrator**

The Data Mining generated rule helps us to construct a model in which a system can infer. In this research we contrived an automatic construction of knowledge base aligned with the data mining task.

In order to achieve construction of knowledge base, the application is developed based on the above figure 4-2. First the data set collected, processed and given to the PART mining algorithm. The rules generated from the algorithm are represented in PROLOG format. To accomplish the mapping, the integrator is designed and implemented comprising of three modules. These modules are discussed as follows.

#### **PART Module**

This module undertakes the data mining task along with selected classification algorithm. In this process the researcher integrate weka.jar, remoteExperimentServer.jar and weka-src.jar files to eclipse library including the classification algorithm and others. In addition this package contains all the algorithms and preprocessing components that are available in the weka 3.7 explorer application.

The `weka.classifiers.rules.PART` is imported to the class to undertake mining involving PART which is best performing classifier algorithm as we discuss earlier. The `BufferedReader` opens the dataset called `Fdataset5.arff`, then following the path specified and instances are declared from the file read by `BufferedReader`. After that, `data.setClassIndex(data.numAttributes()-1)` makes the last attribute as class attribute which is in our case `POLICY_TYPE` of the car. The string array options hold set of default option for PART classifier which is a default parameters and values for the algorithm it is illustrated in the above Table 3-1. `part.buildClassifier (data)` the object that is created from PART class imported from weka performs the mining task. Following successful completion of rule extraction, the rules saved in text file (`out.txt`) project folder.

### rulePreprocessor

The `out.txt` file which we saved in the project folder can be preprocessed for further use. This module removes some special characters, remove unwanted tokens and replace some logical operator by another. The replacement and removal of special characters and logical operator is done based on the tokenization process shown in the above Figure 4-4.

The PART rules unwanted characters are removed and some characters are replaced by some PROLOG equivalent tokens. The `rulePreprocessor` module continues its task of reversing right hand side to left hand side from the tokenized rules.

### factGenerator

After the rule is preprocessed, the next step is integrating rules and facts following the PROLOG syntax for creating the knowledge base needed to enhance reasoning process. The representation of rule in a prolog is different, it needs reversing the rule from the IF...THEN to THEN...IF. The module exchanges the position of the left hand side and right hand side of PART rules.

The module `rulePreprocessor` starts iterating from end of a given preprocessed rule, it brings the conclusion to the beginning after that `‘:-’` is concatenated which means *IF* in PROLOG. The module yields heads such as `Fraud:-` and `Non-Fraud:-`. To complete the rule, the body part (antecedents) must be concatenated with the heads.

For example:

```
Fraud: - POLICY_TYPE = THIRD_PARTY, Cause_of_Loss = Self_Ignition,  
Claim_Clsd_Rsn = settled.
```

Here, the attribute `POLICY_TYPE`, `Cause_of_Loss` and `Claim_Clsd_Rsn` is used as a predicate and `POLICY_TYPE = THIRD_PARTY`, `Cause_of_Loss = Self_Ignition`, `Claim_Clsd_Rsn = settled` is constructed as antecedents or fact. The body of the rule comes after ‘:-’. It holds one or more than one facts joined by comma ‘,’. The statement `revb.append(words[i])+revb.append(" :-" );` concatenates antecedent at the beginning of a rule by making name of an attribute as predicate with the head of the rule. The statement `revb.append(words[i])+ ","+ revb.append(" ")` is used whenever antecedent or condition of a rule is somewhere in the middle of the rule. Here, first places coma (,) to concatenate previous condition with the current one.

Abdulkerim [7], states that each rule in the rule base is built as a combination of facts. While we inference the inference engine refers to the facts and rules to decide whether a certain condition is true or false in its way of identifying a certain accident is Fraud or Non-Fraud.

The statement `if (result! = "") {revb.append (".\n");} reverse += revb.toString();` here simply collects facts in each iteration. After joining the head of the rules with the body and at the end of all facts and rules period (.) is placed to tell PROLOG it’s the end of statement. `Data += reverse.toLowerCase ()+ ".\n" ;` here PROLOG considers all rules which are in uppercase as a variable, so normalization should be applied.

After, the rules are mined from the dataset, and represented as facts, finally we develop asker clause which helps us to interact with the user. This can be separately saved in a “Frule2.pl” formats so as possible to further process in SWI-Prolog.

To develop the Data mining and knowledge base integrator Java Eclipse 3.7 is used and PROLOG code is used whole constructing the knowledge base.

## CHAPTER FIVE

### IMPLEMENTATION AND EXPERIMENTATION

To make the knowledge discovered from the data set applicable to the end user we have to implement the findings. Here, the knowledge base is constructed automatically as rules and facts through knowledge representation. So, this system tries to detect claim which are fraud or not. In addition, it provides advice for the user about the results of diagnosis. Using SWI-PROLOG we create a knowledge base which contains validated rules and facts about fraud detection and its characteristics.

The fraud detection process is done through presenting questions for the user. The system displays a series of questions that help to identify fraud suspicious claims. The user is expected to reply for the questions or ask explanation for the questions.

Once the system detects whether the claim is fraud or Non-fraud, it gives technical support and some recommendations to the user in decision making. The decision made by the user is either to accept or deny the result based on the rules and the displayed information.

Learning should be incorporated in to the knowledge base system because it is the only way for KBS to incorporate new rules or facts or both into its knowledge based system [7]. It changes as the number of instances in the samples data set changes. The existing knowledge base system must be able to accommodate the changes and use them in the detecting fraud. This fraud detection system has three components these are: - Knowledge base, Inquire component and fraud description component.

**Rule base:** - this component contains automatically generated rules by the integrator application. There are 54 rules generated by the classification algorithm. The rule base contains 32 Fraud types and 22 of them are Non-Fraud types.

**Inquire component:** -this component helps us to create interaction with user. The designed system presents question to the user using Inquire component. This is designed in a manner to accommodate changes in the rules and facts. Whenever changes happen in rule and fact, the question asked also changes accordingly.

**Fraud description component:** -the goal of this component is to provide advice and information on fraud type. This gives the user of the system full information and technical support.

## 5.1 Suspicious Fraud diagnosis

The aim of this diagnosis is to identify whether the claim is Fraud or Non-Fraud. It is based on rules and facts in knowledge base. The system asks question the user via displaying attributes with their respective values in the rule set. Then, the user replies with ‘yes’ or ‘no’ by comparing, based on values displayed.

|                                                                                                                                                   |
|---------------------------------------------------------------------------------------------------------------------------------------------------|
| <pre> fraud:-  policy_type    =   third_party    ,cause_of_loss    =   road_condition ,branch_unit = finfinnee_main_br . </pre>                   |
| <pre> fraud:-  policy_type = private ,cause_of_loss = road_condition ,branch_unit = bole_main_br ,veh_type = private_cars_fm . </pre>             |
| <pre> non_fraud:-  policy_type = private ,cause_of_loss = road_condition ,peril_name = motor_own_damage ,branch_unit = awash_head_office . </pre> |

Figure 5-1 sample rules from the rule base

```

PrologAsker [Java Application] C:\Program Files\Java\jre1.8.0_60\bin\javaw.exe (Feb 6, 2016 12:53:07 PM)
% C:/Users/Bullet_Proff/Documents/Prolog/test4.pl compiled 0.00 sec, 7 clauses
consult('C:/Users/Bullet_Proff/Documents/Prolog/test4.pl') succeeded
% C:/Users/Bullet_Proff/Documents/Prolog/test8.pl compiled 0.00 sec, 14 clauses
consult('C:/Users/Bullet_Proff/Documents/Prolog/test8.pl') succeeded
% C:/Users/Bullet_Proff/Documents/Prolog/f_description.pl compiled 0.00 sec, 9 clauses
consult('C:/Users/Bullet_Proff/Documents/Prolog/f_description.pl') succeeded
WEL COME TO INSURANCE FRAUD DETECTION KNOWLEDGE BASE SYSTEM
*****
SYSTEM IS DESIGNED and DEVELOPED
BY
ABDI CHALI
2015
©Awash insurance company
*****

=====
go. true
TYPE < claim> CLIAM TO CHEACK WHETHER THE CLAIM IS FRAUD SUSPECIOUS OR NON_FRAUD SUSPECIOUS
claim.
Is the policy type third_party (y/n) ?
y.
Is claim Closed Reason Settled/ Withdrwal (y/n) ?
y.
Is Vehicle Type Private Cars/ Land Cruisers (y/n) ?
y.
Is Claim Closed Reason Self Ignition/ Malicious Damage/Road Condition (y/n) ?
y.
THE CLAIM IF FRAUD
goo . true

```

Figure 5-2 Integrator application screen shot for Fraud detection system

The examples above are four rules from the rule base. While the system identifies it displays question by using body of the rules starting from Rule1.

```
PrologAsker [Java Application] C:\Program Files\Java\jre1.8.0_60\bin\javaw.exe (Feb 6, 2016 12:32:54 PM)
=====
go. true
TYPE < claim> CLIAM TO CHEACK WHETHER THE CLAIM IS FRAUD SUSPECIOUS OR NON_FRAUD SUSPECIOUS
claim.
Is the policy type third_party (y/n) ?
y.
Is claim Closed Reason Settled/ Withdrwal (y/n) ?
y.
Is Vehicle Type Private Cars/ Land Cruisers (y/n) ?
y.

Is Claim Closed Reason Self Ignition/ Malicious Damage/Road Condition (y/n) ?
y.

THE CLAIM IF FRAUD
goo . true
What do you know about FRUAD SUSPECIOUS CLIAMS ....,

Please enter your choice of action( 1 up to 4)
1. GENERAL INFORMATION
2. TECHINICAL SUPPORT
3. FURTHER INVESTIGATION
4. Exit FRAUD description
```

Figure 5-3 Question and answer in fraud detection

As illustrated above the system starts asking question from first rule. User must respond to question displayed accordingly. Here the user replied ‘yes’ for the first question Is the policy type third\_party (y/n) ? This question asks the user what is the policy type whether commercial or private, Following that the next question Is claim Closed Reason Settled/ Withdrwal (y/n) ? is displayed. These ask whether the claim is resolved or not.

There are a number of criteria for a rule to be true and the claim to be Fraud, the answer for the three questions should be ‘yes’. If the user replied ‘no’ for the 3<sup>rd</sup> question, the system continues asking for the 4<sup>th</sup> question which might be in the Rule 2. The next rule is about Non-Fraud, the system ask conditions in rule2 to the user if it meet with it. Unless again go for the 3<sup>rd</sup> rule until it meets all conditions are yes. After identifying the type of attack the system provide technical support and General information for further investigation.

```

THE CLAIM IF FRAUD
goo . true
What do you know about FRUAD SUSPECIOUS CLIAMS ....,

Please enter your choice of action( 1 up to 4)
1. GENERAL INFORMATION
2. TECHINICAL SUPPORT
3. FURTHER INVESTIGATION
4. Exit FRAUD description
3.
*****Further Investigation*****

Insurance investigator examines matters pertaining to insurance claims that are
suspicious or otherwise in doubt for some reason. Insurance company has their own
in - house investigation teams and also if the claim is fraud suspicious they
use more investigators again and again. If the accident happened have argument
between two insurance companies, the investigation goes to third private
investigation experts to settle the claim. Although such investigations are usually
conducted to combat fraud, very often.
What do you know about FRUAD SUSPECIOUS CLIAMS ....,

```

Figure 5-4 Prolog screen shot for Further Investigation description

## 5.2 Explanation facility

The explanation facility provides information to user for the questions asked by the system. This facility is helpful to have clarity while answering to the questions asked by the system. In steps towards indentifying claims as Fraud or Non-Fraud the system displays questions. The user can get explanation of the question asked if he/she has no clarity with it. For example, Is POLICY\_TYPE = THIRD\_PARTY:? (yes/no) Here the user can get clarification about fraud cause by typing 'cause'.

## 5.3 Recommendation for Fraud detection

This system is able to give advice for the user so as to assist what type of action the claim analyst should take. In this case the system gives the claim analyst, description of the fraud, technical support to identify whether the claim is fraud suspicious or not.

### 5.3.1 General Information Provider

This fraud detection system provides general description about the fraud type. It includes technical support done by the claim department and further investigation should be done. Below figure shows PROLOG code that display in menu.

```

d_fraud :- write('What do you know about .....,'),nl,nl,

write('Please enter your choice of action( 1 up to 4)'),nl,

write('1. GENERAL INFORMATION'),nl,

write('2. TECHINICAL SUPPORT'),nl,

write('3. FURTHER INVESTIGATION'),nl,

write('4. Exit FRAUD description'),nl,

read(Choose),(Choose ==1->    general_info;Choose==2->tech_support;Choose==3    -
>f_invest,Choose==4->exit_frd),nl.

```

Figure 5-5 Prolog code for menu for Fraud claim

### 5.3.2 Technical Support and Further Investigation

As the system identifies the claim whether it's Fraud suspicious or Non-Fraud, general information is provided about the claim fraud type. But this is not enough for the claim analyst in deciding what actions to take about the claim detection process and for further investigation. Here, the system provides technical support and Further Investigation. Fraud detection by its nature it is very complex and hard to trace. So, insurance companies use their different internal investigation team if it is fraud suspicious. What makes detecting fraud more difficult is fraud is done intentionally by the customer for such reason fraud suspicious claims need more investigation by the experts. The researcher acquire the knowledge about the description of the Fraud, technical support should be given to the claim department and what further investigation should be done by the claim analyst.

```

=====
go. true
TYPE < claim> CLIAM TO CHEACK WHETHER THE CLAIM IS FRAUD SUSPECIOUS OR NON_FRAUD SUSPECIOUS
claim.
Is the policy type third_party (y/n) ?
Y.
Is claim Closed Reason Settled/ Withdrwal (y/n) ?
Y.
Is Vehicle Type Private Cars/ Land Cruisers (y/n) ?
Y.

Is Claim Closed Reason Self Ignition/ Malicious Damage/Road Condition (y/n) ?|
Y.

THE CLAIM IF FRAUD
goo . true
What do you know about FRUAD SUSPECIOUS CLIAMS ....,

Please enter your choice of action( 1 up to 4)
1. GENERAL INFORMATION
2. TECHINICAL SUPPORT
3. FURTHER INVESTIGATION
4. Exit FRAUD description
2.
***** Technical Support*****

First, to identify suspicious claims that have a higher possibility of being
being fraudulent. This can be done by computerized statistical analysis and
if any using fraud detecting Knowledge base system or by referrals from claims
adjusters or insurance agents.

```

Figure 5-6 Prolog screen shot for Technical Support description

## 5.4 Testing and Evaluation

Before the deployment of the system to actual use, it has to be tested. Testing is the investigation conducted to evaluate whether a complete and integrated software system complies with its specified requirements [60] . For evaluating the performance of the knowledge base system test cases are prepared and given to the system. The outputs of the system are compared against domain experts' judgment.

The other one is user acceptance test which is taken following system performance testing basically testing the outputs of knowledge base system with user acceptance. However, accurate in performance means how the system knowledge base system is complete, it is difficult if the system doesn't meet the user requirements or is not accepted by users. This testing helps to test the new system in the operational environment whether it fits with user requirement.

### 5.4.1 User Acceptance Testing

During the user acceptance testing, actual system users test the system to make sure that the system is accepted and usable by users.

Two domain experts are selected to test the system from Awash Insurance Company branch claim analyst. The experts get training how the system works and the test cases given to evaluate the system.

The evaluators assessed system by using the following criteria

- Is the system efficient in time?
- Is the user interface of system interactive
- The accuracy of the system in detecting Fraud suspicious claims.
- Adequacy and clarity of the system in recommendation to the user
- Ease of use
- Fitness of the advisory system to the new claims
- Significance of the system in domain area

The user acceptance testing ensures how the users or domain experts view the system on the bases of the above criteria's. Different researchers used different types of user acceptance testing evaluation criteria suggested by Biazen [61] and Abdulkirim [7] both adopt from other researches, customized and used to ease the evaluation process, the weight scale suggested by [7] has been used such that Excellent =5, Very Good =4, Good =3, fair =2 and Poor =1.

Table 5-1 summarizes the domain expert's evaluation of the system. The values indicate numbers of evaluators who evaluate the system as Poor, Fair, Good, Very Good and Excellent with respect to evaluation criteria.

Thus, 50% of evaluators replied poor for **Ease of use** and 50% respondents replied Good for it. This is due to the command line interface used for interaction with the user such that users are expected to type commands and replies. With regard to second question, **is the user interface of the system interactive** the two evaluators agreed (100%) in Fair. The third question is about **Efficiency of the system in time** 50% agreed on Good and the others 50% reply Very Good. The

fourth evaluation criteria is about **The accuracy of the system in detecting Fraud suspicious claims** and agreed in 100% in Fair. Although, they recommend improvements in this area. The fifth question is about **Adequacy and clarity of the system in recommending user** both Experts agreed in Good which is 100%. The sixth question is about **Fitness of the advisory system to the new claims** 50% agreed on Good and the others 50% reply Very Good. The other criterion is about **Significance of the system in domain area** both agreed Very Good which is 100% and they suggested it essential in the Insurance Domain.

| No | Criteria of evaluation                                           | Poor | Fair | Good | Very Good | Excellent | Average |
|----|------------------------------------------------------------------|------|------|------|-----------|-----------|---------|
| 1  | Ease of use                                                      | 1    | 0    | 1    | 0         | 0         | 2       |
| 2  | Is the user interface of the system interactive                  | 0    | 2    | 0    | 0         | 0         | 2       |
| 3  | Efficiency of the system in time                                 | 0    | 0    | 1    | 1         | 0         | 3.5     |
| 4  | The accuracy of the system in detecting Fraud suspicious claims. | 0    | 2    | 0    | 0         | 0         | 2       |
| 5  | Adequacy and clarity of the system in recommending user          | 0    | 0    | 2    | 0         | 0         | 3       |
| 6  | Fitness of the advisory system to the new claims                 | 0    | 0    | 1    | 1         | 0         | 3.5     |
| 7  | Significance of the system in domain area                        | 0    | 0    | 0    | 2         | 0         | 4       |

Table 5-1 User acceptance evaluation

#### 5.4.2 System Performance Testing

After the user acceptance testing is completed then System performance testing is done by preparing test cases. The test cases include samples from fraud detection data set. The test cases, which are unlabelled data set are given to the domain expert to label them as Fraud and Non-Fraud.

About 16 instances including 10 attributes are taken for the test. Few numbers of instances are taken for testing by considering the time it takes for the domain experts to label each instance. Attributes of instances with their respective values describe the behavior of certain claim. The same set of test instances are provided to the system and the outputs are compared to the domain expert's judgment. System performance testing is basically used to measure how accurate the system is. True positive, F-measure, Recall and Precision measure how accurate the system is. Confusion matrix is used for comparing the performance of the system with domain experts. Confusion matrix helps us how similar is the domain experts judgment with the system.

|                       | System KBS<br>Recommendations |       |           |       |
|-----------------------|-------------------------------|-------|-----------|-------|
| Domain                | Class                         | Fraud | Non-Fraud | Total |
| Experts<br>suggestion | Fraud                         | 7     | 1         | 8     |
|                       | Non-Fraud                     | 2     | 6         | 8     |
|                       | Total                         | 9     | 7         | 16    |

Table 5-2 Confusion matrix for evaluation of the system KBS compared to experts judgment

The above confusion matrix shows test cases evaluation by the system KBS and domain Experts suggestion. The column shows the result of System KBS and the row shows evaluation of domain experts.

The entry under column Fraud indicates out of 9 instances the system identified 7 as Fraud correctly and 2 instances classified incorrectly as Non-Fraud. The entry under column Non-Fraud indicates out of 7 instances 6 correctly classified as Non-Fraud and 1 classified incorrectly as Fraud. Identification of each class of claims to their correct class is important to provide proper advice to claim analyst so that they can take appropriate measures.

The system has correctly detected or identified 13 test instances out of 16 to their correct class. This means the system has 81.2 % detection accuracy.

As clearly discussed in Table 5-3, the system's performance is evaluated in terms of TP rate, Precision, Recall and F-measure which enables us to view in detail how accurate is the system in

identifying Fraud suspicious claims and Non-Fraud suspicious claims. The system has registered a TP rate of 87.5% for identifying Fraud suspicious claims correctly but 12.5% of them are incorrectly classified claims.

|              | TP Rate | Precision | Recall | F-measure | class     |
|--------------|---------|-----------|--------|-----------|-----------|
|              | 0.875   | 0.77      | 0.875  | 0.82      | Fraud     |
|              | 0.75    | 0.85      | 0.75   | 0.796     | Non-Fraud |
| Avrg. weight | 0.812   | 0.81      | 0.812  | 0.808     |           |

Table 5-3 Performance evaluation based on TP, Precision, Recall and F-measure

The system is also evaluated using Precision, Recall and F-measure. The system has registered its highest 87.5% Recall in Fraud suspicious claims and 75% in Non-Fraud suspicious claims. Since the system is expected of identifying each claim incidents correctly, its Precision is expected to be high. The system performed well in identifying claim incidents accurately, scored 77% and 85% for Fraud and Non-Fraud claims respectively.

Generally, the System has accuracy 81.2%. This result is promising for using the system for Fraud suspicious detection and providing respective advice for claim analyst.

## CHAPTER SIX

### CONCLUSION AND RECOMMENDATION

#### 6.1 Conclusion

The need for the integration of data mining technology with Knowledge based system is realized and is increasingly becoming very popular and proved to be relevant for many sectors such as insurance, airline, telecommunications, banking, and healthcare industries. Particularly in the insurance industry, DM technology has been applied for fraud detection and the integration of the knowledge base system makes it more interesting. Data mining has proven to induce hidden knowledge from large collection of dataset. PART classifier is employed for knowledge acquisition step since it has performed best among the chosen classifiers with accuracy of 98.61%. As a matter of fact insurance fraud is the most challenging problem in today's motor insurance business.

In this Experimental research, an attempt has been made to integrate the DM technology and knowledge base system in support of detecting and predicting fraudulent insurance claims in the insurance industry and it's performed simultaneously by automatic integrator application. The six step Cios et al. (2000) process model has strictly been followed while undertaking the experimentation. This process model embraces understanding the problem domain, understanding the data, preparation of the data, evaluation of the discovered knowledge. Finally, given rules generated the integrator application has enabled building facts describing about fraud suspicious claims and advisory knowledge base system. In addition, as the number of instances in the data set and attribute combination change in each rule, the system has the ability to accommodate these changes and update the knowledge based system.

The data used in this research has been gathered from the Awash Insurance Company S.C database. Once the data has been collected, it has been preprocessed and prepared in a format suitable for the DM tasks. This phase took considerable time of the study. The study was then conducted in two sub phases, first the clustering then classification phase followed by representing the rule as fact in PROLOG. Second, to design advisory Knowledge base system using integration application.

In this research, unsupervised approaches (K-Means clustering) have been used to segment the insurance data into clusters for supervised approaches (PART rule based algorithms is used). PART classifier algorithm generates 54 rules and 32 of them are Fraud types and the rest 22 are Non-Fraud, from the result shown the size of class labels are unbalanced for model building process. To keep the originality of the model building, the researcher take the class labels as it is.

In the insurance practice there are patterns that characterize the given claim Fraud or not, claims reported quickly with private policy type and fast accident report date are rated as suspicious of non-fraudulent claims. Apart from this, those claims with a slow accident report date, Third Party policy type, and with very high regain amount of the insured are also rated as suspicious of fraudulent claims.

After Data mining findings are represented in Knowledge base as a rule, fraud detection and advising Knowledge base system was built. System performance testing is used to measure the accuracy of the system in detecting fraud. System performance testing shows that the system has 81.2% accuracy with very good precision and recall.

In addition, user acceptance was tested to check whether the system is valid in the eyes of the user based on different criteria. Despite unavailability of data from different stake holders the fraud detection system can update its rule base as the number of data size changes. This makes the fraud detection system to cope with new changes.

Moreover, this study forwards the usage of induced knowledge's and on using automatic knowledge acquisition techniques for the development of knowledge base system. Also it motivates to apply this approach in different areas.

## **6.2 Recommendations**

This research is mainly conducted for an academic purpose. However, the results of this study are found to be promising to be applied to address practical problems of insurance fraud. This research work can contribute a lot towards a comprehensive study in this area in the future, in the context of our country. The results of this study have also shown that the uses of DM technologies with knowledge base system are well applicable in the efforts of insurance fraud detection. Challenges have been faced while building the system. The first one is bringing the integrator and

Knowledge base together under one interface. The other challenge encountered is on using knowledge base has already been used previously before re running the integrator application when change happen in instances in the dataset.

Hence, based on the result of this study the researcher believes further researches have to be done to increase the benefits of integrating data mining with knowledge bases system and the following are recommendation for future study.

- ❖ The model building process in this investigation was carried out in two sub phases. For clustering, the researcher uses the simple K-Means algorithm whereas for classification PART rule algorithm and to build Knowledge base PROLOG is used. So, further investigation needs to be done in representing Data Mining findings on Knowledge base like using ONESQL or other declarative languages.
- ❖ Increase the limited number of possible attributes that are available with their values in the database of the company. There are inconsistency and missing values in the database. There is no automated record from related stakeholders, and hence the need to include traffic police report, age, sex, health status, driving experience, and marital status of driver and the place where and when the accident has occurred, which are important for fraud factors. Since data is the most important component in advising knowledge base system, the company has to design a data warehouse where operational and non-operational data can be kept.
- ❖ Apply integration of data mining with knowledge base system in other domain areas than insurance fraud detection because fraud is not only occur in the insurance claims of the insured, it can also occur within the company by experts, governors and other staffs. These can also be taken as another area for further research.
- ❖ Improve and build knowledge base system with Graphical user interface which makes simple to use and attractive to the user. Using different programming languages like Java, C# and so...on.

## Reference

- [1] IBM Corporation, "Using Data Mining to Detect Insurance Fraud, IBM Corporation," Chicago, 2010.
- [2] H.Lookman Sithic and T.Balasubramanian, "Survey of Insurance Fraud Detection Using Data Mining Techniques," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. Volume-2, Issue-3, 2013.
- [3] Guo, L 2003. Applying Data Mining Techniques in Property & Casualty Insurance,' USA. [Online]. HYPERLINK "<http://www.casact.org/pubs/forum/03wforum/03wf001.pdf>Access%20date:"  
<http://www.casact.org/pubs/forum/03wforum/03wf001.pdf>Access date:
- [4] Han, J and Kamber, M. *Data Mining: Concepts and Techniques*, 2nd ed. San Francisco: Morgan kufman Publishers, 2006.
- [5] Woodfield, J , *Compensation Insurance Fraud Using SAS® Enterprise Miner™ 5.1 and SAS® Text MineData Mining and Predictive Modeling*. USA: SAS Institute Inc, 20 05.
- [6] Pedro Domingos, "Toward knowledge-Rich data mining," " *Data mining and knowledge discovery, springer*, vol. 15, no.1, pp. 21-28, April 2007.
- [7] Abdulkерim Mohammed, *Towards Integrating Data Mining With Knowledge Based System: The Case of Network Intrusion Detection*, M.A. Thesis ed. Addis Ababa: Addis Ababa University, 2013.
- [8] Tesfaye, H. *Predictive Modeling Using Data Mining Techniques in Support of Insurance Risk Assessment: The Case of Nyala Insurance Corporation*, M.A. Thesis ed. Addis Ababa: Addis Ababa University, 2002.
- [9] Tariku Adane, *Mining Insurance Data for Fraud Detection: The Case of Africa Insurance Share Company*, M.A. Thesis ed. Addis Ababa: Addis Ababa University, 2011.
- [10] Bingru, Y and Wei, S , ‘, Oct. 28-30, Jinhong, "A New Data Mining Process Model for Aluminum Electrolysis," in *Proceedings of the International Symposium on Intelligent Information Systems and Applications (IISA'09) Qingdao*, P. R. China , 2009, pp. 193-195.
- [11] Witold.K, P Roman, S and Kurgan A Cios, *Data Mining: A Knowledge Discovery Approach*, Springer. New York, USA, 2007.
- [12] Gorka Sadowski and Philip Rathle. (2014) Fraud Detection: Discovering Connections with Graph Databases. [Online]. HYPERLINK "Internet:%20www.neotechnology.com" [Internet: www.neotechnology.com](http://www.neotechnology.com)
- [13] Ernst & Young. (2011) Fraud in insurance on rise. [Online]. HYPERLINK "www.ey.com" [www.ey.com](http://www.ey.com)
- [14] B. D, Passamb, H. C, Sideridisb, A. B Mahamana, "Yialouris, C. P. DIARES- IPM," *a diagnostic advisory rule-based expert system for integrated pest management in Solanaceous crop systems*.

- Agricultural Systems*, vol. 76, pp. 1119–1135, 2003.
- [15] Pomykalski, J.J. D.E. and Brown, "Reliability Estimation during Prototyping of Knowledge-Based Systems," *IEEE Trans. on Knowl. Data Eng*, vol. 7 (3), pp. 378-390, 1995.
- [16] Linoff and M, G Berry, *Mastering Data Mining: The Art of Science of Customer Relationship Management*. New York: John Willy and Sons Inc, 2000.
- [17] Witten E, Witten I and Frank, *Data mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd ed. San Francisco: Morgan Kaufmann publishers, 2000.
- [18] Kantardzic M, *Data Mining: Concepts, Models, Methods, and Algorithms.*: Wiley – IEEE Press, 2002.
- [19] KATE SMITH<sup>1</sup> & ROSS GAYLER<sup>2</sup> CLIFTON PHUA<sup>1\*</sup> VINCENT LEE<sup>1</sup>, *A Comprehensive Survey of Data Mining-based Fraud Detection Research*". Wellington Road, Clayton: Monash University Victoria 3800, LP0454077, 2010.
- [20] Oded Maimon • Lior Rokach, *Data Mining and Knowledge Discovery Handbook*, 2nd ed. Ramat Aviv, Israel : Springer New York Dordrecht Heidelberg London, 2010.
- [21] Phyu. N, "Survey of Classification Techniques in Data Mining," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2009.
- [22] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, 3rd ed. Potomac, U.S.A: Two Crows Corporation, 2005.
- [23] Ardakani. E, Shahrabi and D, Hajizadeh, "Application of Data Mining Techniques in Stock Markets;," *A Survey*, *Journal of Economics and International Finance*, vol. 2(7), pp. 109-118, 2010.
- [24] Daniel Mamo, *Application of Data Mining Technology to Support Fraud Protection: The Case of Ethiopian Revenue and Custom Authority*, M.A. Thesis ed. Addis Ababa: Addis Ababa University, 2013.
- [25] Quinlan J. R, *C4.5: Programs for Machine Learning*. Los Altos: Morgan Kaufmann, 1993.
- [26] Chauhan and Y, Singh, "Neural Networks in Data Mining," *Theoretical and Applied Information Technology*, pp. 37-42., 2005.
- [27] Bishop and I, Jordan .M, "Neural Networks," *ACM Computing Surveys, CRC Press*, vol. 28(1), pp. 73-75, 1996.
- [28] Chauhan and Y, Singh, "Neural Networks in Data Mining," *Journal of Theoretical and Applied Information Technology*,., pp. 37-42, 2005.
- [29] CALVIN G. MESSERSMITH, and JANET D. JINGKAI ZHOU, "MAstering Data Mining The Mining The and Scince of Customer Relationship Management," *A computer- Based herbicide*

*Injury Diagnostic Expert System*. *Weed Technology*, vol. 19, pp. 486-491, may 2005.

- [30] Berhanu Aebissa, *Developing a Knowledge based system for coffee disease diagnosis and treatment*, M.A. thesis ed. Addis Ababa, Ethiopia: ADDIS ABABA UNIVERSITY, 2012.
- [31] Merriam-Webster Collegiate Dictionary, 10th Edition. Springfield, MA: Merriam-Webster, 1993.
- [32] Ton de Jong, "Types and Qualities of knowledge," *EDUCATIONAL PSYCHOLOGIST*, vol. 31(2), pp. 105-113, 1996.
- [33] Tripathi K P, "A Review on Knowledge-based Expert System: Concept and Architecture," *Artificial Intelligence Techniques - Novel Approaches & Practical Applications*, 2011.
- [34] Fensel, D. Landes, and R. Studer J. Angele. (1997) Developing Knowledge-Based Systems with MIKE. [Online]. HYPERLINK "e-mail:%20landes@dbag.ulm.DaimlerBenz.COM" [e-mail: landes@dbag.ulm.DaimlerBenz.COM](mailto:landes@dbag.ulm.DaimlerBenz.COM)
- [35] Abraham. A, *Rule-based Expert Systems*, P. Hand Thorn, R Sydenham, Ed.: Handbook of Measuring System Design. In, 2005.
- [36] Priti Sajja Rajendra Akerkar. (2013, Feb 18) Knowledge Based Systems. [Online]. HYPERLINK "http://books.google.com.et/books?id=mQZnd4zmZsoC&printsec=frontcover" \ "V=onepage&q&f=false" <http://books.google.com.et/books?id=mQZnd4zmZsoC&printsec=frontcover#V=onepage&q&f=false>
- [37] Nikolopoulos. C, *Expert Systems*. New York: Marcel Dekker, Inc, 1997.
- [38] Mansaf Alam Aankita Kaur, "Role of Knowledge Engineering in the Development of a Hybrid Knowledge Based Medical Information System for Atrial Fibrillation," *American Journal of Industrial and Business Management*, vol. 3, pp. 36-41, 2013.
- [39] James R. Heatherton, *A Comparison of Knowledge Acquisition Techniques Used In the Development of Expert Systems*, B.A. Thesis ed. Ohio: Air Force Institute of Technology, 1990.
- [40] G"unter Neumann, *Intelligence, Programming Languages in Artificialia*, German Research Center for Artificial Intelligence (LT-Lab, DFKI).
- [41] Matthew Huntbach, "AI 1 Notes on semantic nets and frames," Dept of Computer Science, Queen Mary and Westfield College, London, 1996.
- [42] Henock W, *Application of Data Mining Techniques to Support Customer Relationship Management at Ethiopian Airlines*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2002.
- [43] Denekeew A, *The Application of Data Mining to Support Customer Relationship Management at Ethiopian Airlines*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2003.

- [44] Kumneger F, *Application of Data Mining Techniques to Support Customer Relationship Management for Ethiopian Shipping Lines (ESL)*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2006.
- [45] Shegaw A, *Application of Data Mining Technology to Predict Child Mortality Patterns: The Case of Butajira Rural Health Project (BRHP)*, M.A Thesis ed. Addis Ababa, Ethiopia : Addis Ababa University, 2002.
- [46] Tilahun M, *Possible Application of Data Mining Techniques to Target Potential VISA Card Users in Direct Marketing: The Case of Dashen Bank S.C*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2009.
- [47] Melkamu G, *Applicability of Data Mining Techniques to Customer Relationship Management: The Case of Ethiopian Telecommunications Corporation (ETC) Code Division Multiple Access (CDMA Telephone Service)*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2009.
- [48] Little M, B and Lovell, C Rejesus, "Using Data Mining to Detect Crop Insurance Fraud: Is There a Role for Social Scientists?," *Journal of Financial Crime*, vol. 12 (1), pp. 24-32, 2004.
- [Aditi Mahajan, Anita Ganpati." Performance Evaluation of Rule Based Classification Algorithms". International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),vol.3, Issue 10, October 2014]
- [49] Gervais and C, G Koh, *Fraud Detection Using Data Mining Techniques:Applications in the Motor Insurance Industry*. Singapore, 2010.
- [50] Mesfin F, *Predictive Data Mining Technique in Insurance: The Case of Ethiopian Insurance Corporation*, M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2005.
- [51] Awash Insurance Company S.C. (2015, september) Awash Insurance. [Online]. HYPERLINK "http://www.awashinsurance.com/index.html" <http://www.awashinsurance.com/index.html>
- [52] Thomas Finley and Thorsten Joachims, *Supervised k-Means Clustering*, 2005.
- [53] Claire Cardie, Seth Rogers and Stefan Schroedl Kiri Wagstaff, "Constrained K-means Clustering with Background Knowledge," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, pp. 577-584.
- [54] Kardi Teknomo. (2007, July ) K-Means Clustering Tutorials. [Online]. HYPERLINK "http:\\\\people.revoledu.com\\kardi\\%20tutorial\\kMean\\" <http:\\\\people.revoledu.com\\kardi\\tutorial\\kMean\\>
- [55] Jim Prentzas and Ioannis Hatzilygeroudis, *Categorizing Approaches Combining Rule-Based and Case- Based Reasoning*.

- [57] Yuni Xia, Sunil Prabhakar and Yicheng Tu Biao Qin, A Rule-Based Classification Algorithm for Uncertain Data.
- [58] Quinlan J.R, "Induction of decision trees," *Machine Learning 1*, pp. 81-106, 1986.
- [59] Renu, Deepti Gaur Tanu Verma, "Tokenization and Filtering Process in RapidMiner," *International Journal of Applied Information Systems (IJAIS)*, vol. Volume 7– No. 2, April 2014.
- [60] Aniket Mitra, *Stochastic Modeling and Optimal Control for System Testing of Software*, dr. Sandjai Bhulai, Ed. 1081 HV Amsterdam: VU University, 2013.
- [61] Biazen Getnet. “*Application of Case Based Recommender System to Advise Students in Field of Study Selection at Higher Education in Ethiopia* ” M.A Thesis ed. Addis Ababa, Ethiopia: Addis Ababa University, 2013.

## Appendix I: Rules generated from PART data mining algorithm

| Rule# | Rule                                                                                                                        | Class     |
|-------|-----------------------------------------------------------------------------------------------------------------------------|-----------|
| 1     | Cause_of_Loss = Others AND POLICY_TYPE = THIRD_PARTY: Fraud (2687.0)                                                        | Fraud     |
| 2     | Cause_of_Loss = Others AND VEH_TYPE = Private_Cars_FM: Fraud (766.0)                                                        | Fraud     |
| 3     | Cause_of_Loss = Others AND POLICY_TYPE = COMMERCIAL: Fraud (1420.0/1.0)                                                     | Fraud     |
| 4     | <b>POLICY_TYPE = THIRD_PARTY AND Claim_Clsd_Rsn = settled AND VEH_TYPE = Private_Cars_FM: Fraud (489.0)</b>                 | Fraud     |
| 5     | POLICY_TYPE = THIRD_PARTY AND Branch_Unit = Bole_Main_Br: Fraud (461.0)                                                     | Fraud     |
| 6     | Cause_of_Loss = Others AND VEH_TYPE = Land_Cruisers: Fraud (278.0)                                                          | Fraud     |
| 7     | POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Break_Failure: Fraud (185.0)                                                  | Fraud     |
| 8     | Cause_of_Loss = Others AND Peril_Name = Motor_TP_Passengers: Fraud (589.0)                                                  | Fraud     |
| 9     | <b>POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Self_Ignition AND Claim_Clsd_Rsn = settled: Fraud (171.0)</b>              | Fraud     |
| 10    | POLICY_TYPE = THIRD_PARTY AND Branch_Unit = Finfinnee_Main_Branch: Fraud (105.0/2.0)                                        | Fraud     |
| 11    | POLICY_TYPE = THIRD_PARTY AND Branch_Unit = Gotera_Br: Fraud (49.0/1.0)                                                     | Fraud     |
| 12    | POLICY_TYPE = THIRD_PARTY AND Branch_Unit = Nefas_Silk_Br: Fraud (49.0)                                                     | Fraud     |
| 13    | <b>POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Malicious_Damage: Fraud (140.0/2.0)</b>                                    | Fraud     |
| 14    | <b>POLICY_TYPE = THIRD_PARTY AND Claim_Clsd_Rsn = WITHDRWAL: Fraud (119.0/1.0)</b>                                          | Fraud     |
| 15    | <b>POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Road_condition AND Branch_Unit = Finfinnee_Main_Br: Fraud (145.0/16.0)</b> | Fraud     |
| 16    | <b>Cause_of_Loss = Others AND VEH_TYPE = Pickup_OVR_2350cc: Non-Fraud (138.0/17.0)</b>                                      | Non-Fraud |

|    |                                                                                                                                                   |           |
|----|---------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 17 | Cause_of_Loss = Others: Fraud (529.0)                                                                                                             | Fraud     |
| 18 | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Peril_Name = Motor_TP_Passengers: Non-Fraud (199.0)                                  | Non-Fraud |
| 19 | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Peril_Name = Motor_Own_Damage AND Branch_Unit = Awash_Head_Office: Non-Fraud (157.0) | Non-Fraud |
| 20 | POLICY_TYPE = PRIVATE AND VEH_TYPE = Pickup_OVR_2350cc: Non-Fraud (181.0)                                                                         | Non-Fraud |
| 21 | POLICY_TYPE = PRIVATE AND VEH_TYPE = General_Cartage_Tr: Non-Fraud (83.0/2.0)                                                                     | Non-Fraud |
| 22 | POLICY_TYPE = PRIVATE AND Branch_Unit = Awash_Head_Office: Non-Fraud (188.0)                                                                      | Non-Fraud |
| 23 | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Branch_Unit = Finfinnee_Main_Br: Non-Fraud (111.0)                                   | Non-Fraud |
| 24 | POLICY_TYPE = PRIVATE AND VEH_TYPE = SOLELY_PVT__AUTHO: Non-Fraud (75.0/5.0)                                                                      | Non-Fraud |
| 25 | POLICY_TYPE = PRIVATE AND VEH_TYPE = General_Cartage: Non-Fraud (47.0/2.0)                                                                        | Non-Fraud |
| 26 | POLICY_TYPE = PRIVATE AND VEH_TYPE = Mercedes&BMW: Non-Fraud (41.0)                                                                               | Non-Fraud |
| 27 | POLICY_TYPE = PRIVATE AND VEH_TYPE = General_Cartage_NPR: Non-Fraud (37.0/1.0)                                                                    | Non-Fraud |
| 28 | POLICY_TYPE = PRIVATE AND VEH_TYPE = CAR_HIREandTOUR: Non-Fraud (28.0/1.0)                                                                        | Non-Fraud |
| 29 | POLICY_TYPE = PRIVATE AND VEH_TYPE = Buses_25_seats: Non-Fraud (27.0)                                                                             | Non-Fraud |
| 30 | POLICY_TYPE = THIRD_PARTY AND Cause_of_Loss = Over_turning: Fraud (125.0/1.0)                                                                     | Fraud     |
| 31 | POLICY_TYPE = PRIVATE AND VEH_TYPE = OG_TP AND CLAIM_REP_DT_G = Fast: Non-Fraud (19.0)                                                            | Non-Fraud |
| 32 | POLICY_TYPE = PRIVATE AND VEH_TYPE = Land_Cruisers AND Claim_Clsd_Rsn = settled AND Branch_Unit = Hayahult_Mazoria_Br: Non-Fraud (56.0)           | Non-Fraud |
| 33 | Cause_of_Loss = Theft AND VEH_TYPE = Private_Cars_FM: Fraud (15.0)                                                                                | Fraud     |
| 34 | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition AND Branch_Unit = Bole_Main_Br AND VEH_TYPE = Private_Cars_FM: Fraud (49.0)              | Fraud     |

|    |                                                                                                           |           |
|----|-----------------------------------------------------------------------------------------------------------|-----------|
| 35 | POLICY_TYPE = PRIVATE AND Cause_of_Loss = Road_condition: Non-Fraud (488.0)                               | Non-Fraud |
| 36 | VEH_TYPE = Pickup_OVR_2350cc AND Cause_of_Loss = Road_condition: Non-Fraud (232.0)                        | Non-Fraud |
| 37 | Branch_Unit = Awash_Head_Office AND POLICY_TYPE = COMMERCIAL: Non-Fraud (608.0/49.0)                      | Non-Fraud |
| 38 | VEH_TYPE = Private_Cars_FM: Fraud (657.0)                                                                 | Fraud     |
| 39 | POLICY_TYPE = PRIVATE AND Branch_Unit = Bole_Main_Br: Fraud (71.0)                                        | Fraud     |
| 40 | POLICY_TYPE = PRIVATE: Non-Fraud (349.0)                                                                  | Non-Fraud |
| 41 | Cause_of_Loss = Break_Failure: Fraud (101.0/7.0)                                                          | Fraud     |
| 42 | Cause_of_Loss = Tyre__damage: Fraud (15.0/1.0)                                                            | Fraud     |
| 43 | Cause_of_Loss = Malicious_Damage AND VEH_TYPE = SOLELY_PVT__AUTHO: Fraud (12.0)                           | Fraud     |
| 44 | Cause_of_Loss = Malicious_Damage AND VEH_TYPE = General_Cartage_Tr: Fraud (11.0)                          | Fraud     |
| 45 | Cause_of_Loss = Self_Ignition: Fraud (255.0/28.0)                                                         | Fraud     |
| 46 | Cause_of_Loss = Over_turning: Fraud (178.0/17.0)                                                          | Fraud     |
| 47 | Cause_of_Loss = Malicious_Damage: Fraud (90.0/9.0)                                                        | Fraud     |
| 48 | POLICY_TYPE = COMMERCIAL AND Cause_of_Loss = Road_condition AND Branch_Unit = Bole_Main_Br: Fraud (141.0) | Fraud     |
| 49 | POLICY_TYPE = COMMERCIAL AND Cause_of_Loss = Road_condition: Non-Fraud (703.0)                            | Non-Fraud |
| 50 | Claim_Clsd_Rsn = settled AND Branch_Unit = Lideta_Br: Fraud (72.0/1.0)                                    | Fraud     |
| 51 | Claim_Clsd_Rsn = settled AND Branch_Unit = Awash_Head_Office: Non-Fraud (57.0/1.0)                        | Non-Fraud |
| 52 | Claim_Clsd_Rsn = settled AND POLICY_TYPE = THIRD_PARTY: Fraud (375.0)                                     | Fraud     |

|    |                                                                                          |           |
|----|------------------------------------------------------------------------------------------|-----------|
| 53 | Branch_Unit = Awash_Head_Office AND Cause_of_Loss = Road_condition:<br>Non-Fraud (149.0) | Non-Fraud |
| 54 | Fraud (254.0/41.0)                                                                       |           |

## Appendix II: Initial list of original attributes with their description

| No. | Attribute Name                                         | Data Type | Description                                                                      |
|-----|--------------------------------------------------------|-----------|----------------------------------------------------------------------------------|
| 1   | Class of Business                                      | String    | The type of service we get from the insurance                                    |
| 2   | Risk Name                                              | Varchar   | Name of the item in which the accident happened                                  |
| 3   | Customer Name                                          | Varchar   | Name of the claimant                                                             |
| 4   | Policy Number                                          | Varchar   | The policy number of the claim                                                   |
| 5   | Period of Insurance                                    | Date      | Policy effective end date                                                        |
| 6   | Claim Number                                           | Varchar   | Claim serial number                                                              |
| 7   | Claim Notification Date                                | Date      | The date the insured reported to the company about the accident                  |
| 8   | Date of Accident                                       | Date      | The date that the accident has occurred                                          |
| 9   | Peril Name                                             | String    | The policy type coverage                                                         |
| 10  | Proximate Cause of Loss                                | String    | Cause of the accident                                                            |
| 11  | Claim Amount (Birr)                                    | number    | The amount of money that is estimated to be paid for the claim in local currency |
| 12  | Claim Paid Amount(Birr)                                | Number    | The amount of money paid for the claimant in local currency                      |
| 13  | Claim Status                                           | Varchar   | Status of the claim                                                              |
| 14  | Claim Closed/Re-Open Date                              | Date      | The data that the claim has been closed                                          |
| 15  | Claim Closed/Re-Open Reason                            | String    | The reason that the claim has been closed                                        |
| 16  | Claim Paid (Gross) (A)                                 | Number    | Amount of money estimated without further processing on the claim                |
| 17  | Net of recoveries (B)                                  | Number    | The net amount of money payable to the customer                                  |
| 18  | Total Claims Expenses Paid (Net of recoveries) (A-B=C) | Number    | The difference between claim estimated and claim paid                            |
| 19  | Claim Paid (Gross) (A)                                 | Number    | The amount of money paid for the claimant.                                       |
| 20  | Branch name                                            | Varchar   | The branch unit in which the customer get the service                            |

### Appendix III : Sample values of the final selected attribute

| VEH_T<br>YPE               | CLAI<br>M_RE<br>P_DT_<br>G | Peril_Name                                    | Cause<br>_of_L<br>oss    | CLM_<br>REG_<br>AMT | Clai<br>m_S<br>tatus | Claim<br>_Clsd<br>_Rsn | Bra<br>nch<br>_Un<br>it | Poli<br>cy_S<br>tatus | POLI<br>CY_T<br>YPE | Pur_<br>Veh           |
|----------------------------|----------------------------|-----------------------------------------------|--------------------------|---------------------|----------------------|------------------------|-------------------------|-----------------------|---------------------|-----------------------|
| General<br>_Cartag<br>e    | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | G_<br>Car<br>tag<br>e |
| Private<br>_Cars_<br>FM    | Fast                       | Motor_Own_Da<br>mage                          | Winds<br>hield           | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | PRIV<br>ATE         | Prv<br>_U<br>se       |
| General<br>_Cartag<br>e_Tr | Fast                       | Motor_Own_Da<br>mage                          | Malici<br>ous_D<br>amage | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | G_<br>Car<br>tag<br>e |
| GC_TR<br>UCKS              | Fast                       | Motor_TP_Com<br>pulsory___Prop<br>erty_Damage | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | THIR<br>D_PA<br>RTY | G_<br>Car<br>tag<br>e |
| Pick_u<br>p2350c<br>c      | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | Ow<br>n_s<br>erv      |
| Buses_<br>25_seat<br>s     | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | Ow<br>n_s<br>erv      |
| Buses_<br>25_seat<br>s     | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | Ow<br>n_s<br>erv      |
| Pickup<br>_OVR_<br>2350cc  | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | G_<br>Car<br>tag<br>e |
| General<br>_Crtg_<br>FSR   | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | Ow<br>n_s<br>erv      |
| Pickup<br>_OVR_<br>2350cc  | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Very_<br>High       | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | COM<br>MER<br>CIAL  | SP<br>V               |
| Combin<br>e_Harv<br>estors | Fast                       | Motor_Own_Da<br>mage                          | Road_<br>conditi<br>on   | Low                 | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | THIR<br>D_PA<br>RTY | G_<br>Car<br>tag<br>e |
| GC_TR<br>UCK               | Fast                       | Motor_TP_Com<br>pulsory___Prop<br>erty_Damage | Road_<br>conditi<br>on   | High                | Pend<br>ing          | settle<br>d            | Ada<br>ma               | R                     | THIR<br>D_PA<br>RTY | G_<br>Car<br>tag<br>e |

## Appendix IV: User Acceptance Test

Dear Evaluator

First and foremost, thank you for your collaboration to fill this form. The evaluation form is prepared aiming at measuring to what extent does the system is useable and acceptable by claim analyst in the area of Fraud detection. Therefore, you are kindly requested to evaluate the system by labeling (X) symbol on the space provided.

Note: - the value for all attributes in the table are rated as Excellent = 5, Very good = 4, Good =3, Fair =2 and Poor = 1.

| No | Criteria of evaluation                                           | Poor  | Fair | Good | Very Good | Excellent | Average |
|----|------------------------------------------------------------------|-------|------|------|-----------|-----------|---------|
| 1  | Ease of use                                                      |       |      |      |           |           |         |
| 2  | Is the user interface of the system interactive                  |       |      |      |           |           |         |
| 3  | Efficiency of the system in time                                 |       |      |      |           |           |         |
| 4  | The accuracy of the system in detecting Fraud suspicious claims. |       |      |      |           |           |         |
| 5  | Adequacy and clarity of the system in recommending user          |       |      |      |           |           |         |
| 6  | Fitness of the advisory system to the new claims                 |       |      |      |           |           |         |
| 7  | Significance of the system in domain area                        |       |      |      |           |           |         |
|    |                                                                  | Total |      |      |           |           |         |

## Appendix V: Integrator java code

```
import java.io.*;
import weka.classifiers.rules.PART;
import java.io.IOException;
import java.util.Arrays;
import weka.core.Instances;
import javax.swing.JTextArea;

public class DataMining {
    String rep = "";
    String reverse = "";
    String result = "";
    String r = "";
    public static String m_result= "";
    public int c= 0;
    public String Data = "%% Here comes the rules for detecting of fraud\n"
+
        " :- reconsult ('test3.pl').\n)";
    public static String file_path = "";
    public String temprarr= "";
    public static void main (String[] args) throws
FileNotFoundException,IOException, Exception{
        DataMining f = new DataMining();
        f.factAndGenerator(null);
        String m_result;
        // String fileName = "Frule.txt";
        try{
            //JFileChooser JFileChooser1 = new JFileChooser();

            //file_path =
PFChooser1.getSelectedFile().getAbsolutePath();
            //String file_path =
JFileChooser1.getSelectedFile().getAbsolutePath();
            //file_path = new
File(C:\\Users\\Bullet_Proff\\Desktop\\Fdataset5.arff);
            //BufferedReader reader = new BufferedReader (new
FileReader( C:\\Users\\Bullet_Proff\\Desktop\\Fdataset5.arff));
            //File file = new File("Frule.txt");
            //File file = new File("Frule.txt");
```

```

        //file.createNewFile();
        //BufferedReader reader= new BufferedReader(new
FileReader(Fdataset5.arff));
        BufferedReader reader = new BufferedReader(new
FileReader(new File("C:\\Users\\Bullet_Proff\\Desktop\\Fdataset5.arff")));
        //System.out.println("File open successful!");
        //FileWriter fileWriter = new FileWriter(fileName);
        //BufferedWriter bufferedWriter = new
BufferedWriter(fileWriter);

        Instances data = new Instances(reader);
        //reader.close();
        data.setClassIndex(data.numAttributes() -1);

        PART part = new PART();
        String [] options = new String[6];
        options [0]= "-M";
        options [1]= "2";
        options [2]= "-C";
        options [3]= "0.25";
        options [4]= "-Q";
        options [5]= "1";
        part.setOptions(options);
        part.buildClassifier (data);
        m_result = part.toString();

        //System.out.print(m_result);
        //System.out.print ("Mining done ....");
        JTextArea pTextArea2 = new JTextArea();
        pTextArea2.setText (part.toString());
//    PrintStream console = System.out;
        File file = new File("out.txt");
        FileOutputStream fos = new FileOutputStream(file);

        PrintStream ps = new PrintStream(fos);

        System.setOut (ps);

        System.out.println(m_result);

```

```

        //bufferedWriter.newLine();
        //PrintStream out = new PrintStream(new
FileOutputStream("Frule5.txt"));
        //System.setin(part.toString());
        //bufferedWriter.write(m_result.toString());

    }catch (IOException e)
    {
        System.out.println("File I/O error!");
    }

}

public void factAndGenerator (String prules) throws
FileNotFoundException {

    String Nfact = "";
    String askPROLOG = "";

    int numOfLines =0;
    try{
        BufferedReader sr = new BufferedReader(new
FileReader(new
File("C:\\Users\\Bullet_Proff\\Desktop\\AICFraud\\out.txt")));
        String x= "";

        /**while ((x=bf.readLine())!= null){
            //x=bf.readLine();
            result = rulePreprocessor(x);

            //System.out.println(bf.readLine());
            //System.out.println(i);
            String [] n = result.split(" ");
            if (x.trim().length() != 0){
                if (!(x.contains("PART decision list-----"))){
                    */

```

```

while ((x = sr.readLine()) != null)
{
    if (x != ": Fraud (254.0/41.0)")
    {
        result = rulePreprocesser(x);

        if (result.trim().length() != 0)
        {
            String y = result.trim();
            String[] words = y.split(" "); ;
            String rev = "";
            for (int i = words.length - 1; i >= 0; i--)
            {
                rev += words[i] + " ";
            }

            StringBuilder revb = new StringBuilder();

            revb.setLength(0);

            int z = 0;
            for (int i = words.length - 1; i >= words.length - 1; i--)
            {
                if (i < 0)
                    continue;

                z++;
                if (result != "" && result != "*")
                {
                    revb.append(words[i]);
                    revb.append(":- ");
                }
            }

            for (int i = 0; i <= words.length - 2; i++)
            {
                revb.append(words[i]);
                revb.append(" ");
            }

```

```

        if (result != "")
        {
            revb.append(".\n");
        }

        reverse += revb.toString();
    }

    }

    /*

    numOfLines+=1;
    for ( i = n.length -1; i>=0; i--)
    {
        if(i==n.length -1){

if(n[i].length()>=0){

            reverse+= "("+n[i].toLowerCase()+") :-";

            break;

        }}
    } // end of for loop

    for ( i =0; i < n.length-1;)
    {

        reverse += n[i] + "(" + n[i];// i replace i with j
// System.out.println("N[i] = "+n[i].length());
//System.out.println("n = "+n.length);
        for (int k = 0; k < n.length; k++)
        {

            if (k != 0)
                reverse += n[i + k];

        }*/

```

```

    }
    //reverse.trim();
    Data += reverse.toLowerCase ()+ ".\n" ;
    System.out.println( Data.toString()) ;
    //System.out.println("end of fact base writer");
    File file = new File("robel.pl");
    FileOutputStream Data1 = new FileOutputStream(file);
    PrintStream ps = new PrintStream(Data1);

    System.setOut(ps);
    System.out.println(Data);

    }

catch(Exception e){
    System.out.println("end of fact base writer");
    //throw ex;
    System.out.println(e);
}

}

public String rulePreprocessor (String txt){
    String rep = "";
    if (txt.trim().length()>0)
    {
        rep = txt.replace("PART decision list", "");
        rep = rep.replace("-----", "");
        rep = rep.replace("AND", ",");
        rep = rep.replace(":", "");
        rep = rep.replace("(", "");
        rep = rep.replace(".", "");
        rep = rep.replace(")", "");
        rep = rep.replace("Non-Fraud", "Non_Fraud");
        rep = rep.replace("/", "");
        rep = rep.replace("Number of Rules : 54", "");
        rep = rep.replaceAll("[0-9]", "");
    }
}

```

```
        }  
        return rep;  
    }  
}
```

## Appendix VI: Prolog code for knowledge representation

go:-

```
greeting,  
load_kb,  
repeat,  
write('Enter choice'),nl,  
write(' first type load ( to load the KB'),nl,  
write(' type consult ( to identify the Fraud ) and qiut ( to Exit from the system'),nl.
```

greeting:-

```
write(' WEL COME TO FRAUD DETECTION KBS'),nl,  
write(' FOR FRAUD SUSPECOUS DETECTION & ADVISING SYSTEM'), nl,  
write('*****'), nl,  
write(' SYSTEM IS DESIGNED BY'),nl,  
write(' and DEVELOPED by ABDI CK'),nl,  
write('*****'),nl,nl.
```

do(load):- load\_kb,!.  
do(consult):- solve,!.  
do(quit):- !.  
do(X):-

```
write(X),  
write('This is not correct commmand:'),greeting,nl,  
fail.
```

load\_kb:-

write('files loaded ...'),nl.

solve:-

```
reconsult('test3.pl'),  
top_goal(X),  
% cls,  
write('The type of cliam is:'),  
write(X),  
reconsult('f_description.pl'),  
(X=='fraud' -> d_fraud; X=='non-fraud' -> d_nonfraud),nl,  
abolish(known,3).  
%define(known,3).
```

solve:-

```
write("This is needs further investigation you cant judge using this KBS"),nl.  
d_fraud :- write("What do you know about ...."),nl,nl,  
            write("Please enter your choice of action( 1 up to 4)'),nl,  
            write('1. GENERAL INFORMATION'),nl,  
            write('2. TECHINICAL SUPPORT'),nl,  
            write('3. FURTHER INVESTIGATION'),nl,  
            write('4. Exit FRAUD description'),nl,  
            read(Choose),(Choose ==1-> general_info;Choose==2->tech_support;Choose==3  
->f_invest,Choose==4->exit_frd),nl.
```

```
general_info:- write('Insurance Fraud is any act committed with the intent'), nl,  
write(' to obtain a fraudulent outcome from an insurance process. '),nl,  
write(' This may occur when a claimant attempts to obtain some benefit or advantage '),nl,  
write(' to which they are not otherwise entitled, or when an insurer knowingly '),nl,  
write('denies some benefit that is due. '),nl,nl,  
write(' *****cause/done ***** '),nl,  
read(Choose),(Choose=='cause',(  
write("The main motive in all insurance crimes is financial profit. The causes varies, '),nl,  
write('but are usually centered on greed and holes in the protections against fraud. '),nl,  
write(' Another reason that this opportunity arises is in the case of over-insurance, '),nl,  
write('when the amount insured is greater than the actual value of the property insured. '),nl,  
write(' The most common form of insurance fraud is inflating of loss. '))),nl,nl,d_fraud.
```

```
tech_support:- write(' First, to identify suspicious claims that have a higher possibility of being  
' ),nl,  
write('being fraudulent. This can be done by computerized statistical analysis and '),nl,  
write('if any using fraud detecting Knowledge base system or by referrals from claims '),nl,  
write('adjusters or insurance agents. '), nl,nl,d_fraud.
```

```
f_invest :- write(' Insurance investigator examines matters pertaining to insurance claims that are  
' ),nl,  
write(' suspicious or otherwise in doubt for some reason. Insurance company has their own '),nl,  
write(' in – house investigation teams and also if the claim is fraud suspicious they '),nl,
```

```

write('use more investigators again and again.If the accident happened have argument '),nl,
write('between two insurance companies, the investigation goes to third private '),nl,
write(' investigation experts to settle the claim.Although such investigations are usually'),nl,
write('conducted to combat fraud, very often. '),nl,d_fraud.

```

```

exit_frd :- !.

```

```

d_nonfraud :- write('What do you know about ....'),nl,nl,
write('Please enter your choice of action( 1 up to 2)'),nl,
write('1. GENERAL INFORMATION'),nl,
write('2. Exit NON-FRAUD description'),nl,
read(Choose),(Choose ==1-> dgeneral_info,Choose==2-> exit_Nfrd),nl.
dgeneral_info:-
write(' This is not fraud suspecious. its normal accident'),nl,nl,nl,
write(' These are claims free from fraud suspicious as investigated by the insurance
fraud '),nl,
write(' investigation experts. So as the request filed by a policyholder stating that
an insured'),nl,
write(' event has occurred and that the insurance company should provide
coverage. '),nl,nl,d_nonfraud.

exit_Nfrd :- !.

```