NEW RULE BASED CLASSIFICATION ALGORITHM FOR AUTOMOBILE INSURANCE FRAUD DETECTION

BY

Ahmad Okleh AL_Ali

SUPERVISOR

Dr.Fadi Thabtah

This Thesis Submitted in Partial Fulfillment of the

Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduates Studies

Philadelphia University

January 2013

جامعة فيلادلفيا

نموذج التفويض

انا احمد عقلة علي العلي ، أفوض جامعة فيلادلفيا بتزويد نسخ من رسالتي للمكتبات أو المؤسسات أو الهيئات أو الاسخاص عند طلبها.

التوقيع:

التاريخ: 13-1-2013

Philadelphia University

Authorization Form

I am Ahmad Okleh Ali Al.Ali, authorize Philadelphia University to supply copies of my Thesis to libraries or establishment or individuals upon request.

Signature:

Date: 13-1-2013

NEW RULE BASED CLASSIFICATION ALGORITHM FOR AUTOMOBILE INSURANCE FRAUD DETECTION

BY

Ahmad Okleh AL.Ali

SUPERVISOR

Dr.Fadi Thabtah

This Thesis Submitted in Partial Fulfillment of the

Requirements for the Master's Degree in Computer Science

Deanship of Academic Research and Graduates Studies

Philadelphia University

January 2013

Successfully defended and approved on 3/1/2013

Examination Committee	Signature

Dr., Fadi Abdel Latif, Chairman

Academic Rank: Associate Professor

Dr. Ali Alawneh, member.

Academic Rank: Assistant Professor

Dr. Alaa H. Al-Hamami, External Member.

Academic Rank: Professor

(Name of University) Amman Arab University

DEDICATION

For my family,

who gave me endless love,

support and encouragement

throughout the course of this thesis.

Ahmad Al.Ali

ACKNOWLEDGEMENT

It is a pleasure to thank the many people who made this thesis possible.

I would like to take this opportunity to express my appreciation and respect to my supervisor Dr. Fadi Thabtah, who assisted me with his countless guidance, suggestions and encouragements throughout this thesis. He enthusiastically shared his vision, knowledge and expertise, and generously spent his time to assist me to generate ideas and research concept.

It is my honor to express my thankfulness to my parents and brothers for their support to accomplishing this thesis successfully.

I acknowledge the efforts from all faculty members who have taught me in the faculty of Information Technology.

I am grateful to many colleagues in Philadelphia University that have influenced this thesis.

Ahmad Al.Ali

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACRONYM/SYNONYM	MEANING
AC	Associative Classification
AIF	Automobile Insurance Fraud
ANN	Artificial Neural network
ARFF	Attribute Relation File Format
AUROC	Area Under the Receiver Operating Characteristics
CART	Classification And Regression Tree Algorithm
AUROC	Area Under the Receiver Operating Characteristics
CTC	Consolidated Trees Classification
DT	Decision Tree
FMC	Feature Mapping Categorization
GUI	Graphical User Interface
IG	Information Gain
IR	Information Retrieval
IREP	Incremental Reduced Error Pruning
K-NN	K-Nearest Neighbor
LR	Logistic Regression
MLP-ARD	Multilayer Perceptron-Automatic Relevance Determination
NB	Naive Bayes
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
STBCP	Strength Threshold Based Coverage Prism
SVM	Support Vector Machine
TAN	Tree-Argument Naïve Bayes

ABSTRACT

Automobile Insurance Fraud (AIF) is a significant and costly problem for both policyholders and insurance companies. The fraudulent activities may affect negatively on the profits of automobile insurance companies. Data mining especially rule based classification algorithms can contribute in helping the detection of fraudulent activities. In these algorithms the output is represented in simple interpreted "*If-Then*" knowledge and stored in a knowledge base. However, the problem of rule based classification such as *(PRISM)* generates large number of rules. Since maintaining and understanding these classifier rules depend on classifiers size which is hard by the typical end user. Moreover, some correlation rules in *(PRISM)* that near perfection ones can't be extracted. These disappeared rules in competitive environment are considered very significant in the prediction phase. On the other hand, induction rule based algorithm i.e. Repeated Incremental Pruning to Produce Error Reduction *(RIPPER)* have small size classifiers with often low accuracy, these rules is not feasible regarding to the (AIF) classification problem, because some knowledge are undetected. This thesis investigates the applicability of strength threshold based covering method on the problem of detection the accident type in order to make balance in producing the number of generated rules without impacting on the classification rate. The new algorithm named Strength Threshold Based Coverage Prism *(STBCP)* that makes balance, (as a result on average size classifiers) in producing the rules. This balance is accomplished by producing a new rule based classification algorithm (STBCP) that utilized a new learning, pruning and prediction procedures based on different strength threshold values (2%, 3%, **4%**, 5%, and 6%) against "autos" data set using significant and complete features (More details in Chapter Three). Based on those threshold values (2-6%), the experimental results found that, the (STBCP) algorithm produced the highest accurate classifier than *PRISM, RIPPER* and *J.48* decision tree algorithms. We chose (**4%** as average of threshold values) and we found that, STBCP algorithm produced the highest accuracy compared with *PRISM, RIPPER* and *J.48* decision tree algorithms. In general, *the STBCP* algorithm produces neither in large nor in small numbers of rules

(classifiers), but it make balance between them (as a result on average size). These allow end user and decision makers to maintain and understand the produced rules with a clear representation without impacting on the classification rate (accuracy).

Chapter One: Introduction

1.1 Introduction

Fraud refers to the misuse of the earnings of an insurance company system without necessarily leading to direct legal consequences. Fraud is a crime and can be committed by consumers and providers even the employees of insurance companies (Phua at el., 2005). There are several types of insurance fraud; we focus on the Automobile Insurance Fraud (AIF) classification problem, because this problem is considered very significant for insurance companies to handle the fake claims as well as decreasing the cost of compensations that paid by automobile insurance companies.

Furthermore, the impacting of fraudulent activities through illegal procedures may affect on the revenue of insurance companies each year (Wilson, 2009). Since the cost of fabricated activities in general are very large which caused huge drain on financial resources in the insurance companies (Ngai et al. ,2011) and (Phua at el., 2005). Most of the current researches utilized data mining as a business process for exploring a large amount of data and extraction useful information from huge numbers of the data sets.

In any classifications problem the data set divided into labeled training data and unlabeled test data. Since the training data records were used to construct the classification model, whereas the unlabeled test data records are used in validating the model. Then the model is used to predict and classify a new test case to label their type. The users utilized several important features of "autos" data set related to the fraudulent characteristics cases such as car make, wheel-base, height, and length.

Therefore, it’s a typical classification problem where the rule based classification algorithms can contribute in detection the accident type. These algorithms in simple interpreted chunks of knowledge (IF-THEN) and stored in knowledge base. In data mining traditional rule based classification algorithms especially induction and covering algorithms such as RIPPER (Cohen, 1995) and PRISM (Cendrowska, 1987) considered to be popular approaches and play major roles in dealing with the problem of detection of the accident type.

Since these algorithms is not suitable regarding to the AIF detection, because some correlations rules were disappeared. In addition, these hidden rules denote useful knowledge and can't exploit in the prediction purpose. In this thesis, We investigates the applicability of hybrid threshold based covering method named (STBCP) algorithm on the problem of detection the accident type in order to make balance in producing the rules (neither in large nor in small numbers of classifiers) without impacting on the classification rate.

1.2 Motivation

In recent years AIF is considered as an economic problem for insurance companies, due to increase the annual compensations cost that paid by these companies resulting from the fraudulent claims (Wilson, 2009). The goal of automobile insurance company is to compensate an insured who sustained a loss or to restore an insured to the same financial situation before loss. As well as reducing the cost paid by automobile insurance companies to fraud accidents and their claims that frequently happens every year.

Our motivations in this thesis come from the problems of rule based classification algorithms such as induction algorithm like RIPPER and covering algorithm like PRISM in order to detect the accident type either to be fraud or legitimate, we summarized our motivation as under:

- To gain additional knowledge (classifiers) that missed by induction algorithm such as RIPPER and covering algorithm like PRISM. Since perfect knowledge in competitive environments such as (AIF) detection problem is not feasible, because some correlation rules that near perfection ones cant detected. In addition, we need these rules and exploited them in the prediction purpose.

- To make balance in producing the rules which end user and decision makers can understand and maintain them easily with a clear representation, because understanding and maintaining theses rules depends on the classifier size. The number of producing these classifiers are: neither in large nor in small, but we make balance between them without impacting on the classification rate.

- To provide accurate classifiers than those classifiers which generated by RIPPER PRISM algorithms and J.48 Decision Tree (DT), in order to detect accident characteristics cases to be fraud or legitimate.

1.3 Problem Statement

In competitive environments such as (AIF) detection problem, the researchers used several and common data mining and machine learning techniques to assist the automobile insurance companies in order to detect the fraudulent cases. Since the construction of rule induction based RIPPER and covering based PRISM are in greedy fashion. These algorithms suffering from several things such as: A) PRISM algorithms try to get perfect rules and generated very large number of rules with (high accuracy 100%). B) Correlated rules can't detect in RIPPER and also in PRISM. C) Limited (small) number of generating rule is not usefulness regarding to the AIF detection problem, because some important rule may not extracted and denoted useful rules that may used later in the prediction step such as RIPPER algorithm. In order to gain additional knowledge that missed by RIPPER and PRISM, we propose a new algorithm that can help the end user to get accurate classifiers and to make balance with respect to their size of generated rules. This is in order to allow the decision makers to understand and maintain the classifiers in easy way. The new algorithm named (STBCP) generates rules not only (100% accuracy) but also near perfection ones, where the rules that have strength larger than or equal to the user initial strength are produced. After rules are sorted the STBCP algorithm utilized a new pruning to kicks negative rules without impacting on the accuracy. The promising algorithm can retain all information in a clear representation, and interpretability.

In this thesis, we investigates the applicability of (STBCP) algorithm on the problem of detection the accident type in order to make balance in producing the rules without impacting on the classification rate, the number of generating rule are: neither in large nor in small, but we make balance between them.

Some research questions that this thesis try to answer them which as under:

- Can we make balance in generated numbers of rules (classifiers) by the (STBCP) algorithm without negatively impacting on the classification rate?
- How to extract the most effective features related to auto data set?

1.4 Some Definitions Related to Classification

Some definitions used in our thesis, these definitions were frequently used in the field of data mining and machine learning approaches. We used them exactly in Chapter Three of the proposed model, the definitions are listed below:

- **Learning (Training):** is the process of discovering knowledge (rules) based on their confidence.
- **Model (Classifier):** is a set of derived rules that utilised in the prediction phase.
- **Itemset:** is a set of attributes together with their specific values for each attribute in the dataset.
- **Classification Accuracy**: is the number of cases where the predicted class of each test data matches actual class of test case for all cases in the test data.

1.5 Thesis Objectives

The main aim of this thesis is to investigate the applicability of STBCP algorithm on the problem of detection the accident type in order to make balance in producing the number of generated rules without impacting on the classification rate. We used this algorithm in order to classify and predict the accidents cases either to fraud or legitimate using significant features related to the "autos" data set[1] . The STBCP algorithm utilized Chi-Square pruning method that kicks the useless rules without negatively effecting the prediction rate.

this research aims to meet the following objectives:

[1] http://archive.ics.uci.edu/ml/datasets/Automobile

- An extensive and critical study on the aspects of rule based classification algorithms especially induction and covering algorithms.
- The development of a new STBCP algorithm for automobile insurance fraud detection.
- Determine the significant features related to "autos" data set.
- An experimental study to contrast the proposed model with other common based classification algorithms performance against relevant and complete features of "autos" data set.

1.6 Thesis Contributions

Our contribution in this thesis derived from an extensive study on the insufficiencies of the rule based classification algorithms especially induction based classification algorithms i.e. RIPPER and covering based PRISM. We summarized our contributions as follows:

- The detection of fraud cases using a new algorithm named STBCP in automobile insurance industry. The contribution in this point is divided in three folds:

1. Learning of the rule: when we use strength threshold to produce not only perfect rules with (100% accuracy) but also near perfection one in rule induction strategy.
2. Building the classifiers (pruning): The STBCP utilized pruning method in cutting down redundant rules and to prune negative correlated rules in order to decrease the size of the classifiers, the results using our algorithm are in high accurate classifiers and balance in producing numbers of rule. The pruning method based Chi-Square testing which presented in Chapter Three.
3. Prediction: After building the model (classifiers) we use the classifier to predict the class of test data instances. The STBCP algorithm produces a new prediction procedure. More details of the prediction phase are presented in Chapter Three exactly in Section 3.2.5.

- Experimental study against relevant and complete features of "autos" data set using STBCP algorithm and other common data mining based classification algorithms such as RIPPER, PRISM and J.48 DT.

1.7 Thesis Methodology

The methodology in our thesis is summarized as under:

- We conduct a comprehensive literature review on AIF detection as well as common based classification Algorithms as well as induction rule based like RIPPER and covering rule based like PRISM towards fraud detection.

- Quantitative approach:

- Quantitative approach will be exercised for analyzing the experimentation results derived by the classification induction algorithm and it's compared with a new model based data mining.

- Critical analyses of the generated results with respect to different evaluation measures, such as predictive accuracy, number of rule derived.

- Experimental studies in "autos" data using the new model and other common data mining rule based classification algorithms that found the new model have higher accuracy and results in medium size classifier (medium number of generated rules).

The research project steps (activities) are representing in the Figure 1.1.

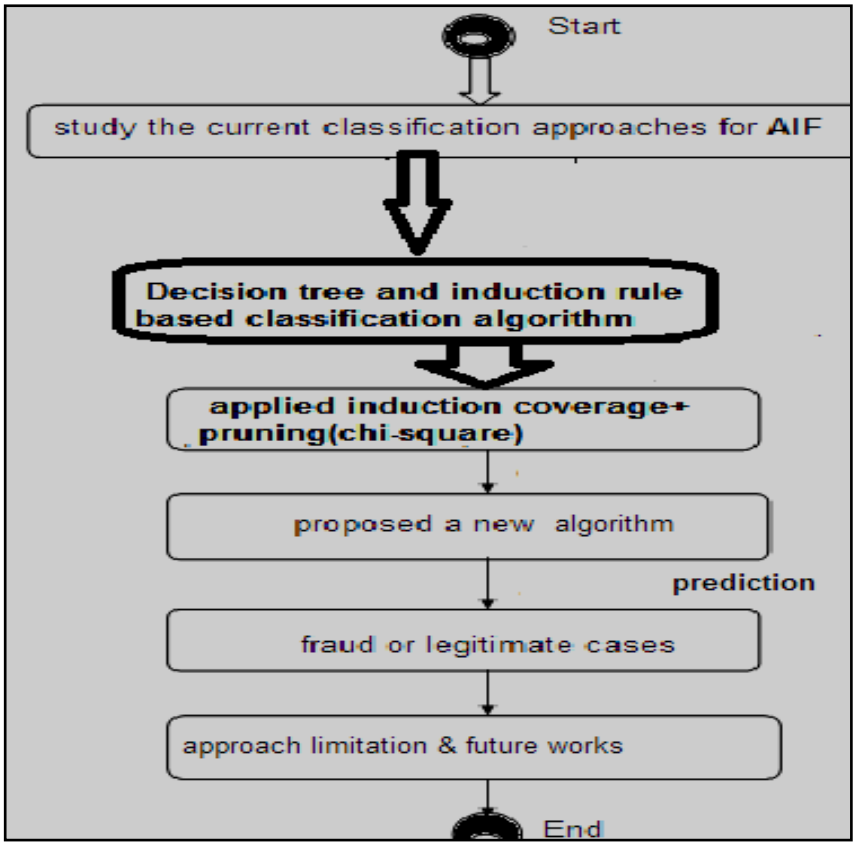


Figure 1.1 Research Project steps.

1.8 Thesis Outline

This thesis consists of four chapters. Chapter two reviews the most supervise data mining techniques and their related works regarding to the automobile insurance fraud detection problem. Furthermore, we present the state-of-the-art rule based classification approaches such as decision tree and covering algorithms regarding to the automobile insurance fraud problems. Chapter three presents the proposed model and its main steps including learning strategy, building the model and prediction procedures, we highlight on the evaluation measures for our results that generated by the proposed model with and without features selection method and compared with other rule based classification algorithms. The last chapter, Chapter four, summarizes the main achievements of this thesis, presents the general conclusions and suggests further research directions**.**

Chapter Two: Literature Review

2.1 Introduction

In this chapter we present several supervised data mining classifications techniques to handle the problem of automobile insurance fraud detection. We give a description and related works of these algorithms. We notice that some of these classifications algorithms have not been investigated exactly in the automobile insurance fraud industry to classify and predict claims cases characteristics into "fraud" or "legitimate". In addition, there is a little works on unsupervised data mining techniques for automobile insurance fraud such as (Brockett et al. 2002). The unsupervised learning data mining techniques is out of scope for this thesis and our aims. Thus, we concentrate mainly on supervised learning data mining techniques.

This Chapter is devoted to explore academic sources on the field of automobile insurance fraud detection problem based on common data mining and machine learning techniques in order to meet objectives of our thesis. The structure of this Chapter as follows: Section 2.2 presents the common supervised data mining classification approaches followed by rule based classification models in form of "IF-THEN" rule which presented in Section 2.2.2, this Section divided on two classification models: Section 2.2.2.1 discusses the Divide and Conquer model and common algorithms like C4.5 and CART DTs algorithms as well as their related works. Section 2.2.2.2 discusses the separate and conquers model and common algorithms like PRISM, RIPPER and IREP algorithms. Some of common evaluation measures in classification which presented in Section 2.3, and finally the Chapter Summary in Section 2.4.

2.2 Supervised Data Mining Techniques

Supervised learning methods which will be presented in this thesis use labeled training data, in which class to which a training sample belongs is known during the learning process. This data is used to build the predictive model and unlabeled data is used to test the model. Numerous classifications of data mining techniques related to supervise learning are surveyed in this chapter; all of them output a classifier that can be used for prediction and classification of any type of classification problem. Also, most of them have been applied most extensively to provide primary solutions to the problems within classification of fraudulent data in automobiles

insurance to handle fraud detection. Example of classification techniques that are used in the application of financial fraud are (Ngai et al. ,2011), (Sudjianto et al., 2010) and (Li et al .,2007).

Here we discuss only techniques for the detection of fraudulent claims in the automobile insurance. This problem not only affected on the revenue and profit of insurers, but also the cost resulting from the fraud maybe reach to millions in poor countries and billions dollars in developed societies. The subsequent sections shed the light on the learning strategy, advantages and disadvantages of them.

2.2.1 Probabilistic Approaches

2.2.1.1 Naive Bayes (NB) Approach

In (NB), there is no influence of an attribute value on a given class of the attributes values (Viaene et al., 2002) . This means that a class of a given attribute is independent of the values of other attributes. Basically, in order to predict the class of a given attribute(s) using NB technique, the probability of this evidence with each class is calculated. The class that have the highest probability value is selected as the class of that evidence (Bhowmik, 2011).

Regarding to the automobile insurance fraud problem, the equation (2.1) used to determine the probability of claims using NB a classifier to predict the claims as to be legitimate or fraud cases.

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})} \quad (2.1)$$

Since "p(x)" is constant for all classes, thus it is ignored. In general, let T be a training set of tuples and their classes, and each tuple is represented by an n-T attribute vector X = (x1, x2, …, xn) and m classes C1, C2, …, Cm. Moreover, the classification in NB is to derive the maximum posteriori, i.e., the maximal P(Ci|X). Since the equation (2.1) became (2.2).

$$P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i) \quad (2.2)$$

(Bhowmik, 2011) applied NB and DT-based algorithms such as "C4.5" and "C5.0" in order to predict the fraudulent data of automobile insurance. Since both techniques used the same data in order to analyze the classifier predictions. The researcher used

subset of attributes to apply their techniques. The study found that the learning phase and classification phase are very fast in DT. Also, when applied C4.5 for huge dataset, the performance of C4.5 are minimized. C5.0 provides an enhancement to DT induction, on other hand the NB classifier can assign a new data case to the class that has the highest probability. Also NB is effective and do well with respect to the accuracy when it's compared to C4.5 DT and backpropagation algorithms. The results of these techniques were evaluated using confusion matrix against test data set and found that the accuracy was equal to 78%, as well as the recall and precision were equal to 86% and 70% respectively.

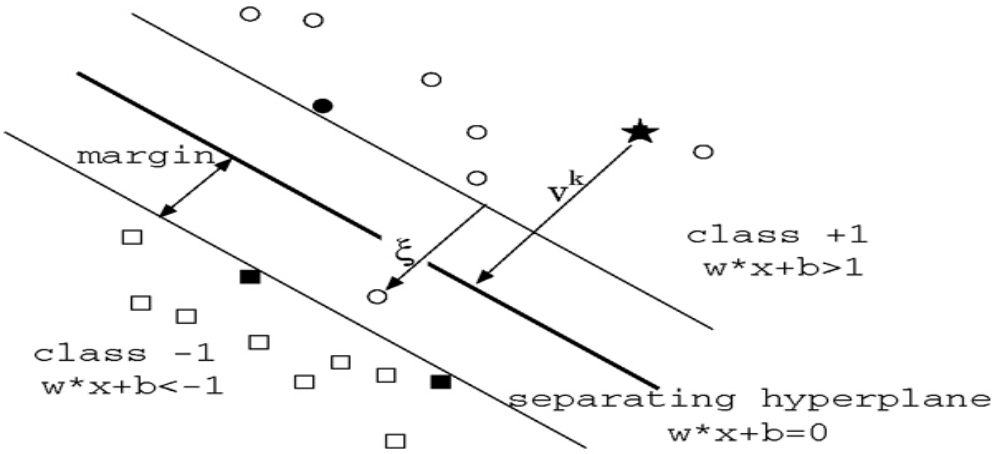
Moreover,(Viaene et al., 2004) applied a weight of evidence reformulation of AdaBoosted NB to the problem of diagnosing insurance claim fraud. The data sets were obtained from accidents in "Massachusetts" during 1993, and gathered by "Automobile Insurance Bureau of Massachusetts". These data related to the personal injury protection automobile insurance claims. The data covered some characteristics of input indicators. These indicators are equal to (48) for accident (12 input), claimant (5 inputs), insured driver(6 inputs), injury (11 inputs), treatment (7 inputs), and lost wages (7 inputs). Compared to the smoothed NB, and adaBoosted AB, the results of the study shown that "adaBoosted weight of evidence" outperformed classic NB and AB in the rate correctly classified and (AUROC) which are 0.8443 and 0.8919 respectively. At the same time the limitation of NB and adaboosted classifier is that the real data may not satisfy the independence assumption between attributes and make the accuracy of the NB classifier highly sensitive to the correlation attributes. Assumption of class conditional independence usually does not hold, as well as dependencies among these cannot be modeled by NB.

According to the (Viaene et al.,2002), the researchers applied different classification data mining techniques such as TAN, LR, C4.5 decision tree, K–NN, (MLP)Bayesian learning multilayer perceptron neural network, and Least-squares (SVM) and evaluated regarding to the automobile insurance fraud detection. The data sample consisted of 1,399 personal injury protection claims, gathered by Automobile Insurers Bureau of Massachusetts. The study used red-flag predictors and non-flag predictors. The performance measures found in this study using Ten-Fold Cross Validation that

the accuracy and "area under the receiver operating characteristics" (AUROC) for NB are better than TAN (tree-argument NB) which equal to 84.7% and 85.2% respectively.

2.2.1.2 Support Vector Machine (SVM) Approach

In machine learning SVM is a supervised learning technique used for classification and regression analysis developed by "Vapnik", to separate two different categories of data using specific rules, and to solve Quadratic equations (Burges, 1997). SVMs carry out non-linear classification by using "kernel trick". SVM doesn't have a prior knowledge about the problem. This technique has one limit when the size of data is enormous. Figure (2.1) show the space of SVM.



Figure(2.1) : The space of SVM (Burges, 1997)

According to the (Viaene et al.,2002) the researchers applied several state-of-art classification data mining techniques where one of these techniques is the Least-squares SVM and evaluated regarding to the automobile insurance claim fraud detection. The data sample consisted of 1,399 personal injury protection claims. The study used red- flag predictors and non-flag predictors. The study using ten-fold cross validation, and found that the accuracy of SVM (Viaene et al.,2002) is better then those of NB, k-NN, polynomial SVM, C4.5 and TAN.

2.2.1.3 Artificial Neural Network (ANN) Approach

The main goal ANN to provide efficiently scalable parameterized nonlinear mappings A amongst set of inputted variables and a set of outputted variables. Figure 2.2 shows a simple three-layer neural network (an input, hidden and an output layer). ANN

consisted of a three layers. These layers have a link and connected together by modifiable weight. In ANN, the function of a processing unit play a major roles to accept signals along with incoming connections and (nonlinearly) transforming a weighted sum of these signals, into a single outputted signal.

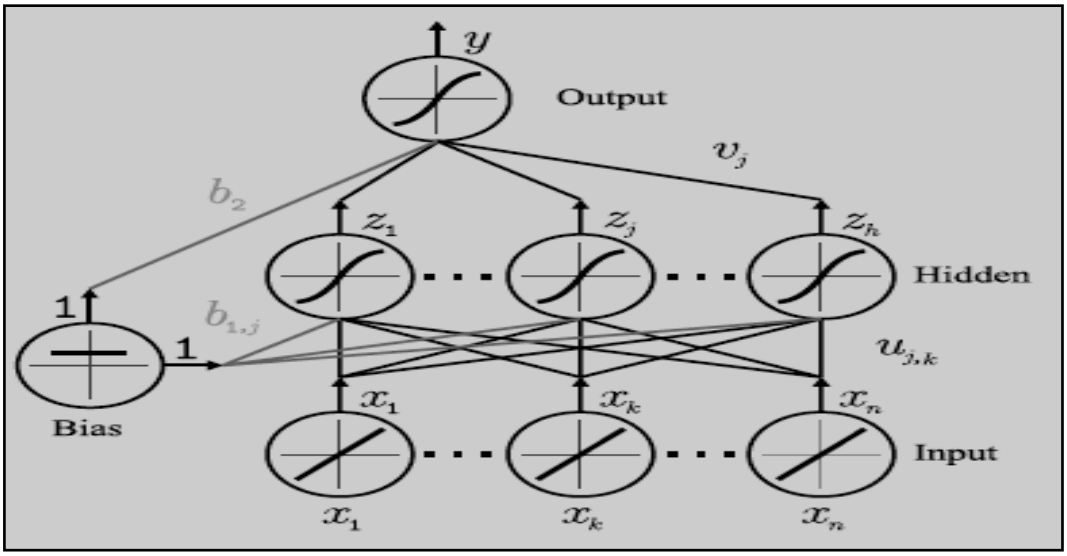


Figure (2.2): a simple three-layer neural network (Viaene et al. ,2005).

(Gepp et al., 2012) presented a comparison of several computational data mining techniques, one of these techniques is ANN with Monte Carlo methodology. The size of the data set that used in this study are 1000 data points, 700 for training set and 300 for testing . The results of this model showed that the ANN was able to detect the underlying pattern in the training data set of the 700 well enough to get 88% correct prediction. However, the test set predictions were not impressive with the ANN correctly predicting only 169 (about 56%) of the 300 test data point.

(Viaene et al. ,2005) presented the capabilities of NN with "automatic relevance determination weight" and applying this algorithm for personal injury protection automobile insurance claim fraud detection which presented in (Viaene et al.,2002). The sample based on a data set of 1,399 automobile insurance claim of bodily injury. The researchers found that "Bayesian Multilayer Perceptron-Automatic Relevance Determination" (MLP-ARD) NN better than other classifiers such that decision trees and Logit Model. MLP-ARD suffers from sensitivity to changing the training data set, especially when the data sets very large and less attractive from practical perspective due to the intensity of its computational power.

In (Brockett et al. , 1998), the researchers introduced "feature map" to classify automobile bodily injury claims fraud based on the fraud suspected. The study used feed forward NNs and a back propagation algorithms to check the validity of the feature map approach. The data set used in this study consisted of 127 claims selected from among 387 claims for accidents in 1989 in USA . In total, there are 65 fraud indicators which have been divided into 6 categories based on practice used in automobile insurance claims processing: characteristics of the accidents, the claimant, the insured, the injury, the treatment and the lost wages.

The results of "Feature Mapping Categorization" FMC are better than insurance adjusters fraud assessment and insurance investigators fraud assessment in term of average correct classification rate (accuracy). The average accuracy of FMC was equal to the (64%), and higher than those produced by domain experts by 22%. Moreover, Experts based method normally required time to discover these fraudulent claims, especially when the number of claims are increasingly as well as causing additional cost to the insurance companies.

2.2.1.4 Logistic Regression (LR) Approach

LR is a mathematical model deal with a dichotomous dependent variable (Wilson, 2009). The dependent variable is either "0" (legitimate claim) or "1" (fake claim). The equation of LR as under:

$$Pi=P(Y=1|Xik)=\exp(bkXik \div (1+\exp(bkXik)) \quad (2.3)$$

where, "Pi" is the probability that "Y=1" (fake claim) given set of characteristics for the set of independent variables(Xik), and "exp" represents "exponentiation", since exp(2) means [exp(2) = e^2 = 2.718^2 = 7.388]. The natural logarithm of the odds (probability) is called the logit of "Y". The logit of "Y" is calculated statistically then its converted back to the odds by "exp" as follows:

$$(P(Y=1|Xik)=e^{logit(Y)}) \quad (2.4)$$

Regarding to the automobile insurance fraud detection,(Wilson, 2009) presented supervised statistical techniques (LR model) for AIF to detect fraudulent claims which occurs in both auto physical damage and injury claims. The data sets consisted of 98 observations and 6 independent variables. Since the researchers applied two measures: "Chi Square" and the "–2 log likelihood".

The results of this model found that the accuracy is equal to 70.4%. Since 81.6% for the percentage of legitimate claims and 59.2% for the percentage of fraudulent claim. This model not be tested on hold out data.

(Bermúdez, et al., 2008) contrasted "Bayesian Asymmetric Skewed Logit Model", and "symmetric Logit Model" techniques and used the most common asymmetric link functions "log–log and complementary log–log" regarding to the automobile Spanish insurance. The study applied a Monte Carlo Bayesian for fitting an insurance fraud database using a dichotomous model. The data sets consisted of 10, 000 automobile claims. All claims were classified as honest (9899 cases), or fraudulent (101 cases). The same variables have been used in another studies (Artıs et al., 2002), (Caudill et al., 2005) and (Pinquet et al., 2007).

The same data sets were used for both techniques (symmetric logit model and Asymmetric Skewed Logit model) and found that the accuracy using "symmetric logit model" was equal to 60.7% while the percentage of fraudulent claims are 85.2%,and legitimate claims are 60.4%. So this indicates that the auditing cost for the percentage of legitimate claims(which its low) incorrectly classified as fraud (false negative).

On the other hand, by using asymmetric logit model the accuracy was equal to 99.5% and the percentage of legitimate claims equal to 99.7%, as well as the percentage of fraudulent claims was equal to 85.2.

Therefore asymmetric model (skewed models) is better than symmetric model (non-skewed models). Moreover, both of them had a drawback especially in case of the incorrect classification of zero situations.

(Viaene et al., 2007) applied logistic model for several scenarios, each scenario had a certain assumptions based on the existing information of each cost on in the claim screening. LR used to predict the probability of claim fraud using real data from accidents that happened in Spain. Since the claims were classified in two categories: "honest" or "fraudulent", after the investigation phase. The dataset consisted of 2403 claims: 2229 were legitimate and 174 were fraudulent. This means that about 7.24% of the claims in this sample are fraudulent, moreover the study focused on cost as a profitable approach.

The results of this study were based on the variety of scenarios, for scenario "1": suppose that no existing information that being available to the insurance company at the time of claim classification. The result of this scenario revealed that accuracy was equal to 65.5% and the percentage of "fraud" and "honest" cases were equal to 54.0% and 66.4% respectively. For scenario "2": the researchers assumed that the insurance company contained all information of each existing cost.

The result of this scenario showed that accuracy was equal to 99.42%, and the percentage of "fraud" and "honest" cases were equal to 99.43% and 99.42% respectively.

For scenario "3": the researchers assumed that the insurance company contained the average claim amount and average of each existing cost. The result of this scenario found that accuracy was equal to 23.51%, and the percentage of "fraud" and "honest" cases were equal to 89.66% and 18.35% respectively. For scenario "4": the researchers assumed that the insurance company contained claim amount information and an average of existing audit cost for every incoming. The result of this scenario found that accuracy was equal to 62.63% , and the percentage of "fraud" and "honest" cases were equal to 66.67% and 62.31% respectively.

For scenario "5": it was assumed that the insurance company contained an existing claim amount information and the predicted of audit cost. The result of this scenario found that accuracy was equal to 58.18%, and the percentage of "fraud" and "honest" cases were equal to 71.26% and 57.16% respectively, and lastly, for scenario "6": the existing claim amount and multiple-model predicted of audit cost were utilized. The result of this scenario found that accuracy was equal to 58.59%, and the percentage of "fraud" and "honest" cases were equal to 70.69% and 59.47% respectively.

(Pinquet et al., 2007) also applied statistical model called "bivariate probit model" to assess and to address selection bias. The claims datasets was from automobile insurance company in "Spain", and consisted of 64,587 claims, 80% for usual auditing strategy and 20% for holdout samples. The fraud rate are without selection bias in hold out samples.

The results of this study by using this model found that the average estimated "unconditional" fraud probability for suspicious claims is 8.4%. further, the estimated coefficient ranges between 0.36 and 0.64, and it depends on the number of regression

variables. Therefore, the "average unconditional" fraud range between 6.9% and 10.8%. From these results within set of regression variables. The drawback of this selection models is that estimation results highly depend on the regression components set.

(Caudill et al.,2005) used logit estimation approach which presented in (Artis 2002) and modified logit model based on misclassified claims. The study aimed to estimate the model of "artis, ayuso, and guillen" based on a logit model with missing information using the expectation–maximization algorithm. The data set contained information about the accident characteristics, the insured driver and vehicle, these data used for estimating the model of the year 1995 claims for car damages from a Spanish insurance company. All claims classified as "fraudulent" or "honest", 50% of the claims are legitimate, and the other half is fraudulent. The results of this study revealed that the new modified model predict about 50 claim reclassified as fraudulent from those of 998 claims were categorized as honest, which resulted in a fraud probability of 0.05. This value was closed in Artis, Ayuso, and Guillen model, therefore the new modified model found that 5% of the fraudulent claims not been detected.

(Tennyson and Salsas-forn, 2002) presented a logistic model estimation in an automobile insurance to predict the probability of a claim that being audited, by given complex claims characteristics. The data sets consisted of 1,091 automobile personal injury protection claims. Each claim classified into three categories: "confirmed" , "doubted" ,and "refuted" as well as two measures of auditing were used, one called (all audits) and the others called (investigative audits).

The results of this research found that: as in (all audits) within lowest range probability (0-0.25) of the 73 claims, 86.3% were "confirmed" by audit, and 13.7% were "doubted" or "refuted". In the middle range probability (0.25-0.5) of the 149 claims, 71.1%were "confirmed" by audit and 28.9% were " doubted" or "refuted". In the highest probability range (>0.5) of the 135 claims, 56.3% were "confirmed" by audit, and 43.8% were "doubted" or "refuted". In "investigative audits" within lowest range probability of the 61 claims, 65.6% were "confirmed" and 34.4% were "doubted" or "refuted". Of the 91 claims predicted to be investigated with probability (0.25-0.5), 57.1% were "confirmed" and 42.9% were "doubted" or "refuted".

(Viaene et al.,2002) contrasted the performance of data mining techniques using real-life auto insurance fraud data, where one of these techniques is logit model. The data sample consisted of 1,399 personal injury protection claims. The study used red- flag predictors and non-flag predictors. By using ten-fold cross validation, the study found that the accuracy of logit equals to 84.92%, the logit model classifier has the best prediction rate when compared to the others classifiers such as NB, TAN, least-squares SVM, C4.5 and K-NN .

(Weisberg and Derrig , 1998) used "Tobit regression model" to discover the potential for reducing unwarranted claims payments by applying quantitative methods to detect fraudulent of automobile bodily injury claims. The data sets used in this study consisted of 127 claim that were selected from among 387 claims. The data sets includes 62 suspect claims within a random sample of 65 claims, since all claims were divided into 6 categories based on practice used in automobile insurance claims processing.

This study are complementary of the others related works for the same researchers (Weisberg and Derrig, 1991), but with increased use of claim handling techniques in fighting fraudulent claims and analyzing the effectiveness of those methods against certain accidents data. By using regression model for the 5 models used in this study within different variables indicators as inputted of these models, the researchers found that (R^2) by the predictors of adjuster suspicion was equal to (0.65), followed by investigator which was equal to (0.56) and the lowest value among them was by fraud vote which was equal to (0.46). In addition, the study applied "Tobit regression model" within set of indicators.

The study found that the most powerful predictor with respect to the total compensation was the (claimed) medical charges and calculated by "chi-Square" which was equal to (79.0) and "P-Value" that was equal to (0.0001). Since the study used different quantitative methods for handling the claims, such as "special investigations", "medical audits and wage verifications", these techniques reduced the total settlement to (18%).

Finally, the study presented the benefit of investigative techniques by adjusters and investigator. In general, the usefulness increased with the degree of suspicion. Therefore the usefulness by the adjuster found that a check of the "Central Index

Bureau" was equal to 92.3%, an independent medical examination was equal to 92.3%, recorded statements of the claimant and insured was equal to 92.3%.

2.2.1.5 K-Nearest Neighbor (K-NN) Approach

K-NN is supervised classification method that used to classify and predict the object cases based on majority of its neighbors. K-NN algorithm is only performed locally. In K-NN all calculation process were applied until classification an object is classified by a majority of its neighbors. Moreover, all training examples are ranked based on "Euclidean distance". Sometimes K-NN called Memory-Based Classification, because induction is delayed based time to run. Therefore this algorithm required intensive computation on the training data (Cunningham and Delany, 2007). There is a little work on the K-NN classifier regarding to the automobile insurance fraud detection which is presented below.

(Viaene et al.,2002) contrasted the performance of data mining techniques using real-life auto insurance fraud data, one of these techniques is K-NN. The data sample consisted of 1,399 personal injury protection claims from 1993 accidents.. The study used red- flag predictors and non-flag predictors. By using ten-fold cross validation, the study found that the accuracy of 500-NN were equal to 83.7o% and for 1-NN was equal to 80.77%. In this study K-NN classifier have the worst performance when compared to the others classifiers such as NB, TAN,LS-SVM, MLP-NN, Logit model, and C4.5.

2.2.2 Rule Based Classification Approaches

Classification is used to determine class membership of data samples, and it's one of the most important tasks in data mining. Since the class of training samples is known beforehand, this learning is called supervised learning. The actual classification is done on the basis of the learnt classification model and it comprises of assigning a class label to test samples. The aim of a classification task is to find a classifier that can determine the class of any instance of the object with high accuracy. The performance of the classifier in predicting an "unseen" object is evaluated by applying the classifier to the testing data set. The classifier is usually represented by a set of rules. These rules have an "if-then" format: if "conditions" then "class", with a conjunction of attribute terms in the rule antecedent and a class label in the rule consequent.

There are two kinds of classification models for discovering knowledge from data. These classification model "Divide-and Conquer" (Quinlan, 1987) which presented in Section 2.2.2.1 and "Separate-and-Conquer" (Furnkranz, 1996) in Section 2.2.2.2.

In Section 2.2.2.1.1 we present a general description of DT approach that belonged to the divide and conquer model. Followed by common rule based classification algorithms such as C4.5 algorithm (Quinlan, 1993) which presented in Section 2.2.2.1.1.1, and Classification And Regression Tree CART (Breiman et al., 1984) algorithm which presented in Section 2.2.2.1.1.2. We highlight on the advantages and disadvantages of these algorithms together with their related works in Section 2.2.2.1.2 regarding to the AIF detection classification problem.

Moreover, Section 2.2.2.2.1 discusses the general description of covering and induction approaches. Followed by common rule based covering algorithm such as PRISM which presented in Section 2.2.2.2.1.1, and common rule based induction algorithms such as RIPPER (Cohen, 1995), and "Incremental Reduced Error Pruning" IREP (Furnkranz and Widmer, 1994) together which presented in Section 2.2.2.2.1.2 . We notice that some algorithms have not been investigated regarding to the automobile insurance fraud detection problem such as CART, PRISM, RIPPER and IREP. Finally the Chapter summary which presented in Section 2.4.

2.2.2.1 Divide-and Conquer Model

The construction of "divide and-conquer" model begins by choosing an node as a root node, after that this model construct a branch for each possible level of that node. This led to split the training data into subsets, one for each possible value of the node. The same mechanism will be invoked again till all data instances that belong to one branch that have the same class or the remaining data instances cannot be split. All nodes that connected between the root to the leaf called intermediate nodes. There are several algorithms that used this model in discovering the knowledge (rules) such as C4.5 (Quinlan, 1993) and CART algorithms (Breiman et al., 1984).

2.2.2.1.1 Decision Trees Approach

Decision trees (DTs) are one of the most popular data mining techniques for knowledge discovery. The roots of this technique back to the late 1970s and early 1980s when Ross Quinlan developed and introduced the first algorithm, which is called ID3 (Quinlan, 1986). Quinlan enhanced the ID3 and introduced new algorithm

called C4.5 (Quinlan, 1993). Each path from the root to leaf node becomes a rule, this path contains a root node, middle nodes and leaf nodes which corresponds to the class.

In general, the process of building any decision tree is illustrated according to (Witten and Frank, 2005) as follows: the learning method starts by selecting an attribute as a root node (A) and constructs a single branch for every possible value (A=1 and A=2). Consequently, the data set will be split into two subsets (B and C). B node will be separated into two nodes according to its value (B=1 or 2), Class X if the value was 1 and C node otherwise. The same process is repeated recursively for each branch until all data examples in the training data set at the node level have a similar classification". Figure (2.3) depicts a sample of decision tree.

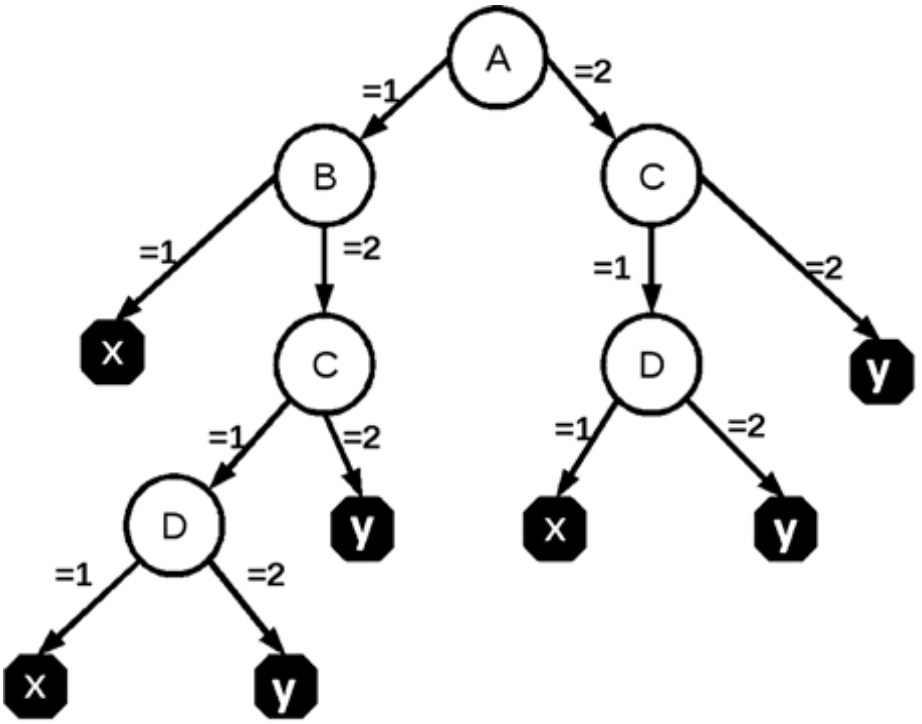


Figure (2.3) :Decision Trees and IF-THEN Transformation (Bhukya and Ramachandram, 2010).

Sometimes, the size of the constructed tree is large, so that the tree will be complex and difficult to understand (Kantardzic, 2003). In order to reduce the size of the tree and make it less complex and more understandable, there are two pruning techniques

are used; pre-pruning and post-pruning (Han et al., 2006). Normally, each single path starts from the root node going through the middle nodes to reach the leaf node is transformed in IF-THEN rules, where IF part contains all the nodes of the path excepting the leaf node (class) and THEN parts includes the leaf node

Figure 2.4.and Figure 2.5 presented the construction and pruning strategy in DTs respectively.

```
Construct_Tree (data set S )
If all  row in S belong to the same class,
Return;
For each attribute Ai
Compute splits on attribute Ai;
Apply best split found to partition S  into S1 and S2;
Construct _Tree (S1);
Construct _Tree (S2);
EndConstructt_Tree;
```

Figure 2.4: The construction of decision tree (Lokanatha and Venkatadri, 2008).

```
Prune_Tree (node t)

If  t  is  leaf return C(S) +1 / * C(S) is the cost  of  encoding  the classes  for  the rows
in set S */

Min_Cost1:= Prune_Tree (t1);

Min_Cost2:= Prune_Tree (t2); /* t1, t2 are the children of   node  t  */

Min_Cost  t := Min{ C(S)+1, C_split(t)+1+Min_Cost1+Min_Cost2};

Return Min_Cost t ;   /* C_split : cost of encoding a split  */

EndPrune_Tree;
```

Figure 2.5: Decision tree pruning (Lokanatha and Venkatadri, 2008).

There are several algorithms DTs classification algorithms belonged to the divide-and conquer approach such as C4.5 DTs, and CART algorithms.

2.2.2.1.1.1 C4.5 Algorithm

C4.5 algorithm is one of the popular classification algorithm in data mining. C4.5 algorithm is an improved version of ID3 algorithm (Quinlan, 1986). Since it's a statistical algorithm and can handled both continuous and discrete attributes, by using the concept of information gain as a heuristic for choosing the attribute that best separates the training samples on the basis of their classes, each internal node represents a test condition on an attribute and leaf nodes represent classes. Since the construction of this classifier based on top down approach. The goal of the algorithm is to build a tree that best fits the training data. The tree starts with a single node, and the attribute that has highest information gain is selected. The selected attribute become a test condition or node. A branch is created for every value of the attribute.

The same process of the algorithm is used recursively on each branch to form a decision tree. Once an attribute has appeared in a node, it is not considered again in any of the node's decedents. After building the tree, every path from the root node to leaf node becomes a rule. The leaf represents the class of the rule. C4.5 can deal with datasets that have patterns with un-known attribute values. As well as it deals with continuous attributes by discretization. C4.5 handling missing values. Moreover C4.5 has an improved method to deal with the tree pruning that minimized misclassification errors due to noise or too much detail in the training data set.

2.2.2.1.1.2 Classification and Regression Tree (CART) Algorithm

Another DTs rule based classification algorithm called Classification and Regression Tree (CART). CART classification algorithm is restricted to the binary split and belongs to the divide-and conquer approach, it constructs a binary trees, and containing exactly two branches for each decision node. CART recursively partition the records into subsets with similar values for the class attribute, as well as it deals with continuous and nominal attributes.

CART algorithm also used for regression purpose with the assistance of regression trees in predicting the results, given a set of predictor variables over a given period of time. CART uses "gini Index" for splitting purpose. Basically CART cant use stopping rule and pruning mechanism that performed back which their construction are based on cost-complexity. The CART algorithm is utilized to produce not only one, but a sequence of nested pruned trees. CART handles missing attributes value.

Further, Both CART and C4.5 algorithms are available in Weka mining tool as "J48" and "simple CART".

Since CART and C4.5 algorithms, first grow the full tree and after that prune it back. The tree pruning accomplished by checking the performance of the tree on a holdout dataset, and comparing it to the performance on the training set. The tree is pruned till the performance is similar on both datasets this is an indicating that there is no over-fitting of the training set. Another difference between them: C4.5 use a single dataset to arrive at the final tree, while CART uses a training set to construct the tree and a holdout set to prune it.

Since most of studies that used different data set showed that the accuracy and time for construction the rule of CART classifiers were outperformed C4.5 classifiers (Lokanatha and Venkatadri, 2008). Some differences between C4.5 and CART algorithms, for instances the tree construction of C4.5 classification algorithm is differs in several aspect from CART:

- Testing in CART algorithm is binary split , whereas C4.5 allows two or more branches.
- "Gini index" in CART algorithm used to make rank of the test, whereas C4.5 uses information gain.
- Pruning in CART tree based on a cost-complexity but in C4.5 is based on one pass algorithm.

2.2.2.1.2 Related Works

Regarding to the automobile insurance fraud detection, there is a little works on the DTs algorithms such as CART. Recently, (Gepp et al., 2012) presented DTs such as C5.0 within computational data mining techniques, the data sample in this study are a real-life automotive insurance fraud data from US based and consisted of 98 (49 fraudulent and 49 legitimate) the researchers used the same data set that presented in (Wilson,2009) within the most important features (variables) that extracted from the data set.

These computational techniques are: (logit analysis), (discriminant analysis), (survival analysis) and (C5.0 decision tree) respectively, the result of these comparisons found that the accuracy of logit analysis slightly outperformed other techniques which were

(70.4%), whereas (discriminant analysis, survival analysis, and C5.0) together were equal to 68.4%. Also the percentage of detection the fraudulent claims by logit analysis is better than (discriminant analysis, survival analysis and C5.0 decision tree) which it was equal to 59%. 88% is the percentage of legitimate claims by (discriminant analysis, survival analysis and C5.0 decision tree) together but in (logit analysis) was equal to 82% , at the same time the (logit analysis) was better than (discriminant analysis, survival analysis and C5.0 decision tree) together in classifying the fraudulent claims.

(Bhowmik, 2011) applied decision tree-based algorithms such as C4.5, C5.0 and NB classification to predict the fraudulent data of automobile insurance. The study found that the learning phase and classification steps are very fast in DT rather than NB, since C4.5 when utilized for huge data sets, their performance is reduced. C5.0 shows marginal enhancement to DT induction. NB algorithm is very effective and can perform well with respect to the accuracy. As a result NB can do well rather than C4.5 DT and backpropagation algorithms.

(Basak and Lim, 2009) presented a feasibility study on automating the automobile insurance claims processing and applied C4.5 decision trees to classify and to estimate the probability of which claims were fast-tracked (not exaggerated claims), and claims not fast- tracked (exaggerated claims). The database consisted of 35000 entries with (30) independent attribute within the derived variables, the data sets contained from various different region across the Indian. The results of this study found that the accuracy of the decision tree model on all labeled samples was equal to (62%), as well as the decision trees collectively for every region separately, provide an accuracy of approximately (80%) on the labeled samples, since all the rule generated were validated from the decision trees with domain experts.

Some models - software- are outperformed the logistic model by (Derrig and Francis, 2008), the outcomes of these models were compared to LR as in (Viaene et al.,2002). (Derrig and Francis, 2008) used a variety of these models, such as Classification And Regression Tree(CART) as a software program , TreeNet, Iminer Ensemble, Iminer Tree, Random Forest,and SPLUS Tree were compared to (2) of prediction methods which are LR and NB. The study applied (8) model to closed claim data. The data set consisted of the 162,76 claims, (21) variables were selected to use in the models. (4) target categorical variables were selected to the model. In

addition, the study used software implementations of six classification and regression tree methods with the benchmark procedure of NB and LR.

The result of this study were based on the "IME" and "SIU" request decision and outcome favorable for each investigation, therefore the study found that, the values AUROC for both TreeNet and Random Forest outperformed than the LR model, as well as both of them performing well better for all trees models..

(Pérez et al., 2005) presented a new method based on the Consolidated Trees Classification (CTC) vice versa (C4.5) decision trees classifiers with different class distributions. The study used several performance evaluations to measure the effectiveness of the classifiers on basis of the threshold such as recall, precision, and ROC curve for both data mining techniques. The researchers used database consisted of 108,000 examples, and just 7.40% of them are fraudulent cases, 31 independent variables about the accidents, since the dependent variables or class have two categories (fraud, and not fraud). 75% of database for training data and 25 for testing, as well as the experimentation for both model are repeated 10 times.

The results of this study found that the CTC trees classifier outperformed than the C4.5 trees classifier with respect to the many factors: accuracy, error rate, recall, precision, the structural stability or explanation, and the ROC curve. The error rate of CTC classifiers was smaller than the C4.5 trees classifiers for all threshold used (CTC was stable classifiers and small complexity in average) than the C4.5 trees classifiers, the common variable (is calculated starting from the root and covering the tree, level by level) which measure the structural stability or explanation and found that by CTC was equal to 49.36% in average but in C4.5 equal to 10.31%, the average AUC for all the analyzed class distributions are : 68.87% for CTC and 60.71% for C4.5 classifiers.

There is a little works for using hybrid model regarding to the automobile insurance fraud detection problem, such as (Phua et al., 2004) presented a new fraud detection method (meta-learning approach) for the skewed data distributions problem, the researchers attained accuracy improvements by merging C4.5, Back-Propagation(BP) artificial neural networks and Naïve Bayes (NB) model when applied to 15,421 cases of automobile insurance fraud, and it is interesting to note that C4.5 was a

very important predictor as part of the hybrid model (stacking-bagging model) on the data partition.

The sample contained 15,421 cases of automobile insurance fraud, 11.338 (training data), and 4083 instances (score data), 6% were fraudulent and 94% were legitimate distribution. The original data set has 6 numerical attributes and 25 categorical attributes, including the binary class label (fraud or legitimate).

The study found that the C4.5 is the best classifiers, followed by NB, and BP, but when merging these classifiers (stacking-bagging), the hybrid approach(stacking-bagging model) achieved the highest cost savings and the optimum success rate is 60% for highest cost savings in the skewed data set.

Several advantages and disadvantages of the DTs amongst the other data mining techniques. Hereunder we listed these advantages:

- DTs are Simple to understand for normal users, since it based on chunk of knowledge (If-Then-Rule).
- DTs can deal with nominal and numeric attributes.
- DTs representations are rich to represent any discretization value classifier.
- DTs can handle the datasets that may have errors.
- DTs can handle the datasets that may have missing values.
- DTs have no assumptions about the space distribution and the classifier structure.
- Data in DT need little preparation. Than other techniques that needs normalization steps.
- DT Performs well with huge data set with respect to the time typically in short.

On the other hand, some disadvantages of DTs (limitations) as under:

- Most of the algorithms of DTs such as C4.5 require that the target attribute (class) will have only discrete values. Since C4.5 can handling missing of attributes values and its performed well with missing values of existing attributes.

- Most DTs divide the instance space into regions, in some cases the tree should contain several duplications of the same sub-tree in order to represent the classifier. In addition, the greedy characteristic of DTs leads to over-sensitivity to the training set, to irrelevant attributes and to noise (Quinlan, 1993).
- Finally DTs do not generalize the data well, and its caused overfitting. A pruning are very significant to mitigate this problem.

2.2.2.2 Separate-and Conquer Model

Separate-and-Conquer model begins by constructing the rule in greedy design. After a rule is existed, all data instances covered by the rule will be discarded and this mechanism is invoked again till the best rule existed has a large error rate. Since in classification rules, there is only one pr-identified class. There are several algorithms that used this model in discovering the knowledge (rules) such as PRISM (Cendrowska, 1987), RIPPER (Cohen, 1995) and IREP (Furnkranz and Widmer, 1994).

In the last few years, rule based classification become a popular approach in data mining where the output is represented in "If-Then" knowledge and stored in the knowledge base, whereas the problem of traditional rule based classification especially induction algorithms like RIPPER is limited size classifiers (limited number of generating rule) with often low accuracy. On the other hand, PRISM algorithm generates large numbers of rules and always tries to get perfect rules. PRISM beside not treating numeric attributes.

2.2.2.2.1 Covering Approaches

The Separate-and-Conquer approach (Furnkranz, 1996) started by taking every class and creates rule that cover several instances of this class as possible, while excluding a few instances of other classes as possible. The covering approach examines only and only one class at a time.

This approach is building a rule, for each class by choosing and adding tests to the rule till the subset of instances covered by the rule are "pure". Then all instances covered by the rule will be discarded from any further processing, since the rule

generation process continues till no more unclassified instances are left in the data sets.

Building the classification rules are described into direct method and indirect method. Direct methods are those that extract rules directly from data such as RIPPER. Indirect methods are those that extract rules from other classification model like decision tree such as C4.5. There are several classifiers derived from these approaches such as PRISM, RIPPER and IREP.

Since the advantages of covering approach is time efficiency for creating knowledge of rule directly without inducing an intermediate DT as well as it immediately ignored instances covered by the new rule from further induction.

For example PRISM is a simple and straightforward covering algorithm. Its works by choosing a class from the data set to create a new rule having the target attribute (class) and its conclusion. Basically, the PRISM adding tests to the condition of the rule, to get a maximum number of instances covered as well as to arrive the 100% accuracy (higher accuracy). Details on these algorithms are presented in the following Section.

2.2.2.2.1.1 PRISM Algorithm

PRISM is a simple and straightforward covering algorithm. It works by choosing a class from the data set to create a new rule having the target attribute and its conclusion. Basically, the PRISM adding tests to the condition of the rule, to get a maximum number of instances covered as well as to arrive 100% accuracy (higher accuracy). The accuracy of the test is measured by (p/t) where (p/t) is a ratio of the number of positive instances (p)to the total number of instances covered by the rule (attribute being used) , after that the positive instance covered by the new rule are deleted from the data set for further rule generation. The criteria used for test selection and standard " purity". Figure 2.6 showed the Pseudocode of standard PRISM algorithm.

```
For every class C
Initiate E to the instances set
While E contains data instances in class C
Construct  rule R with an empty condition that predict class C
Until   R is perfect (or further attributes to use) DO
FOR each attribute X not mentioned in R, AND each value v, consider
add the condition X=v to the condition side of R
choose X and v to maximize the accuracy p/t
Break ties by choosing the condition with the largest p
Add  X=v to R
delete the  data instances covered by R from E
```

Figure 2.6: Pseudocode of PRISM Algorithm (Cendrowska, 1987).

In real world applications, removing other rules that PRISM discarded them didn't applicable especially in large database and 100% accuracy can't be benefit in order to assess a certain situation, these hidden rules were considered very important for prediction used.

Therefore, in the proposed model which presented in Chapter three, we modify standard PRISM in order to discovery more rules than PRISM algorithm, resulting in medium size classifiers this is done, by the new model based strength threshold that inputted through the Graphical User Interface GUI by the end-user.Regarding to the automobile insurance fraud detection problem, PRISM algorithm have not been investigated in this domain.

Here is a simple example of PRISM algorithm in the Table 2.1. Assume we want to derive a rule for "recommendation = hard", based on the following dataset which belongs to the "lenses "dataset ".

Table (2.1): "Lenses dataset". (Cendrowska, 1987)

"Age"	"Spectacle prescription"	"astigmatism"	"Tear production rate"	"Recommended /class"
"Young"	"Myope"	"No"	"Reduced"	"None"
"Young"	"Myope"	"No"	"Normal"	"soft"
"Young"	"Myope"	"Yes"	"Reduced"	"none"
"Young"	"Myope"	"Yes"	"Normal"	"hard"
"Young"	"Hypermetrope"	"No"	"Reduced"	"none"
"Young"	"hypermetrope"	"No"	"Normal"	"soft"
"Young"	"Hypermetrope"	"Yes"	"Reduced"	"none"
"Young"	"Hypermetrope"	"Yes"	"Normal"	"hard"
"pre- presbyopic"	"Myope"	"No"	"Reduced"	"none"
"pre- presbyopic"	"Myope"	"No"	"Normal"	"soft"
"pre- presbyopic"	"Myope"	"Yes"	"Reduced"	"none"
"pre- presbyopic"	"Myope"	"Yes"	"Normal"	"Hard"
"pre- presbyopic"	"Hypermetrope"	"No"	"Reduced"	"none"
"pre- presbyopic"	"Hypermetrope"	"No"	"Normal"	"soft"
"pre- presbyopic"	"Hypermetrope"	"Yes"	"Reduced"	"none"
"pre- presbyopic"	"Hypermetrope"	"Yes"	"Normal"	"none"
"presbyopic"	"Myope"	"No"	"Reduced"	"none"
"presbyopic"	"Myope"	"No"	"Normal"	"none"
"presbyopic"	"Myope"	"Yes"	"Reduced"	"none"
"presbyopic"	"Myope"	"Yes"	"Normal"	"hard"
"presbyopic"	"Hypermetrope"	"No"	"Reduced"	"none"
"presbyopic"	"Hypermetrope"	"No"	"Normal"	"soft"
"presbyopic"	"hypermetrope"	"Yes"	"Reduced"	"none"
"presbyopic"	"Hypermetrope"	"Yes"	"Normal"	"none"

Next Figure 2.7 presented all the candidate tests and their accuracies after choosing the recommendation = hard as a class label.

"Age = young"	accuracy= 2/8
"age=pre-presbyopic"	accuracy= 1/8
"age=presbyopic"	accuracy= 1/8
"spectacle prescription=myope"	accuracy= 3/12
spectacle prescription=hypermetrope	accuracy= 1/12
"astigmatism=no"	accuracy=0/12
"astigmatism=yes"	accuracy= 4/12
"tear production rate=reduced "	accuracy= 0/12
"tear production rate=normal"	accuracy= 4/12

Figure 2.7: Candidate tests and their accuracies after choosing the" recommendation = hard"

After calculated the accuracy (p/t) of the each attribute values, among the 9 candidates, the following two have the highest accuracy "astigmatism=yes" and "tear production rate=normal". So we randomly choosing the "astigmatism=yes", so the first intermediate rule is: "If astigmatism=yes then recommendation=hard". This process are repeated until To get the perfect rules by applied more tests, finally we find that "spectacle prescription=myope" has higher accuracy (perfect rule), so the rule become : "If astigmatism=yes and tear production rate=normal and spectacle prescription = myope then recommendation = hard". Since the rule that we derived covered 3 out of 4 instances that have "recommendation = hard". Therefore, we delete these 3 instances and start the process over again. finally, the complete rules list for "Recommendation = hard", as the following:

R1:"If astigmatism=yes and tear production rate= normal and spectacle prescription = myope then recommendation = hard".

R2:"If age=young and astigmatism=yes and tear production rate=normal then recommendation = hard**".**

2.2.2.2.1.2 RIPPER and IREP Algorithms

RIPPER (Cohen, 1995) and IREP (Furnkranz and Widmer, 1994) are another well known rules based classification that belongs to the separate and conquer approach. RIPPER is improved version of IREP. Both classifiers divided the dataset into 2 sections: the growing set and the pruning set, since the growing set is applied to form the over-fitted rules, while the pruning set used to prune and evaluate these rules. These algorithms splits the training data set randomly into growing set "two-third" of samples) and pruning set "one-thrid" of samples). The rule growing step is carried out to form over-fitted rule. Then the rule is immediately pruned by deleting conditions in the reverse order till no deletion enhance the prediction of the rule.

IREP is a greedy rule induction which learn a rule at a time, since the rule covered a maximum number of instances in its current training data, all the instances that correctly labeled by the resulting rule are terminated from the training set. This process is repeated till a predetermined stopping conditions satisfied or the training set become empty. Among the candidate sequences of condition from the rule,

In Ripper the basic sets of rules were taken by using growing set. It again adds conditions by testing each probable value for each attribute and choosing the value with maximized information gain. After that it prunes the rules by using the pruning set. Also it uses incremental reduced error pruning technique for pruning rules. Ripper slightly modifies the mechanisms that used in IREP to form and prune individual rule. Ripper provide additional optimization step to enhance the prediction accuracy of the rule set by revising, deleting or replacing the pruned rules. In general, the final rule set presented by Ripper are more accurate than those produced by IREP, and competitive with those of C4.5 without seriously effected on the algorithms efficiency (cohen,1995).

Regarding to the automobile insurance fraud detection problem, and our knowledge the RIPPER and IREP algorithms have not investigated for this kind of problem.

2.3 Common Evaluation Measures In Classification

2.3.1 Precision and Recall

Precision is an evaluation measure in automobile insurance fraud problem. It has been originated with another measure called recall. These two measures have been adopted from the IR(Information Retrieval) field (Van Rijsbergan, 1979). In order to compute the precision and recall of a classifier, the confusion matrix must be constructed which contains the total number of the following: true positive (A), false positive (B), true negative (D) and false negative (C) as illustrated in Table(2.10). Where "A" represents the total number of the positive claim that were classified correctly, "B" gives the total number of incorrect hits of positive claim, "C" denotes the number of the negative claim that were classified correctly and "D" is the total number of incorrect hits of the negative claim.The following table (2.10) show the confusion matrix as follows:

Table (2.2): Confusion Matrix Automobile Insurance Claim Fraud Problem.

	Classified Positive	**Classified Negative**
Actual Positive	**A(TP)**	**B(FP)=False Alarm**
Actual Negative	**C(FN)**	**D(TN)**

Precision Rate =A/A+C= TP/TP+FP (2.5)

Recall Rate TP=A/A+B = TP/TP+FN (2.6)

Accuracy Rate=A+D/A+B+C+D (2.7)

Error Rate=Complementary of Accuracy= 1-Accuracy Rate (2.8)

F1= 2*P*R / P+R (2.9)

2.4 Chapter Summary

Automobile insurance fraud detection is a global problem that affected to the profit of insurance companies. Data mining techniques that conducted from several studies can contribute to detect and classify whether the accident type is fraud or legitimate. Since the researchers used several rule based classification approaches that presented in this Chapter such as DTs, induction and covering approaches. However, these algorithms can detect the fraudulent cases but also have limitation. For instances covering algorithms suffered from their construction in building the rules as we mentioned before. Furthermore covering algorithms depends on separate and conquer approach, whereas decision tree based on divide and conquer approach.

In general, the limitation of probabilistic approaches in that when it's dealing with large data set its required additional extensive computation. In this Chapter we have reviewed the most important supervise data mining techniques, as well as rule based induction and covering classification algorithms in the form of "If-Then" knowledge. This form is necessary for end-user to understand and maintain the rules with a clear representation.

Furthermore, the result of rule based algorithms is an "IF-THEN" rules that used to predict the class label of their test cases. In addition, the process of construction induction algorithms such decision tree and covering algorithm is accomplished in multiple steps including rule generation, rule pruning, model construction and prediction. We notice that the covering algorithm like PRISM generates large numbers of rules. Moreover, PRISM algorithm tries to get perfect rules (high accuracy 100%). However perfect rule in competitive environment is not feasible due to the some rule that near perfection ones cant detected. These disappeared rules in standard PRISM considered very significant for future prediction.

Also in PRISM algorithm that generates large number of rules, the end user can't understand and maintain these rules produced by PRISM algorithm. Because understanding and maintaining rules in PRISM depends on the classifiers size which is hard to the end used to achieve his works in easy way. In addition, induction algorithm like RIPPER algorithm generates small numbers of rules. Since the few rules generated by RIPPER are not suitable because some rules are undetected, this is very important for AIF detection problem.

The next Chapter is devoted to represent a new algorithm named STBCP as an enhancement on standard PRISM. In order to get accurate rules and to make balance with respect to the numbers of rules in order the end user can deal and maintain them in a clear representation without effecting on the accuracy rate. The STBCP algorithm divided into Three step: 1) a new rule learning are produced .2) a new pruning method to kicks useless rule and removed them from the classifiers set.3) a new prediction method for predicting the test case to label their type. More detail of our algorithm which presented in Chapter Three. The new algorithm is applied a against "autos" data set and compared with other well known algorithms like PRISM, RIPPER and J.48 DT algorithm.

Chapter Three: The proposed Algorithm (STBCP)

3.1 Introduction

Nowadays, Automobile Insurance Fraud (AIF) has attracted a great deal of concern in whole world. Automobile insurance companies suffering from increasing the value of their compensations paid resulting from automobile accidents claims that fabricated by individuals (internal or external parties) (Ngai et al. ,2011) and (Wilson,2009), this is in order to obtain money through legal or illegal procedures, which will reflect negatively on the income and profits of these insurance companies each year. In the fraud detection process, the user utilizes different important features against the "autos" data set such as: car make, price, wheel-base, height, and length etc. to classify whether the accidents type is fraud or legitimate, therefore it is a typical classification problem where rule based classification algorithms can contribute in helping the detection of these accidents cases**.**

Rule based classification algorithms are a popular approaches in data mining where the output is represented in simple interpreted chunks of knowledge "If-Then" rules. One problem of induction classification algorithms is the limited size classifiers (small number of generated rules) i.e. RIPPER. Moreover, the mechanism of building the classifiers in the induction algorithms is based on a greedy fashion, and the construction of these classification algorithms depend on exhaustive search to get perfect rules i.e. PRISM Beside not treated numeric attributes. However, in some competitive environment, rule based classification algorithms that tries to get perfect rules is not suitable or feasible due to some rules that are near perfection (high accuracy) and can't be extracted. These hidden

Therefore, one primary goal of this thesis is to investigates the applicability of (STBCP) algorithm on the problem of detection the accident type in order to make balance in producing the rules without impacting on the classification rate, we mean balance for the number of generating rule are- neither in large nor in small numbers but between them- in order the end user and decision makers can handle and deal with these rules and maintain them in easy way.

The searching for rules is similar to coverage approach but with the strength constraints. Once rules are found a Chi-Square pruning is applied to reduce the

number of rules generated during learning by removing redundant and useless rules. This normally results in perfect and high strength rules. As we know, understanding and maintaining the classifier rules by the end-users depend on the classifier size (number of rules). Thus, using STBCP algorithm the end-users are able to understand and maintain the rules easily.

Before applied STBCP algorithm on the hard problem of fraud detection in automobile insurance application and compared with other classification algorithm like PRISM, RIPPER and J.48 DT, we prepare the features of "autos" data set, in order to use the same features for the proposed algorithm and other classification algorithms. So we converted continuous features to categorical features after that we used Chi-square scoring method that available in WEKA intelligent tools to make evaluation on these features in order to determine the relevant features that used in our experiment. Then the STBCP algorithm is firstly processed and trained on a real "autos" data set. Then the output is used to classify test instances.

In this Chapter we explain several steps of (STBCP) algorithm including: a new rule learning, a new pruning method (classifier builder) based Chi-Square testing method that reduces the size of the classifier and a new prediction procedure. Each step is discussed in the next section of the proposed algorithm without negatively effecting on the prediction rate. The structure of this chapter is as follows: the proposed model is described in Section 3.2 in which we focus on preprocessing phase and feature assessments in Section 3.2.1. Rules discovery (learning), rule pruning (classifier builder), and prediction phases have been covered in Sections 3.2.3, 3.2.4, 3.2.5 respectively. Experimental results on real auto data set are presented in section 3.3 and finally the chapter summary is given in Section 3.4.

3.2 The Proposed Model

The general steps of the proposed model are shown in Figure 3.1. The first step that should be done is the pre-processing of the input data where all irrelevant attributes and duplicated instances are removed before the learning starts. More details of the pre-processing step which is presented in Section 3.2.1. After that, the learning algorithm starts discovering rules based on coverage search. Once all rules are derived, pruning kicks into remove unnecessary rules and finally all remaining rules form the classifier. The classifier is then tested on unseen autos cases to label their type.

These steps are briefly summarized as follows:

1) Learning of the rules using a strength threshold to produce near perfect rules in rule induction fashion. The strength of a rule is presented in equation 3.1. Assume we have a rule: A $\rightarrow$ C, the confidence of this rule as under:

$$|(A \cup C)| / |A| \qquad 3.1$$

This means the frequency (<**A,C**>) together in the training data divided by the frequency of attribute value A.

2) Building the classifiers (pruning): pruning methods are employed to cut down redundant rules aiming to decrease the size of the classifiers. We used coverage based strength and Chi-Square methods for this purpose.

3) Predicting Test Data: After building the model (classifiers), we choose the best rules to represent the classifier in order to predict the class of test data. A new prediction procedure is presented in Section 3.2.5. Figure 3.1 shows the general structure of our model.

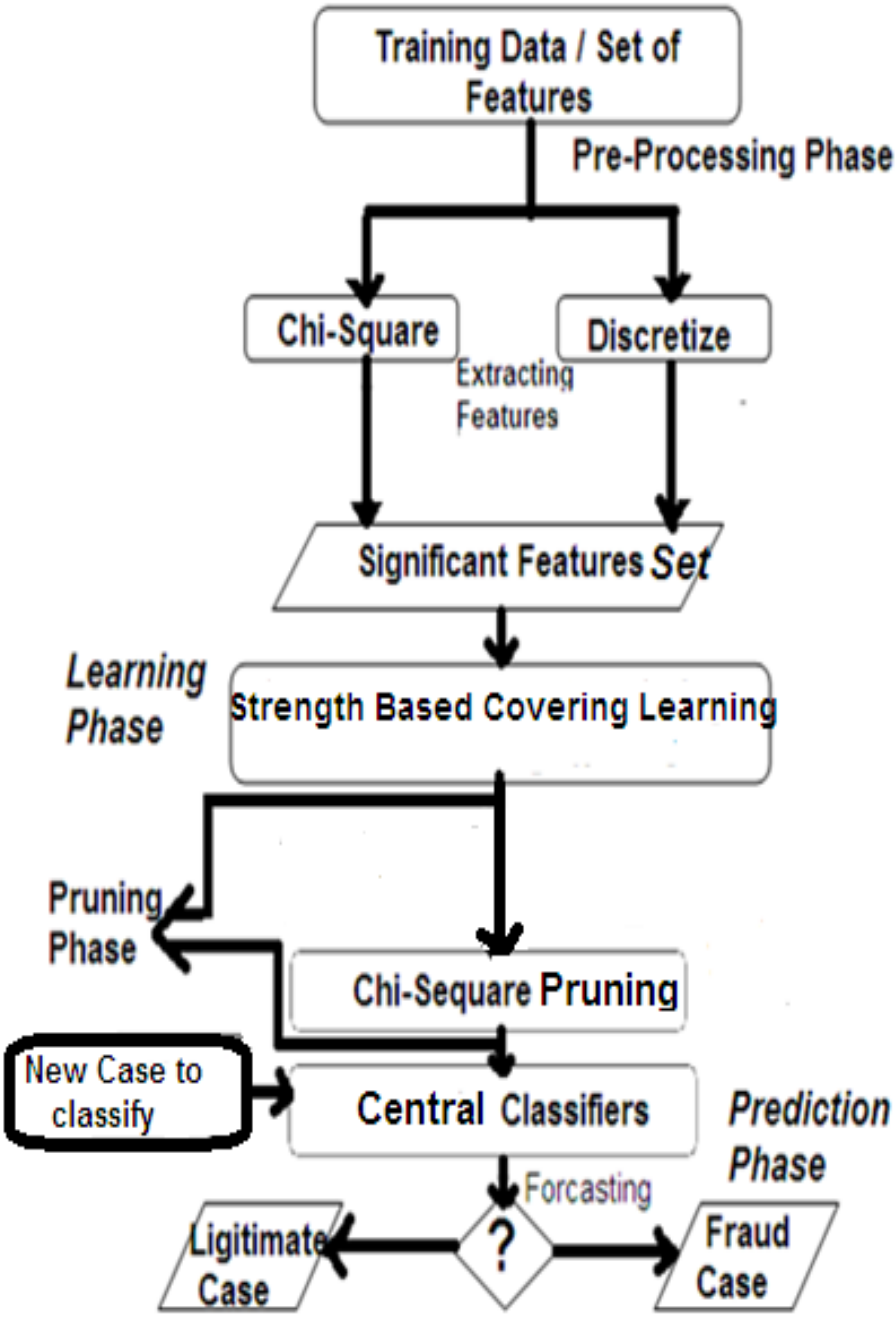


Figure (3.1) The general structure of proposed model(STBCP).

Data used by the proposed model contain attribute names and its values, the class attribute must be identified. Missing values in the training data set are treated by computing the average values of existing attributes (Al Shalabi et al., 2006). Details on our model phases (learning, classifier construction, prediction) are given in Sections 3.2.3, 3.2.4 and 3.2.5 respectively.

3.2.1 Data Preprocessing and Feature Assessment

Data preprocessing is a step that can be considered crucial for any learning algorithm (Kotsiantis et al., 2006). Pre-processing phase deals with the preparation and transformation of initial data set to a suitable format (Agrawal and Srikant, 1994). The input data may contain noise such as records redundancy, missing values, etc. (Kantardzic, 2003). Thus, the quality of the output is significantly impacted by the quality of its source. So, preparing the input data for mining is considered important in classification problem for automobile insurance fraud detection. Since the preprocessing phase is performed before the learning algorithm starts.

Our STBCP algorithm, and other classification algorithms (PRISM, RIPPER and J.48) traced on "autos" data set that obtained form the UCI **(**University of California, Irvine) *Machine Learning Repository* ***(http://archive.ics.uci.edu/ml/datasets/Automobile)*** and it consisting of 26 attributes with the class and 205 instances. The data set have 16 continuous attributes, and 10 nominal attributes. The continuous attributes have been converted into categorical attributes by using multi-interval discretisation techniques (Fayyad et al. ,1996), in Weka(Witten and Frank, 2005). The discretisation of numeric attributes starts by sorting in ascending order with the class values associated with the instance belonging to it.

After that, breaking points are placed whenever the class values changes to compute the information gain for each possible breaking point. The information gain represents the amount of information claimed to an attribute value with respect to its gain. Then the breaking point that minimizes the information gain over all possible breaking points is choosed and the algorithm is triggered again on the lower rang of that attribute. In order to calculate the information gain for the data set (S) by given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the entropy after partitioning is

$$E(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2) \qquad 3.2$$

In addition the Entropy(S) = - p_1*$\log_2(p_1)$ – p_2*$\log_2(p_2)$, therefore: 3.3

Information gain of the split, Gain(S,T) = Entropy(S) - E(S,T) 3.4

After converting the continuous features to categorical features, we find "3" features that have no gain, so we removed them. These attributes are (stroke, compression-

ratio, and peak-rpm) since the values of these features have not interval (ALL). Figure 3.2 represent a snapshot of three features that have not gain using discretisation technique by WEKA intelligent tool.

bore Nominal	stroke Nominal	compression-ratio Nominal	horsepower Nominal	peak-rpm Nominal
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(87-inf)'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'
'(-inf-3.255]'	'All'	'All'	'(-inf-87]'	'All'

Figure 3.2: a snapshot of three features that have not gain using discretisation technique by WEKA.

In this thesis, we analzed 23 different features before using them in the experiment. This feature analysis depends on the Chi-Square (Liu and Setiono, 1995). The advantages of this method is that it is easier and effective method in statistics (Suryakumar et al.,2012). It deals with data that has been measured on nominal-categorical scale. In addition, it assumes that there are no assumptions about the distribution of the population.

Other statistics methods have some characteristics about the distribution of the population such as normality (Christopher, 2011). Also in "autos" data set some attributes have missing values. We dealt with missing values found in "autos" data set, by calculating the average values of these attributes and then this value replaced in each attributes that have missing values (Al Shalabi et al., 2006).

Chi-square scoring method is utilized for extracting the significant features of autos data set using Weka software. Weka is an open source business intelligence tool that implements different machine learning and data mining methods such as (Arora and Suman, 2012) (Bouckaert et al. , 2010). Table 3.1 shows the ranking attributes of the "autos" data set after applying Chi-Square method.

From Table 3.1, we noticed that all features (25) are ranked from the highest to the lowest rank when associated with the class. Also both the ranked values of features "stroke" "compression-ratio" and "peak-rpm"are zero. So we removed them and excluding from the calculation process, since both of them does not have significant effect on the accuracy in our algorithm. Figure 3.3 shows the ranked values of 25 features by applied Chi-square evaluartor attributes using WEKA.

Table 3.1 The ranked attributes of the "autos" data set using Chi-Square method.

Feature Id	Ranked (values) for each features	Features name
1.	**423.5488**	**Height**
2.	**398.9092**	**Length**
3.	**301.4669**	**Make**
4.	**282.35**	**wheel-base**
5.	143.2358	normalized-losses
6.	109.7952	body-style
7.	105.8046	fuel-system
8.	100.8373	Width
9.	96.372	num-of-doors
10.	83.3499	engine-size
11.	79.5273	engine-type
12.	68.0995	curb-weight
13.	62.0716	Bore
14.	59.2196	Price
15.	55.8558	num-of-cylinders
16.	51.7611	highway-mpg
17.	47.3478	Horsepower
18.	44.4226	city-mpg
19.	38.7012	drive-wheels
20.	20.0715	engine-location
21.	14.662	fuel-type
22.	11.9877	aspiration
23.	0	stroke
24.	0	peak-rpm
25.	0	compression-ratio

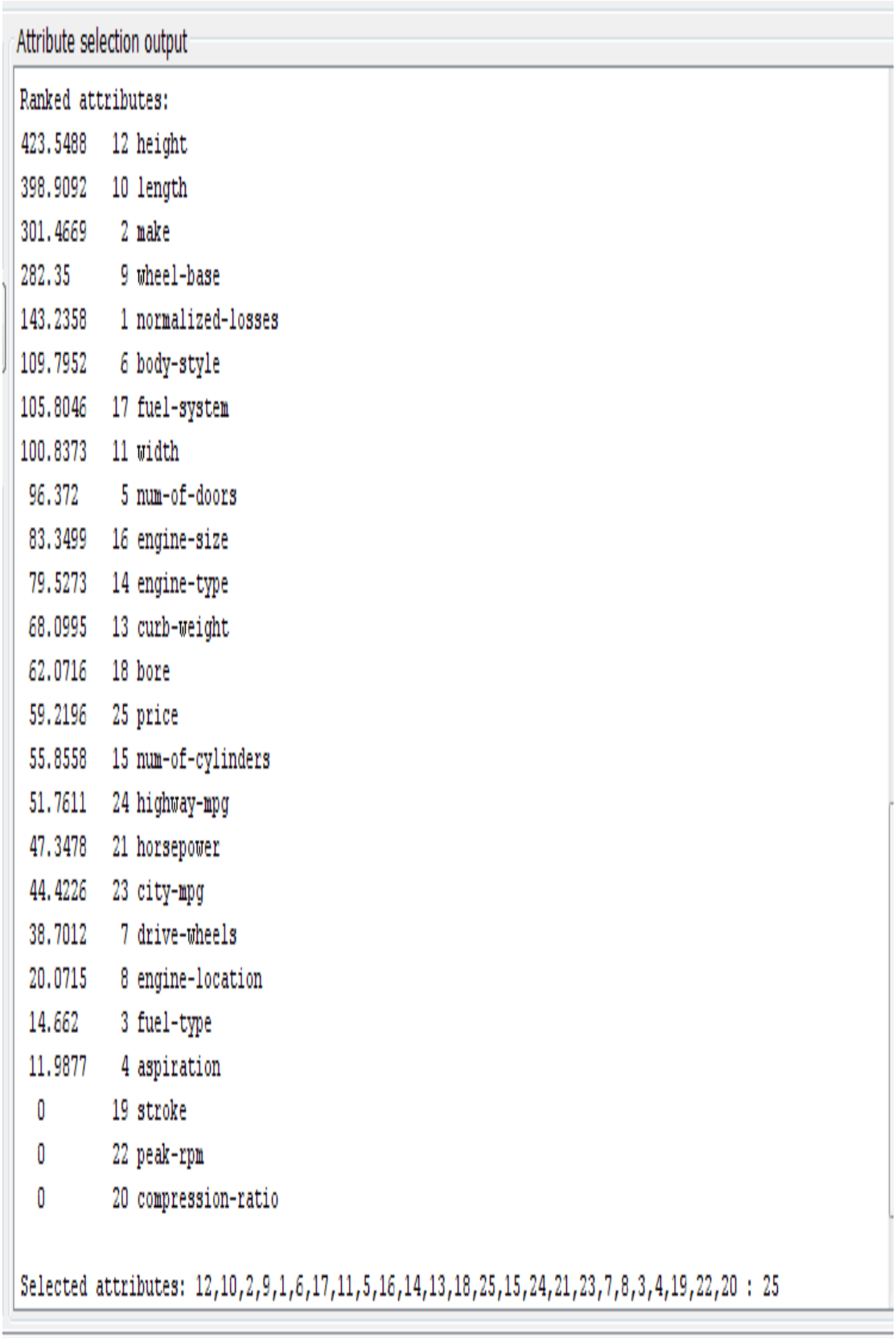


Figure 3.3: The ranked attributes of the "autos" data set using Chi-Square evaluator prduced by WEKA.

Figure 3.3 shows the ranked attributes of the "autos" data set using Chi-Square evaluator attribute prduced by WEKA. In this Figure the features "stroke", "peak-rpm", and "compression-ratio" have not ranked values so we dose not considered them in the experiment results. So, these features are removed and cant affect on the accuracy results of our algorithm and other classification algorithms.

In Table 3.2 presents the eliminated features using Chi-Square scoring and ranker searching method. Furthermore, in order to determine the most relevent features in Table 3.1, we apply equation (3.5 and 3.6) to extract features that can be used in the experiment. In the fact, there are different methods to extract the relevent features like scoring. Here, we used equation 3.5 to choose the best features (significant) for testing purpose in our algorithm and in other classification algorithms such as PRISM, RIPPER and J.48 DT.

Assuming that the highest ranked value of the features is : (Hv), the lowest ranked value of theses features is:(Lv), and R: represent the results, therefore the normalized equation become as the following:

$$R=Hv-Lv/2 \qquad 3.5$$

R= ((423.5488) - (11.9877) /2) = 205.78055

Another way for compute R is as under:

$$\sum_{i=1}^{j=n} xi + xj / n \qquad 3.6$$

where $xi + xj$: is the sumation of all ranking values of these features, and n: is the total numbers of features which equal to 23. Assume that 3.5 equation is used in our experiment. Therefore we calculate each ranked value of these features and compared to the R. If the ranked value of a single feature greater than or equal $\geq R$ (we choose this feature and inserted to the list of important features that, otherwise these features are discarded.

Table 3.2: The eliminated features using chi-square scoring and ranker searching method by WEKA.

Feature Id	Ranked (values) for each features	Feature Name	Ignored Reasons
1	143.2358	normalized-losses	Ranked feature value < R
2	109.7952	body-style	Ranked feature value < R
3	105.8046	fuel-system	Ranked feature value < R
4	100.8373	Width	Ranked feature value < R
5	96.372	num-of-doors	Ranked feature value < R
6	83.3499	engine-size	Ranked feature value < R
7	79.5273	engine-type	Ranked feature value < R
8	68.0995	curb-weight	Ranked feature value < R
9	62.0716	Bore	Ranked feature value < R
10	59.2196	Price	Ranked feature value < R
11	55.8558	num-of-cylinders	Ranked feature value < R
12	51.7611	highway-mpg	Ranked feature value < R
13	47.3478	Horsepower	Ranked feature value < R
14	44.4226	city-mpg	Ranked feature value < R
15	38.7012	drive-wheels	Ranked feature value < R
16	20.0715	engine-location	Ranked feature value < R
17	14.662	fuel-type	Ranked feature value < R
18	11.9877	aspiration	Ranked feature value < R
19	0	stroke	Ranked feature value < R
20	0	peak-rpm	Ranked feature value < R, therefore, both of them are discarded when we used the equation 3.1 and 3.2, as a result just 19 features are considered in the calculation process.
21	0	compression-ratio	

Table 3.3 contains the most important features that extracting using Chi-Square scoring method, in order to prepare the autos data set to used in the experiment results and thus these relevant features are refering to the fraud. These ranked values for each features satisfy (≥R). Furthermore, out of 25 features we extrated 4 features. Since the

number of the features extracted vary and depends on the used techniques for evaluating the attributes and the searching methods. Section 3.2.2 discusses the mechanisim that used for representing these features (data representation).

Table 3.3: significant features of autos data set using ranker searching method and chi-square filter evaluator.

Feature Id	Ranked (values) for each features (≥R)	The significant features
1	437.173	length
2	423.5488	height
3	306.8962	wheel-base
4	301.4669	make

Since Weka software provides various techniques for evaluating the attributes using different methods. Our choice of Chi-Square and ranker methods for each features in that, Chi-Square has been utilized by many scholars successfully to identify the significant variables (Christopher, 2011) and (Liu and Setiono, 1995). Also, Chi-Square method is used for large data set effectively (Suryakumar et al.,2012).

3.2.2 Data Representation

Horizontal data format inherited from association rule mining (Thabtah et al., 2005). The training data set represents as a group of rows (records) where each row has an integer number, which is considered as unique identifier, followed by attribute values. The features are represented in column as well as the last column is the class and all instances (attributes values) represented in rows. Table 3.4 depicts the horizontal data format for significant features that extracted from "autos" data set, these features considered to be important as indictor for fraud detection.

Table 3.4: Horizontal format of significant features for autos data set.

Id	make	wheel-base	length	height
1	alfa-romero	'\'(88.5-92.15]\"	'\'(168.75-169.05]\"	\'(-inf-50.35]\"
2	alfa-romero	'\'(88.5-92.15]\"	'\'(168.75-169.05]\"	'\'(-inf-50.35]\"
3	alfa-romero	'\'(92.15-95.2]\"	'\'(169.05-172.8]\"	'\'(50.35-52.45]\"
4	audi	'\'(97.25-100.1]\"	'\'(175.65-177.55]\"	'\'(52.45-54.85]\"
5	audi	'\'(97.25-100.1]\"	'\'(175.65-177.55]\"	'\'(52.45-54.85]\"
6	audi	'\'(97.25-100.1]\"	'\'(175.65-177.55]\"	'\'(52.45-54.85]\"

Horizontal format has been used by the majority of association rule mining and Associative Classification (AC) algorithms (Thabtah et al., 2005) (Liu et al., 1998). In the last few years, horizontal data format is considered more competitive than other data representation format. There are several advantages of horizontal data representation format in that, the execution time is almost constant, whereas the vertical data representation increase linearly which depends on the number of columns projected. Moreover in the database there is a crossover point which horizontal format do better than vertical data representation (Witten and Frank (2005).

3.2.3 Rule Learning Phase

After invoking the pre-processing methods (discretisation technique and Chi-Square scoring method) to prepare the training data set. The mining process starts. Rule learning depends on strength threshold, because it produces not only perfect rules (100% strength) but near perfect ones as well, and the learning strength is like PRISM. Though, our coverage search (separate & conquer) algorithm normally generates more several rules than standard PRISM.

All the rules in this phase are derived based on the induction strategy and then are stored according to strength and size. This has enhanced PRISM rule learning. Since a rule must satisfies user strength to be produced. The strength threshold is entered by the end user through Graphical User Interface (GUI) in order to specify the border of rule success or failure. However these additional near perfect rules that satisfied the user strength threshold in some competitive environment such as in automobile fraud detection problem are considered very significant for end user, since they denotes useful knowledge used in prediction step.

In the rules discovery step of the our algorithms, the new strength threshold plays crucial role to find out the hidden correlation rules that are eliminated by classic covering algorithm such as PRISM. Figure 3.4 shows the enhancement on the PRISM algorithm (rule learning) by using rule strength. In addition, the new algorithm utilized a pruning method based Chi-square to build the classifier by discarding any useless rules. This method is presented in Section 3.2.4.

Input: strength user threshold, Training Data T
For each class C
While T contains cases belonging to C Do
Create rule R with an empty body for C
Until R is perfect or R > = Strength threshold DO
FOR each attribute A not mentioned in R, and each value v, consider adding the condition A=v to the condition side of R
Select A and v to maximize the Strength |A U B| / |A|
Break ties by choosing the condition with larger than or equal to Strength TH, STH= ….
Add A=v to R
Else
Remove the instances covered by R from T
End

Figure3.4 Pseudocode of rule learning of proposed algorithm by author.

3.2.4 Rule Pruning Phase

Chi-Square testing is a statistical method test for discrete data hypothesis (Christopher, 2011). Moreover, after rules are sorted, our algorithm filters the rules based Chi-Square to get rid of useless rules in order to keep accurate rules in the classifier.

This method is used to discard negatively correlated rules. Precisely, if the result of the test in a rule r is larger than certain constant which equal to the **3.8415** (rule body and its class). Then this is an indication on positive correlation. A rule r will become a classifier. Otherwise r is discarded (negative correlated).

Suppose we have the following rules, after applied Chi-Square testing based on equation 3.8

R1: (a, b, c, d) →Fraud **accuracy=100% , X^2 > 3.8414**

R2: (a, b, e, f) → Legitimate **accuracy=100%, X^2 > 3.8414**

R3: (h, i, c,d) → Legitimate **accuracy=90, X^2 > 3.8414**

R4: (h, i, e,f) → Fraud **accuracy= 80%, X^2 < 3.8414**

R5: (h, i, e, d) → Legitimate **accuracy= 50%, X^2 < 3.8414**

R1 and R2 is the highest strength rules than other rules, both of them are generated by the standard PRISM, in our model, After rules are sorted, the new algorithm filters these rules based on Chi-Square to get rid of useless rules in order to keep accurate rules in the classifier. Since if the result of the test in a rule r is larger than constant value > 3.8415. Then this is an indication on positive correlation. A rule r will become a classifier. Otherwise r is discarded (negative correlated). Figure 3.5 illustrated the Pseudocode of building the classifier. In the above example, the Rules (R1,R2,R3) were inserted to the classifier list, since all these rule are positive correlated > 3.8414 (the test of chi-square for each rule larger than threshold constant). So these rules become classifiers and will be used them to predict the test cases. (R4,R5) are removed, since the test of chi-square for both rules less than constant threshold. The prediction steps are invoked. Details on this step are presented in Section 3.2.5.

Figure 3.5 presents the Pseudocode of classifier builder of our model. Chi-Square method used in AC (Christopher, 2011). Hereunder the equation of Chi-Square.

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}, \qquad 3.7$$

where e_i is the expected frequencies as well as the f_i is the observed frequencies. When e_i and f_i are different, the hypothesis that they are correlated is refused. In order to calculate the X^2 value, suppose we have a rule: A⟶C. The calculation of this rule using X^2 as follows: let sup(A)=a, sup(C)=c, sup(A U C)=z and the number of the instances in the data set = N. The calculation of X^2 using a, c, z and N: as follows

$$X^2 = \frac{N(z-ac)^2}{ac(1-a)(1-c)} \qquad 3.8$$

Given a set of generated rules Rs, and the training data set T. The classifier builder works as follows:

Output: Classifiers Cl

for each rule ri in Rs Do

if correlation (A⟶C) > 3.8415 // ***Threshold constant value of Chi-Square***

insert (A ⟶C) into classifier

else

remove (A⟶C) // ***Negative Correlated***

End

Figure 3.5: Pseudocode of building the classifier of the proposed model by author.

3.2.5 Prediction Phase

Prediction in data mining is the process that predict the class of unseen test data case . In predicting test data as shown in Figure 3.6, our algorithm gives the test the class of the rule that its body matches the test data. Otherwise, each rule(R) in the classifier have a weight which represents number of corresponding items between the test case and the rule(R) body over the total number of example's items. This weight of each rule *R* can be computed by the following equation:

$$Weight(R)=X/N \quad (3.9)$$

This is a new prediction procedure that utilized in our STBCP algorithm. Since X

Represents the number of corresponding items between *R* and the test case, and *N* is the total number of example's items.

The test case is assigned the class of the rule *R* that holds the highest weight. If there is more than one rule holding the same highest weight, the test data is assigned the class of the rule that holds highest strength. Finally, in cases when no rules in the classifier are applicable to the test data, the default class (majority class in the training dataset) will be assigned to that case.

```
Input: Classifier and test data
Output: predicted class
give a test case and the classifier, the prediction process works as follow:
for each test data Do
 for each rule  in c1 Do
        if test data  matches  ri
Assign ri class  to test data
else
begin
find all applicable rules that partly match test data
compute the weight  of the rules
assign the class of the highest weight  to test data
if two or more rules have identical weight
assign the class of rule with the highest strength to test data
else
assign the default class to test data.
end if

end
```

Figure 3.6: The prediction algorithm of the proposed algorithm by author.

Here is a simple example to show new prediction procedure in STBCP algorithm, for instances, suppose that we have the following rules:

R1: (a, b, c, d) →Fraud

R2: (a, b, e, f) → Legitimate

R3: (h, i, c,d) → Legitimate

R4: (h, i, e,f) → Fraud

R5: (h, i, e, d) → Legitimate

We want to predict the class of the following test cases:

Case1: (a, b, c,d) →??

Case2: (h, i, c, f) →??

Based on the above model, we find that case"1" matches the body of R1. Consequently, the class label of case"1" is "fraud" which is the class of R1. For case"2", we find that case"2" corresponds with "R1" and "R2" in one item, "R5" in two items, "R3" and "R4" in three items. The weight of R4 and R3 is 3/4 which is the highest weight of the other rules. Thus, we consider "R4" and "R3" to predict case2 class. The category of "R4" is "fraud" where R3's is "legitimate". In this case we choose the class of the rule which it's strength greater than the other. In our example, suppose that "R3" precedes "R4" in strength, the class of R3 will be chosen as the category of case"2".

3.3 Data Set and Experimental Results

Different rule based classification algorithms are compared with our algorithm according to classification accuracy, and number of rules that generated by these algorithms. In order to evaluate our algorithm, it has been compared with other traditional classification techniques such as: covering algorithm like PRISM, induction algorithm like RIPPER and J48 DT algorithm. The same "autos" data sets were used in our algorithms and other classification algorithms, we evaluate the results of our algorithm and other classification algorithms against the complete (26 features) and relevant features (4 features) extracting from autos data set.

The reason behind selecting these algorithms is the different training strategy they use in discovering the rules. For instances, PRISM belong to separate and conquer approach, and didn't utilize any pruning. Since RIPPER uses extensive pruning. DT like J48 belong to divide and conquer approach and it adapted additional pruning. Furthermore, in our experiments we used cross validation as a testing technique to produce the classifiers. This method often divides the training data set into (n+1) folds arbitrary and the rules get learned from n folds at each iteration and then evaluated on the remaining hold out fold. The process is repeated n+1 times and the results are averaged and produced.

3.3.1 Autos Data Set

Autos data set were obtained from the UCI Machine Learning Repository (University of California, Irvine**)** and it consisting of 26 features with the class and 205 instances (records).(16 are continuous features and 10 are nominal). *(http://archive.ics.uci.edu/ml/datasets/Automobile).* The results of this thesis and other compared classification algorithms were based on the "autos" data set. Before using autos data set were treated the missing values that found on them, by calculated the average values of these attributes and then this value are replaced in each attributes that have missing values (Al Shalabi et al., 2006). Figure 3.7 depicts some information of the significant features that extracted from autos data set that used in our algorithm and other compared algorithms.

We make fair treatment for preparing the autos data set to be used when derived the results in our algorithm and other classification algorithm that used in our experiment. For example PRISM beside not treating numeric features and can't handle the missing value, compared to the other classification algorithms, After that, we applied Chi-Square in order to extract the significant features. The Chi-Square method extracted only four significant features that used in our experiments out of 26 features.

These features considered very important to predict the fraudulent cases. We evaluate the results of our algorithm and other compared algorithms using the same features of "autos" data set with significant and complete features. The derived results by our algorithm using significant features and all features were compared with other classification algorithms with respect to the numbers of rule that been produced and the average accuracy.

Since the pre-processing steps which applied on these data including discretisation techniques and chi-square scoring method were producing by the same WEKA

No.	make Nominal	wheel-base Nominal	length Nominal	height Nominal
1	alfa-romero	'(88.5-92.15]'	'(168.75-169.0...	'(-inf-50.35]'
2	alfa-romero	'(88.5-92.15]'	'(168.75-169.0...	'(-inf-50.35]'
30	dodge	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
50	jaguar	'(100.1-inf)'	'(188.9-inf)'	'(-inf-50.35]'
56	mazda	'(95.2-95.5]'	'(168.75-169.0...	'(-inf-50.35]'
57	mazda	'(95.2-95.5]'	'(168.75-169.0...	'(-inf-50.35]'
58	mazda	'(95.2-95.5]'	'(168.75-169.0...	'(-inf-50.35]'
59	mazda	'(95.2-95.5]'	'(168.75-169.0...	'(-inf-50.35]'
81	mitsubishi	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
82	mitsubishi	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
83	mitsubishi	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
84	mitsubishi	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
85	mitsubishi	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
105	nissan	'(88.5-92.15]'	'(169.05-172.8]'	'(-inf-50.35]'
106	nissan	'(88.5-92.15]'	'(169.05-172.8]'	'(-inf-50.35]'
107	nissan	'(97.25-100.1]'	'(177.55-179.3...	'(-inf-50.35]'
125	plymouth	'(95.5-97.25]'	'(172.8-173.3]'	'(-inf-50.35]'
126	porsche	'(92.15-95.2]'	'(168.75-169.0...	'(-inf-50.35]'
3	alfa-romero	'(92.15-95.2]'	'(169.05-172.8]'	'(50.35-52.45]'
10	audi	'(97.25-100.1]'	'(177.55-179.3...	'(50.35-52.45]'
20	chevrolet	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
21	chevrolet	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
22	dodge	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
23	dodge	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
24	dodge	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
25	dodge	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'
26	dodge	'(92.15-95.2]'	'(-inf-168.75]'	'(50.35-52.45]'

Figure 3.7: Some characteristics of the significant features from autos data set using WEKA Tool

In autos data set it contained three types of entities: (1) The specification of an auto in terms of various characteristics like in Figure 3.7, (2) Its referred to insurance risk rating, (3) Its normalized losses in use as compared to other autos. The second rating corresponds to the degree to which the auto is more risky than its price indicates like Figure 3.8. The third factor is the average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports /speciality, etc...), and represents the average loss per auto per year.

Also From Figure 3.7 we notice that, all features have the same nominal type, for instances column "1": Represent the car make feature, this feature contained several types of autos such as : Alfa-Romero, Audi, Bmw, Chevrolet, Dodge, Honda, Isuzu, Jaguar, Mazda, Mercedes-Benz, Mercury, Mitsubishi, Nissan, Peugot, Plymouth, Porsche, Renault, Saab, Subaru, Toyota, Volkswagen, and Volvo).

Figure 3.8 shows the 25 features information and their range before applied the descretization technique and Chi-Square method.

Attributes Information

	Attribute:	Attribute Range:
1.	normalized losses:	continuous from 65 to 256.
2	make:	alfa-romero, audi, bmw, chevrolet, dodge, honda, isuzu, jaguar, mazda, mercedes-benz, mercury, mitsubishi, nissan, peugot, plymouth, porsche, renault, saab, subaru, toyota, volkswagen, volvo
3.	fuel-type:	diesel, gas.
4.	aspiration:	std, turbo.
.	num-of-doors:	four, two.
.	body-style:	hardtop, wagon, sedan, hatchback, convertible.
.	drive-wheels:	4wd, fwd, rwd.
.	engine-location:	front, rear.
.	wheel-base:	continuous from 86.6 120.9.
.	length:	continuous from 141.1 to 208.1.
.	width:	continuous from 60.3 to 72.3.
.	height:	continuous from 47.8 to 59.8.
.	curb-weight:	continuous from 1488 to 4066.
.	engine-type:	dohc, dohcv, l, ohc, ohcf, ohcv, rotor.
.	num-of-cylinders:	eight, five, four, six, three, twelve, two.
.	engine-size:	continuous from 61 to 326.
.	fuel-system:	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi.
.	bore:	continuous from 2.54 to 3.94.
.	stroke:	continuous from 2.07 to 4.17.
.	compression-ratio:	continuous from 7 to 23.
.	horsepower:	continuous from 48 to 288.
.	peak-rpm:	continuous from 4150 to 6600.
.	city-mpg:	continuous from 13 to 49.
.	highway-mpg:	continuous from 16 to 54.
25.	price:	continuous from 5118 to 45400.

Figure 3.8 : Attribute information and their range

3.3.2 Compared Classification Algorithms

Our choosing these algorithms is depend on different training strategy in discovering the rule, for instances:

- C4.5 algorithm is one of the most popular and powerful decision tree classifiers. This algorithm is available In WEKA tool as J48. Since it used additional pruning for building the rule. Moreover it belongs to the divide and conquers approach.
- PRISM algorithm is a classification rule which can only deal with nominal attributes and doesn't do any pruning. PRISM belongs to separate and conquer approach, and was used in many fields of science.
- RIIPPER algorithm also belongs to the separate and conquer approach. Since this algorithm utilized extensive pruning.

3.3.3 Results and Analysis

In this section, we present the results of our algorithm not only on complete "autos" data set (26 features) but also with the significant features (4 features) that extracted by Chi-Square. The results of PRISM, RIPPER and J48 DT were derived from WEKA Miner Tool, and it compared by our algorithm using the complete and significant features with respect to the number of rules produced and average accuracy.

In the experiments, we have set the number of folds in cross validation to 10 similar to other research studies, (Liu, et al., 1998) and (Yin and Han, 2003). In the experimentation results, the results of PRISM, RIPPER and J48 DT algorithms derived from WEKA within console version 3.6. WEKA is an open source machine learning software (Bouckaert et al., 2010), and the experimentation results of our proposed algorithm (STBCP) were implemented using java on Pentium 4 (2.8 GHz) with 512 RAM.

In order to determine the powerful of our algorithm for achieved highest average accuracy than other compared algorithm such as PRISM, RIPPER, and J48 DT, we inputted several threshold values (2%, 3%, 4%, 5%, and 6%) and run our algorithm based on these values (5 times) and compared the results of our algorithm to other classification algorithms against the same autos data set with complete and significant features.

By using these threshold values (2%,3%,4%,5%) we notice that the proposed algorithm with/without relevant features has achieved the highest average accuracy compared with PRISM, RIPPER and J.48 DT algorithms. However, the average accuracy of PRISM algorithm with/without features selection slightly outperformed proposed algorithm when the inputted threshold value is equal to 6%. Therefore, we choose 4% as the average threshold inputted values and we found that the average accuracy of the proposed algorithm with/without relevant features outperformed PRISM, RIPPER, and J.48 DT which are equal to 83.11% and 84.59% respectively.

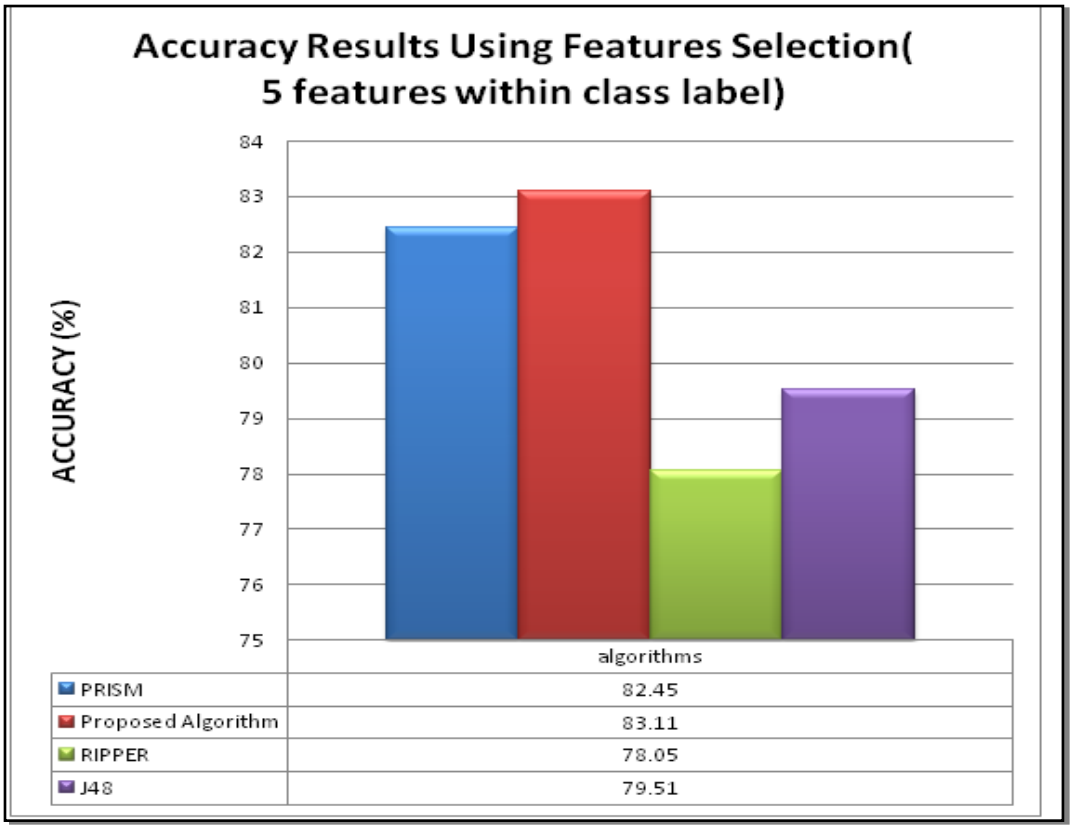


Figure 3.9: The average accuracy of the proposed algorithm with average threshold values (4%) and other classification algorithms using the most significant features against autos data set.

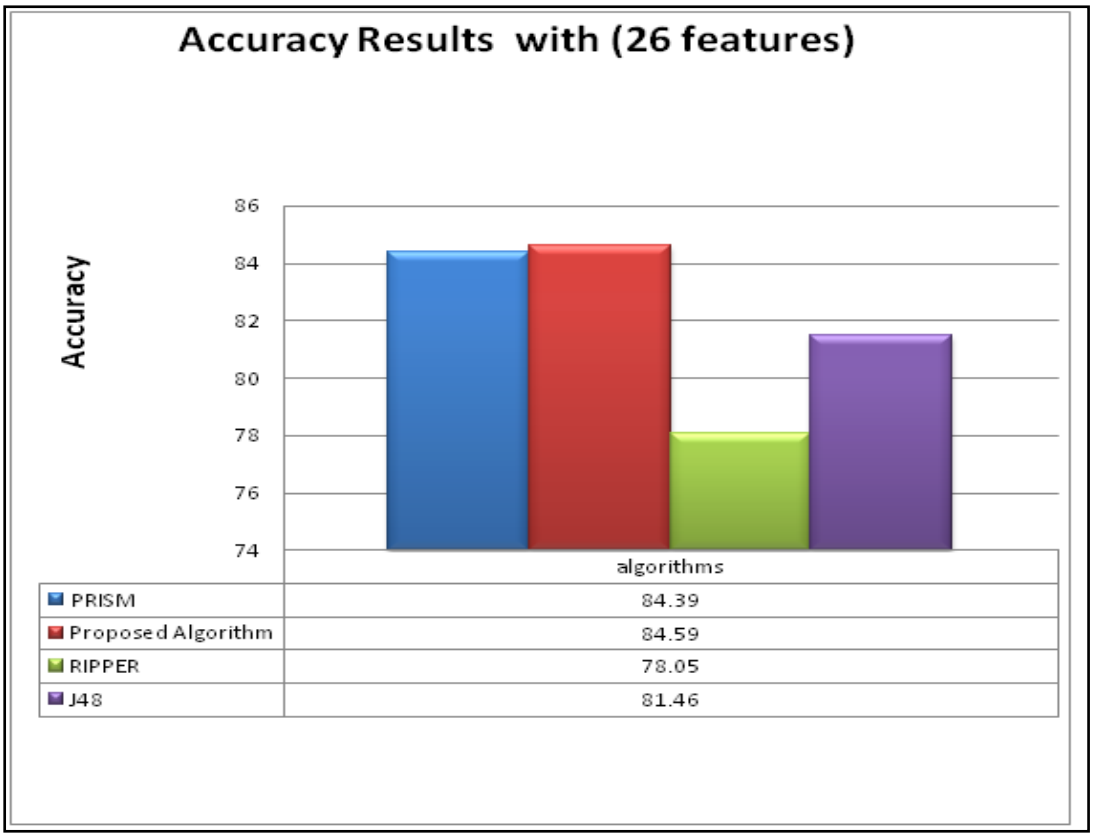


Figure 3.10: The average accuracy of the proposed algorithm and other classification algorithms with average threshold values (4%) using complete features against autos data set.

From the figure 3.9 and figure 3.10 we found that the average accuracy of the proposed algorithm (STBCP) outperformed PRISM, RIPPER and J48 DT when using the significant and complete features of autos data set, because our proposed algorithm utilized Chi-Square correlation analysis in order to build accurate classifiers for prediction purpose. After applied the Chi-Square correlation analysis, the data became correlated, useless and redundant rules are removed. Therefore the average accuracy of our proposed algorithm better than standard PRISM, RIPPER and J48 DT algorithms.

Also, from Table 3.5 and Table 3.6 we make 5 different run based on several inputted threshold values (2 - 6%) to the proposed algorithm, and we notice that the highest accuracy are produced by proposed algorithm when the strength threshold value is equal to 2% compared with the other classification algorithm and the accuracy got decreased when the strength threshold value is equal to 6%, we choose (4%) as the average of threshold values that inputted by the user, and we found that, the accuracy of average threshold values (4%) of our proposed algorithm better than PRISM,

RIPPER, and J48 DT algorithms which is equal to (**83.11**) against significant features and with complete features which equal to (**84.59)**.

Table 3.5 The accuracy results of the proposed algorithm against the most important features (relevant) of "autos" data set using several inputted threshold values.

Accuracy Results of Proposed Algorithm (%)	**Threshold Values**
84.25	2%
83.73	3%
83.11	**4% (average)**
82.67	5%
82.31	6%

Table 3.6 The accuracy results of the proposed algorithm against the complete features of "autos" data set using several inputted threshold values.

Accuracy Results of Proposed Algorithm (%)	**Threshold Values**
85.55	2%
85.12	3%
84.59	**4% (average)**
84.45	5%
82.9	6%

Figure 3.11 and Figure 3.12 revealed that our proposed algorithm reduced the number of rules in the classifier if it compared with PRISM, RIPPER and J48 DT algorithms, the number of rules that our proposed algorithm generated them neither in large nor in small but it make balance (between them). Since large number of rule generated by PRISM and J48 DT compared with the proposed algorithm. This is because PRISM always searches for the perfect rules with high accuracy and covering these rules in the training data are very limited, since it maximizes the accuracy of the rule body. On the other hand, J48 DT adapted additional pruning but generate large number of

rules. The number of rules by J48 DT algorithm is not differ than PRISM, since the data set here is small. Moreover small number of rules generated by RIPPER, because RIPPER utilized extensive pruning.

The proposed algorithm generates not only 100 accuracy rule but also near perfection one than PRISM. At the same time the proposed algorithm utilized Chi-Square testing to decrease the size of the classifiers by kicks useless and redundant rules(negative correlated).Consequently, this made our proposed algorithm more understandable and controllable than PRISM, RIPPER and J48 DT algorithms to the end user

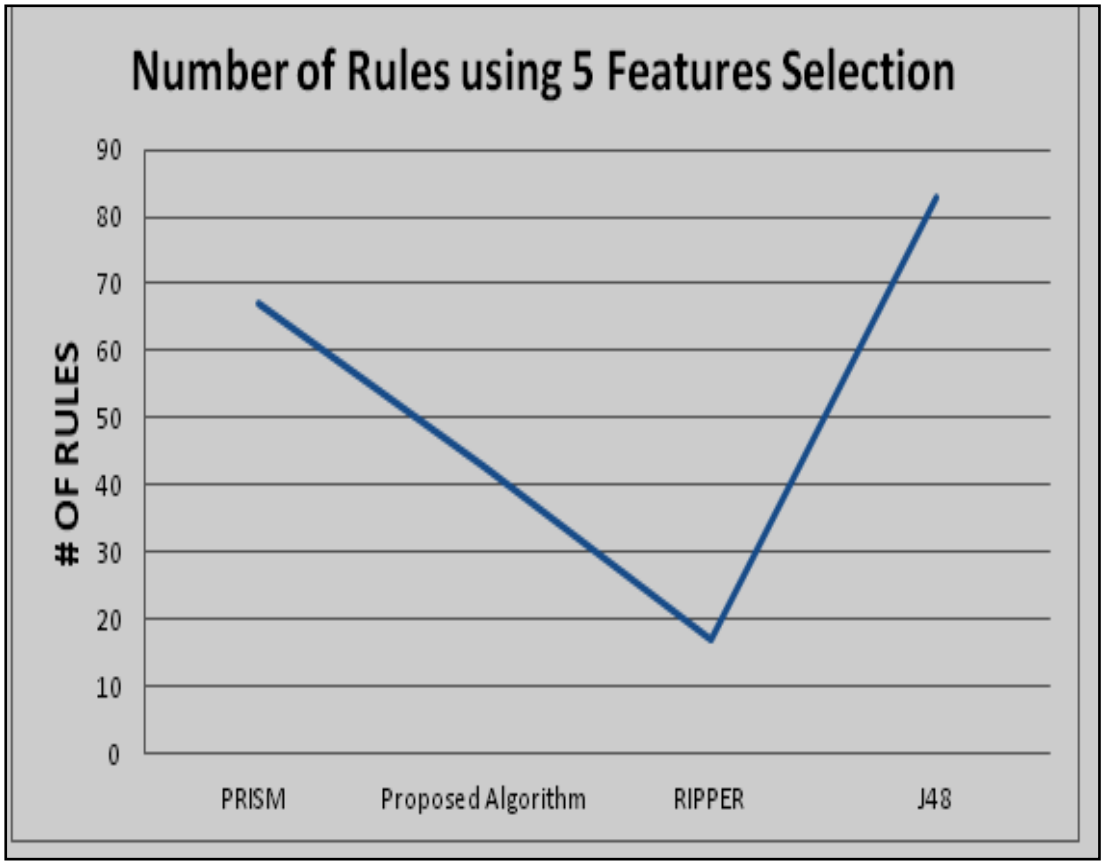


Figure 3.11: The number of rules by the proposed algorithm and other classification algorithms using the significant feature against autos data.

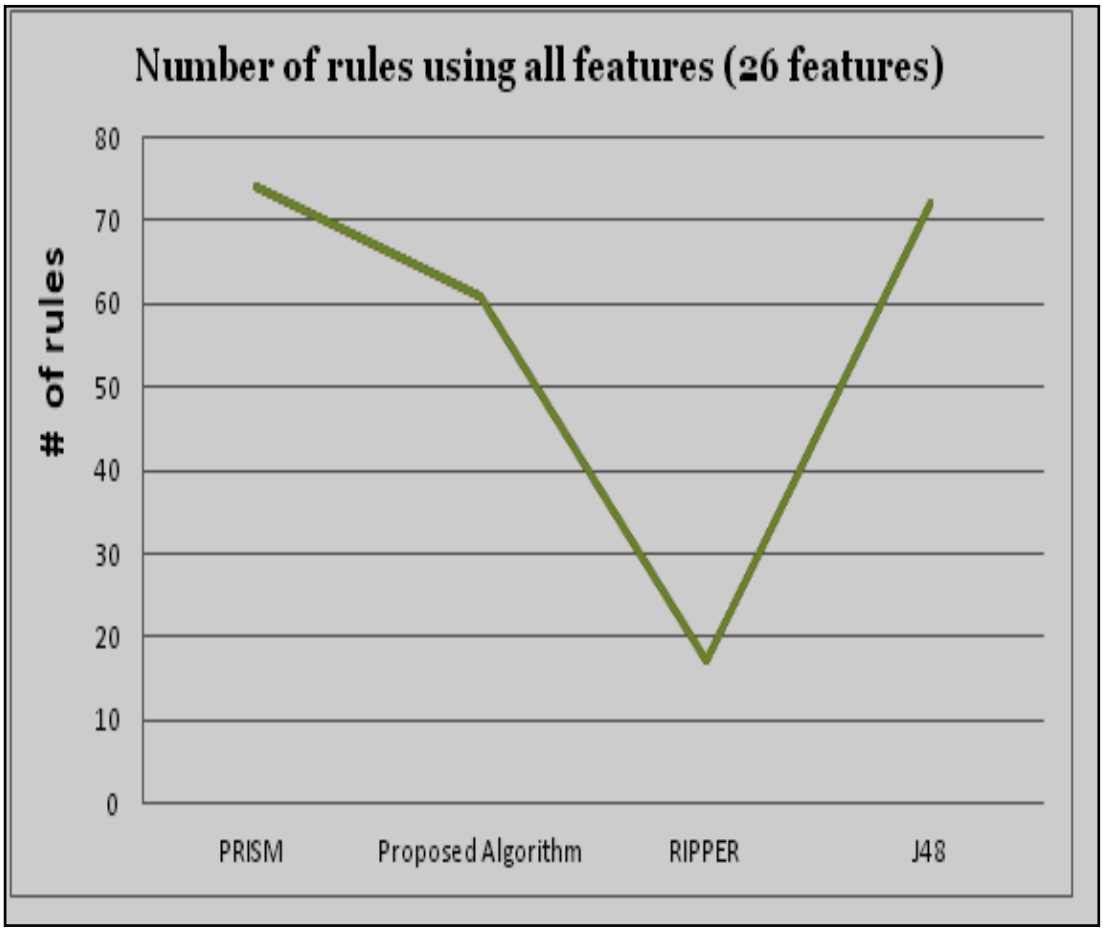


Figure 3.12: The number of rules by the proposed algorithm and other classification algorithms using complete feature against autos data set.

Overall, our proposed algorithm can make balance with respect to the size of the classifiers, resulting in near to perfect rules. The goal of our thesis is achieved through producing a new algorithm based on Strength Threshold and a new pruning method based Chi-Square. By Using Strength threshold values, additional knowledge is gained. This has enhanced PRISM rule learning. Since a rule must satisfies user strength to be produced. However these additional near perfect rules that satisfied the user threshold in some competitive environment such as in automobile fraud detection problem are considered very significant for end user, since they denotes useful knowledge used in prediction.

3.4 Chapter Summary

In this chapter, we proposed a new algorithm called (STBCP), in order to investigate the applicability of strength threshold based covering method on the problem of detection the accident type in order to make balance in producing the number of generated rules without affecting on the accuracy, we notice that (STBCP) produced not only perfect rules (100% strength) but near perfect ones as well. This has enhanced PRISM rule learning. The new Strength plays crucial role to find out the hidden correlation rules that are eliminated by classic covering algorithm such as PRISM and RIPPER. In addition, (STBCP) algorithm utilized anew a pruning method based Chi-square testing to build the classifier by discarding any useless rules. Also a new prediction procedure was produced in order to predict the class "label" of test case instances. We used WEKA as a tool to prepare the autos data set and to determine the significant features related to the fraudulent cases. This is achieved by applied descretization technique and Chi-Squares scoring method using WEKA. Since the experimental results of PRISM, RIPPER and J48 DT were generated from WEKA intelligent tool as open source in java, whereas our algorithm using java, with the same machine (computer).

We evaluated our algorithm and other compared algorithms (PRISM, RIPPER and J48 DT) on the number of rules and the average accuracy against autos data set with significant and complete features. The experimental results found that (STBCP) algorithm that used different threshold values (2-6%), produced the highest accuracy than PRISM, RIPPER and J48 DT). In addition the number of rules (classifiers) that generated by (STBCP) algorithm were on average, not in larg nor in small numbers of rules, but it make balance between them, in order the user can maintain and understand those rules with easy way and high interpretability.

Chapter Four: Conclusions and Future Works

4.1 Conclusions

The key aim of this thesis is to investigate the applicability of STBCP algorithm on the problem of detection of the fraudulent and legitimate cases of automobile insurance in order to make balance in producing the number of generated rules without impacting on the classification rate. We used this algorithm in order to classify and predict the accidents cases either to fraud or legitimate using significant features related to the "autos" data set. The outcome of this research is a new rule based classification algorithm named STBCP. Since this algorithm produced not only perfect rules (100% strength) but near perfect ones as well.

.

Therefore, the main contributions of this thesis is to achieve this balance,so we propose: A) A new learning procedure in which the rules that have strength greater than or equal to the user initial strength value are generated. B) A new rule pruning based *Chi-Square* testing that decreases the size of the classifiers by prune useless rules (negative correlated). C) A new prediction procedure to predict the class of unseen test data instances, especially in partially matches between test case and the rule in the classifier set by calculated the weight of each rule in the classifier set (More details in Chapter 3).

The results showed that: *(STBCP)* algorithm that used different initial strength values (2%, 3%, 4%, 5%, and 6%) produced - not only in *average* (4% strength value) but also with using different strength values (2-5%) - higher accurate classifiers than *PRISM, RIPPER* and *J.48* decision tree algorithms with/without relevant features of "autos" data set. Since PRISM algorithm slightly outperformed *(STBCP)* when the user inputted 6% strength value against complete and relevant features of "autos" data set. Also, *STBCP* algorithm produces neither in large nor in small numbers of rules (classifiers), but it make balance between them. This is achieved by STBCP algorithm in order to allow end user and decision makers to understand and maintain them easily.

4.2 Future Works

Most of current rule based classification data mining and machine learning algorithms utilized the pre-processing step before the discovery of rules (learning step). In pre-processing steps, continuous attributes are converted to categorical attributes (discretisation process). This is achieved before the learning step in order to prepare the data collection (data sets). Therefore, continuous attributes may build and converted during the learning step rather than in pre-processing step which have not carefully studied. We believe that handling the continuous attributes during the learning of rules become a challenging problem in data mining and machine learning algorithms.

Furthermore, most of data mining classification algorithms generated rules one by one based on their criterion and the algorithms nature in generating these rules. And thus, it required time in generating these rules, especially when these rules utilized for building the classifier, and the classifier used in prediction unseen test case. Since if these algorithms produced more than rule which have the same (confidence). The problem is: what are the best mechanisms to select these rules? Why choose random rule?.

We believe that toward to parallel generating these rules are best mechanisms to handle this problem in order to decrease the time, and thus led to increasing the efficiency and accuracy of these algorithms. Therefore, this problem which has not carefully studied by the researchers in classification data mining algorithms.

References

Agrawal. R and Srikant R. (1994). Fast algorithms for mining association rule. Proceedings of the 20th International Conference on Very Large Data Bases (pp. 487-499), Santiago, Chile.

Arora.R and Suman (2012). Comparative Analysis of Classification Algorithms on Different Datasets Using WEKA. International Journal of Computer Applications, 54(13), 21-25. doi:10.5120/8626-2492

Artıs.M., Ayuso.M., Guillen. M., 2002. Detection of automobile insurance fraud with discrete choice models and misclassified claims. Journal of Risk and Insurance 69 (3), 325–340.

Al Shalabi. L, Najjar.M, Al Kayed .A (2006). A framework to Deal with Missing Data in DataSets. Journal of Computer Science 2(9), 740-745. DOI: 10.3844/jcssp.2006.740.745

Basak. J., and Lim.D (2009). A Feasibility Study on Automating the Automotive Insurance Claims Processing. International Conference On Management of Data-COMAD(15th).Mysore,India.

Bermúdez.L, Pérez J.M., Ayuso M., Gómez E., Vázquez F.J. (2008) .ABayesian dichotomous, Model with asymmetric link for fraud in insurance, Insurance: Mathematics and Economics vol 42 (2) ,pages779–786 . doi:10.1016/j.insmatheco.2007.08.002

Bhowmik.R. (2011). Detecting Auto Insurance Fraud by Data Mining Techniques. Journal of Emerging Trends in Computing and Information Sciences , vol 2(4), 156-162.

Bhukya, D. P., and Ramachandram, S. (2010). Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. International Journal of Computer and Electrical Engineering, 2(4), 660-665.

Breiman L., Friedman J. H., Olshen R. A., & Stone C. J.(1984). Classification and Regression Trees, Belmont, CA: Wadsworth International Group, 1984.

Brockett P.L, Derrig .R.A., Golden. L.L, (2002) Fraud classification using principal component analysis of RIDITS, The Journal of Risk and Insurance 69 (3) 341–371.

Brockett .P.L, Xia. X, & Derrig. R.A(1998).Using Kononen's self-organizing feature map to uncover automobile bodily injury claims fraud, The Journal of Risk and Insurance, 65 (2). 245–274

Bouckaert, R., Frank, E., & Hall, M. (2010). WEKA Experiences with a Java open-source project. Journal of Machine Learning Research 11, 2533-2541. Retrieved from http://dl.acm.org/citation.cfm?id=1953016

Burges,C.J.C.(1997). A Tutorial on Support Vector Machines for Pattern Recognition, 43, 1-43

Caudill.S, Ayuso. M, and Guillen, M, (2005). Fraud detection using a multinomial logit model with missing information. Journal of Risk and Insurance 72 (4), 539–550.

Cendrowska.J.(1987). PRISM: An algorithm for inducing modular rules. International Journal of Man-Machine Studies, Volume 27, Issue 4, pages 349-370.

Christopher. J. J. (2011). A Statistical Approach for Associative Classification. European Journal of Scientific Research, 58(2), 140-147.

Cohen. W. W. (1995.) Fast effective rule induction. In the Proceeding of the 12th International Conference on Machine Learning, pp. 115-123, Morgan Kaufmann..

Cunningham. P. and Delany.S. (2007). K-Nearest neighbor classifiers. Multiple Classifier Systems, Volume: 34, issue: 8.1-17.

Derrig. R. A, and Francis, L. (2008.). Distinguishing the Forest from the TREES A Comparison of Tree- Based Data Mining Methods, Variance 2(2), 184-208.

Fayyad, U., Piatetsky-shapiro, G. and Smyth.P (1996) ‘The KDD process for extracting useful knowledge from volumes of data’. ACM, Vol. 39, Iss. 11, PP.27-34

Furnkranz. J.(1996) Separate-and-conquer rule learning. Technical Report TR-96-25, Austrian Research Institute for Artificial Intelligence, Vienna

Furnkranz. J and Widmer.G (1994): Incremental reduced error pruning (IREP),in Cohen.W and Hirsh(eds.), proceeding of the 11th international conference on machine learning (ML94),70-77,Morgan Kaufmann,1994.

Gepp. A., Wilson. J. and Kumar, K. (2012). A Comparative Analysis of Decision Trees Vis a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection. Journal of Data Science, vol. 10. no. 3, pp. 537-561 Retrieved from http://www.jds- online.com/file_download/366/JDS-1056.pdf

Han J., Kamber .M and Pie.J. (2006). Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann Publishers.

Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms, John Wiley & Sons.

Kotsiantis S, Kanellopoulos. D, and Pintelas, P. (2006). Data preprocessing for supervised leaning. International Journal of Computer Science (2), 111-117.

Li.Y, Wu.X, Wang. Y., Li. Y, Chu .C (2007). A review of data mining-based financial fraud detection research, international conference on wireless communications, Networking and Mobile Computing, 5519–5522.

Liu B., Hsu W. and Ma Y. (1998). Integrating classification and association rule mining. Proceedings of the KDD, (pp. 80-86). New York, NY

Liu, H., and Setiono, R (1995). “Chi2: Feature selection and discretization of numeric attributes”, Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, 1995, 338-391. DOI: 10.1109/TAI.1995.479783

Ngai. E. W. T., Hu. Y., Wong, Y. H., Chen, Y., and Sun. X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559-569.

Pérez, J. M., Muguerza. J., Arbelaitz.O, Gurrutxaga I, and Martín. J. I. (2005). Consolidated Tree Classifier Learning in a Car Insurance Fraud Detection Domain with Class Imbalance, Lecture Notes in Computer Science: Pattern Recognition and Data Mining; Proc. Intl. Conference on Advances in Pattern Recognition. pages 381-389.

Phua. C., Alahakoon. D., and Lee.V. (2004) Minority Report in Fraud Detection Classification of Skewed Data, SIGKDD Explorations 6(1), 50-59.

Phua C., Lee.V, Smith.K, Gayler.R. (2005). A comprehensive survey of data mining-based fraud detection research, Artificial Intelligence Review 1–14.

Pinquet, J., Ayuso, M., Guillen, M., 2007. Selection bias and auditing policies for insurance claims. . Journal of Risk & Insurance, 74(2), 425-440. doi:10.1111/j.1539-6975.2007.00219.x

Quinlan. J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Quinlan. J.R. (1987) Generating production rules from decision trees. In Proceedings of the l0th Int. Joint Conferences on Artificial Intelligence, pp. 304-307, Morgan Kaufmann

Quinlan. J. R. (1986). Induction of Decision Trees. Machine Learning. 1(1), 81-106. DOI: 10.1080/10629360600933723.

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., and Cela-Díaz, F. (2010). Statistical Methods for Fighting Financial Crimes. Technometrics, 52(1), 5-19. doi:10.1198/TECH.2010.07032 .

Suryakumar. D., Sung. A. and Liu. Q. (2012). Critical Dimension in Data Mining. eKNOW 2012, The Fourth International Conference on Information, Process, and Knowledge Management 97-100

Tennyson, S., and Salsas-forn, P. (2002). Claims auditing in automobile insurance: detection and deterrence objectives, The Journal of Risk and Insurance 69 (3) 289–308.

Thabtah.F, Cowling.P, and Peng.Y. (2005). MCAR: multi-class classification based on association rule. In: Proceedings of the 3rd ACS/IEEE International Conference on Computer Systems and Applications, Cairo, Egypt, January. doi:10.1109/AICCSA.2005.1387030

Van Rijsbergan C. (1979). Information retrieval, Buttersmiths, London, 2nd Edition.

Viaene. S., Ayuso. M., Guillen. M., Gheel . D. Van, Dedene. G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research 176 (1). 565-583. doi:10.1016/j.ejor.2005.08.005

Viaene. S, Derrig R.A, Baesens.B, and Dedene.G (2002). A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection, The Journal of Risk and Insurance 69 (3) ,373–421.

Viaene, S., Dedene, G., & Derrig, R. (2005). Auto claim fraud detection using Bayesian learning neural networks. Expert Systems with Applications, 29(3), 653-666. doi:10.1016/j.eswa.2005.04.030

eneaiV. S., Derrig. R. ,Dedene. G. (2004). A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis. IEEE Transactions on Knowledge and Data Engineering 16(5): 612-620.

Weisberg.H. I, Derrig. R. A. (1998). Quantitative Methods For Detecting Fraudulent Automobile Bodily Injury Claims, Risques 35 . 75–101

Weisberg.H. I, Derrig. R. A. (1991) , Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Claims in Massachusetts," Journal of Insurance Regulation, 9: 427-541

Wilson, J. (2009). An Analytical Approach to Detecting Insurance Fraud Using Logistic Regression. Journal of Finance and Accountancy, vol(1). pages 1-15.

Witten I. and Frank E. (2005). Data mining: practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann.

Yin .X. and Han. J. (2003). CPAR: Classification based on predictive association rule. Proceedings of the SDM (pp. 369-376). San Francisco, CA.

ملخص

احتيال تأمين المركبات (AIF) هو مشكله مهمه لكل من حاملي الوثيقه وشركات التأمين. الفعاليات المزوره ممكن تؤثر سلبا على أرباح شركات تأمين المركبات. التنقيب عن البيانات (Data Mining) خصوصا خوارزميات التصنيف (Classification Algorithms) المبنية على القواعد (Based Rules) ممكن تساهم في المساعدة في كشف الفعاليات المزوره. في هذ ه الخوارزميات فان الناتج تمثل معرفه مفسره بسيطه اذا-ثم (If-Then) وتخزن في قاعدة المعرفة. ومع ذلك فان مشكلة خوارزميات التصنيف المبنيه على القواعد مثل (PRISM) تولد عدد كبير من القواعد حيث فهم وصيانة (الحفاظ على) تلك المصنفات (Classifiers) يعتمد على حجم المصنفات وهذا صعب بالنسبة للمستخدم العادي. علاوة على ذلك بعض القواعد المترابطه في (PRISM) والتي قريبه من الكمال لا يتم استخراجها. اختفاء هذه القواعد في البيئات التنافسية تعتبر مهمة جدا في مرحلة التنبؤ. من ناحية أخرى, خوارزمية استقراء المبني على القواعد Induction Rule Based Algorithm أي التقليم المتزايد المتتكرره لانتاج الحد من الخطأ (RIPPER) تمتلك حجم صغير وغالبا دقه منخفضه. هذه القواعد ليس مقبوله فيما يتعلق بمشكلة التصنيف في (AIF) بسبب بعض من المعرفه لم تكتشف. هذة الاطروحه تتحقق من قدرة تطبيق قوة عتبة المبنيه على طرق التغطيه (Strength Threshold Based Covering Method) على مشكلة اكتشاف حالات الحادث من اجل عمل توازن في عدد القواعد المتولدة بدون تأثير على معدل التصنيف (Classification Rate). الخوارزميه الجديده تدعى قوة العتبة المبنيه على تغطيه (PRISM) (Strength Threshold Based Coverage Prism) اختصارها (STBCP) والتي تعمل توازن (كنتيجه حجم مصنفات متوسط) (As a Result Average Size Classifiers) في انتاج القواعد. ويتم إنجاز هذا التوازن من خلال إنتاج خوارزمية جديدة للتصنيف المبني على القواعد (STBCP) والتي استخدمت اجراءات تعليم وتقليم وتنبؤ جديده (New Learning, Pruning, Prediction Procedures) استنادا إلى القيم قوة العتبه المختلفة (2%، 3%، 4%، 5%، و6%)اتجاه (ضد) بيانات "السيارات" باستخدام كامل واهم الخصائص .استنادا على قيم قوة العتبه المختلفه (2% - 6%) ، النتائج التجريبيه وجدت ان خوارزمية (STBCP) انتجت اعلى دقه مقارنه بخوارزميات (RIPPER ,PRISM و شجرة القرار جي 48 J48DECSION TREE) اخترنا (4٪ كمتوسط للقيم العتبة) ووجدنا أنه STBCP خوارزمية تنتج أعلى دقة مقارنة مع PRISM، RIPPER والخوارزميات J.48 شجرة القرارات. بشكل عام، تنتج الخوارزمية STBCP ليس في اعداد كبيرة ولا في أعداد صغيرة من القواعد (المصنفات)، لكنها تعمل التوازن بينهما (كنتيجة لحجم متوسط). هذه تسمح للمستخدم النهائي وصناع القرار للحفاظ وفهم القواعد المنتجة مع تمثيل واضح دون التأثير على معدل تصنيف (الدقة).

خوارزمية جديدة للتصنيف المبني على القواعد في اكتشاف احتيال التأمين على المركبات

من قبل

احمد عقلة علي العلي

بإشراف

د.فادي فايز

قدمت هذه الرسالة إستكمالا لمتطلبات الحصول على درجة الماجستير في علم الحاسوب

عمادة البحث العلمي والدرسات العليا

جامعة فيلادلفيا

كانون الثاني 2013