A Comparative Study to Predict Insurance Fraud Detection using Data Mining Algorithms

[1]Betelhem Zewdu

Abstract-- The insurance industry has been factually a growing industry. It plays an important role of insuring the economic aspects of a county. The annual income of the insurance company is increase time to time but as the same interval, the scale of insurance fraud also increase and it's difficult to face fraud problem. Workmen's compensation insurance fraud occurs when someone knowingly, with intent to cheat, makes a false, material statement or it may employees exaggerate or even fabricate injuries. Data Mining has been widely used to predict insurance fraud. It is a technique that used to extract hidden patterns from large amount of data set and, used to for make decision. In this paper, it is conducted to a comparative study for different algorithms to select the final best performance accuracy of the insurance fraud detection in the case of workmen compensation. For this purpose, four classification algorithms such as J48, Jrip, PART and Naïve Bayes were implemented and compared to know the best prediction result. The outcome showed that J48 decision tree is outperforming in the particular case since it has higher accuracy rate of 96.81% in 90% percentage split test option.

Key words- comparative, data mining, fraud detection, performance, prediction

1. Introduction

The insurance industry has been factually a growing industry. It plays an important role in ensuring the economic aspects of a county. Insurance is an agreement or policy in which an individual or entity receives financial protection or compensation against losses from an insurance company. It is also a way of managing risks. However, frauds occur many faces in workmen's compensation insurance. People who exaggerate or misrepresent a claim or submit a false claim are committing fraud. It may also health care providers who submit false medical reports or inflated bills are committing fraud [1]. Insurance companies lose a lot of money through fraudulent claims each year [2]. Researches indicates that the annual insurance fraud cost for the property and casualty insurance industry is over 25 billion dollars. From this Workers' compensation insurance alone accounts for a sizable portion of this total cost [3]. Now a day's insurance company's store large amount of data in their databases. But the stored data's is not indicates their customers behavior and preferences. So it needs to extract useful information from such evenly growing data. Therefore, data mining is an important tool to extract useful information from such databases.

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns of data [4]. There are many data mining techniques, methods and tools to extract a useful information from a large dataset. From this the one data mining techniques is classification. It is one of the most commonly used techniques, to develop models that can datasets as large. And it maps data into their predefined groups or segments [5]. It is commonly used in insurance fraud detection. In this research work, J48, JRip, PART and Naïve Bayes classification algorithms that are used for insurance fraud detection. This current study were compare these algorithms performance and to select the best performance in workmen's insurance fraud detection.

2. Literature Review

2.1. Naïve Bayes Algorithm

Naïve Bayesian Classifier is a machine-learning algorithm that maps (classifies) a data set into one of several predefined classes. It attempts to maximize the posterior probability in determining the class. Observations show that Naïve Bayes performs consistently before and after reduction of number of attributes [6].

The Naive Bayes algorithm can be used for both binary and multiclass classification problems [7]. It builds a model to classify new cases based on observed probabilities and supporting evidence from the training data. Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks [8].

[1] *Betelhem Zewdu is currently pursuing masters degree in Information Technology, Wachemo University, Ethiopia, +251 912747335. Email zewdubetty9@gmail.com*

Bayes rule is a way to go from P(X|Y) to find P(Y|X)

$$P(X|Y=\frac{P(X \text{ n } Y)}{P(Y)}$$ P(Evidence | Outcome)(known from training dat

$$P(Y|X) = \frac{P(X \text{ n } Y)}{P(X)}$$ P(Outcome | Evidence)(to be predicted to test

$$\text{Bayes Rule} \quad P(Y|X)=\frac{P(X|Y)\, P(Y)}{P(X)} \qquad (1)$$

2.2. JRip Rule Based Algorithm

The basic form of a rule is the following: if<conditions> then <conclusion>. Where <conditions> represents the situations of a rule, whereas <conclusion> represents its conclusion. The conditions of a rule are connected between each other with logical connectives such as AND, OR, NOT, etc., thus forming a logical function [9]. When sufficient conditions of a rule are satisfied, the conclusion is derived and the rule is trigger. Rules represent general knowledge regarding a domain.

JRip implements a propositional rule learner. A Repeated Incremental Pruning to Produce Error Reduction (RIPPER). It is an inference and rules-based learner (RIPPER) that can be used to classify elements with propositional rule [10]. The RIPPER algorithm is a direct method used to extract rules directly from the data. JRip (Weka's implementation of the RIPPER rule learner) is a fast algorithm for learning condition-conclusion rules. Like decision trees rule learning algorithms are popular because the knowledge representation is very easy to interpret.

2.3. PART Rule Based Algorithm

PART as a rule based algorithm that produces a set of if-then rules that can be used to classify data. PART is a modification of C4.5 and RIPPER algorithms and draws strategies from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5. PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as a new rule. Though unlike that of C4.5 the trees built by PART for each rules are partial and incomplete, PART is advantageous because of its simplicity and its capability of generating sufficiently strong rules [11].

2.4. J48 Decision Tree Algorithm

Decision Tree Classifier is a simple and widely used classification technique to solve the classification problem. Decision trees are machine learning techniques that express independent attributes and a dependent attribute in a tree-shaped structure, that represents a set of decisions for inductive inference over supervised data [12]. A decision tree is a procedure for classifying categorical data based on their attributes. It is also efficient for processing large amount of data, so is often used in data mining application.

Decision tree algorithms such as ID3, J48 and NB Tree can be applied on a large amount of data, and produced valuable predictions to evaluate future behavior of a problem. Decision tree are preferred because they can evaluate information more accurately than other methods [13].

3. Methodology

Methodology is a step or procedures that a research work follows from its inception to a conclusion, to achieve a specific tasks or goals. To do this research work, it uses red flag variables to scores the claims, which provides an indication of a claim being fraud or not and a hybrid model is implemented. To detect insurance fraud the research has follow a clustering techniques then followed by a classification technique. The Clustering algorithms used to discover the natural grouping of insurance claims as fraud and non-fraud. For this purpose K-mean clustering algorithm is used. Then followed by a classification technique, which helps to predict insurance fraud suspicious claims. For this case, J48, JRip, PART and Naïve Bayes algorithms are used, to predict the insurance fraud. Finally the study compares the performance of the algorithm that predict fraud.

The data that used for this research work is get from Ethiopia Insurance Corporation main branch in Addis Ababa. There are different insurances offered by the company, from this, the study is considered in workmen's compensation insurance. The original dataset which gets from the INSIS database is 17296 records and 27 attributes. This study comprises three steps to preprocess the data that is to comfortable for the selected technique. The first step is data cleaning, the second is attribute selection, and the last one is derived a new attribute which, is used to necessary to identify fraud suspicious claims. Data cleaning is an important thing in a data mining research to get a good result and it perform before analysis like to identify missing value, noisy and inconsistent data. The whole dataset that get from the insurance company is not important for fraud detection. So, the next step is to select the important attributes that necessary to detect fraud. For this purpose, gain ration attribute selection method is used. Finally, derived a new attribute from the existing data is processed. CLAIM_REPORT_LENGTH_ DATE (the length of the accident report date since its occurrence) is derived from the EVENT DATE and NOTIFICATION DATE columns of the existing data set.

CLAIM_REPORT_LENGTH_DATE=EVENT_DATE-NOTIFICATION_DATE

The final selected attribute for the study is CLAIM_REPORT_LENGTH_ DATE, claim office, risk type, claim state, cover type, notification date, event place, sales channel, object type, profession and claim amount. After the pre-processing step is performed, the total records in the study is 17275 records and 12 attributes with the final class are used for classification purpose.

4. Result and Discussion

As the nature of dataset that get from the insurance company is not divided as fraud and non-fraud, it is necessary to develop a clustering technique. To implement this technique, a k means algorithm is used to set a different value for the distance function, seed value and 2 as a number of cluster. After this the first group contains the claim that the claim report length date is slow, the claim amount is high, the notification date is Monday and the claim state annulled is considered as a fraud suspicious claims. On the other hand, the claim report length date is fast, the claim amount is low, the notification date is Tuesday and the claim state paid is considered as a non-fraud suspicious claims. Next to this, to develop a classification model it uses the output of clustering model. J48, JRip, PART and Naïve Bayes algorithms are used for classify insurance claims as fraud and non-fraud for their predefined classes. In order to classify the records, 10-fold cross-validation and the percentage split test options are used. The following table table1 shows, the correctly and incorrectly classified insurance claim instance with their performance is described.

Table 1 the comparison of different algorithms in 10 fold cross validation and 90% percentage split

Algorithm	Correctly Classified		Incorrectly Classified		Time taken /sec	Mean Absolute Error	Root Mean Square error
	In %	Instance	In %	Instance			
J48	96.27	16631	3.72	644	0.07 sec	0.05	0.181
JRip	96.15	16661	3.84	664	40.78 sec	0.06	0.184
PART	94.8	16387	5.14	888	0.72 sec	0.06	0.19
Naïve Bayes	89.78	15511	10.21	1764	0.02 sec	0.14	0.26
			10 fold cross validation				
Algorithm	Correctly Classified		Incorrectly Classified		Time taken /sec	Mean Absolute Error	Root Mean Square error
	In %	Instance	In %	Instance			
J48	96.81	1672	3.18	55	0 sec	0.054	0.172
JRip	96.35	1664	3.64	63	0 sec	0.057	0.183
PART	94.90	1639	5.09	88	0.02 sec	0.064	0.199
Naïve Bayes	89.57	1547	10.42	180	0 sec	0.147	0.270
			90% percentage split test option				

As seen in the above table, 10 fold cross validation and 90% percentage split and test options are described. And it were done a different experiments in percentage split like the default 66%, 70%, 80%, and 90%. But the one which is a higher accuracy is described in this paper i.e. 90% percentage split. So from the experimental result, as it can be seen in the table, J48 decision tree algorithm in a 90 % percentage split is the highest accuracy of the prediction model than the other algorithms. It scores an accuracy of 96.81%, which is a highest value and the final selected algorithm for the study.

The following figure fig1 shows the comparison of different algorithms in 90% percentage split.

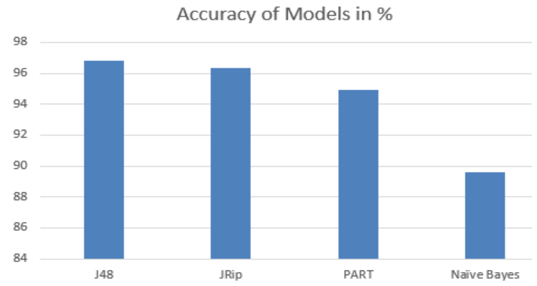

Fig 1 the comparison of algorithms in 90% percentage split

The following table table 2, shows that the confusion matrix of the final selected algorithm which is J48 in 90% percentage split is described. From the total testing datasets of 1727 records, 1672 records are correctly classified, while 55 records are incorrectly classified. Which means that, from the total 443 fraud suspicious claim 410 instance is correctly classified as fraud suspicious and 33 instances is incorrectly classified as non- fraud suspicious. On the other hand, from the total 1284 non-fraud suspicious claims, 1262 instances are correctly classified as non-fraud suspicious and 22 instances is incorrectly classified as fraud suspicious claims.

Table 2 the confusion matrix of the final selected algorithm J48 in 90% percentage split

Actual	Predicted		Total
	Cluster 0 (fraud)	Cluster 1 (non-fraud)	
Cluster 0 (fraud)	410	33	443
Cluster 1 (non-fraud)	22	1262	1284
Total	432	1295	1727

For this study, a number of the experiment was done but it present only the finally selected ones in the above table table 2. In addition to this, to know the best algorithm for the given case, it is possible to compare the mean absolute error and root mean square error. So as it can be seen from the above table table 1, the mean absolute error and root mean square error is lower in J48 algorithm with 90% percentage split test option than the others. Therefore, it is credible to conclude that J48 decision tree algorithm is the best prediction performance than the other for in workmen's compensation insurance fraud detection.

5. Conclusion

As a conclusion insurance companies store large amount of data from their customers in the day to day activities. These data are invaluable information about their customers'. However, to extract valuable information from this enormous data it takes time and manpower. So, it needs an advanced technique and tool to extract meaningful patterns and rules from this data. Data Mining is a powerful techniques have been applied for workmen's insurance fraud detection. In the study, a clustering technique is followed by a classification technique to detect and predict fraud. Then, J48 decision tree is the best algorithm to predict insurance fraud whether a given claim is fraud or non-fraud in workmen's compensation. It scores an accuracy of 96.81% in 90% percentage split test option. In addition to this, it is proved that when the number of testing set is increase the performance of the algorithm is also increase.

References

[1] "Workers Compenssation Board," 2016. [Online]. Available: http://www.wcb.ny.gov/content/main/TheBoard/WhatIsFraud.pdf. [Accessed 31 March 2017].

[2] "IBM Corporation," *Using Data Mining to Detect Insurance Fraud,* 2010.

[3] T. J. Woodfield, "Predicting Workers' Compensation Insurance Fraud Using SAS® Enterprise Miner™ 5.1 and SAS® Text Miner SAS Institute Inc.," 2005.

[4] P. K. Krishna, DATABASE MANAGEMENT SYSTEM ORACLE SQL AND PL/SQL., 2013.

[5] A. B. Devale and D. R. V. Kulkarni, ""Applications of Data Mining Techniques in Life Insurance"," *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.2, No.4,* pp. 31-40, 2012.

[6] "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm," *International Journal of Engineering Science and Technology,* 2010.

[7] A. Tariku, "Mining insurance data for fraud detection: the case of Africa insurance share company," *Master Thesis Addis Abeba University,* june 2011.

[8] S. Prabhakaran, "Machine Learning Plus," [Online]. Available: https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/. [Accessed 22 July 2020].

[9] J. Prentzas and I. Hatzilygeroudis, "Categorizing approaches combining rule-based and case-based reasoning," *Journal Compilation,* vol. 24, no. 2, pp. 97-120, May 2007.

[10] Veeralakshmi and Ramyachitra, "Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset," *International Journal of Computer Science Engineering (IJCSE),* vol. 4, no. 03, pp. 79-85, 03 May 2015.

[11] A. Chali, "An Integration of Prediction Model with Knowledge Base System for Motor Insurance Fraud Detection: The Case of Awash Insurance Company S.C," February 2016.

[12] R. Bhowmik, "Data Mining Techniques in Fraud Detection," *Journal of Digital Forensics, Security and Law,,* vol. 3(2), pp. 35-54, 2010.

[13] P. Rekha and P. Saurabh, "Application of Data Mining Techniques in Health Fraud Detection," *International Journal of Engineering Research and General Science,* vol. 3, no. 5, 2015.