International Journal of Computer Architecture and Mobility (ISSN 2319-9229)
Volume 7 -Issue 7, July 2019

Multi-Level Clustering Machine Learning Algorithm (Mlcmla) For Fraud Detection in Banking Transactions

Sanjana Kukreja, Shweta Negi, Laya Gyanee

Abstract: Fraud detection methods are continuously developed to defend criminals. They allow us to identify quickly and easily the frauds. In this work, we will focus on the problem of fraud detection in banking transactions. A single algorithm may not be suitable for every problem. Therefore, selecting an algorithm that performs best in a given situation is very crucial the number of frauds in daily life is increasing in sectors like banking, finance, and government. Accurate detection of fraud is a challenge. Data mining techniques help in anticipation and detection of fraud. Data mining tools can be used to spot patterns and detect fraud transactions. Through data mining, factors leading to fraud can be determined. The performance is analysed based on the parameters of the Total Running Time and the Accuracy. The results proved that the multi-level clustering machine learning algorithm gave the best results and the simple anomaly detection algorithm gave the worst results.

Keywords : Data mining techniques , Fraud detection , machine learning algorithm , multi-level clustering.

I. INTRODUCTION

Financial fraud could be a growing concern with way reaching consequences within the government, company organizations, and finance trade. In Today's world high dependency on web technology has enjoyed magnified banking transactions. However, banking sector fraud had conjointly accelerated as online and offline transaction. As transactions become widespread mode of payment, focus has been given to recent procedure methodologies to handle the fraud downside. There are many fraud detections solutions and software system which prevents frauds in businesses like credit card, retail, e-commerce, insurance and industries. Data mining technique is one notable and common methods employed in determination banking sector fraud detection downside. It's not possible to be sheer sure concerning actuality intention associated right behind an application or transaction. In reality, to hunt out doable evidences of fraud from the accessible data using mathematical algorithms is the best effective possibility. Fraud Detection Data Mining Techniques and Models can be found useful in the banking sector, mostly for the purpose of detection of fraud happening through the various techniques used by the fraudsters. By using all the different types of data mining models and techniques, the ever-increasing fraudulent activities, which are of a major concern for the business as well as the customers and banks, happening can be detected and also reported. Although, there are two processes through which the patterns of the frauds can be detected by the help of data mining [5]. The first process is the one in which the bank approaches different data warehouses which contain transaction information and implements its data mining codes to determine the frauds in it. Then, they can refer these patterns with their own collection of information of how frauds take place and determine the amount of trouble in it. Whereas, in the second procedure, the determination of the fraud pattern is done on the own personal information of the banks. The method which most of the banks use is the "Hybrid" approach to detect a fraud.

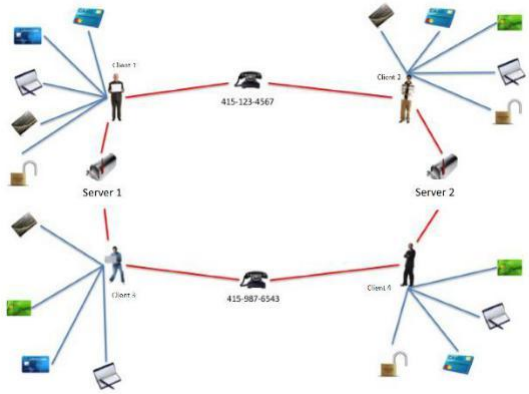

Figure 1: Financial fraud detection process

Data Mining is not only the single factor which will facilitate the banking systems to achieve new customers, but also it can facilitate them to retain their existing customers. Customer accession and retentiveness are vital issues for any business, particularly the sector. Now-a-days, customers have such a lot of opinions with relevance where they will

prefer to do their business. Executives within their banking system, therefore, should remember that if the bank employees are not giving every customer their full recognitions, the client will just search for the other bank that may notice. Data mining is also used to facilitate in marking new clients for merchandise and all the service and in finding or exploring a client's precious buying patterns in order that the banks are going to be ready to keep old clients by providing reason on individual basis customized to every client's requirements. Considering financial fraud bring huge property damage to investors, a large number of researches have been conducted on the this area using machine learning methods such as ANN[1-3] ,DT[4] , SVM[5, 6], and text mining[7]. Meanwhile, other fraud problems like credit card fraud[8, 9], internal transaction fraud[10] and insurance fraud[11] have also been investigated. Given the different characteristics of each type of financial fraud, specific methods have been developed[12]. This paper puts forward a hybrid detection model for financial statement fraud, and this model have the advantages of (1) combining the financial and non-financial data, (2) using two feature selection methods, and (3) easy to explain. The rest of the paper is organized as follows. The first section is the state of the art, we present a review of various works that focus on using machine learning techniques for fraud detection.. In Section III, we describe the four machine learning algorithms used in our experimental study which are the Simple Anomaly detection algorithm, multi-level clustering machine learning algorithm for fraud detection in banking transactions. Section IV is devoted to experimental evaluation and results. Finally in section V, we present concluding remarks and perspectives.

II. FRAUD DETECTION TECHNIQUES

Fraud detection is a technique which is used to find the fraud. Fraud detection techniques helps organizations simpler and with efficiency by facultative them to run deep analyses on their complete information [5][2]. Fraud detection techniques help in investigation departments like police department, CBI and other investigation departments. The detection scheme gives added value because insight in trends and developments, enabling the organization to react quickly and effectively. This makes tracking down criminals more efficient and more focused [1]. Advanced data analysis technologies create finding relationships, patterns and trends a fast and simple job, this provides users additional insight and permits them to from higher forecasts. The foremost necessary functionalities in fraud Detective are: predicting

analysis, clustering techniques, discovery relationships, profiling, association techniques, network analysis, fuzzy, matching, making graphs, making maps shaping picks and making cross tables [5][3][2]. Fraud Detection actively supports the user in applying the inbuilt data processing techniques, thereby replacement technology-oriented work by task-oriented work. Applicable techniques square measure chosen and designed during a approach that most accurately fits the info to be analyzed [5][4]. This makes the software system accessible for an outsized cluster of individuals and ensures economical application of information mining. The present fraud detection system needs more domain knowledge to execute the system and analysis process. In this paper, expectation of present detection system needs more domain knowledge from the analysis part. These issues make the current system to bigger gap during fraud detection process. This issue is addressed in this proposed paper using user friendly framework with less domain knowledge [5][1][3]. With this proposed system, any user with partial domain knowledge or new user can also able to execute the proposed system easily to find variety of patterns from the user side and previous history of other users using intelligent agent.

III. RELATED WORK

This research work has been initiated with the literature study. After reviewing relevant literature, we get to know about research efforts made to overcome the targeted problem and to determine what all data mining techniques have been applied to achieve high accuracy in insurance fraud detection of health care data. In the past, people used to use expert analysis to find fraududent financial statements. In this way people may not fully analyzed the report data for its huge amount and wild range, which caused many shortcomings in judgment. In recent years, data mining method has been widely used in fraud detection to reduce the errors caused by experts' judgment, including Internet fraud detection

Yao, J. et al[1] Financial statement fraud has been a difficult problem for both the public and government regulators, so various data mining methods have been used for financial statement fraud detection to provide decision support for stakeholders. The purpose of this study is to propose an optimized financial fraud detection model combining feature selection and machine learning classification. The study indicated that random forest outperformed the other four methods. As to two feature selection methods, Xgboost performed better. And according to our research, 2 or 5 variables are more acceptable for models in this paper.

Panigrahi, P. K et al[2] The proposed framework provides a systematic process for the auditors in discovering internal financial frauds. The auditors can use their own experience and investigation skills

and integrate with tools and techniques available in different software. The suggested data structures of fraudulent transactions assist the auditors in preparing the data for application of various techniques using software.

Chen, Y.-J. et al[3] This study considers the characteristics of variety and value of big data used in finance and economics to develop a big data-based fraud detection approach for the financial statements of business groups to more precisely detect the financial statement fraud of business groups, and thus reducing investment losses and risks and enhancing investment decisionmaking benefits for investors and creditors.

Rawte, V., et al[4] Fraud is widespread and very costly to the healthcare insurance system. Fraud involves intentional deception or misrepresentation intended to result in an unauthorized benefit. It is shocking because the incidence of health insurance fraud keeps increasing every year. In order to detect and avoid the fraud, data mining techniques are applied. This includes some preliminary knowledge of health care system and its fraudulent behaviors, analysis of the characteristics of health care insurance data. Data mining which is divided into two learning techniques viz., supervised and unsupervised is employed to detect fraudulent claims. But, since each of the above techniques has its own set of advantages and disadvantages, by combining the advantages of both the techniques, a novel hybrid approach for detecting fraudulent claims in health insurance industry is proposed.

Verma, A., et al[5] Fraudulent insurance claims increase the burden on society. Frauds in health care systems have not only led to additional expenses but also degrade the quality and care which should be provided to patients. Insurance fraud detection is quite subjective in nature and is fettered with societal need. This empirical study aims to identify and gauge the frauds in health insurance data. The contribution of this insurance claim fraud detection experimental study untangle the fraud identification frequent patterns underlying in the insurance claim data using rule based pattern mining.

Jayabrabu, R et al[6] Frauds are happened based on instance or incidents, but they are repeated offences using some methods (old and new), instances are more similar in content and appearance but they are non - identical while comparing. Fraud deduction is one of difficult process not only technology, but also in crime investigations.

Lucas, Y. et al[7]In this paper we propose a strategy to quantify the covariate shift in a temporal dataset.

This strategy consists in classifying the transactions of each day against every other days: If the classification is efficient then the days are different and there is a covariate shift between them. On the other hand, if the classification is not efficient, the days are similar. This strategy allows us to build a distance matrix characterizing the covariate shift between days.

IV. PROPOSED METHODOLOGY

Firstly, to identify period based claim anomalies we analyse the statistical decision rules to detect the short-term outliers which helps in detecting the frauds. Then k-means clustering is used to group similar patterns in one cluster and rest in other respectively as clustering simplify the fraud detection process because clusters having more chances of mismatch with the standard policies would have to be checked rather than comparing all records i.e. to facilitate the mining process. Then elbow test is applied to improve the performance of k-means clustering as it helps in computing the optimal value of k for existing data. Secondly, to identify disease based anomalies, we discover all the significant association rules to identify the frequent patterns like disease, policy provider, policy holder etc according to the type of time slab and payment slab. Then for each type of slab outliers like irregular period outliers and irregular payment outliers are detected for fraud detection as provider may explain the benefits of insurance claim to the user very well and then insurance carrier may accept or reject the claims depending on defined policies but how much trustworthy is the carrier is a big question like whether user will be getting reimbursement for the services claimed.

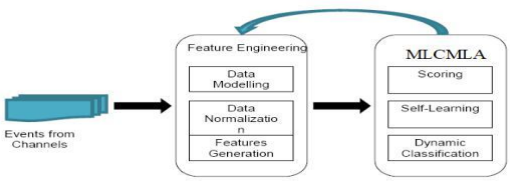

Figure 2: proposed system

To ensure transparency in the system, fraudulent patterns are identified using data mining technology. In this research we proposed a cost effective fraud detection framework for health care. The detection of frauds in health care is highly challenging task so effective techniques are needed to detect the frauds in this area. Broadly, we classify the fraudulent behaviour into two categories period based claim anomalies and disease based anomalies. The period based claim anomalies are investigated by analysing the statistical decision rules which helps in detecting the outliers and hence frauds and then clustering is performed to simplify the fraud detection process. The disease based anomalies are identified by discovering the association rule mining and identifying the frequent patterns. The overall objective of research in this area is to get maximum benefit out of medi-claim coverage justifying the investment and reimbursement of claimed services which should be provided. The proposed framework has been evaluated on real world medical data. The performance of the proposed framework is assessed using experimental analysis by involving all the entities like policy providers, policy holders, diseases etc. to get clear of frauds at every level. The results show that our proposed approach is efficient to identify the fraudulent claims from the existing data using data mining techniques. Though existing research has achieved the objectives but still it can be further explored to identify newly emerging frauds; effective strategies for fraud prevention can be developed; novel data analysis techniques can improve the state-of-art of fraud detection systems in health care; fraudulent patterns in health care data may change over time so health care fraud detection methods have to be dynamic enough to adapt these changes. Hence, future researchers can attempt to develop self-evolving fraud detection methods.

Proposed algorithm

Step 1: Start

Step 2: Navigation of user profile using user profile agent

Step 3: Design a suitable procedure for the intelligent agents to interact with user profile and clustering techniques.

Step 4: Execution of pre – processing process using pre – processing agent

Step 5: Selection of appropriate multi-level clustering using machine learning agent

Step 6: Decision making process using decision agent

Step 7: visualization of outputted result using appropriate Visualization tool

Step 10: Store the results of step 5 to step 7 for future reference in archival

Step 11:Stop

Various data mining techniques are considered in this proposed model for bettering clustering, prediction, association and outlier detection with respect to selected attributed types. Each mining technique is considered by data mining agent to mine meaningful clusters at various instances by considering the nature of data set selected for mining process. The natures of data set are numerical, categorical and mixed data types. The new patterns after applying various mining techniques is further validated with the help of decision agent. The decision agent checks whether the produced new patterns by data mining agent are based on the pre-processing or not. If the pattern produced by data mining agent is based on the pre – processing agent, it means that produced clusters are good in nature and also it can be recommended for further usage decision agent. Finally, the recommended patter from decision agent is given to the visualization agent by selecting appropriate visualization tool based on the nature of the data types present within the recommended cluster for better and meaningful visualization with the help of visualization agent. Thus, the final results are easily understood by the non-expert users using various intelligent agents. The methods and techniques proposed and discussed in various literatures in past on discovering internal financial frauds are very technical in nature from auditors point of view. Auditors find difficulties in understanding and applying the same in some context and unable to utilize and integrate their vast experiences and knowledge base with these methods and techniques. Even they find difficulties in applying simple techniques such as outlier analysis, link discovery, Benfords's law, trend analysis, and matching. The techniques are not suitable for discovering all types of frauds, known and unknown. The auditors are not well versed with data and its structure and hence find difficulties in preparing the data and transferring the same for further analysis. When there exists no framework, available of software with various advanced techniques is little help to auditors. This section implements the fraud detection techniques for financial statements of business groups developed in Section 3 using Python 2.7 and Matlab R2014a.

Additionally, the feasibility and validity of the proposed method is also demonstrated using the financial statements of business groups in Taiwan.

Furthermore, the detection accuracy is evaluated using a comparison with other detection models to prove the effectiveness of the proposed method. these business groups had similar total assets to those belonging to the fraudulent business group. Relevant data of both types of business groups are then retrieved to demonstrate and evaluate the effectiveness of the approach proposed in this study. In this section, the financial statement fraud for business groups "Irrational Balance Sheet through ECB" is used to explain the feasibility of our method. Based on reports to shareholders, stocks trading volume, debt structure indicators,. The established datasets are input into the clustering multi-level clustering machine learning algorithm for clustering training and testing. The clustering accuracy is compared with the accuracy of other six clustering models, including decision tree, logistic, neural network, KNN, GA-SVM, and multi-level clustering machine learning algorithm. revealing that the adopted clustering model has a higher accuracy than the other six clustering models.

V. CONCLUSION

In the data mining technique all the important information from such a huge amount of knowledge and changes all the high level decision making in the banks and retail sectors. From different databases they mix the varied data and store the mixed data using data storage in acceptable format that the data mining can be done for it. Analysis of data is done further and thus the captured information is used in banking sector or in any organisation to support decision-making. In banking sector the data mining techniques are a huge help to them for targeting and exploit new clients, most useful in fraud interference, providing phase based mostly merchandise, fraud detection in real time, risk management, analysis of the customers. Data Mining is a most important tool for detecting the fraud activities happening in the banks related information and prevent the frauds happening in our daily life due to the fraudsters. Data mining operates to provide the security to database and to enhance and the choice creating power, taking right decisions at the right time, and selecting the correct options. It fetches the important pattern from the large database which helps us in improving the quality of the database. Hence, this research paper includes lots of problems associated with the banking information security and ways to overcome the problems of the banking system frauds easily through the techniques provided my data mining.

Reference

[1].Yao, J., Zhang, J., & Wang, L. (2018). A financial statement fraud detection model based on hybrid data mining methods. 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). doi:10.1109/icaibd.2018.8396167.

[2]. Panigrahi, P. K. (2011). A Framework for Discovering Internal Financial Fraud Using Analytics. 2011 International Conference on Communication Systems and Network Technologies. doi:10.1109/csnt.2011.74

[3]. Rambola, R., Varshney, P., & Vishwakarma, P. (2018). Data Mining Techniques for Fraud Detection in Banking Sector. 2018 4th International Conference on Computing Communication and Automation (ICCCA). doi:10.1109/ccaa.2018.8777535.

[4]. Rawte, V., & Anuradha, G. (2015). Fraud detection in health insurance using data mining techniques. 2015 International Conference on Communication, Information & Computing Technology (ICCICT). doi:10.1109/iccict.2015.7045689.

[5]. Verma, A., Taneja, A., & Arora, A. (2017). Fraud detection and frequent pattern matching in insurance claims using data mining techniques. 2017 Tenth International Conference on Contemporary Computing (IC3). doi:10.1109/ic3.2017.8284299.

[6]. Jayabrabu, R., Saravanan, V., & Tamilselvi, J. J. (2014). A framework for fraud detection system in automated data mining using intelligent agent for better decision making process. 2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE). doi:10.1109/icgccee.2014.6922411

[7]. Chen, Y.-J., & Wu, C.-H. (2017). On Big Data-Based Fraud Detection Method for Financial Statements of Business Groups. 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI). doi:10.1109/iiai-aai.2017.13.

[8].B.Rajasekhar, B. Sunil Kumar, Rajesh Vibhudi, "Quality of Cluster Index Based on Study of Decision Tree", International Journal of Research in Computer Science, Vol 2, Issue 1, pp 39-43, eISSN 2249-8265, 2011.

[9]. Xu, R. Xu, J. ; Wunsch, D. C. "A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions, Volume: 99, Issue: 99, pp: 1 - 14, ISSN : 1083-4419, 2012

[10]. Kumar, S.P. Ramaswami, K.S. "Fuzzy K- Means Cluster Validation for Institutional Quality Assessment", Communication and Computational Intelligence (INCOCCI), 2010 International Conference, pp: 628 – 635, E-ISBN : 978-81-8371-369-6, 2010.

[11]. Min, X., & Lin, R. (2018). K-Means Algorithm: Fraud Detection Based on Signaling Data. 2018 IEEE World Congress on Services (SERVICES). doi:10.1109/services.2018.00024.
[12]. Lucas, Y., Portier, P.-E., Laporte, L., Calabretto, S., He-Guelton, L., Oble, F., & Granitzer, M. (2019). Dataset Shift Quantification for Credit Card Fraud Detection. 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE). doi:10.1109/aike.2019.00024.