ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2020.0312

A NOVEL APPROACH TO DETECTION OF EMERGING FRAUD USING MINING TECHNIQUES

Preeti Rathi and Nipur Singh
Department of Computer Science, Kanya Gurukul Campus Dehradun, India

Abstract

Fraud means accessing of information by unauthorized users. In today's scenario, online fraud plays a vital role in many internet applications. Web mining is the application of data mining which detect and extract information from web documents using mining techniques. These techniques also determine the behaviour of the user, i.e. authorized or unauthorized. For the detection of fraud, fraud detection systems are used. Credit card fraud is increase day by day. In this paper, we used phishing tank database to detect emerging fraud using clustering, classification and regression techniques. We collect the data from database and apply pre-processing techniques to remove irrelevant data and proposed architecture and approach to detect fraud and find the following metrics i.e. accuracy, error rate, memory consumption and search time.

Keywords:

Emerging Fraud, Web Mining, Clustering, Classification, Association Rule Mining, Pop-Up Windows

1. INTRODUCTION

Fraud is an activity which is play by an unauthorized person. Now a day's usage of credit card increased because of online transactions. Online fraud [7] can be inferred as phishing and steal useful information. Fraud is carried out by mysterious the internet user, false clicks and pop-up ads etc. There are lots of fraud done using a credit card. According to Global Payment Report 2015, a credit card is mostly used for the transaction is compared to other methods like e-wallet, transfer money etc. [8]. Credit card fraud detection is the process to identify whether the transaction is unaffected or affected. There are various approaches of data mining to detect fraud from weblogs, and also detect credit card fraud activities.

Data mining is the process of acquisition interesting and useful pattern using predictive and descriptive models from large data sets [3]. Data mining focused on discovered interesting pattern and developing model using classification, clustering and regression techniques.

There are various types of fraud, i.e. credit card fraud, intrusive data etc. In today's era credit card fraud increase day by day. It is an activity performed by an unauthorized user. Due to web advertising fraud can be increased. Multiple pop-up windows open on one click.

We have summarized the rest of paper into following sections- Related work discuss into section 2. In section 3 we discuss the dataset description. In section 4 discuss architecture of fraud detection and algorithm for emerging fraud. In section 5 discuss the result of proposed work and calculate the metrics i.e. accuracy, error rate, memory consumption and search time, and last in section 6 we conclude this paper and in section 7 we write the references of our paper.

2. RELATED WORK

In this section, we survey the fraud detection techniques for how to detect fraud. Fraud means wrongful deception to retrieve gain. In fraud detection, we detect the fraud i.e. unauthorized access. There are lots of work done in the detection of fraud. In today's scenario, credit card fraud is popular. In credit card fraud, unauthorized person use card by stole card information. There are many research paper on detection on fraud and many authors gave the solution of how to detect fraud. We discuss the techniques, methodology and tools to detect fraud. Contribution of each author discuss in this section:

Jeslet [4] discuss the application of web mining in fraud. Web mining is the application of data mining which deals with mining techniques. Fraud is a violation of existing law, and it is also a crime. The fraud detection system used to detect fraud using mining techniques, i.e. classification, clustering etc. In this paper, the author discusses credit card fraud, which is increase day by day.

Vimala and Sharmili [5] discuss credit card fraud detection using data mining techniques. These detection techniques based on methods like decision tree, clustering techniques, neural network, and hidden Markov model. These techniques include various credit card fraudulent transactions. Decision tree algorithm is an induction technique that recursively shares a set of records. Clustering means a group of similar types of data. Hidden Markov model is a set of states associated with the probability distribution.

Tripathi [1] discuss a novel web fraud detection technique using association rule mining and for experimental result using phish tank database. Linear search algorithm used to match the frequent patterns with datasets. Calculates accuracy, error rate, memory consumption, and search time through phish tank database and achieve low error rate and high accuracy and less time-consuming. In this technique, if data size is increase then searching time also increase.

Kurien [2] introduce detection and prediction of credit card fraud transaction using machine learning. In this paper, the author describes the probability of fraudulent transaction occurrence of credit card usage. Experimental metrics include precision, recall and f1 score, achieve high accuracy and low false alarm rate. The author used the classification algorithm to find classifier and the data set contains fraud and non -fraud data samples and analysed both data sets.

Mekteroviv [3] discuss a systematic review of data mining approaches of credit card fraud detection and finds the solution of the problem. In today's scenario, the credit card fraud problem plays a vital role, an unauthorized person, hacks the information through card detection. For performance measure recall, precision, f-measure parameters used with high accuracy.

Yee [6] discuss credit card fraud detection using machine learning as data mining techniques. Online based transaction raises the fraud case all over the world and causes huge losses to individual and industries. In this paper author discuss the supervised learning, i.e. classification using Bayesian network classifier, tree augmented naive Bayes, naive Bayes classifier and J48 classifiers. The pre-processing technique used to remove unwanted data from the data set and normalize the data set. In experimental result achieve more than 95% accuracy after pre-processing techniques and comparison of various classification algorithms.

Amanze and Onukwugha [9] discuss data mining application in credit card fraud system. In this paper, the author discusses the various fraud detection techniques, and different types of the fraudster, i.e. pre-planned fraudster, immediate fraudster, and slippery slope fraudster that commit online credit card fraud. Adaptive data mining is help full to detect online credit card fraud detection.

Folake and Kolawole [10] discuss credit card fraud detection system in commercial sites. In this paper author designed a web-based application whose adopted transition state model and detect fraudulent operation on the credit card. This system designs for security purpose. It also measures system performance and manages data integrity.

3. DATASET DESCRIPTION

In this section, we discuss the data set which is used in our research work to generate experimental results. There are other sources of phishing datasets which are analysing and recorded phishing data like APWG (Anti Phishing Working Group). It is an organisation which is same as phishing tank data set, it also consists of phish data.

For fraud detection machine learning techniques are used because these techniques are provide more accurate result to detect fraud. There are basically three factors of machine learning algorithms:

- Speed
- Scalability
- Efficiency

Phish tank database is used to detect fraud from log files. It is a database which contains both fraud and phishing URL. This database help to find authentic URL and fraud URL. Phish tank database created by historical web access log and frequently used access URL.

Phish tank database launched by David Ulevitch in San Francisco, California, October 2006. It is a community-based phish verification system where user submit suspected phishes and other user opinions whether it is suspected or not. This database consists of 154554 approx. phishing websites with its URL and time with date. This dataset updates time to time. Phishing data set help to determine whether the web is fraud or not.

Phish tank data set is used for the training data set. This dataset is help full to detect fraudulent data from the web dataset. For result analysis, we collect data from the UCI repository. It contains a large amount of data.

4. PROPOSED ARCHITECTURE OR MODEL FOR WEB FRAUD DETECTION

In this section we discuss the architecture and proposed algorithm for web fraud detection. The Fig.1 show the architecture of web fraud detection. There are many users whose access www and each user entry manage in file i.e. log file, then apply pre-processing technique to remove irrelevant data and receive the relevant data for further processing.

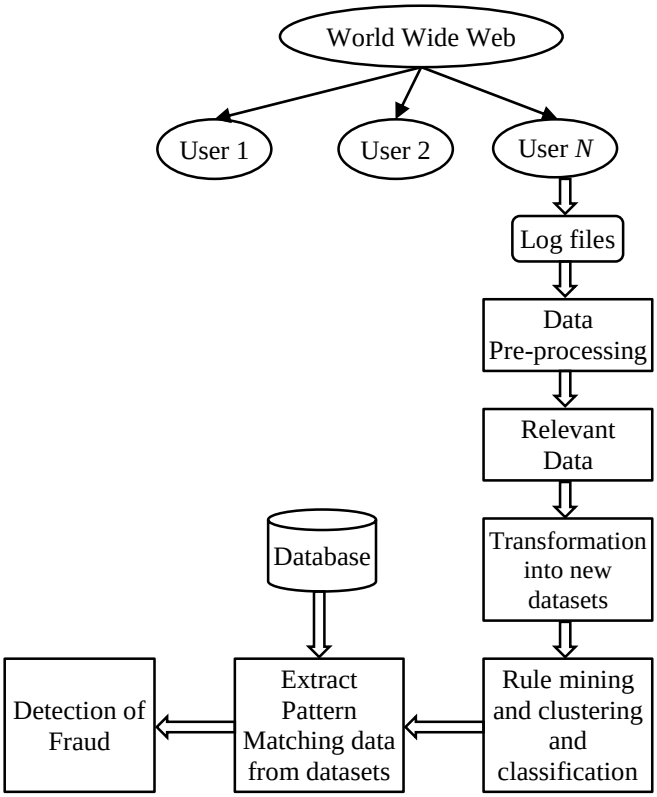

Fig.1. Architecture or Model of Web Fraud Detection

In pre-processing techniques, we apply data cleaning algorithm to remove unwanted data from data sets. For data cleaning used extension of files whose remove from data sets, and retrieve relevant data for next phase.

Relevant data transform into new database and apply mining techniques like rule mining, clustering, classification. In this research work we used frequent pattern growth technique to find the exact pattern from data sets, and matching these dataset to exiting datasets as well as matching patterns from dataset. If pattern is matching then data is accurate else if pattern is not matching then data consist inconsistency and it is fraud. We used threshold value for extracted useful patterns from datasets, this threshold value is based on analyser.

5. ALGORITHM FOR EMERGING FRAUD

Input: Log files from server logs

Generate frequent access patterns

Output: Fraud Detection

Step 1: Collect the data c logs and create a log table L.

Step 2: Transform the table and named as transformed log table *T*, with parameters such as IP address, time stamp, request and response, where, *T* is the subset of *L*.

Step 3: Select an IP address consist of request, response, date and time.

Step 4: Set minimum support with threshold value between 0-1. (If data set is large then minimum support is less otherwise high), where, support is an indication of how frequently the items appear in the data.

Step 5: Generate frequent patterns using FP growth with minimum threshold value.

Step 6: Interval search (binary) techniques to apply phish tank dataset with generated frequent patterns.

Step 7: If frequent pattern is match existing dataset then pattern is accurate otherwise it is fraud.

Above algorithm is used to detect fraud from log files using Apriori algorithm and generate frequent item set. Frequent item sets are those sets which are used frequently i.e. IP addresses which are mostly access.

6. EXPERIMENTAL RESULTS

In this section we discuss the results of proposed algorithm and compute memory consumption and time required to execute algorithm.

Accuracy shows the ratio between correctly identified requests and total requests of pattern analysis i.e.

Accuracy = correctly identified request/*total requests*

Error rate shows the ratio between incorrectly identified request and total requests of pattern analysis i.e.

ER = incorrectly identified request/*total requests*

where, *ER* = *Error Rate* or *Error Rate* = 100 - *Accuracy*

Memory consumption defines the how much space or memory used by the proposed algorithm. Search time indicates the searching time to find URL's from the datasets.

Above parameters shown in following graphs-

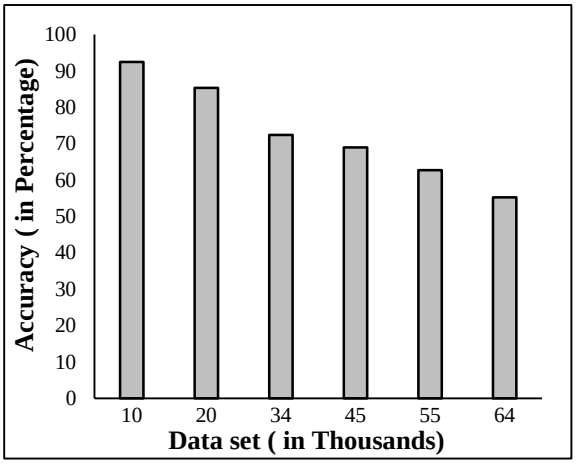

Fig.2. Accuracy of proposed model

In Fig.2, x-axis show the size of data set in thousands and y-axis show the accuracy of system in percentage. If data size is increase accuracy should be decrease.

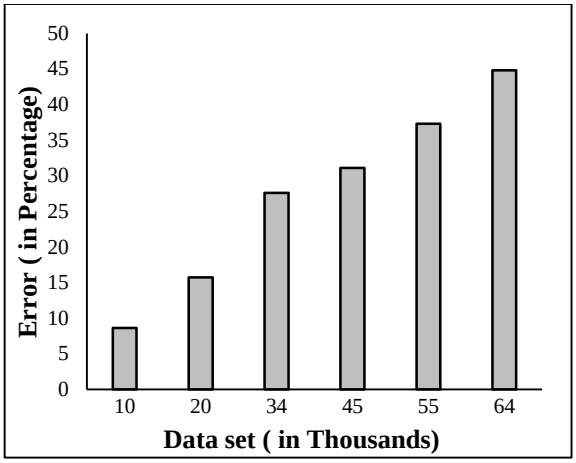

Fig.3. Error rate of proposed model

In Fig.3, x-axis show the size of data set in thousands and y-axis show the error rate of the system. Error rate define the incorrectly identification during the detection of fraud. Error rate increase if data size will be increase.

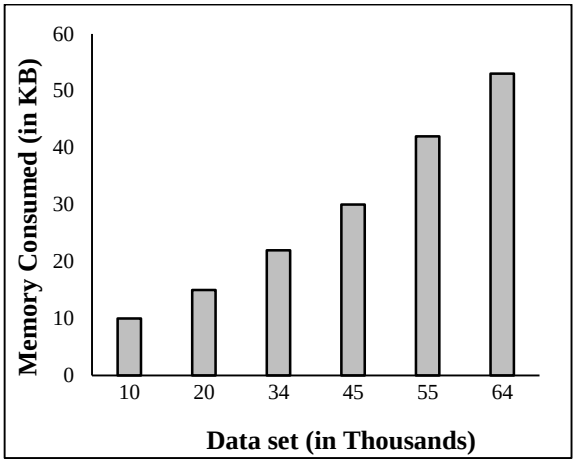

Fig.4. Memory consumption of proposed model

In Fig.4, x-axis show the size of dataset in thousands and y-axis show the memory or space consumed by the proposed algorithm. If data size is increase space also increase. It occur due to how much data loaded into memory for analysis.

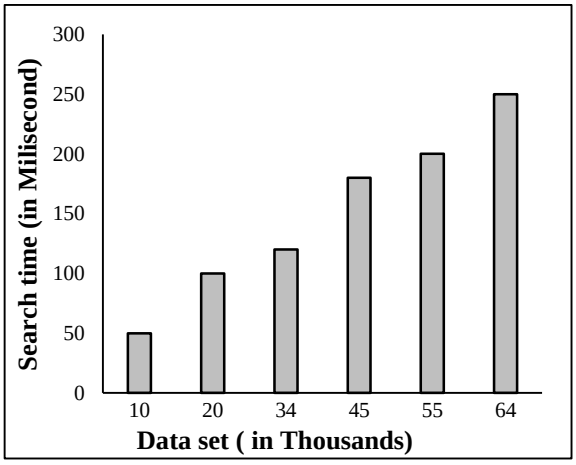

Fig.5. Searching time of proposed model

In Fig.5, x-axis show the size of data set in thousands and y-axis show the time consumption in millisecond. Search time indicates the time spend to verify URL's are fraudulent or not. If data size for analysis is increase then searching time also increase.

For comparison we used parameter such as accuracy, error rate and search time with existing algorithms. Proposed algorithm emerging fraud detection gives accurate results in compare to other existing fraud detection algorithms.

The Table.1 shown the results of existing and proposed algorithms with parameters as follows. For compare we used data size is 1MB or above.

Table.1. Comparison of Existing Algorithms with Proposed Algorithm

Algorithms / Parameters	Decision tree	Random forest	Apriori	Emerging Fraud Detection
Memory Consumption	65 KB	60 KB	55 KB	50 KB
Searching Time	350 ms	300 ms	250 ms	200 ms
Accuracy	85%	89%	91%	95%

The Fig.6 shown the results of existing and proposed algorithms with parameter accuracy.

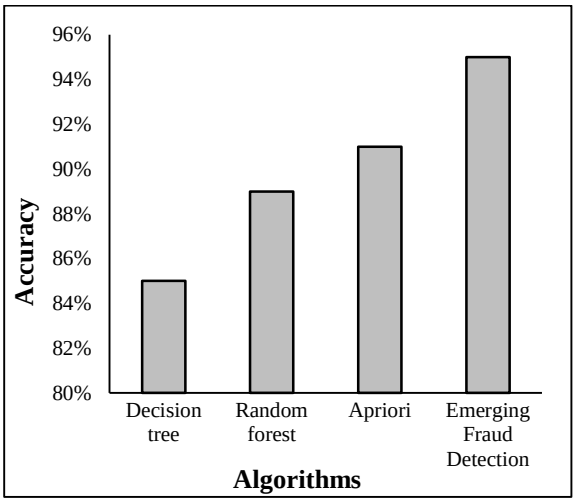

Fig.6. Comparison of Existing Algorithms with Proposed Algorithm with parameter accuracy

7. CONCLUSION AND FUTURE WORK

It is the last section of our research work. We conclude our research paper. Fraud is an action which is done by an unauthorized person. We proposed an algorithm to detect web-based fraud through phish tank datasets. Calculate evaluation parameter like accuracy, error rate, memory consumption and searching time.

In future, we extended the proposed algorithm with clustering and classification technique to enhance performance. Combination approach is useful to find patterns frequently, and reduce searching time used multi-label indexing and another searching time.

REFERENCES

[1] Diwakar Tripathia, Bhawana Nigamb and Damodar Reddy Edlaa, "A Novel Web Fraud Detection Technique using Association Rule Mining", *Proceedings of 7th International Conference on Advances in Computing and Communications*, pp. 274-281, 2017.

[2] Kaithekuzhical Leena Kurien and Ajeet Chikkamannur, "Detection and Prediction of Credit Card Fraud Transactions using Machine Learning", *International Journal of Engineering Sciences and Research Technology*, Vol. 8, No. 3, pp. 199-208, 2019.

[3] Igor Mekterovic, Ljiljana Brkic and Mirta Baranovic, "A Systematic Review of Data Mining Approaches to Credit Card Fraud Detection", *WSEAS Transactions on Business and Economics*, Vol. 15, No. 2, pp. 437-444, 2018.

[4] D. Santhi Jeslet, "Application of Web Mining in Fraud Detection", *International Journal for Research in Applied Science & Engineering Technology*, Vol. 6, No. 1, pp. 51-54, 2018.

[5] S. Vimala and K.C. Sharmili, "Survey Paper for Credit Card Fraud Detection Using Data Mining Techniques", *International Journal of Innovative Research in Science, Engineering and Technology*, Vol. 6, No. 11, pp. 357-364, 2017.

[6] Ong Shu Yee, Saravanan Sagadevan and Nurul Hashimah Ahamed Hassain Malim, "Credit Card Fraud Detection using Machine Learning as Data Mining Techniques", *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 10, No. 1, pp. 23-27, 2018.

[7] Internet Fraud, Conton Police Department, "Crime Prevention Unit 743", Available at: http://www.cantonpublicsafety.org/Documents/safety_tips/pst_internet_fraud.pdf._sharing, Accessed at: 2017.

[8] Haddadi, Hamed, "Fighting Online Click-Fraud using Bluff Ads", *ACM SIGCOMM Computer Communication Review*, Vol. 40, No.2, pp. 21-25, 2010.

[9] B.C. Amanze and C.G. Onukwugha, "Data Mining Application in Credit Card Fraud Detection System", *International Journal of Trend in Research and Development*, Vol. 5, No. 4, pp. 23-26, 2018.

[10] Akinbohun Folake and Atanlogun Sunday Kolawole, "Credit Card Fraud Detection System in Commercial Sites", *European Journal of Engineering Research and Science*, Vol. 3, No. 11, pp. 1-5, 2018.