ANVESAK
ISSN: 0378–4568
UGC Care Group1 Journal

ANALYSIS ON FRAUD IDENTIFICATION AND DETECTION OF CREDIT CARD USING MACHINE LEARNING ALGORITHMS

[1]R. Sagar, [2]Dr. N. Chandra Mouli, [3]B.Adithya
[1]Assistant Professor, [2]Associate Professor, [3]Student, [1,2,3] Dept. of Computer Science Engineering,
[1,2,3]Vaageswari College of Engineering, Karimnagar, Telangana.
E-Mail: [1]sagarrachuri@gmail.com, cmnarsingoju@gmail.com, sateesh.singireey@gmail.com

ABSTRACT
In order to create an appropriate framework to detect credit card fraud, which is primarily observed in financial services, we apply the Random forest and majority voting methods under the platform[1] of Python in this study. After a payment has been authorised, the process of fraud detection involves locating the fraudulent transactions. Use of the Random Forest Algorithm A machine learning method called Random Forest, sometimes known as Random Foresting [2], is used in conjunction with many other learning algorithms to enhance their effectiveness. This algorithm produces a prediction for each test sample and then uses a weighted sum to combine the results to produce the combined output of the booster classifier.
Application of Majority voting Method: Majority Voting is most repetitively used in data classification, which involves a combined with at least two algorithms. Each algorithm makes its own prediction for every test sample and the final output is for the one that receives the majority of the votes.

1. INTRODUCTION

One of the emerging technologies to change how credit card fraud is created and spread is the Random Forest Algorithm and majority voting techniques, which run on the Python programming language. In a similar vein, these python-based algorithm technologies recognise the manipulation and authentication of source's credit card information. These algorithms' traceability, transparency, and decentralisation make it possible to successfully address the issue of credit card fraud. Online readers can have a trustworthy means of confirming the scam with information about its source thanks to these algorithms and a Python-enabled platform. Our system's capacity to identify the source or origin of fraud will assist stop the spread of credit card fraud. In the proposed framework, Random forest algorithm stores each and every minute detail of source's credit card detail shared or uploaded on the proposed platform in the form of a transaction performed by registered users. Because of the transparent and traceable nature of these algorithms, it is possible to verify the source of any information that is shared on such a platform. Tracing the fraud by using these algorithms can be achieved with the help of time stamping and the chain connection between these methods.

2. LITERATURE REVIEW

[3] The Use of Predictive Analytics Technology to Detect Credit Card Fraud in Canada. "Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavlov Zavarsky."
This paper mainly deals with the implementation of score card from the appropriate evaluation criteria, features and the capabilities of predictive analytics vendor solutions mostly being used to identify the credit card fraud. So in this published paper the implemented scorecard provides a side by side comparision of five credit card predictive analytics vendor solutions adopted in Canada.
[4] BLAST-SSAHA Hybridization for Credit Card Fraud Detection. "Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar".
In this research paper, it uses the two-stage sequence procedure in which the profile Analyser (PA) mainly detects the similarity of an incoming sequence of transactions on the given credit card with the genuine card holder's past spending sequences. In this case the unusual transactions made by the profile analyser are

passed on to the deviation analyser (DA) for possible alignment with past fraudulent behaviour. The final outcome on the nature of the transactions is considered on the basis of the observations made by the above two analysers. In order to achieve online response time for both PA and DA, we suggest a new approach for combining two sequence alignment algorithms BLAST and SSAHA.

[5] Credit card fraud detection using the Distance Sum. "Wen-Fang YU, Na Wang".

In these days with the rapid rise in the credit cards and the increasing in the trade percentage in the china, a credit card fraud rises sharply. So in order to detect the fraud and prevent the credit card fraud this research paper proposes the credit card fraud detections model by using the outlier detection which is based on the distance sum according to the in frequency and unconventionality of the fraud in the credit card transaction data, applying the outlier mining method to fraud in the credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.

[6] Credit card fraud detection using SVM & Decision Tree. "Vijayshree B.Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande".

Due to the increasing advancements in the electronic commerce field, frauds are increasing in the world, effecting major financial losses due to credit card fraud. So normally Decision tree, Genetic algorithm, Meta learning strategy, neural network, HMM are the main methods for the detection of fraud and also artificial intelligence concept of Support Vector Machine (SVM) & decision tree are also being used to solve this problem. Hence by the implementation of this hybrid approach, financial losses are reduced to the greater extent.

[7]Credit card fraud detection by the Support Vector Machine. "Sitaram patel, Sunita Gond".

This research paper talks about the SVM (Support Vector Machine) based methodology with the multiple kernel involvement which mainly involves the several fields of the user profile instead of the only spending profile.

[8]Credit card fraud detection by Decision Trees and Support Vector Machines. "Y. Sahin and E. Duman"

In this research paper, it mainly focuses on the classification models which are based on the decision trees and the support vector machines (SVM) are synthesized and applied on to the credit card fraud detection problem.

3.EXISTING SYSTEM

In the existing thesis, the research on the case study involving fraud detection on the credit cards, in this first the data normalization is applied before the cluster analysis and with the results attained due to the usage of the cluster analysis and the Artificial neural networks on the fraud detection has been resulted that by changing the clustering features and neuronal inputs could be minimized.Somainly this research paper was entirely based on the unsupervised learning method approach. The data set used in this existing method is real life transactional data on the very big European company and the details of the employees are kept private and after applying this cluster analysis on this large European dataset, the accuracy of the algorithm was found to be around 50%.Since the accuracy of the unsupervised learning algorithm was very low, we need to find the new approaches for the detection of the frauds on the credit cards and also mainly to increase the accuracy of the results and also to reduce the cost measure. By considering all this affects and also based on the simultaneous researches concludes that the promising results can be obtained by using the normalized data and this data should be MLP trained. The significance of the MLP training on the normalized data is that the supervised learning algorithms give the best accuracy results.

4. PROBLEMS IN EXISTING SYSTEM

The main problem in the existing system is to trace whether the transaction is fraud or not with 100% accuracy because once the fraud occurs, it becomes hard to find out the solution with 100% guarantee and also the approach used is cost sensitive approach.

OBJECTIVE OF THE PROPOSED WORK

In this work we propose a framework to analyze and identify the fraud. We examine how the fraud takes place and transmits repeatedly by using majority voting method which tells how many times fraud has occurred and Random forest algorithm which predicts that the fraud has occurred. This whole framework uses an evolution decision tree modelling method to examine the credit card fraud over the python enabled online system. The main goal of the system is to improve the output of the system.

PERFORMANCE EVALUATION ENVIRONMENT

In this project we are using the random forest algorithm for classification of the credit card dataset which implies that based on the attributes of the dataset random forest algorithm classifies the target value whether it has undergone the fraud or not since the random forest algorithm mainly used for the classification and regression analysis. So first in this evaluation we first split the dataset into train data set and train data set and then we give the training data set to the random forest model which in turn it samples each of the attributes of the data and builds the decision that is whether the credit card got fraud or not. After this training we use the test dataset on random forest to predict the fraud scenario correctly.

PROPOSED SYSTEM

In this proposed system, we implement the random forest algorithm for the classification of the credit card dataset because it is an algorithm used for the classification and regression data sets. Mainly the random forest algorithm is the collection of decision tree and also random forest has the advantage over the decision trees as it corrects the habit of over fitting to the resultant training dataset. In this model some portion of the training set is selected randomly in order to train each of the individual trees and then at last the decision tree is built, and after that each node then splits on to attribute which is selected from the random portion of the dataset. Even though for the large datasets with the more number of the attributes and the data nodes training is very fast in the random forest it's because each tree is trained independently to one another. So that's the reason the random forest algorithm has found to give the good accuracy with the less error and also resistant towards the over fitting.

ADVANTAGES OF PROPOSED SYSTEM

- The random forest algorithm has the ability to work accurately even though when the data set has the null values or it has not at all scaled properly.
- This algorithm is very much stable, even if the new data point is introduced into the existing dataset, since the random forest algorithm contains very large number of decision trees because even if a new point is introduced it doesn't affect the all tress.
- And the last and the most important one that is random forest algorithm is not biased, because it contains a very large number no. of tress and every tree works on each subset of the data. Therefore the overall biasedness of the algorithm is reduced.

SYSTEM ARCHITECTURE

In the process first the credit card data set is taken from the source and then it has to undergo into the cleaning and then the validation is performed on this dataset which mainly includes the partition of redundancy, filling empty space in columns, and then converting the required variable into the factors or the classes and then the data is divided into the two parts, one is used for the training of the dataset and the another one is used for the testing of the data set. So basically what we are trying to do is dividing the dataset into the two parts and they are test and train dataset.

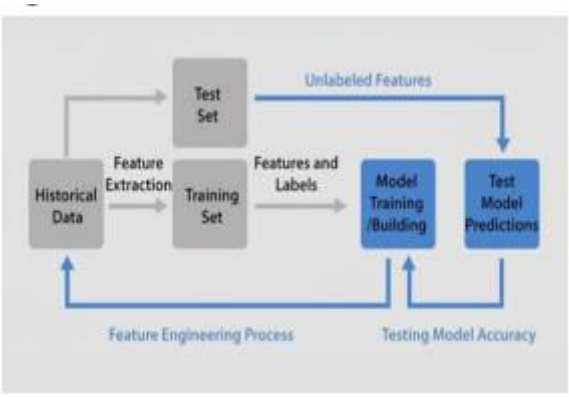


Fig.1. Architecture of System

STEP-1
DATA COLLECTION
According to the architecture of the proposed system first the raw data is collected from the required sources and after collecting this raw data we have to analyse the data set, so the data we have collected and using in this paper is a set of the people's credit card transactions on the different products in the different retail stores. The first step in this process is selecting the subset of all the available data in order to work with the machine learning algorithm. The basic machine learning problem starts with the data mainly, because there will be lots of data to analyse.

STEP -2
DATA PRE-PROCESSING
In this data pre-processing step first we have to arrange the data in a ordered manner by formatting, cleaning and sampling from it.
They are three common steps in the data pre-processing and they are:
FORMATTING:
In this step the data we have picked might not be in a format which is suitable for us to work with. The data may be in a relational database and we would like it in a flat file, or the data may be in a proprietary file format and we would like to be it in a relational database or a text file.

CLEANING
In the second process we have to clean the data, cleaning the data means it's the removal or fixing of the missing data. In the data set there may be data instances that are incomplete and do not carry the data which seems that you need to handle the problem. These instances might need to be removed. And also additionally, there might be the sensitive information in some of the features and these features might need to be removed from the data entirely.
SAMPLING
In the third process there might be very far more selected data available than we need to concentrate on. The more data for the understanding results in the much longer running times for the algorithms and the larger computational and the more memory requirements. So in order to avoid this problem we have to take small quantities of the selected data that might give the better solutions with the fast rate of exploring before considering the whole dataset.
FEATURE EXTRACTION
In this step we have to do is feature extraction, it's an attribute reduction technique. Its not about the selecting the features, it mainly ranks the existing features according to their predictive significance, and it mostly transforms the attributes. And these transformed features are the linear combination of the original attributes. At last our models are trained using the classifier algorithm.

EVALUATION MODEL

Model evaluation is the important step which is an integral part of the model building process. Basically it helps us to find the best model that actually represents our data and how good the selected model will perform in the future. By calculating model performance with the data used for the training is not at all feasible in data science because it could easily generate the more over optimistic and over fitted models. But there are two methods in order to evaluate models in the data science and they are Hold-Out and Cross-Validation methods. To avoid the over fitting we use both the methods and these methods are used on a test data to evaluate the model performance. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

ALGORITHM UTILIZED

RANDOM FOREST

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

WORKING OF RANDOM FOREST

The following are the basic steps involved in performing the random forest algorithm

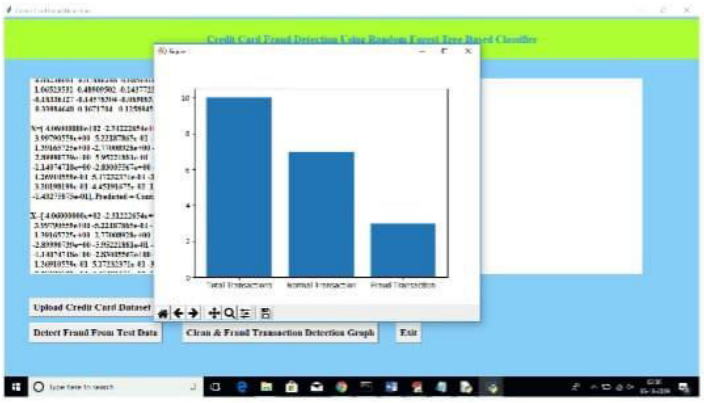
1. Pick N random records from the dataset.
2. Build a decision tree based on these N records.
3. Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
4. For classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

ADVANTAGES OF RANDOM FOREST

1. The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore, the overall biasedness of the algorithm is reduced.
2. This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.
3. The random forest algorithm works well when you have both categorical and numerical features.

SCREENSHOTS OF THE OUTPUT

In the below graph we can see the total test data and number of normal and fraud transaction detected. In below graph the x-axis represents type and Y-axis represents the count of the clean and fraud transactions.



CONCLUSION

Here Random Forest Algorithm and majority voting methods are used as one of the emerging technologies to transfigure the way in which the credit card fraud is generated and propagated, and in a similar way the these python based algorithm technology recognizes the manipulation and authentication of source's credit card details. Due to the traceability, transparency and decentralization nature of these algorithms, the problem of credit card fraud is handled successfully. These algorithms and python enabled platform provided online readers with a reliable way of verifying the fraud with its source details. Our system has the ability to trace the root or origin of fraud so that it helps in refraining the propagation of credit card fraud.

REFERENCES

1. Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2017, pp. 0630-0633. doi: 10.1109/KBEI.2017.8324876
2. A. Charleonnan, "Credit card fraud detection using RUS and MRN algorithms," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, 2016, pp. MIT-73-MIT-76. doi: 10.1109/MITICON.2016.8025244
3. LI Changjian, HU Peng: Credit Risk Assessment for ural Credit Cooperatives based on Improved Neural Network, International Conference on Smart Grid and Electrical Automation vol. 60, no. - 3, pp 227-230, 2017.
4. Wei Sun, Chen-Guang Yang, Jian-Xun Qi: Credit Risk Assessment in Commercial Banks Based On Support Vector Machines, vol.6, pp 2430-2433, 2006.
5. Amlan Kundu, Suvasini Panigrahi, Shamik Sural, Senior Member, IEEE,"BLAST-SSAHA Hybridization for Credit Card Fraud Detection", vol. 6, no. 4 pp. 309-315, 2009.
6. Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi Conference of Engineers and Computer Scientists, vol. I, 2011.
7. Sitaram patel, Sunita Gond , "Supervised Machine (SVM) Learning for Credit Card Fraud Detection, International of engineering trends and technology, vol. 8, no. -3, pp. 137- 140, 2014.
8. Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar," Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April- 2015, pg. 92-95
9. Dahee Choi and Kyungho Lee, "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System", vol. 5, no. - 4, December 2017, pp. 12-24.