International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Credit Card Fraud Detection using Random Forest Algorithm

S. Monika[1], K. Venkataramanamma[2], P. Pritto Paul[3], M. Usha[4]
[1,2]*UG Student, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India*
[3,4]*Assistant Professor, Dept. of Computer Science and Engg., Velammal Engineering College, Chennai, India*

***Abstract*: Credit risk is one of the main functions of banking. Banks classify risk according to their profile. Although many algorithms came into existence still the issue is yet to solve. In existence, data normalization is applied before Cluster Analysis and the obtained results from Cluster Analysis and Artificial Neural Networks on fraud detection has shown by clustering attributes and the neuronal inputs can be minimized. Significance of the paper is to find an algorithm to reduce the cost measure. The result obtained was 23% and the algorithm used was Minimum Bayesian-Risk (MBR). In proposed system, Random Forest Algorithm is used for classification and regression. Random forest has the advantage over decision tree as it corrects the habit of over fitting to their training data sets. It has been found to provide a good estimate of generalization error and resistant to over fitting. In credit card fraud detection, credit card data sets are collected for trained data sets and user credit card queries are collected for testing data sets. After classification process, Random Forest Algorithm is used for analysing data sets and current data sets. Finally, the optimization is done and the accuracy obtained by Random Forest is 99.9%.**

***Keywords*: Support Vector Machines (SVM), Random Forest Algorithm and algorithm.**

1. Introduction

Billions of losses are caused every year by the fraudulent credit card transactions. Fraud is old as humanity itself and can take an unlimited variety of different forms. The PWC global economic crime survey of 2017 suggests that approximately 48% of organizations experienced economic crime [3]. Therefore, there's positively a requirement to resolve the matter of credit card fraud detection. The use of credit cards is prevalent in modern day society and credit card fraud has been kept on growing in recent years [2]. Hugh financial losses have been fraudulent affects not only merchants and banks, but also individual person who is using the credits. Fraud may also affect the reputation and image of a merchant causing non-financial losses that, though difficult to quantify in the short term, may become visible in the long period [4]. For example, if a cardholder is victim of fraud with a precise company, he might no longer trust their business and opt for a rival.

2. Literature survey

Along with increasing credit card and growing trade volume in china, credit card fraud rises sharply. How to enhance the detection and bar of credit card fraud becomes the main target of risk management of banks. This paper proposes a credit card fraud detection model victimization outlier detection supported distance add consistent with the scarceness and unconventionality of fraud in credit card dealing information, applying outlier mining into credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud [1].

With growing advancement within the electronic commerce field, fraud is spreading all over the world, causing major financial losses. In current scenario, Major cause of financial losses is credit card fraud. It not only affects trades person but also individual clients. Decision tree, Genetic algorithm, Meta learning strategy, neural network, HMM are the presented methods used to detect credit card frauds. In contemplate system for fraudulent detection, artificial intelligence concept of Support Vector Machine (SVM) & decision tree is being used to solve the problem. Thus by implementation of this hybrid approach, financial losses can be reduced to greater extend [2].

In this paper, we tend to proposing the SVM (Support Vector Machine) primarily based methodology with multiple kernel involvement that additionally includes many fields of user profile rather than solely of only spending profile. The simulation result shows improvement in TP (true positive), TN (true negative) rate, & also decreases the FP (false positive) & FN (false negative) rate [3].

In this study, classification models supported on decision trees and Support Vector Machines (SVM) are developed and applied on credit card fraud detection problems. This study is one of the first to compare the performance of SVM and decision tree methods in credit card fraud detection with a real data set [4].

A new cost-sensitive decision tree approach which reduces the sum of misclassification costs while selecting the splitting attribute at each non-terminal node is advanced and the act of this approach is compared with the well-known ancient classification models on a true world credit card data set. This analysis is completely involved with master card application fraud detection by performing arts the method of asking security queries to the persons byzantine with the transactions and as well as by eliminating real time data faults [5].

3. Proposed model

The proposed System uses Random Forest Algorithm for classify the credit card data set. Random Forest is an algorithmic program for classification and regression. Summarily, it is a set of decision tree classifiers. Random Forest has advantage over decision tree because it corrects the habit of over fitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is build, each node then splits on a feature selected from a random subset of the total feature set. Even for large data sets with many features and data instances training is extremely fast in Random Forest and because each tree is trained independently of the others. The Random Forest Algorithm has been found to produce a good estimate of the generalization error and to be resistant to over fitting.

A. System architecture

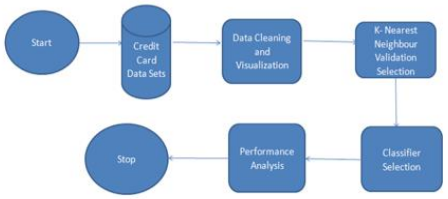

Fig.1. System architecture diagram

Descriptive analysis is done and the target variable is determined. It explores how many classes were there in the target and selects high cardinality variables. Then the high cardinality variables were dropped during this step as a precursor to the pre-processing step. The data set has 31 features, 28 of which have been label from V1 to V28. The remaining three features are the time, amount and the class. The target variable was removed from the entire data sets and transforms the categorical variable into a model matrix with one-hot encoding. This is required to process the data in a sparse matrix format. The missing values are imputed in the data to 0. The partitioned data sets are pre-processed into a training and test data set. Then the K-Nearest Neighbour classifier model is built, using neighbour classes. The classifier on unseen testing data sets was scored and calculates the R squared values for both the training and testing data sets. The results were evaluated and obtained 99.9% through random forest algorithm

B. Algorithm

Random Forest could be a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning could be a type of learning where different types of algorithms or same algorithm with multiple times to form a more powerful prediction model. The Random Forest combines multiple algorithms of the same type i.e. multiple decision trees, leading to a forest of trees, thus the name "Random Forest". The Random Forest can be used for regression and classification tasks.

4. How random forest works

The following are the essential steps concerned in performing the Random Forest Algorithm:

- Pick N random records from the data set.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- 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 class and that wins the huge vote.

5. Conclusion

The Random Forest Algorithm will perform better with a larger number of training data, and the result obtained is 99.9%. The SVM algorithm can be used instead of Random Forest, but it still suffers from the imbalanced data set problem and requires more pre-processing to give better results.

6. Future enhancement

In future, privacy preserving techniques can be applied in distributed environment which will resolve the security related issues preventing private data access.

References

[1] W. Yu and N. Wang, "Research on Credit Card Fraud Detection Model Based on Distance Sum," *2009 International Joint Conference on Artificial Intelligence*, Hainan Island, 2009, pp. 353-356.
[2] Vijayshree B. Nipane, Poonam S. Kalinge, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande, Fraudulent Detection in Credit Card System Using SVM & Decision Tree.
[3] Sitaram patel, Sunita Gond, Supervised Machine (SVM) Learning for Credit Card Fraud Detection.
[4] Y. Sahin and E. Duman, Detecting Credit Card Fraud by Decision Trees and Support Vector Machines.
[5] Snehal Patil, Harshada Somavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar, Credit Card Fraud Detection Using Decision Tree Induction Algorithm.
[6] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A neural network based database mining system for credit card fraud detection," in Proc. IEEE/IAFE Computat. Intell. Financial Eng., Mar. 1997, pp. 220–226.
[7] C. Alippi, G. Boracchi, and M. Roveri, "A just-in-time adaptive classification system based on the intersection of confidence intervals rule," Neural Netw., vol. 24, no. 8, pp. 791–800, 2011.
[8] C. Alippi, G. Boracchi, and M. Roveri, "Hierarchical change-detection tests," IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 2, pp. 246–258, Feb. 2016.
[9] C. Alippi, G. Boracchi, and M. Roveri, "Just-in-time classifiers for recurrent concepts," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 4, pp. 620–634, Apr. 2013.
[10] B. Baesens, V. Van Vlasselaer, and W. Verbeke, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. Hoboken, NJ, USA: Wiley, 2015.
[11] A. C. Bahnsen, D. Aouada, and B. Ottersten, "Example-dependent cost-sensitive decision trees," Expert Syst. Appl., vol. 42, no. 19, pp. 6609–6619, 2015.

[12] A. C. Bahnsen, D. Aouada, A. Stojanovic, and B. Ottersten, “Detecting credit card fraud using periodic features,” in Proc. 14th Int. Conf. Mach. Learn. Appl., Dec. 2015, pp. 208–213.
[13] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, “Data mining for credit card fraud: A comparative study,” Decision Support Syst., vol. 50, no. 3, pp. 602–613, 2011.
[14] A. Bifet and R. Gavalda, “Learning from time-changing data with adaptive windowing,” in Proc. SDM, vol. 7. 2007, pp. 443–448.
[15] R. Bolton and D. Hand, “Statistical fraud detection: A review,” Stat. Sci., vol. 17, no. 3, pp. 235–249, 2002.