Eleyon
SOUTH ASIAN JOURNAL OF ENGINEERING AND TECHNOLOGY

Full Length Article

CREDIT CARD FRAUD DETECTION USING BOOSTED STACKING

E.Prabhakar[a*], M.Naresh Kumar[b], K.Ponnar[b], A.Suresh[b], R.Jayandhiran[b]

[a] *Assistant Professor, Department of Computer Science and Engineering, Nandha College of Technology, Erode – 638 052, Tamilnadu, India*.

[b] *Student, Department of Computer Science and Engineering, Nandha College of Technology, Erode – 638 052, Tamilnadu, India*.

***Corresponding Author**
prabhakarit10@gmail.com
(E.Prabhakar)
Tel.: +91 8056795606

ABSTRACT: Credit card fraud is one of the severe issue in financial field. Huge amount of money is lost because of the credit card fraud. There are many studies related to credit card fraud detection. But most of the studies failed to analyse different set of attributes. This research concentrates on identifying the best machine learning algorithm for credit card fraud detection. Existing system used the hybrid methods. This method integrate AdaBoost and majority voting methods. But many of the existing system failed to obtain the higher level accuracy. In the proposed approach, combination of stacking and AdaBoost is considered. Model efficiency is calculated based on the accuracy. Results shows that the proposed approach provides better detection of credit card fraud.

.

Keywords: Credit Card Fraud Detection, Stacking, Boosting, Majority Voting.

1. Introduction

The significance of the Machine Learning and the Data Science are not able to be overstated. In case, you have interested in learning past trends and also train the machines for the purpose of find along with time as for define scenarios, to identify and the label events, or helps to predict value in future or present, the data science is of an essence. It is an important to learn underlying data and that would modelled by means of select an appropriate mechanism in the sense of dealing any of those use case. The different control parameters of algorithm are needed to tweak for fitting data set. Finally, developed application has been found as that would enhance the usage and found as highly efficient in solving issues.

In this paper, attempt is made to demonstrate modelling the data set by means of machine learning paradigm classification, along the Credit Card Fraud Detection that is being a base. Classification is one of the machine learning paradigms which includes deriving function that would separate the data as categories, or as classes, and characterize by training set of the data that contains observations in other words instances whereas, category membership has been known. This particular function is using for identify categories of new observation that belongings.

The rest of the paper is organized as follows: Section 2 discusses the literature review. Section 3 overviews the proposed methodology. Experimental results of the proposed scheme are presented in Section 4. Concluding remarks with future work are covered in Section 5.

2. Literature Review

Sharmila Subudhi, Suvasini Panigrahi in [1] introduced hybrid approach with the help of applying the Genetic Algorithm based on Fuzzy C-Means clustering and different models of Supervised Classifiers. This System is used to identify fraudulence in the automobile insurance claims. The test has been extracted from Insurance dataset and on train set, the clustering has applied for an under Sampling to divided the instances as malicious, suspicious or genuine and to eliminate fraudulent, genuine records. By individually using Support Vector Machine, Decision Tree, Multi-Layer Perceptron and Group Method of Data Handling, suspicious records are analyzed fatherly.

Vladimir Zaslavsky and Anna Strizhak in [2] presenting an approach that will helps to detect the fraud activities in credit card, since it has been considered as significant one for various financial institutions. This system adopts neural network technology in the sense of introducing an automatic credit card detection system. Self-Organizing Map mechanism is used for creating the model for typical cardholder's behavior and for analyzing deviation in transactions to find suspicious transactions.

Roberto Saia and Salvatore Carta in [3] introduced an approach which brings out benefits of novel evaluation criterion on frequency domain in spectral pattern of data. This helps to acquire highly stable system to represent information which is respect to canonical ones that minimize both issues of the imbalance and the heterogeneity of data. Finally, the proposing system has been compared with state-of-the-art competitor, and finds that the proactive strategy can have ability for contrast cold-start problem.

Jon T. S. Quah and M. Sriganesh in [4] bring out the real-time fraud detection system to understand spending patterns for decipher the potential fraud cases. Yusuf Sahin , Serol Bulkan , Ekrem Duman[5] developed an approach that can useful for identify the fraudulence transactions in the sense of reduce the losses in finance.

Neda Soltani Halvaiee,, Mohammad Kazem Akbari in [6] address the credit card fraud detection by means of Artificial Immune Systems that presents new model nammed AIS-based Fraud Detection Model that enhance the fraud detection. Ekrem Duman , M. Hamdi Ozcelik in [7] presenting a mechanism that enhances the credit card fraud detection solution. This can be obtained by scoring the transaction and depends on such scores, legitimate and fraudulent transactions are classified.

M. Bakopoulos in [8] developed a system that employs on online algorithms which is optimally aggregate the statistical data from raw data and also apply a number of the pre-specified checks that against the fraud scenarios is known. Teymur Rahmani et al., in [9] paper concentrating on effectiveness of adopting hybrid intelligent based system which combines few classifiers in the aim of detecting corporate tax evasion for Iranian National Tax Administration. E.Prabhakar and K.Sugashini in [10] proposed new ensemble approach to improve the accuracy. In [11] the authors introduced new type of improved version of boosting approach to obtain the better accuracy and to handle imbalanced data. The authors in [12], [13] highlighted the importance of public opinion mining for the purpose of improvement in today's world.

3. Proposed Methodology

The Credit Card Fraud Detection Problem involves modelling the past credit card transactions along knowledge of one which turned out as fraud. This model is been then used for identifying whether the new transaction can be fraudulent or not. Our intent is to identify 100% of fraudulent transactions that minimize fraud classifications that are incorrect.

The proposed methodology consists of following steps:

- ✓ Dataset Collection, Data Pre-processing
- ✓ Applying Machine Learning Techniques
- ✓ Performance Analysis
- ✓ Identify the Best Model
- ✓ Apply the Model

Data Set

Credit Card Fraud Detection is one of typical example for classification. In this particular process, we have been focused highly on analyze feature that modelling and additionally possible business use cases of algorithm's result than algorithm by itself.

Credit card fraud dataset is picked from the Kaggle. The data set is been skewed highly, which consists of 492 frauds as total of about 284,807 observations. This can found as resulting of about 0.172% fraudulent cases. This skewed set has been justified by low amount of the fraudulent transactions.

The dataset consisting of the numerical values from 28 'Principal Component Analysis (PCA)' is transformed features ranging from V1 to V28. Additionally, there can be no more metadata which are relating to original features which is provided, so the pre-analysis or the feature study that could not done. The 'Time' and 'Amount' features aren't transformed data. There is no missing value in dataset.

Inferences drawn from Dataset

Owing to that imbalance in the data, an approach which doesn't do any of feature analysis and also predicts all transactions as one of non-fraud will helps to achieve accuracy of about 99.828%. Thus, accuracy isn't correct measure of the efficiency in our own case. We also need some of other standard of the correctness that would classify transactions which can be as fraud or as non-fraud.

The 'Time' feature doesn't indicating actual time of transaction and high of list of data in the chronological order. Thus, we are assuming that 'Time' feature possess little or some no significance in the classifying fraud transaction. Therefore, we are eliminating this particular column from the further analysis.

Data Pre-processing

Due to the high non-fraudulent to fraudulent ratio, displayed within the dataset, predictions made from the initial training set, which had a normal to fraudulent ratio of 49:1, were greatly skewed. Many of the utilized algorithms classified the test data with ninety-eight percent accuracy by predicting every transaction as normal, with only true negative and false negative cases.

Popular class balancing techniques are Random Under Sampling (RUS), Random Over Sampling (ROS) and SMOTE. For undersampling, RUS is preferred, as it is considered both simple yet effective. ROS and SMOTE were picked as oversampling methods since of it's widely usage.

Boosting

Boosting refers to algorithms family which is able to converting weak learners as strong learners. Main principle of the boosting is to fit sequence of all weak learners– models which are only slightly better when compared to random guessing, namely small decision trees– to a weighted versions of data. More weight has been given for examples which have been misclassified in the earlier rounds.

The predictions are made to combine through weighted majority vote in terms of classification or a weighted sum as a regression in the sense of producing final prediction. The principal difference that in between the boosting and committee methodologies namely, bagging, are base learners who are trained in a sequence on weighted version of data.

The algorithm below gives detailed view on widely using form of the boosting algorithm is said to be as AdaBoost that stands as adaptive boosting.

Algorithm AdaBoost

1: Init data weights $\{w_n\}$ to $1/N$
2: for $m = 1$ to M do
3: fit a classifier $y_m(x)$ by minimizing weighted error function J_m:
4: $J_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n]$
5: compute $\epsilon_m = \sum_{n=1}^{N} w_n^{(m)} 1[y_m(x_n) \neq t_n] / \sum_{n=1}^{N} w_n^{(m)}$
6: evaluate $\alpha_m = \log\left(\frac{1-\epsilon_m}{\epsilon_m}\right)$
7: update the data weights: $w_n^{(m+1)} = w_n^{(m)} \exp\{\alpha_m 1[y_m(x_n) \neq t_n]\}$
8: end for
9: Make predictions using the final model: $Y_M(x) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m y_m(x)\right)$

Stacking

Stacking is one of an ensemble learning technique which combines the multiple classification or the regression models through meta-classifier or meta-regressor. Base level models are been trained which is

based on complete training set, meta-model has been trained on outputs of base level model by means of features.

The base level often could consist of various learning algorithms and thus the stacking ensembles are also often heterogeneous. The algorithm presented below demonstrate the stacking.

Algorithm Stacking

Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
Ouput: ensemble classifier H
Step 1: learn base-level classifiers
for $t = 1$ to T do
    learn h_t based on D
end for
Step 2: construct new data set of predictions
for $i = 1$ to m do
    $D_h = \{x_i', y_i\}$, where $x_i' = \{h_1(x_i), ..., h_T(x_i)\}$
end for
Step 3: learn a meta-classifier
learn H based on D_h
return H

Boosted Stacking

This proposed methodology integrates the boosting and stacking ensemble. It considers the advantages of both the algorithms.

Step 1: Import the dataset
Step 2: Give 75% data for training and remaining data for testing.
Step 3: Apply "n" different base learners for AdaBoost
Step 4: Assign train dataset to the models
Step 5: Update the weight based on the misclassification rate
Step 6: Choose the best learners among "n" different base learners
Step 7: Apply the selected learners for AdaBoost
Step 8: Now create the best combination for AdaBoost
Step 9: Create the model using Stacking
Step 10: Make predictions for test dataset and calculate accuracy

4. Experimental Results

Performance Metrics

The following are the essential definitions – in current problem's context.

True Positive: The fraud cases that a model which predicted as a 'fraud.'

False Positive: The non-fraud cases that a model that predicting as a 'fraud.'

True Negative: The non-fraud cases that the model predicting as the 'non-fraud.'

False Negative: The fraud cases that the model is predicted as a 'non-fraud.'

Confusion Matrix: Merely tabulates confusion matrix that won't provide clear understanding over performance of data. This is due to total amount of the fraud cases which is highly less, and the variation in confusion matrix will also be very small which would be equivalent to justified error in the balanced dataset is probably even less!. So, this measurement is can be ruled out.

Accuracy: Measurement of correct predictions could be made by a model – which is, ratio of the fraud transactions that classified as a fraud and a non-fraud classified as a non-fraud to total transactions in test data.

Results

Table 1: Performance analysis for three different algorithms

Algorithm	**Accuracy**
Logistic regression	87.2
Decision tree	89.1
Random Forest	90.0
AdaBoost	91.4
Stacking	92.6
Boosted Stacking	94.5

In this paper, 6 machine learning algorithms are compared to identify the best model to detect the fraud in credit card system. To evaluate the algorithms, 75% of the dataset is used for training and 30% is used for testing. Accuracy is used to evaluate the performance of the algorithms. The accuracy result is shown for logistic regression, Decision tree, random forest, boosting, stacking, boosted stacking.

5. Conclusion

A study on credit card fraud detection using machine learning algorithms has been presented in this paper. A number of standard models which include logistic regression, Decision tree, random forest, boosting, stacking have been used in empirical evaluation. A publicly available credit card data set has been used for evaluation using individual (standard) models and hybrid models using AdaBoost and stacking methods. The comparative results show that the proposed boosted stacking performs better than the other techniques.

References

[1] Sharmila Subudhi, Suvasini Panigrahi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection", Elsevier, DOI.org/10.1016/j.jksuci.2017.09.010, (2017).

[2] Vladimir Zaslavsky and Anna Strizhak, "Credit card fraud detection using self- organizing maps", Information & Security, An International Journal, (2006).

[3] Roberto Saia and Salvatore Carta, "Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach", DOI: 10.5220/0006425803350342, In Proceedings of the 14th International Joint Conference on e-Business and Telecommunications, (2017).

[4] Jon T. S. Quah and M. Sriganesh, "Real Time Credit Card Fraud Detection using Computational Intelligence", Elsevier – Expert Systems with Applications, https://doi.org/10.1016/j.eswa. 2007.08.093, (2007).

[5] Yusuf Sahin , Serol Bulkan , Ekrem Duman, "A cost-sensitive decision tree approach for fraud detection", Elsevier, https://doi.org/10.1016/ j.eswa.2013.05.021, (2013).

[6] Neda Soltani Halvaiee,, Mohammad Kazem Akbari, "A novel model for credit card fraud detection using Artificial Immune Systems", Elsevier – Applied Soft Computing, https://doi.org/10.1016/j.asoc.2014.06.042, (2014).

[7] Ekrem Duman , M. Hamdi Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search", Elsevier – Expert Systems with Applications, https://doi.org/10.1016/ j.eswa.2011.04.110, (2011).

[8] I.T. Christou, M. Bakopoulos, "Detecting fraud in online games of chance and lotteries", Elsevier – Expert Systems with Applications, https://doi.org/10.1016/j.eswa.2011.04.124, (2011).

[9] Teymur Rahmani,, Mehdi Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran", Elsevier, International Journal of Accounting Information Systems, https://doi.org/10.1016/ j.accinf.2016.12.002, (2017).

[10] E. Prabhakar and K. Sugashini, "New Ensemble Approach to Analyze User Sentiments from Social Media Twitter Data", The SIJ Transactions on Industrial, Financial & Business Management (IFBM), Vol. 6, No. 1, (2018).

[11] E. Prabhakar, "Enhanced AdaBoost Algorithm with Modified Weighting Scheme for Imbalanced Problems", The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 6, No. 4, (2018).

[12] E.Prabhakar, R.Parkavi, N.Sandhiya, M.Ambika, "Public Opinion Mining For Government Scheme Advertisement", International Journal of Information Research and Review, Volume 3, Issue 4, Page No.2112-2114, April 2016.

[13] E.Prabhakar, G.Pavithra, R.Sangeetha, G.Revathy, "Mining Better Advertisement Tool for Government Schemes", International Journal for Technological Research in Engineering, ISSN (Online): 2347 – 4718, Volume 3, Issue 5, Page No.1023-1026, January 2016.