Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 1, January 2020, pp.1-5
Available online at: http://pen.ius.edu.ba

Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers

Admel Husejinović[1]
[1]Central Bank of Bosnia and Herzegovina

ABSTRACT

Growing problem of card payment fraudulent abuse is a main focus of banks and payment Service Providers (PSPs). This study is using naive Bayes, C4.5 decision tree and bagging ensemble machine learning algorithms to predict outcome of regular and fraud transactions. Performance of algorithms is evaluated through: precision, recall, PRC area rates. Performance of machine learning algorithms PRC rates between 0,999 and 1,000 expressing that these algorithms are quite good in distinguishing binary class 0 in our dataset. Amongst all algorithms best performing PRC class 1 rate has Bagging with C4.5 decision tree as base learner with rate of 0,825. For prediction of fraud transactions with success of 92,74% correctly predicted with C4.5 decision tree algorithm.

Keywords: Card payments, Fraud transactions, Naive Bayes, C4.5 decision tree, Bagging

Corresponding Author:
Admel Husejinović,
Dr. Silve Rizvanbegović 24B, 71210, Ilidža, BiH
Email: hadmel@hotmail.com

1. Introduction

Card payments are very popular payment method in nowadays. Card payments are quite easy to perform on merchant side by presenting credit card or on internet by announcing credit card details: number, expiring date and security code. As result of low level of security card payments are influence of fraudulent abuse. Also, another important reason is increase in mobile devices use for initialization of payments. Globally general-purpose payment cards reached $21,84 billion losses as result of fraud transactions in 2015 [1]. European Banking Authority (EBA) issues guidelines on the security of internet payments. Payment Service Directive (PSD 2) recommends to the Payment Service Providers (PSPs) at card payments or e-money transfers strong customer authentication. According to the PSD 2 effective process for authorizing transactions, as well as monitoring transactions and system should be implemented by PSPs.

In this study naive Bayes, C4.5 decision tree, and bagging ensemble machine learning classifiers are performed on the dataset that contains transactions made by credit cards in September 2013 by European cardholders [2]. This dataset contains 284.807 transaction where 492 are fraud. In dataset there are 31 features where 31st is a binary variable with 0 regular transaction and 1 as fraud transaction. The performances of algorithms are evaluated through following performance matrices: precision, recall and precision-recall (PR) curve area rate. Our dataset is imbalanced as result of low rate of fraud transaction dataset for that reason better indicator for algorithm performance is PR curve than receiver operating characteristic (ROC) rate [3], [4].

2. Literature review

Many authors [5] were dealing with card payment fraud detection problem using machine learning techniques. In 2016 Shimpi made survey on different machine learning, data mining and artificial intelligence methods

used in credit card fraud detection [6]. Also [7] performed comparative study in credit card fraud detection techniques used. Kumar and Assistant introduced concept of tree levels of security in credit card fraud detection using Hidden Markov Model (HMM) [8]. Yadav and Siddartha also used HMM for their study of card payments fraud detection [9]. Nami and Shajari [10] used cost recall customer behavior in two stages of payment card fraud detection employing dynamic random forest and k-nearest neighbor algorithms. Brause et al [11] showed how advanced data mining techniques and a neural network algorithm can be combined successfully tested on real credit card date. Xuan et al [12] used two kinds of random forests methods to train the behavior features of normal and abnormal transactions. Authors [13] parallelize the negative selection algorithm on the cloud platform using apache Hadoop and MapReduce decreasing the training time of basic algorithms. Authors [14] tried to increase accuracy of highly imbalanced real-world datasets by ensemble learning methods. To solve highly imbalanced problem Ghobadi and Rohani [15] used Meta Cost procedure. Using machine learning methods [16] approaches to minimize the false alarm rates and increase the fraud detection rate. Duo and Gui-Yang Li [17] showed how neural network algorithm and receiver operating characteristic (ROC) analysis technology can be combined successfully to perform credit card fraud detection in two levels of credit card fraud detection. Patil, Nemade and Soni [18] used Big data analytical framework to test performances of different machine learning techniques by processing large volume of data.

3. Methodology

In this study we investigate performance of naive bayes, C4.5 decision tree machine and bagging ensemble methods to test performance through recall, precision and precision-recal curve (PRC) area rates.

3.1. Naive Bayes

Bayesian classifier is a statistical method calculating probability that feature belongs to class based on applying Bayes' theorem. Because it assumes that the probabilities of individual features are independent of each other which is quite hard to happens in real world it is reason to be called naive. Considering that another event has already occurred to calculate the likelihood that an event will occur. It can be written as:

$$\Pr\,(c/X) = \frac{(\Pr\,(X/c)*\Pr(c))}{(\Pr(X))} \quad (1)$$

Where posterior probability of target class c P(c|X) is calculated from P(c), P(X|c) and P(X).

3.2. C4.5 Decision tree

R. J. Quilan in early 1980s developed ID3 (Iterative Dichotomiser) machine learning classifier which he later used to develop C4.5 decision tree as an inheritor of ID3 [19]. Decision trees are considered as simplest among machine learning methods. Because of they are a completely transparent method of classifying observations that look like a series of if-then statements arranged into a tree [20]. Classification starts from topmost node called root node down the tree to the leaves based on the outputs. Leaves are the target classes of our dataset. Classification can be easy applied by answering question down the decision tree. Basically, what algorithm does is splitting data in two or more sets. Splitting is done built on most significant attributes to make distinct groups as possible using information gain and entropy. Entropy measures impurity of result class in subset with ps attributes in some D dataset as shown in formula:

$$H(p1, p2, \ldots ps) = \sum_{i=1}^{s}(p_i \log(\frac{1}{p_i})) \quad (2)$$

Information gain is calculated as difference between entropy of whole dataset and entropy of splitting attribute as shown in formula:

$$Gain(D,S) = H(D) - \sum_{t=1}^{s} p(D_i)H(D_i) \quad (3)$$

3.3. Bagging ensemble learner

Bagging (Bootstrap aggregating) was presented by Leo Breiman in 1994 to increase classification accuracy by joining classifications of randomly generated training sets. Bagging is ensemble learning algorithm used for reducing variance and avoiding overfitting in models. Randomly generated training sets make ensemble of "weak learner" to come together and form "strong learner". How the algorithm works could be explained with Han and Kamber's patient diagnostics [21] that is based on their symptoms. Instead of one doctor's we can choose to have more doctors' diagnostics. So, as some of diagnostics start to appear more than one's final diagnosis would be made based on majority vote of equally weighted each doctor's diagnosis. Instead of doctors we have classifiers and rest of explained process is basic idea of how the bagging algorithm works [21].

3.4. Confusion matrix

Confusion matrix summarizes performance of algorithm. Idea of what is algorithm doing correct and what is doing incorrect can be understood from it. Confusion matrix rows represent predicted class, while rows represent actual class.

Table 1. Confusion Matrix

Confusion matrix		Predicted class	
		0	1
Actual class	0	True positive (TP)	False negative (FN)
	1	False positive (FP)	True negative (TN)

The **precision** represents ratio of true positives (TP) and actual positives (TP + FP).

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

The **recall** or true positive rate (TPR) is ratio of true positives (TP) and actual positives (TP + FN). Measures the fraction of actual positives that are correctly recognized so.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

The area under a **PR** curve measures the overall ability of the test to distinguish between binary classes. The PR curve is fundamental tool for model evaluation of imbalanced datasets with binary target class. Graphical visualization of PR Curve is represented with recall on X axis and precision on the Y axis. Higher numbers mean better model performance.

4. Results

Original dataset is highly imbalanced as only 0,17% of data is classified as fraud. If we predict all data inputs to be classified as class 1, we would be 99,83% accurate.

Table 2. Performance Rates

	Precision		Recall		PRC Area	
	Class 0	**Class 1**	**Class 0**	**Class 1**	**Class 0**	**Class 1**
Naive Bayes	0,999	0,061	0,978	0,829	1,000	0,080
C4.5	1,000	0,927	1,000	0,778	0,999	0,745

	Precision		Recall		PRC Area	
Bagging	1,000	0,916	1,000	0,797	1,000	0,825

According to the results of following machine learning algorithms: naive bayes, C4.5 decision tree and bagging with C4.5 decision tree as a base learner performed in Waikato Environment for Knowledge Analysis (Weka) software precision, recall and PRC rates are presented in table 2.

5. Discussion

As result of performance of machine learning algorithms PRC Area rates for 0 class are between 0,999 and 1,000 telling us that these algorithms are very good in distinguishing binary class 0 in our dataset. For PRC rates for class 1 results are 0,080 for naive Bayes, 0,745 for C4.5 decision tree and 0,825 bagging ensemble learner telling us that: naive Byes algorithms is poor performing while C4.5 and bagging are good in distinguishing binary class 1. This is important indicator as we were testing algorithms to predict result class as normal transaction or fraud. If we talk about precision rate of class 1, we talk about negative predicted value or accuracy of alarm rate of class 1. For all predicted fraud transactions 92,74% would be correctly predicted as result of best performing C4.5 decision tree algorithm.

6. Conclusion

Overall according to the PRC Area overall best performing algorithm is bagging with C4.5 decision tree as base learner with the rate of 1,000 for class 0 and 0,825 for class 1. Highest recall rates of 0,978 for class 0 and 0,829 for class 1 are recorder in performance of naive bayes model. Highest precision rates of 1,000 for class 0 and 0,927 for class 1 are recorder in performance of C4.5 decision tree model. If we remain that our dataset is quite imbalanced PRC rates of 1,000 for class 0 and 0,825 for class are quite promising.

References

[1] D. Robertson, 'Top Card Issuers in Asia-Pacific', *The Nilson Report*, p. 12, 2016.

[2] Worldline and the Machine Learning Group, 'Credit Card Fraud Detection at Kaggle', *Credit Card Fraud Detection Dataset*, 2013. [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud/home. [Accessed: 24-Nov-2018].

[3] T. Saito and M. Rehmsmeier, 'The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets', *PLoS One*, vol. 10, no. 3, 2015.

[4] J. Davis and M. Goadrich, 'The Relationship Between Precision-Recall and ROC Curves', in *ICML '06 Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[5] S. N. John, C. Anele, O. O. Kennedy, F. Olajide, and C. G. Kennedy, 'Realtime Fraud Detection in the Banking Sector Using Data Mining Techniques/Algorithm', in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2016, pp. 1186–1191.

[6] P. R. Shimpi, 'Survey on Credit Card Fraud Detection Techniques', *Int. J. Eng. Comput. Sci.*, 2016.

[7] K. Modi and R. Dayma, 'Review on fraud detection methods in credit card transactions', in *2017 International Conference on Intelligent Computing and Control (I2C2)*, 2017, pp. 103–107.

[8] V. Kumar and P. Assistant, 'Method and System for Detecting Fraud in Credit Card Transaction', *ISSN Int. J. Innov. Res. Comput. Commun. Eng.*, 2013.

[9] S. Yadav and S. Siddartha, 'FRAUD DETECTION OF CREDIT CARD BY USING HMM MODEL', *Int. J. Res. Eng. Technol.*, vol. 6, no. 1, pp. 41–46, 2018.

[10] S. Nami and M. Shajari, 'Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors', *Expert Syst. Appl.*, vol. 110, pp. 381–392, Nov. 2018.

[11] R. Brause, T. Langsdorf, and M. Hepp, 'Neural data mining for credit card fraud detection', in *Proceedings 11th International Conference on Tools with Artificial Intelligence*, 1999, pp. 103–106.

[12] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, 'Random forest for credit card fraud detection', in *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, 2018, pp. 1–6.

[13] H. Hormozi, M. K. Akbari, E. Hormozi, and M. S. Javan, 'Credit cards fraud detection by negative selection algorithm on hadoop (To reduce the training time)', in *The 5th Conference on Information and Knowledge Technology*, 2013, pp. 40–43.

[14] S. Dhankhad, E. Mohammed, and B. Far, 'Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study', in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 122–125.

[15] F. Ghobadi and M. Rohani, 'Cost sensitive modeling of credit card fraud using neural network strategy', in *2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS)*, 2016, pp. 1–5.

[16] N. Malini and M. Pushpa, 'Analysis on credit card fraud identification techniques based on KNN and outlier detection', in *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, 2017, pp. 255–258.

[17] Tao Guo and Gui-Yang Li, 'Neural data mining for credit card fraud detection', in *2008 International Conference on Machine Learning and Cybernetics*, 2008, pp. 3630–3634.

[18] S. Patil, V. Nemade, and P. K. Soni, 'Predictive Modelling For Credit Card Fraud Detection Using Data Analytics', *Procedia Comput. Sci.*, vol. 132, pp. 385–395, Jan. 2018.

[19] A. Dželihodžić, D. Đonko, and J. Kevrić, 'Improved Credit Scoring Model Based on Bagging Neural Network', *Int. J. Inf. Technol. Decis. Mak.*, vol. 17, no. 06, pp. 1725–1741, Nov. 2018.

[20] T. Segaran and Toby, *Programming collective intelligence : building smart web 2.0 applications*. O'Reilly, 2007.

[21] J. Han, M. Kamber, and J. Pei, *Data mining : concepts and techniques*, 2nd ed. Elsevier Inc., 2006.