2581-4575

International Journal For Recent Developments in Science & Technology

Crossref

A Peer Reviewed Research Journal

CREDIT CARD FRAUD DETECTION USING ADABOOST AND MAJORITY VOTING

[1]SANKA.LAKSHMI SRAVYA, [2]M.KIRAN KUMAR

[1]STUDENT, DEPT OF CSE, GUNTUR ENGINEERING COLLEGE, GUNTUR, ANDHRA PRADESH

[2]ASSOCIATE PROFESSOR, DEPT OF CSE, GUNTUR ENGINEERING COLLEGE, GUNTUR, ANDHRA PRADESH

Abstract

Credit card fraud is a serious problem in financial services. Billions of dollars are lost due tocredit card fraud every year. There is a lack of research studies on analyzing real-world credit card dataowing to confidentiality issues. In this paper, machine learning algorithms are used to detect credit cardfraud. Standard models are firstly used. Then, hybrid methods which use AdaBoost and majority votingmethods are applied. To evaluate the model efficacy, a publicly available credit card data set is used. Then,a real-world credit card data set from a financial institution is analyzed. In addition, noise is added to thedata samples to further assess the robustness of the algorithms. The experimental results positively indicatethat the majority voting method achieves good accuracy rates in detecting fraud cases in credit cards.

1.INTRODUCTION

Fraud is a wrongful or criminal deception aimed to bring financial or personal gain [1]. In avoiding loss from fraud, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, where it stops fraud from happening in the first place. On the other hand, fraud detection is needed when a fraudulent transaction is attempted by a fraudster. Credit card fraud is concerned with the illegal use of credit card information for purchases. Credit card transactions can be accomplished either physically or digitally [2]. In physical transactions, the credit card is involved during the transactions. In digital transactions, this can happen over the telephone or the internet. Cardholders typically provide the card number, expiry date, and card verification number through telephone or website. With the rise of e-commerce in the past decade, the use of credit cards has increased dramatically [3]. The number of credit card transactions in 2011 in Malaysia were at about 320 million, and increased in 2015 to about 360 million. Along with the rise of credit card usage, the number of fraud cases have been constantly increased. While numerous authorization techniques have been in place, credit card fraud cases have not hindered effectively. Fraudsters favour the internet as their identity and location are hidden. The rise in

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

credit card fraud has a big impact on the financial industry. The global credit card fraud in 2015 reached to a staggering USD $21.84 billion [4]. Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges [5]. Since the merchants need to bear the loss, some goods are priced higher, or discounts and incentives are reduced. Therefore, it is imperative to reduce the loss, and an effective fraud detection system to reduce or eliminate fraud cases is important. There have been various studies on credit card fraud detection. Machine learning and related methods are most commonly used, which include artificial neural networks, rule-induction techniques, decision trees, logistic regression, and support vector machines [1]. These methods are used either standalone or by combining several methods together to form hybrid models. IEEE In this paper, a total of twelve machine learning algorithms are used for detecting credit card fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using both benchmark and realworld credit card data sets. In addition, the AdaBoost and majority voting methods are applied for forming hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set. The key contribution of this paper is the evaluation of a variety of machine learning models with a real-world credit card data set for fraud detection. While other researchers have used various methods on publicly available data sets, the data set used in this paper are extracted from actual credit card transaction information over three months. The organization of this paper is as follows. In Section II, related studies on single and hybrid machine learning algorithms for financial applications is given. The machine learning algorithms used in this study are presented in Section III. The experiments with both benchmark and realworld credit card data sets are presented in Section IV. Concluding remarks and recommendations for further work are given in Section V.

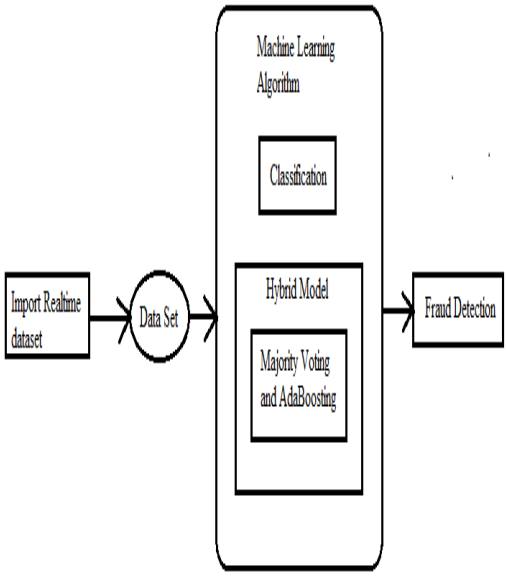
2.RELATED STUDIES

In this section, single and hybrid machine learning algorithms for financial applications are reviewed. Various financial applications from credit card fraud to financial statement fraud are reviewed. A. SINGLE MODELS For credit card fraud detection, Random Forest (RF), Support Vector Machine, (SVM) and Logistic Regression (LOR) were examined in [6]. The data set consisted of one-year transactions. Data under-sampling was used to examine the algorithm performances, with RF demonstrating a better performance as compared with SVM and LOR [6]. An Artificial Immune Recognition System (AIRS) for credit card fraud detection was proposed in [7]. AIRS is an improvement over the standard AIS model, where negative selection was used to achieve higher precision. This resulted in an increase of accuracy by 25% and reduced system response time by 40% [7]. A credit card

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

fraud detection system was proposed in [8], which consisted of a rule-based filter, Dumpster–Shafer adder, transaction history database, and Bayesian learner. The Dempster–Shafer theory combined various evidential information and created an initial belief, which was used to classify a transaction as normal, suspicious, or abnormal. If a transaction was suspicious, the belief was further evaluated using transaction history from Bayesian learning [8]. Simulation results indicated a 98% true positive rate [8]. A modified Fisher Discriminant function was used for credit card fraud detection in [9]. The modification made the traditional functions to become more sensitive to important instances. A weighted average was utilized to calculate variances, which allowed learning of profitable transactions. The results from the modified function confirm it can eventuate more profit [9]. Association rules are utilized for extracting behavior patterns for credit card fraud cases in [10]. The data set focused on retail companies in Chile. Data samples were defuzzified and processed using the Fuzzy Query 2+ data mining tool [10]. The resulting output reduced excessive number of rules, which simplified the task of fraud analysts [10]. To improve the detection of credit card fraud cases, a solution was proposed in [11]. A data set from a Turkish bank was used. Each transaction was rated as fraudulent or otherwise. The misclassification rates were reduced by using the Genetic Algorithm (GA) and scatter search. The proposed method doubled the performance, as compared with previous results

Architecture



3.EXISTING SYSTEM

Three methods to detect fraud are presented. Firstly, clustering model is used to classify the legal and fraudulent transaction using data clusterization of regions of parameter value. Secondly, Gaussian mixture model is used to model the probability density of credit card user's past behavior so that the probability of current behavior can be calculated to detect any abnormalities from the past behavior. Lastly, Bayesian networks are used to describe the statistics of a particular user and the statistics of different fraud scenarios. The main task is to explore different views of the same problem and see what can be learned from the application of each different technique.

2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

4.PROPOSED SYSTEM

Total of twelve machine learning algorithmsare used for detecting credit card fraud. The algorithmsrange from standard neural networks to deep learningmodels. They are evaluated using both benchmark and realworldcredit card data sets. In addition, the AdaBoost andmajority voting methods are applied for forming hybridmodels. To further evaluate the robustness and reliability ofthe models, noise is added to the real-world data set. Thekey contribution of this paper is the evaluation of a varietyof machine learning models with a real-world credit carddata set for fraud detection.

A.Module Implementation

1. **Standard Neural Networks To Deep Learning**

The Feed-Forward Neural Network (NN) uses thebackpropagation algorithm for training as well. Theconnections between the units do not form a directed cycle,and information only moves forward from the input nodes tothe output nodes, through the hidden nodes. Deep Learning(DL) is based on an MLP network trained using a stochasticgradient descent with backpropagation. It contains a largenumber of hidden layers consisting of neurons with tanh,rectifier, and maxout activation functions. Every nodecaptures a copy of the global model parameters on local data,and contributes periodically toward the global model usingmodel averaging.

2. **Forming Hybrid Models**

Adaptive Boosting or AdaBoost is used in conjunction withdifferent types of algorithms to improve their performance.The outputs are combined by using a weighted sum, whichrepresents the combined output of the boosted classifier, AdaBoost tweaks weak learners in favor of misclassifieddata samples. It is, however, sensitive to noise and outliers.As long as the classifier performance is not random,AdaBoost is able to improve the individual results fromdifferent algorithms.Majority voting is frequently used in data classification,which involves a combined model with at least twoalgorithms. Each algorithm makes its own prediction forevery test sample. The final output is for the one thatreceives the majority of the votes,

3. **Evaluate The Robustness And Reliability**

To further evaluate the robustness of the machine learningalgorithms, all real-world data samples are corrupted noise,at 10%, 20% and 30%. Noise is added to all data features.It can be seen that with the addition of noise,the fraud detection rate and MCC rates deteriorate, asexpected. The worst performance, i.e. the largest decrease inaccuracy and MCC, is from majority voting of DT+NB andNB+GBT. DS+GBT, DT+DS and DT+GBT show gradualperformance degradation, but

2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal


their accuracy rates are stillabove 90% even with 30% noise in the data set.

B. Testing methodologies

The following are the Testing Methodologies:

- **Unit Testing.**
- **Integration Testing.**
- **User Acceptance Testing.**
- **Output Testing.**
- **Validation Testing.**

Unit Testing

Unit testing focuses verification effort on the smallest unit of Software design that is the module. Unit testing exercises specific paths in a module's control structure to ensure complete coverage and maximum error detection. This test focuses on each module individually, ensuring that it functions properly as a unit. Hence, the naming is Unit Testing.During this testing, each module is tested individually and the module interfaces are verified for the consistency with design specification. All important processing path are tested for the expected results. All error handling paths are also tested.

Integration Testing

Integration testing addresses the issues associated with the dual problems of verification and program construction. After the software has been integrated a set of high order tests are conducted. The main objective in this testing process is to take unit tested modules and builds a program structure that has been dictated by design.

5.ALGORITHM

1. Machine Learning Algorithm

A total of twelve algorithms are used in this experimentalstudy. They are used in conjunction with the AdaBoost andmajority voting methods.Naïve Bayes (NB) uses the Bayes' theorem with strong ornaïve independence assumptions for classification. Certainfeatures of a class are assumed to be not correlated to others.It requires only a small training data set for estimating themeans and variances is needed for classification.The presentation of data in form of a tree structure is usefulfor ease of interpretation by users. The Decision Tree (DT) isa collection of nodes that creates decision on featuresconnected to certain classes. Every node represents a splittingrule for a feature. New nodes are established until the stoppingcriterion is met. The class label is determined based on themajority of samples that belong to a particular leaf. TheRandom Tree (RT) operates as a DT operator, with theexception that in each split, only a random subset of featuresis available. It learns from both nominal and numerical datasamples. The subset size is defined using a subset ratioparameter.The Random Forest (RF) creates an ensemble of random trees. The user sets the number of trees. The resulting model employs voting of all created trees to determine the final classification outcome. The Gradient Boosted Tree (GBT) is an ensemble of classification or

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

regression models. It uses forward-learning ensemble models, which obtain predictive results using gradually improved estimations. Boosting helps improve the tree accuracy. The Decision Stump (DS) generates a decision tree with a single split only. It can be used in classifying uneven data sets. The MLP network consists of at least three layers of nodes, i.e., input, hidden, and output. Each node uses a non-linear activation function, with the exception of the input nodes. It uses the supervised backpropagation algorithm for training. The version of MLP used in this study is able to adjust the learning rate and hidden layer size automatically during training. It uses an ensemble of networks trained in parallel with different rates and number of hidden units. The Feed-Forward Neural Network (NN) uses the backpropagation algorithm for training as well. The connections between the units do not form a directed cycle, and information only moves forward from the input nodes to the output nodes, through the hidden nodes. Deep Learning (DL) is based on an MLP network trained using a stochastic gradient descent with backpropagation. It contains a large number of hidden layers consisting of neurons with tanh, rectifier, and maxout activation functions. Every node captures a copy of the global model parameters on local data, and contributes periodically toward the global model using model averaging.

6.EXPERIMENTAL RESULTS

In this section, the experimental setup is firstly detailed. This is followed by a benchmark evaluation using a publicly available data set. The real-world credit card data set is then evaluated. All experiments have been conducted using RapidMiner Studio 7.6. The standard settings for all parameters in RapidMiner have been used. A 10-fold crossvalidation (CV) has been used in the experiments as it can reduce the bias associated with random sampling in the evaluation stage [23].In the credit card data set, the number of fraudulent transactions is usually a very small as compared with the total number of transactions. With a skewed data set, the resulting accuracy does not present an accurate representation of the system performance. Misclassifying a legitimate transaction causes poor customer services, and failing to detect fraud cases causes loss to the financial institution and customers. This data imbalance problem causes performance issues in machine learning algorithms. The class with the majority samples influences the results. Under-sampling has been used by Bhattacharyya et al. [6], Duman et al. [24], and Phua et al. [25] to handle data imbalance problems. As such, under-sampling is used in this paper to handle the skewed data set. While there is no best way of describing the true and false positives and negatives using one indicator, the best general measure is the Matthews Correlation Coefficient (MCC) [26]. MCC measures the quality of a two-class problem, which takes into account the true and false positives and

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

negatives. It is a balanced measure, even when the classes are from different sizes.

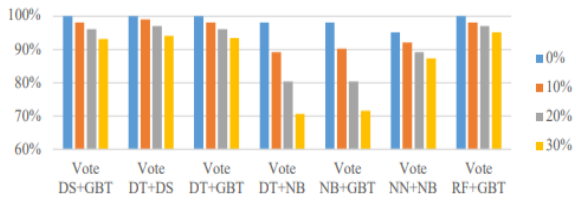


FIGURE 1. Fraud detection rates with different percentages of noise

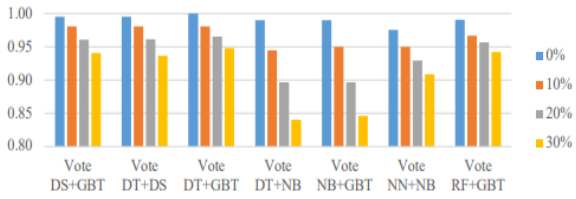


FIGURE 2. MCC scores with different percentages of noise

System Configuration

H/W System Configuration:

Processor - Pentium –III

Speed - 1.1 Ghz

RAM - 256 MB(min)

Hard Disk - 20 GB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

S/W System Configuration:

Operating System : Windows95/98/2000/XP/7

Application Server : Tomcat6.0/8.X

Front End : HTML, Java, Jsp

Scripts : JavaScript,jquery,ajax

Server side Script : Java Server Pages.

Database : Mysql

Database Connectivity : JDBC.

7.CONCLUSION

A study on credit card fraud detection using machinelearning algorithms has been presented in this paper. Anumber of standard models which include NB, SVM, and DLhave been used in the empirical evaluation. A publiclyavailable credit card data set has been used for evaluationusing individual (standard) models and hybrid models usingAdaBoost and majority voting combination methods. TheMCC metric has been adopted as a performance measure, as it takes into account the true and false positive and negativepredicted outcomes. The best MCC score is 0.823, achievedusing majority voting. A real credit card data set from afinancial institution has also been used for evaluation. Thesame individual and hybrid models have been employed. Aperfect MCC score of 1 has been achieved using AdaBoostand majority voting methods. To further evaluate the hybridmodels, noise from 10% to 30% has been added into the datasamples. The

2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

majority voting method has yielded the bestMCC score of 0.942 for 30% noise added to the data set.This shows that the majority voting method is stable inperformance in the presence of noise.

REFERENCES

[1] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," Expert Systems with Applications, vol. 40, no. 15, pp. 5916–5923, 2013.

[2] A. O. Adewumi and A. A. Akinyelu, "A survey of machine-learning and nature-inspired based credit card fraud detection techniques," International Journal of System Assurance Engineering and Management, vol. 8, pp. 937–953, 2017.

[3] A. Srivastava, A. Kundu, S. Sural, A. Majumdar, "Credit card fraud detection using hidden Markov model," IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37–48, 2008.

[4] The Nilson Report (October 2016) [Online]. Available: https://www.nilsonreport.com/upload/content_promo/The_Nilson_R eport_10-17-2016.pdf

[5] J. T. Quah, and M. Sriganesh, "Real-time credit card fraud detection using computational intelligence," Expert Systems with Applications, vol. 35, no. 4, pp. 1721–1732, 2008.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C., "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602–613, 2011. [

7] N. S. Halvaiee and M. K. Akbari, "A novel model for credit card fraud detection using Artificial Immune Systems," Applied Soft Computing, vol. 24, pp. 40–49, 2014.

[8] S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, "Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning," Information Fusion, vol. 10, no. 4, pp. 354–363, 2009.
[9] N. Mahmoudi and E. Duman, "Detecting credit card fraud by modified Fisher discriminant analysis," Expert Systems with Applications, vol. 42, no. 5, pp. 2510–2516, 2015.

[10] D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, "Association rules applied to credit card fraud detection," Expert Systems with Applications, vol. 36, no. 2, pp. 3630–3640, 2009.

[11] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," Expert Systems with Applications, vol. 38, no. 10, pp. 13057–13063, 2011.

[12] P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining

60techniques," Decision Support Systems, vol. 50, no. 2, pp. 491–500, 2011.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

[13] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," Expert Systems with Applications, vol. 32, no. 4, pp. 995–1003, 2007.

[14] F. H. Glancy and S. B. Yadav, "A computational model for financial reporting fraud detection," Decision Support Systems, vol. 50, no. 3, pp. 595–601, 2011.

[15] D. Olszewski, "Fraud detection using self-organizing map visualizing the user profiles," Knowledge-Based Systems, vol. 70, pp. 324–334, 2014.

[16] J. T. Quah and M. Sriganesh, "Real-time credit card fraud detection using computational intelligence," Expert Systems with Applications, vol. 35, no. 4, pp. 1721–1732, 2008.

[17] E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran," International Journal of Accounting Information Systems, vol. 25, pp. 1–17, 2017.

[18] I. T. Christou, M. Bakopoulos, T. Dimitriou, E. Amolochitis, S. Tsekeridou, and C. Dimitriadis, "Detecting fraud in online games of chance and lotteries," Expert Systems with Applications, vol. 38, no. 10, pp. 13158–13169, 2011.

[19] C. F. Tsai, "Combining cluster analysis with classifier ensembles to predict financial distress" Information Fusion, vol. 16, pp. 46–58, 2014.

[20] F. H. Chen, D. J. Chi, and J. Y. Zhu, "Application of Random Forest, Rough Set Theory, Decision Tree and Neural Network to Detect Financial Statement Fraud–Taking Corporate Governance into Consideration," In International Conference on Intelligent Computing, pp. 221–234, Springer, 2014.

[21] Y. Li, C. Yan, W. Liu, and M. Li, "A principle component analysisbased random forest with the potential nearest neighbor method for automobile insurance fraud identification," Applied Soft Computing, to be published. DOI: 10.1016/j.asoc.2017.07.027.

[22] S. Subudhi and S. Panigrahi, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," Journal of King Saud University-Computer and Information Sciences, to be published. DOI: 10.1016/j.jksuci.2017.09.010.

[23] M. Seera, C. P. Lim, K. S. Tan, and W. S. Liew, "Classification of transcranial Doppler signals using individual and ensemble recurrent neural networks," Neurocomputing, vol. 249, pp. 337-344, 2017.

[24] E. Duman, A. Buyukkaya, and I. Elikucuk, "A novel and successful credit card fraud detection system Implemented in a Turkish Bank," In IEEE 13th International Conference on Data Mining Workshops

(ICDMW), pp. 162–171, 2013.

[25] C. Phua, K. Smith-Miles, V. Lee, and R. Gayler, "Resilient identity crime detection," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 533–546, 2012.

[26] M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[27] Credit Card Fraud Detection [Online]. Available: https://www.kaggle.com/dalpozz/creditcardfraud

[28] R. Saia and S. Carta, "Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach," In Proceedings of the 14th International Joint Conference on eBusiness and Telecommunications, vol. 4, pp. 335–342, 2017.

[29] ISO 8583-1:2003 Financial transaction card originated messages [Online]. Available: https://www.iso.org/standard/31628.html