Credit Card Fraud Detection using Adaboost and Majority Voting

Mr. S. Siva Prakash M.E.,(Ph.D.),[1]
[1]Professor,
Department of Computer Science and Engineering.
K.S.R. College of Engineering,
Tiruchengode, India.

Ahubhakumar[2],S.Appash[3], J.Cibiragul[4]
[2,3,4] UG Students,
Department of Computer Science and Engineering.
K.S.R. College of Engineering,
Tiruchengode, India.

***Abstract*: - Credit card fraud is a serious problem in financial services. Billions of dollars are lost due to credit card fraud every year. There is a lack of research studies on analyzing real-world credit card data owing to confidentiality issues. In this paper, machine learning algorithms are used to detect credit card fraud. Standard models are firstly used. Then, hybrid methods which use AdaBoost and majority voting methods 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 the data samples to further assess the robustness of the algorithms. The experimental results positively indicate that 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. 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. 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. 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 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.

Loss from credit card fraud affects the merchants, where they bear all costs, including card issuer fees, charges, and administrative charges. 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.

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 WORK

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. 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. 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%.

A credit card fraud detection system was proposed in, 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.

Simulation results indicated a 98% true positive rate. A modified Fisher Discriminant function was used for credit card fraud detection in. 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.

B. HYBRID MODELS

Hybrid models are combination of multiple individual models. A hybrid model consisting of the Multilayer Perceptron (MLP) neural network, SVM, LOR, and Harmony Search (HS) optimization was used in to detect corporate tax evasion. HS was useful for finding the best parameters for the classification models. Using data from the food and textile sectors in Iran, the MLP with HS optimization acquired the highest accuracy rates at 90.07%. A hybrid clustering system with outlier detection capability was used to detect fraud in lottery and online games. The system aggregated online algorithms with statistical information from the input data to identify a number of fraud types. The training data set was compressed into the main memory while new data samples could be incrementally added into the stored datacubes. The system achieved a high detection rate at 98%, with a 0.1% false alarm rate.

To tackle financial distress, clustering and classifier ensemble methods were used to form hybrid models in. The SOM and k-means algorithms were used for clustering, while LOR, MLP, and DT were used for classification. Based on these methods, a total of 21 hybrid models with different combinations were created and evaluated with the data set. The SOM with the MLP classifier performed the best, yielding the highest prediction accuracy. An integration of multiple models, i.e. RF, DR, Roush Set Theory (RST), and back-propagation neural network was used in to build a fraud detection model for corporate financial statements. Company financial statements in period of 1998 to 2008 were used as the data set. The results showed that the hybrid model of RF and RST gave the highest classification accuracy.

3. PROBLEM STATEMENT

3.1 *EXISTING MODEL*

Three methods to detect fraud are presented.

Firstly, clustering model is used to classify the legal and fraudulent transaction using data parameter value.

Secondly, Gaussian mixture model past behavior and 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.

3.1.1 *Drawbacks*

The high amount of losses due to fraud and the awareness of the relation between loss and the available limit has to be reduced.

Testing credit card FDSs using real data set is a difficult task.

The fraud has to be deducted in real time and the number of false alert.

3.2 *PROPOSED SYSTEM*

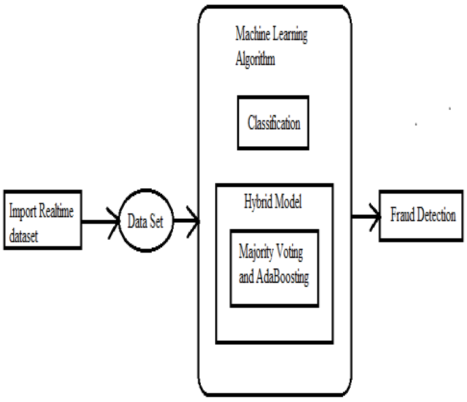
Total of twelve machine learning algorithms are used for detecting credit card fraud.

The algorithms range from standard neural networks to deep learning models.

In addition, the AdaBoost and majority voting methods are applied for forming hybrid models.

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.

SYSTEM ARCHITECTURE DIAGRAM – BASEPAPER



3.2.1 *Advantages Of Proposed System*

The system is very fast due to AdaBoost Technique.

Effective Majority Voting techniques.

4.0 METHODOLOGY

MACHINE LEARNING ALGORITHMS

A total of twelve algorithms are used in this experimental study. They are used in conjunction with the AdaBoos

ALGORITHMS

Naïve Bayes (NB) uses the Bayes' theorem with strong or naïve independence assumptions for classification. Certain features of a class are assumed to be not correlated to others. It requires only a small training data set for estimating the means and variances is needed for classification.

The presentation of data in form of a tree structure is useful

for ease of interpretation by users. The Decision Tree (DT) is a collection of nodes that creates decision on features connected to certain classes. Every node represents a splitting rule for a feature. New nodes are established until the stopping criterion is met. The class label is determined based on the majority of samples that belong to a particular leaf. The Random Tree (RT) operates as a DT operator, with the exception that in each split, only a random subset of features is available. It learns from both nominal and numerical data samples. The subset size is defined using a subset ratio parameter.

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 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.

Linear Regression (LIR) models the relationship between scalar variables by fitting a linear equation to the observed data. The relationships are modelled using linear predictor functions, with unknown model parameters estimated from the data set. The Akaike criterion, a measure of relative goodness of fit for a statistical model, is used for model selection. Logistic Regression (LOR) can handle data with both nominaland numerical features. It estimates the probability of a binary response based on one or more predictor features.

The SVM can tackle both classification and regression data. SVM builds a model by assigning new samples to one
category or another, creating a non-probabilistic binary linear classifier. It represents the data samples as points in the space mapped so such that the data samples of different categories can be separated by a margin as wide as possible. A summary of the strengths and limitations of the methods discussed earlier is given in Table I.

TABLE I
STRENGTHS AND LIMITATIONS OF MACHINE LEARNING METHODS

Model Strengths Limitations	Strengths	Limitations
Bayesian	Good for binary classification problems; efficient use of computational resources; suitable for real-time operations.	Need good understanding of typical and abnormal behaviors for different types of fraud cases
Trees	Easy to understand and implement; the procedures require a low computational power; suitable for realtime operations.	Potential of over-fitting if the training set does not represent the underlying domain information; retraining is required for new types of fraud cases.
Neural Network	Suitable for binary classification problems, and widely used for fraud detection.	Need a high computational power, unsuitable for real-time operations; re-training is required for new types of fraud cases.
Linear Regression	Provide optimal results when the relationship between independent and dependent variables are almost linear.	Sensitive to outliers and limited to numeric values only.
Linear Regression	Provide optimal results when the relationship between independent and dependent variables are almost linear.	Sensitive to outliers and limited to numeric values only.
Logistic Regression	Easy to implement, and historically used for fraud detection.	Poor classification performances as compared with other data mining methods.
Support Vector Machine	Able to solve non-linear classification problems; require a low computational power; suitable for real-time operations.	Not easy to process the results due to transformation of the input data.

6. 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 NB, SVM, and DL have been used in the empirical evaluation. A publicly available credit card data set has been used for evaluation using individual (standard) models and hybrid models using AdaBoost and majority voting combination methods. The MCC metric has been adopted as a performance measure, as it takes into account the true and false positive and negative predicted outcomes. The best MCC score is 0.823, achieved using majority voting. A real credit card data set from a financial institution has also been used for evaluation. The same individual and hybrid models have been employed. A perfect MCC score of 1 has been achieved using AdaBoost and majority voting methods. To further evaluate the hybrid models, noise from 10% to 30% has been added into the data samples. The majority voting method has yielded the best MCC score of 0.942 for 30% noise added to the data set. This shows that the majority voting method is stable in performance in the presence of noise.

For future work, the methods studied in this paper will be extended to online learning models. In addition, other online learning models will be investigated. The use of online learning will enable rapid detection of fraud cases, potentially in real-time. This in turn will help detect and prevent fraudulent transactions before they take place, which will reduce the number of losses incurred every day in the financial sector.

7. REFERENCE

[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_Report_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 techniques,” *Decision Support Systems*, vol. 50, no. 2, pp. 491–500, 2011.

[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.