International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT195274

Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection

Karthik R, Navinkumar R, Rammkumar U, Mothilal K. C.
Computer Science and Engineering, Sri Krishna College of Engineering & Technology, Coimbatore, Tamil Nadu, India

ABSTRACT

Cashless transactions such as online transactions, credit card transactions, and mobile wallet are becoming more popular in financial transactions nowadays. With increased number of such cashless transaction, number of fraudulent transactions is also increasing. Fraud can be distinguished by analyzing spending behavior of customers (users) from previous transaction data. Credit card fraud has highly imbalanced publicly available datasets. In this paper, we apply many supervised machine learning algorithms to detect credit card fraudulent transactions using a real-world dataset. Furthermore, we employ these algorithms to implement a super classifier using ensemble learning methods. We identify the most important variables that may lead to higher accuracy in credit card fraudulent transaction detection. Additionally, we compare and discuss the performance of various supervised machine learning algorithms that exist in literature against the super classifier that we implemented in this paper.

Keywords : Credit Card Fraud, Online Fraud, Cashless Transactions, Neural Network

I. INTRODUCTION

Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card or any similar payment mechanism as a fraudulent source of funds in a transaction. The purpose may be to obtain goods without paying, or to obtain unauthorized funds from an account. Credit card fraud is also an adjunct to identity theft. According to the Federal Trade Commission, while identity theft had been holding steady for the last few years, it saw a 21 percent increase in 2008. However, credit card fraud, that crime which most people associate with ID theft, decreased as a percentage of all ID theft complaints for the sixth year in a row. Despite the large amount of money lost to credit card fraud, it is actually quite rare an occurrence as a result of extensive countermeasures introduced since early 1990s. In 1999, out of 12 billion transactions made annually, approximately 10 million—or one out of every 1200 transactions—turned out to be fraudulent. Also, 0.04% (4 out of every 10,000) of all monthly active accounts was fraudulent. Even with tremendous volume and value increase in credit card transactions since then, these proportions have stayed the same or have decreased due to sophisticated fraud detection and prevention systems. Today's fraud detection systems are designed to prevent a mere one twelfth of one percent of all transactions processed which still translates into billions of dollars in losses.

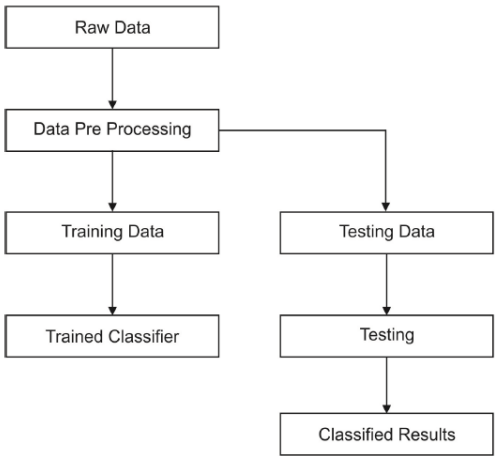

Figure 1. Behavior Based credit card Fraud Detection

The costs of card fraud in 2006 were 7 cents per 100 dollars worth of transactions (7 basis points). Due to the high volume of transactions this translates to billions of dollars. In 2006, fraud in the United Kingdom alone was estimated at £535 million, or US$750–830 million at prevailing 2006 exchange rates. The rest of this paper is organized as follows. Section II describes our existing method deals with street light power minimization. Proposed system design is provided in Section III. Chapter IV describes the result and discussions part. Finally, the conclusion is drawn and future work is proposed in Section IV.

II. PROBLEM IDENTIFICATION

The most important moral issue in the credit card trade is fraud involvement. The main aspires are, primarily, to recognize the different types of credit card fraud, and, secondly, to evaluate unconventional techniques that have been used in fraud detection. The sub-aim is to present, compare and examine recently published discovering in credit card fraud detection. Credit card fraud detection has developed a number of techniques via bunch of investigate interest and, with special importance on, data mining and distributed data mining have been recommended. Also there is a problem is, first the observation data possibly will be missing for a number of intervals. Following that there are multiple observation streams that are not necessarily synchronous to each other and possibly will have different emission distributions" for the same state. So in proposed research we are using multiple observation sequences which are associated with the semi hidden state sequence and these observations may not be synchronized to each other. We divide a large data set of labelled transactions (either fraudulent or legitimate) into smaller subsets by applying distributed data mining techniques to generate classifiers in parallel, and come together the resultant base models by meta learning from the classifiers' performance to produce a meta classifier. in addition extensibility, combining multiple models computed over all available data produces meta classifiers that can counterbalance the loss of predictive presentation that usually occurs when mining from data subsets or sampling. Furthermore, when we use the learned classifiers (for example, during transaction authorization), the base classifiers can carry out in parallel, with the meta classifier then combining their results. So, our approach is highly efficient in generating these models and also relatively efficient in applying them.

III. LITERATURE REVIEW

Ghosh and Reilly [9] used three-layer feed forward Neural network to detect frauds in 1994. The Neural Network was trained on examples of fraud containing stolen cards, application fraud, counterfeit fraud, Non Received Issue (NRI) fraud, and mail order fraud.

Abhinav and Amlan [7] proposed a Hidden Markov Model to detect the frauds in credit cards. Proposed Model does not require fraud signatures and still it can detect frauds by considering a cardholder's spending habit. This system is also scalable to handle large number of transactions.

Y. Sahin and E. Duman [6] proposed approach to detect credit card fraud by decision tree and Support Vector Machine. Performance of classifier models of various decision tree methods (C5.0, C&RT and CHAID) and a number of different SVM methods (SVM with polynomial, sigmoid, linear and RBF kernel functions) are compared in this study. An approach is proposed towards fraud detection in banking transactions in [2] using fuzzy clustering and neural network. In this approach fraud detection is

done in three phase. First phase is initial user authentication and verification of card details.

After successfully completing this phase, fuzzy cmeans clustering algorithm is performed to find out normal usage behavior of user based on past transactions. If new transaction is found to be doubtful in this phase, mechanism based on neural network based is applied to determine whether it was actually fraudulent transaction or not.

Kang Fu, Dawei Cheng, Yi Tu, and Liqing Zhang at [3] proposed a convolutional neural network (CNN) based approach to find fraudulent transactions. Convolutional Neural Network is a part of deep learning and is a type of feed-forward Neural Network that consists of more than one hidden layer. In this paper, for finding more complex fraud patterns and to improve classification accuracy, a new feature trading entropy is proposed. To relieve the problem of the imbalanced dataset, cost based sampling method is used to generate more number of frauds. Generally, CNN is used for image recognition, Character recognition, image processing, video recognition and recommender system. In this paper for the first time, CNN is used to detect frauds.

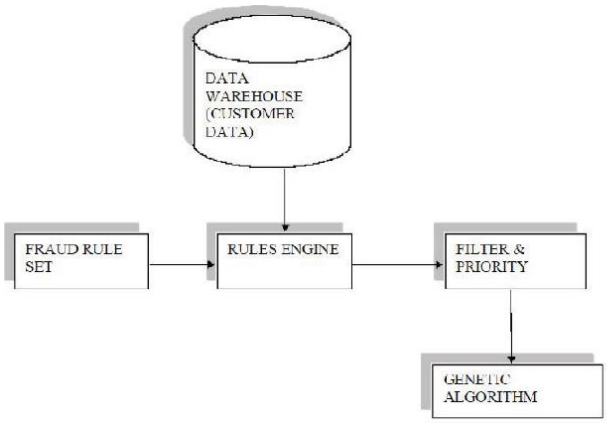

Figure 2. Architecture for fraud detection using genetic algorithm

In this chapter of literature review, we analyzed several papers that are formerly involved in credit card fraud detection. However multiple observation streams of these systems do not have necessity of being synchronous to each other and lack in effective prediction of fraud detection. This detecting efficiency is improved by using three variants of Hidden Markov Model which are mentioned in upcoming chapter.

IV. SYSTEM DESIGN

1. Supervised learning and unsupervised learning

Using supervised method helps to find out the label on past transaction, they tend to not recognized fraud pattern that has occurred in the past [23], [24]. While unsupervised technique helps to find out the class of transactions [22].

2. Unbalanced data

It is quite challenging to learn from an unbalanced dataset and for balancing it, the sampling method used. A publicly available dataset that contains 284,807 transactions made in Sep. 2013 by European cardholders [25], [11]. The dataset includes 492 fraud transactions, which is highly imbalanced. Hence, under-sampling was applied [14].

3. Pre processing

Initially, the attributes used in the dataset are converted into numerical data. Feature selection is a very important stage in fraud detection. The features in the data efficiently portray the usage behavior of an individual. In this model, the features which interpret the behavior of the customer are selected for detection. Adding irreverent features make the classifier inefficient. Transaction amount is the most important behavior it varies from person to person. Frequency of card usage is calculated from the Date and Time Attributes. Average amount of transactions are calculated from each transactions.

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then the input data will be transformed into a reduced representation set of features. Transforming the input data into the set of features is called feature extraction. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced

representation instead of the full sizeinput. Feature extraction is special form dimensionality reduction. Here, the features represent the relevant characteristics of the input data are chosen. Instead of using full size input one may use this reduced representation set. If it is properly chosen then it will give successful task. Best results are achieved when an expert constructs a set of application-dependent features. Nevertheless, if no such expert knowledge is available general dimensionality reduction techniques may help.

4. Fraud Detection Classifier

Logistic Regression can handle the data with theoretical and statistical characteristics. Decision Tree is a supervised learning method that widely uses models for classification and regression tasks [26]. Random Forest method used for classification and regression using the collection of the decision tree, each one is slightly different from each other [8], [6]. With first introduction in 1995 Navies Bayes using Bayes theorem for independence hypothesis[27]. K-Nearest Neighborhood (KNN) is a necessary calculation which stores every single accessible occurrence [28]. The Gradient Boosted Tree Classifier (GBT) is a collection of classification and regression models. Boosting supports improve the tree accuracy.

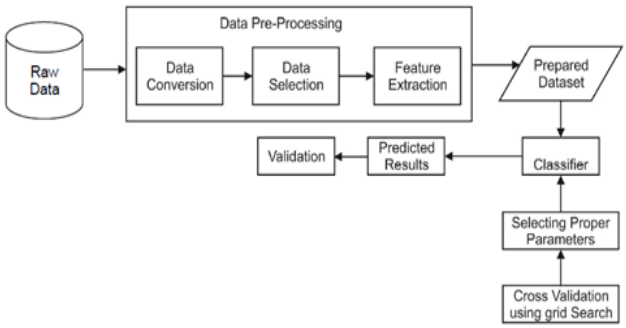

Figure 2. Proposed system architecture

XGB (XG boost Classifier) is the most refined classifier that works with all type of dataset. The support vector machines (SVM) are initially presented in 1995, and they have been observed to be extremely fruitful in an assortment of exemplary classification tasks [6]. Figure 2 and 3 describe the proposed system algorithm implementation.

1. Obtain transaction data
2. Preprocess data to convert categorical attributes to numerical attributes
3. Normalize the numerical data using Min-Max Normalization
4. Create training and testing data files using the SVM format
5. Select an attribute a1
6. Apply K-Means Clustering with respect to the attribute a1
7. Select the second level attribute a2
8. For every cluster c obtained
    a. Apply K-Means Clustering with respect to the attribute a2
9. End for
10. Supply the training file to SVM
11. Set the values for C and □
12. Obtain results using the current C and □ pair
13. Perform step 6 and 7 till satisfactory results are obtained from the training set
14. Test the accuracy using the test file
15. For every malicious transaction obtained,
    a. Find clusters that contains the current transaction
    b. Using collective animal behavior, check for a similar pattern in the clusters
    c. If the pattern similarity exceeds the threshold t consider the transaction as normal
    d. Else
    e. Consider the transaction as malicious
16. End for

Figure 3. Pseudo code for proposed system algorithm

The MLP organize comprises of no less than three layers of hubs, i.e., input, covered up, and yield [26]. Ensemble learning (also known as meta-classifier) helps to improve the results by combining multiple machine learning classifiers to improve the predictive outcomes. Accuracy is one important method to compare the performance of classification models we also look at the other factors like F1-Score, Precision, TPR, FPR, Recall, G-mean and Specificity. All these

evaluations measures adequately reveal validation of the study very well. Proposed system naïve bayes merged with random forest to achieved highest classification accuracy.

V. RESULTS AND DISCUSSION

This process is evaluated with various sets of data containing different number of data items and the obtained values are recorded in a confusion matrix.

		Predicted	
		Positive	Negative
Actual	Positive	TP	FP
	Negative	TN	FN

Table 1 Confusion Matrix

In this work,

- TP shows the number of genuine transactions correctly identified as non fraudulent.
- FP gives the number of genuine transactions incorrectly identified as fraudulent.
- TN is shows the number of fraudulent transactions correctly identified as fraudulent.
- FN is mistakenly considering fraudulent transaction as genuine.

The two performance measures, sensitivity and specificity are used for evaluating the results. Sensitivity is the accuracy on the positive instances.

$$\text{Sensitivity} = TP/(TP+FN),$$

Sensitivity represents the ratio of positive class that was correctly identified. Specificity is the accuracy of the negative instances

$$\text{Specificity} = TN/(TN+FB)$$

Specificity represents the ratio of the negative cases that was incorrectly identified as positive.

We used 70% of the data is used for training and 30% used for the testing set. Data was balanced by using an under-sampling technique. So, we used Accuracy, F1-Score, Recall, Precision, G-Mean, FPR, TRP and specificity are used to compare the models. Table 1 shows all classifier results and comparisons. In table 1, stacking classifier (0.9527 accuracies) is leading the other classifiers, followed by the random forest (0.94594 accuracies) and XGB classifier (0.94594 accuracies) is helpful only when we have a symmetric dataset. Having a high precision is related to the low false rate. In Figure (1) Random forest, stacking and XGB classifier all have the same precision score of 0.95 followed by the Gradient boosting and logistic regression with the precision score of 0.94. We find out recall also developed the same ranking of precision in Figure (2). The F1-score is the weighted median of precision and recall, and its score take false positive and false negative into account F1-score. F1-score also followed the same ranking of Precision and Recall in Figure (3). SVM has the highest ranking with 0.5360 FPR, and stacking classifier has the lowest ranking with 0.0335 in Figure (4). TPR of the logistic regression has the highest ranking followed by the MLP and stacking classifier. We find out the top five features in Features 14 is the essential features and features and got selected by all algorithms. And V4 is decided by four features.

Table 2 Dataset With 100 % Fraud Data

SUBJECT	KNN	SVM	LOGISTIC REGRESSION	HYBRID NB-RF
Accuracy	0.6704	0.9517	0.9429	0.9675
95% CI	0.6437, .6963	0.9383, .9628	0.9287, 0.9551	0.9562, .9766
NoInformation	0.5111	0.5349	0.5182	0.5182
P-Value	<2e-16	< 2e-16	< 2.2e-16	< 2.2e-16
Kappa	0.3399	0.9031	0.8855	0.9348
Test P-Value	0.9609	0.01045	0.0006316	5.806e-07
Sensitivity	0.6595	0.9659	0.9161	0.9391
Specificity	0.6804	0.9393	0.9679	0.9939
Pos PredValue	0.6574	0.9326	0.9637	0.9930
Neg Pred	0.6825	0.9694	0.9254	0.9461
Prevalence	0.4818	0.4651	0.4818	0.4818
Detection Rate	0.3177	0.4493	0.4414	0.4525
Prevalence	0.4834	0.4818	0.4580	0.4556
Accuracy	0.6700	0.9526	0.9420	0.9665
'Positive'Class	Yes	Yes	Yes	Yes
AUC	0.670	0.953	0.945	0.970
Precision	0.6672	0.9523	0.9643	0.9915
F-measure	0.6633	0.9590	0.9394	0.9645

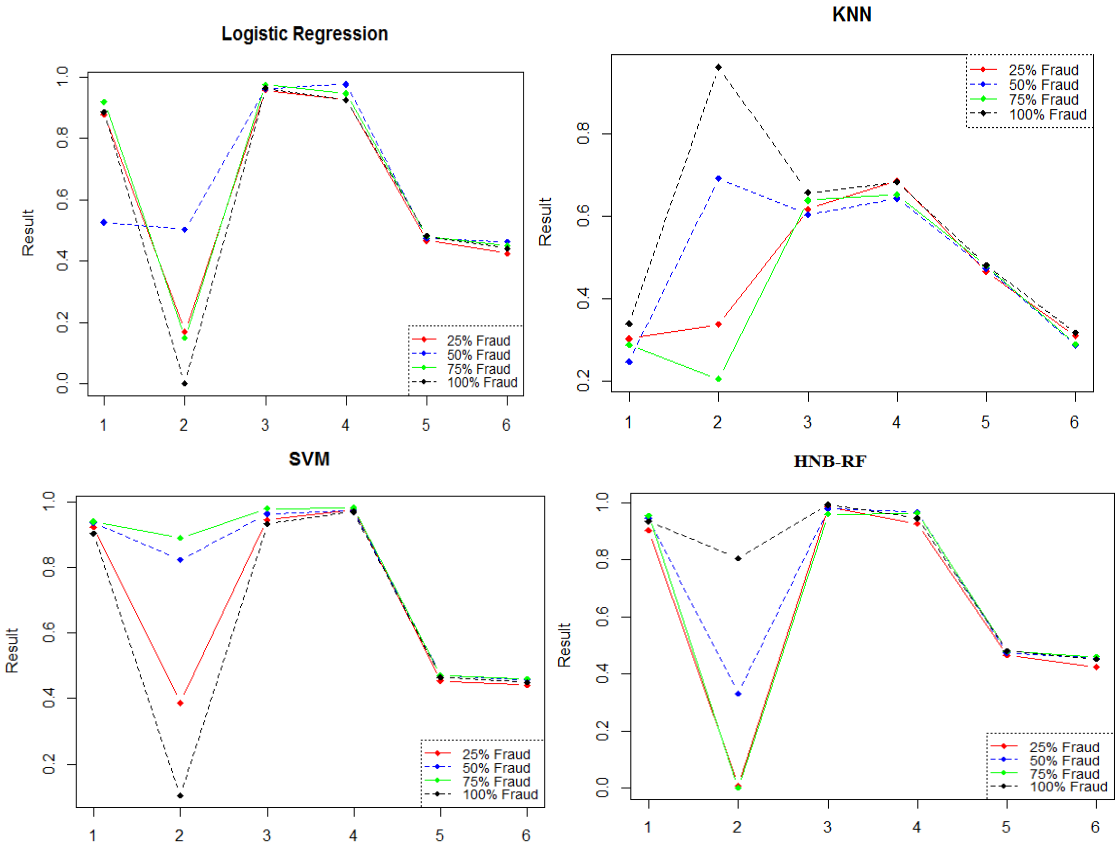

Figure 4. TPR and FPR performance of all the classifier

The hybrid approach. During the initial stages, when the number of entries are minimal, the plots point to 0,0 and 0,1 points. As the number of entries keeps increasing, the plotted points are clustered towards the northwest corner and are above the diagonal. This proves that this process provides a high level of accuracy, almost meeting the perfect standard of 0,1.Here, the values of the F-Measure show a rate of 0.869341figure 4 and table 2 clearly explain Proposed System Achieved Better accuracy rate of 0.96 is obtained. Hence it is proved that this process shows a higher accuracy rate and better performance.

VI. CONCLUSION

Under-sampling is done for balancing the unbalanced dataset. The learning model's evaluation is based on their accuracy, recall, precision, TPR, FPR, specificity and G-mean. The result of all the purposed models were superior in overall performance. Overall results show that stacking classifier which is used LR as meta classifier is most promising for predicting fraud transaction in the dataset, followed by the SVM, LR, KNN and HNB-RF classifier. Future work will be conducting the using the voting classifier and check the performance with other ML learning methods, increase the size of training and testing dataset.

VII.REFERENCES

[1]. Statista the statistic portal (2017, March 14) available https://www.statista.com/topics/871/online-shopping/

[2]. Tanmay Kumar Behera, Suvasini Panigrahi, "Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network", IEEE Computer Society, 2015

[3]. Kang Fu, Dawei Cheng, Yi Tu, and Liqing Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks", Springer International Publishing AG 2016.

[4]. Smt.S.Rajani, Prof.M. Padmavathamma, "A Model for Rule Based Fraud Detection in telecommunications", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 5, July –2012.

[5]. Hidden Markov model (2017, March 15) available https://en.wikipedia.org/wiki/Hidden_Markov_ model [6] Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines", IMECS vol 1, 2011.

[6]. Abhinav Srivastava, Amlan Kundu, Shamik Sural, Senior Member, IEEE, and Arun K. Majumdar, Senior Member, IEEE , "Credit Card Fraud Detection Using Hidden Markov Model" , IEEE transactions on dependable and secure computing, vol. 5, no. 1, january-march 2008.

[7]. Michael Nielsen (2017, March 15), Deep learning available http://neuralnetworksanddeeplearning.com/chap6.html [9] Ghosh, S., Reilly, D.L.: Credit card fraud detection with a neuralnetwork. In: Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, 1994, vol. 3, pp. 621–630. IEEE (1994)

[8]. Emin Aleskerov, Bernd fieisleben and Bharat Rao, "CARDWATCH: A Neural Ntwork based database Mining System for Credit Card Fraud Detection"

[9]. Sam Maes, Karl Tuyls, Bram Vanschoenwinkel, Bernard Manderick, "Credit card fraud detection using bayesian and neural networks", International Naiso Congress on Neuro Fuzzy Technology, 2002.

[10]. Minewiskan, Microsoft Neural Network Algorithm Technical Reference (2017, March 14) available at https://docs.microsoft.com/enus/ sql/analysis-

services/data-mining/microsoft-neural-networkalgorithm- technical-reference

[11]. Krishna Modi, Bhavesh Oza, "Outlier Analysis Approaches in Data Mining", IJIRT vol 3 issue 7. [14] Raghavendra Patidar, Lokesh Sharma, "Credit card fraud detection using Neural Network", IJSCE Volume-1, Issue-NCAI2011, June 2011.

[12]. Alejandro Correa Bahnsen, Djamila Aouada, Aleksandar Stojanovic, Björn Ottersten, "Feature engineering strategies for credit card fraud detection", 0957-4174/ 2016 Elsevier.

Cite this article as :

Karthik R, Navinkumar R, Rammkumar U, Mothilal K. C., "Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 394-401, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT195274 Journal URL : http://ijsrcseit.com/CSEIT195274