Analysis of Various Credit Card Fraud Detection Techniques

Heena Kochhar[1]
[1] Department of Computer Science & Engineering, CT University Ludhiana, Punjab, India

Abstract: Data mining is a technique that is applied to mine valuable information from the rough data. A prediction analysis is an approach that has the potential for forecasting future possibilities based on the recent data. The CCFD is the challenge of prediction in which fraudulent transactions are predicted based on certain rules. There are several stages included in the detection of fraud in credit cards. Various classification algorithms are reviewed with respect to the performance analysis in order to detect fraud in the credit card. The performance is measured with regard to precision.

Keywords: *Naive Bayes, Credit card, Logistic regression, random forest, K-nearest neighbor*

1. Introduction

Credit cards are commonly utilized to purchase several goods and access to several services in our daily lives. From providing a successful payment method, the cards are provided by the cardholder to the merchant during the purchasing method's execution based on a physical-card. By stealing a credit card, a fraudulent attack has been conducted through the attacker. The credit card company can also face a huge loss if the cardholder is not aware of losing it. For performing any fraudulent transaction in online transactions, the attacker requires a significant handle future fewer data. The internet and telephones are used for purchasing products and services online. For reducing the rate of successful credit card fraud, it is important to introduce fraud detection methods. This method is proposed based on the purchase information of the particular cardholder. It is possible to know that the card is stolen with the help of recognizing patterns in which the transactions were done by the user over the past few years [1]. Business organizations today face credit card frauds, which are the biggest threat today. It is important to perceive the actions which have resulted in causing fraud. To commit fraud, the fraudsters utilize different methods. It is usually not possible for the owner to recognize that the card is lost. Only when an unauthorized user uses the credit card for personal reasons, can the authorized user know about the fraud. It is also not possible to track down the individual who is committing the fraud. The various classification techniques have carried out to detect credit card fraud are described below: -

1.1 Logistic Regression (LR): is a kind of generalized linear model. It is not apposite to implement a simple linear regression if the variable is binary, which will be predicted because of normality assumptions. [2].

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021
DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

1.2 Decision Trees (DT) has a structure of a tree in which a test is denoted through every node on a feature, and every branch is employed to reveal a result obtained in testing. Hence, the observations are split into mutually exclusive subgroups by the tree [3].

1.3 Neural Networks (NN): is a mature technology that has an established theory and recognized application areas. Numerous neurons are composed in these networks. The weight is a numerical value that is related to every connection [3].

1.4 Support Vector Machines (SVM): has employed a linear model for applying nonlinear class boundaries. To achieve this, input vectors are mapped in a nonlinear way into a high-dimensional feature space. The development of an optimal separating hyperplane is carried out within a novel space [4].

1.5 Bayesian belief network (BBN) : facilitates to demonstration of the dependencies among subsets of features. It is a directed acyclic graph in which every node is utilized to show an attribute, and a probabilistic dependence is shown by every arrow [5].

1.6 K-nearest neighbor (KNN) : is carried in the systems which are executed for detection. The KNN is proved efficient in CCFD systems with the utilization of supervised learning schemes. [6].

1.7 Hidden Markov Model (HMM) : is different from the normal statistical Markov model as it contains invisible states; however, a visible form is produced through every state at random. A hidden Markov model is represented as the simplest dynamic Bayesian network [7].

1.8 Artificial neural networks (ANN): are initially constructed to impersonate the nature of the human brain. A NN is the association of elementary objects recognized as the simple neuron [8].

2. Literature Review

It is stated that [9] that with the help of CFD techniques, one outcome of the classifier was evaluated. By using the technique named principal component analysis, twenty-eight prime components of the data were obtained. The method of PCA was applied to variable time, amount, and real data. For performing the experiments, four kinds of under-sampling schemes and five types of oversampling methods were used. The evaluation of the different methods was done by using the performance metrics. The tested results showed that the cost-effective support vector machines performed well compared to others during under-sampling. While oversampling, bagging classifiers performed better than others. It is reported that [10] proposed a technique named CCFD by means of ML. After the use of Adaboost and majority voting methods, hybrid models came into the picture. The evaluation of the model efficiency was done based on data sets collected from different financial institutions. Various theoretical results showed that most voting techniques achieved good precision rates for detecting credit card fraud through the experiments conducted. [11] proposed various fraud methods, as well as techniques employed for detecting and

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021

DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

preventing fraud. The major aim is that the techniques and methods have to be recognized to obtain the best results. The results demonstrated that the hybrid forms of detecting fraud are commonly implemented because, in these methods, the potential of various conventional schemes is integrated. The future work would emphasize analyzing the CCF for enhancing the existing approaches, including a hybrid model that is both able to handle imbalanced dataset, and the real-time problem, to have a response during the financial transaction runtime having enhanced precision. With handleFutureJohn [15] stated that for the detection of extremely distorted credit fraud data, LR, NB, and KNN were compared. For the investigation of credit card transactions, the dataset was taken from European cardholders. The technique of oversampling and undersampling applied to the available data. On the unprocessed and preprocessed data, three techniques were used. Python was used for the implementation of this work. These techniques were compared based on several performance parameters. The tested outcomes showed that the performance of k-nearest was better than the performances of the other two techniques

3. Motivation

Credit card fraud is the major issue that comprises the security of the transactions. The attackers do credit card fraud by stealing information on credit cards. The approach of prediction analysis can detect credit card fraud. In the earlier research, the voting classifier has applied for the prediction analysis. The voting approach has integrated various classification algorithms that affect the accuracy and also increase execution time. The models designed earlier for CCFD perform well on small datasets, but the accuracy gets reduced when the dataset size increases. The credit card fraud detection models need to handle a large quantity of historical information to train the model. While handling such a large amount of data, the model's execution time is increased, affecting performance.[10]. The technique is required for credit card fraud detection, which gave high accuracy in the least amount of time

4. Proposed Methodology

The dataset, which is carried out in the base paper, and the classification algorithms named NB, RF, and LR have been defined. These algorithms are produced under various phases in which the data is gathered, pre-processed, and analyzed; classification techniques are trained and tested. The conversion of data is done into a usable format under the pre-processing phase. A hybrid of under-sampling and oversampling is applied for attaining two sets of data distributions. The Principal Component Analysis is employed for performing feature selection and feature reduction in the analysis phase. The training phase includes the construction of classification techniques and is utilized with the processed data. The computation of experiments is done concerning TPR, TNR, FPR, etc. The performance comparison of these techniques is quantified in terms of different evaluation parameters.

4.1Read Dataset

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021

DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

The dataset is available insourced from [15]. The payments of credit card done through European cardholders in September 2013 are comprised in this dataset. The transactions made in 2 days and further include 284,807 transactions have been described in this dataset. The positive class has utilized the the0.172% of data. The dataset is unbalanced and skewed in the positive class. The input variables are available in only numerical form as these values have resulted from the PCA. Therefore, thirty input attributes have been carried out in it. The details and background information regarding the attributes are not defined as there are some confidential issues. The seconds elapsed amid every transaction, and the first transaction in the dataset is included in the time attribute. The amount attribute is the amount of the payment. The feature class is a destination class employs to perform the binary classification and provides the value 1 in the positive case situation and 0 for the non-fraud case.

Data Pre-Processing

Data pre-processing is implemented on the data. The exploitation of a hybrid of under-sampling and over-sampling is done on a too unbalanced dataset for obtaining two sets of distribution to perform the analysis. The stepwise addition and subtraction of a data point are interpolated among existing data points until the over-fitting threshold is reached.

4.2 Classification

1. Naïve Bayes Classifier: This is a Bayes, theorem based statistical algorithm. This algorithm performs decision making according to the maximal probability. Bayesian probability makes use of given values for estimating indefinite probabilities. This algorithm enables past knowledge and logic to be implemented to indefinite descriptions. This algorithm assumes the conditional independence of features within the data. This classification model follows the idea of conditional probabilities of the two classes called fraudulent and non-fraudulent.

$$P\{ci|fk\} = p(fk|ci) * p(ci)$$
$$P\{fk|ci\} = p(fk|ci) k = 1,2 \ldots . n; i = 1,2)$$

Here, n corresponds to the maximal number of attributes (30). $P(ci|fk)$ signifies the probability of feature value fk to be in class. $P(fk|ci)$ refers to the probability that generates the feature value of a known class. $p(ci)$ is the probability of incidence of class. Also, $p(fk)$ depicts the probability of the incidence of feature value fk.

Result is C1 when $P(fk) > P(c2|fk)$

Result is C2 when $P(fk) < P(c2|fk)$

The target class for classification is denoted by Ci. Moreover, C1 and C2 represent non-fraudulent and fraudulent cases, respectively [15]

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021
DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

2. Logistic Regression Classifier: Logistic regression extracts some weighted features from the input, takes, logs, and combines them linearly. This implies that each feature after multiplying with weight is added. The basic difference between naive Bayes and logistic regression is that logistic regression refers to a discriminative classification model, whereas the naive Bayes is a generative classification model. This is a sort of regression models. This model fits data to a logistic function that predicts the probability of the incidence of an episode. Logistic regression uses several predictor variables. These variables can be categorical or numerical

The logistic regression hypothesis is defined as:

$$h_\theta(x) = g(\theta^T x)$$

Where the function g is a sigmoid function defined as:

$$g(z) = \frac{1}{1+e^{-z}}$$

The sigmoid function has unique properties that result in the values in the range [0,1]. The cost function for logistic regression is given as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \quad [-y^{(i)}\ log\ log\ \left(h_e\left(x^{(i)}\right)\right) - \left(1-y^{(i)}\right) log\ (1-\left(x^{(i)}\right))]$$

An integrated function called is used for finding the minimal of this cost function in machine learning. This discovers the optimal metric θ for the logistic regression cost function given a fixed dataset (of x and y values). The parameters refer to the initial values of the parameters that need to be optimized and a function that, when given the training set and a particularθ, computes the logistic regression cost and gradient with respect to θ for the dataset with x and y values. The final θ value is employed for plotting the decision boundary of the training data [15].

3. AdaBoost: Adaptive Boosting or AdaBoost is used in conjunction with different algorithms to improve their performance. The outputs are combined by using a weighted sum, which represents the combined output of the boosted classifier, i.e.,

$F\mathrm{t}(x) = \sum_{t=1}^{t} \quad f\mathrm{t}(x)$ (6)

Where every ft is a classifier (weak learner) that returns the predicted class with respect to input x. Each weak learner gives an output prediction, h(xi), for every training sample. In every iteration t, the weak learner is chosen, and is allotted a coefficient, αt, so that the training error sum, Et, of the resulting t-stage boosted classifier is minimized,

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021
DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

$Et=\sum [FKO0(x'') + \alpha Kh(x'')]$ (7)

In whichFt–1(x) is a boosted classifier that is constructed in the preceding phase, E(F) represents the error function, and ft(x) = αth(x) is a weak learner which is considered for the final classifier. The weak learners are squeezed with respect to misclassified data samples by AdaBoost. But it has susceptibility against noise and outliers.

AdaBoost has the potential for enhancing the individual outcomes from diverse algorithm till the classifier performs randomly [14′]

5. Result and Discussion

There are four basic metrics carried out to compute the experiments based on TPR, TNR, FPR, and FNR rates metrics.

$$\text{TPR} = \frac{TP}{P}$$

$$TNR = \frac{TN}{N}$$

$$FPR = \frac{FP}{N}$$

In which TP, TN, FP, and FN denote the number of true positive, true negative, false positive, and false-negative test cases classified whereas P represents the total positive class cases, and N denotes the total number of negative class cases under test.TPs are the cases that are identified as positive and positive in actual. Whereas, TNs are the cases which are identified as true negative. FP is determined as positive, but they are negative also.FNs are cases which are defined as negative but are truly positive also. The performance of NB, KNN, and LR classifiers is computed based on a variety of evaluation metrics such as accuracy, MCC, etc. These parameters are deployed based on their relevance for the quantification of an imbalanced binary classification issue.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$Specificity = \frac{TN}{FP+TN}$$

$$Precision \frac{TP}{TP+FP}$$

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021

DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

The precision in cases classified as positive is provided through sensitivity. The accuracy of negative cases is obtained from the specificity. The precision provides the accuracy of fraud case classification. The MCC is an evaluation parameter utilized to deal with the issues of binary classification. This metric is carried out basically with the unbalanced datasets as true positive, false positive, true negative, and false negative is composed in its evaluation. It takes the value which lies between -1 and +1. The excellent classification is denoted with the; a +1, whereas, the total distinction between classification and observation is revealed using a -1. The average of sensitivity and specificity that is the portion of negatives categorized as negatives is illustrated through a balanced classification rate [13]

Table 1: Performance Analysis

CLASSIFIERS	ACCURACY	SENSITIVITY	F1
LOGISTIC REGRESSION	99.92%	0.77	0.83
RANDOM FOREST	99.95%	0.77	0.83
ADABOOST	99.91%	0.59	0.68
NAÏVE-BAYES	97.88%	0.79	0.11

6. Conclusion

This work makes use of various classification models for detecting scams related to a credit card. The dataset used in this work is divided into two subsets of training and testing. These models are evaluated by using a training subset that consists of 60% of the entire dataset. The rest of the part of the original dataset is used to validate and test these models. The efficiency of the four classification models is evaluated in terms of certain parameters. Dataset greatly improves the performance of binary classification. In the future, a hybrid classification method will be designed to detect credit card scams.

References

[1] Jain, R., Gour, B., & Dubey, S. (2016). A hybrid approach for credit card fraud detection using a rough set and decision tree technique. International Journal of Computer Applications, 139(10), 1-6.

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021

DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**

[2] Viaene, S., Ayuso, M., Guillen, M., Van Gheel, D., Dedenne, G. (2007). Strategies for detecting fraudulent claims in the automobile insurance industry. European Journal of Operational Research, 176(1), 565-583.

[3] Kirkos, E., Spathis, C., Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. Expert systems with applications, 32(4), 995-1003.

[4] Ravisankar, P., Ravi, V., Rao, G. R., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques.Decision Support Systems, 50(2), 491-500.

[5] Li, C., Poskitt, D. S., & Zhao, X. (2019). The bivariate probit model, maximum likelihood estimation, true pseudo parameters, and partial identification. Journal of Econometrics, 209(1), 94-113.

[6] Hoogs, B., Kiehl, T., Lacomb, C., &Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement

[7] Dai, Y., Yan, J., Tang, X., Zhao, H., & Guo, M. (2016, August). Online credit card fraud detection: A hybrid framework with big data technologies. In 2016IEEE Trustcom/BigDataSE/ISPA (pp. 1644-1651). IEEE.

[8] Mubarek, A. M., &Adalı, E. (2017, October). Multilayer perceptron neural network technique for fraud detection. In the 2017 International Conference on Computer Science and Engineering (UBMK) (pp. 383-387). IEEE.

[9] Sisodia, D., &S.Bhandari, N. r. (2017). Performance evaluation of class balancing techniques for credit card fraud detection. IEEE International conference on power,control,signals and instrumentation ,(Chennai), 2747-2752.

[10] Randhawa, K., loo, c. k., manjeevan seera, c. p. (2018). Credit card fraud detection using Adaboost and majority voting. IEEE, 14277-14284.

[11]. Sadi Gali, I., Sael, N., Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia computer science, 148, 45-54.

[12] majority voting, c. c. (2017). Kuldeep Randhawa; chu king loo; IEEE,mnjeevanseera. IEEE.

[13] awoyemi, j. o., adetunmbi, a. o. (2017). Credit card fraud detection using machine learning techniques.IEEE, 978-15090-3/17.

Received: August 23, 2020 Revised: October 19, 2020 Accepted: Jan 15, 2021

DOI: 10.5281/zenodo.4706137 **Special issue on Computational Intelligence based Techniques for COVID-19 Benefits and Challenges**