International Journal of Pure and Applied Mathematics
Volume 118 No. 20 2018, 825-838
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models

[1]Navanshu Khare and [2]Saad Yunus Sait
[1]Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Kattankulathur Campus,
Chennai, Tamil Nadu.
[2]Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Kattankulathur Campus,
Chennai, Tamil Nadu.

Abstract

Finance fraud is a growing problem with far consequences in the financial industry and while many techniques have been discovered. Data mining has been successfully applied to finance databases to automate analysis of huge volumes of complex data. Data mining has also played a salient role in the detection of credit card fraud in online transactions. Fraud detection in credit card is a data mining problem, It becomes challenging due to two major reasons–first, the profiles of normal and fraudulent behaviors change frequently and secondly due to reason that credit card fraud data sets are highly skewed. This paper investigates and checks the performance of Decision tree, Random Forest, SVM and logistic regression on highly skewed credit card fraud data. Dataset of credit card transactions is sourced from European cardholders containing 284,786 transactions. These techniques are applied on the raw and preprocessed data. The performance of the techniques is evaluated based on accuracy, sensitivity, specificity, precision. The results indicate about the optimal accuracy for logistic regression, decision tree, Random Forest and SVM classifiers are 97.7%, 95.5% and 98.6%, 97.5% respectively.

Key Words: Fraud in credit card, data mining, logistic regression, decision tree, SVM, random forest, collative analysis.

1. Introduction

Financial fraud is a growing concern with far reaching consequences in the government, corporate organizations, finance industry, In Today's world high dependency on internet technology has enjoyed increased credit card transactions but credit card fraud had also accelerated as online and offline transaction. As credit card transactions become a widespread mode of payment, focus has been given to recent computational methodologies to handle the credit card fraud problem. There are many fraud detection solutions and software which prevent frauds in businesses such as credit card, retail, e-commerce, insurance, and industries. Data mining technique is one notable and popular methods used in solving credit fraud detection problem. It is impossible to be sheer certain about the true intention and rightfulness behind an application or transaction. In reality, to seek out possible evidences of fraud from the available data using mathematical algorithms is the best effective option. Fraud detection in credit card is the truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transactions, several techniques are designed and implemented to solve to credit card fraud detection such as genetic algorithm, artificial neural network frequent item set mining, machine learning algorithms, migrating birds optimization algorithm, comparative analysis of logistic regression, SVM, decision tree and random forest is carried out. Credit card fraud detection is a very popular but also a difficult problem to solve. Firstly, due to issue of having only a limited amount of data, credit card makes it challenging to match a pattern for dataset. Secondly, there can be many entries in dataset with truncations of fraudsters which also will fit a pattern of legitimate behavior. Also the problem has many constraints. Firstly, data sets are not easily accessible for public and the results of researches are often hidden and censored, making the results inaccessible and due to this it is challenging to benchmarking for the models built. Datasets in previous researches with real data in the literature is nowhere mentioned. Secondly, the improvement of methods is more difficult by the fact that the security concern imposes an limitation to exchange of ideas and methods in fraud detection, and especially in credit card fraud detection. Lastly, the data sets are continuously evolving and changing making the profiles of normal and fraudulent behaviors always different that is the legit transaction in the past may be a fraud in present or vice versa. This paper evaluates four advanced data mining approaches, Decision tree, support vector machines, Logistic regression and random forests and then a collative comparison is made to evaluate that which model performed best.

Credit card transaction datasets are rarely available, highly imbalanced and skewed. Optimal feature (variables) selection for the models, suitable metric is most important part of data mining to evaluate performance of techniques on skewed credit card fraud data. A number of challenges are associated with credit card detection, namely fraudulent behavior profile is dynamic, that is fraudulent transactions tend to look like legitimate ones, Credit card fraud detection

performance is greatly affected by type of sampling approach used, selection of variables and detection technique used. In the end of this paper, conclusions about results of classifier evaluative testing are made and collated.

From the experiments the result that has been concluded is that Logistic regression has a accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by ULB machine learning.

2. Related Work

In [1] This paper represents an research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

In [2] In this paper a new collative comparison measure that reasonably represents the gains and losses due to fraud detection is proposed. A cost sensitive method which is based on Bayes minimum risk is presented using the proposed cost measure. Improvements up to 23% is obtained when this method and other state of art algorithms are compared. The data set for this paper is based on real life transactional data by a large European company and personal details in data is kept confidential., accuracy of an algorithm is around 50%. Significance of this paper was to find an algorithm and to reduce the cost measure. The result obtained was by 23% and the algorithm they find was Bayes minimum risk.

In [3] Various modern techniques based on Sequence Alignment, Machine learning, Artificial Intelligence, Genetic Programming, Data mining etc. has been evolved and is still evolving to detect fraudulent transactions in credit card. A sound and clear understanding on all these approaches is needed that will certainly lead to an efficient credit card fraud detection system. Survey of various techniques used in credit card fraud detection mechanisms has been Shown in this paper along with evaluation of each methodology based on certain design criteria. Analysis on Credit Card Fraud Detection Methods has been done. The survey in this paper was purely based to detect the efficiency and transparency of each method. Significance of this paper was conduct a survey to compare different credit card fraud detection algorithm to find the most suitable algorithm to solve the problem.

In [4] A comparison has been made between models based on artificial intelligence along with general description of the developed fraud detection system are given in this paper such as the Naive Bayesian Classifier and the model based on Bayesian Networks, the clustering model. And in the end conclusions about results of models' evaluative testing are made. Number of legal truncations was determined greater or equal to 0.65 that is their accuracy was 65% using Bayesian Network. Significance of this paper is to compare between models based on artificial intelligence along with general description of the developed system and to state the accuracy of each model along with the recommendation to make the better model. In [5] Nutan and Suman on review on credit card fraud detection they have supported the theory of what is credit card fraud, types of fraud like telecommunication, bankruptcy fraud etc. and how to detect it, in addition to it they have explained numerous algorithms and methods on how to detect fraud using Glass Algorithm, Bayesian, networks, Hidden Markova model, Decision Tree and 4 more. They have explained in detail about each algorithm and how this algorithm works along with mathematical explanation. Types of machine learning along with classifications has been studied. Pros and cons of each method is listed.

This research is to detect the credit card fraud in the dataset obtained from ULB by applying Logistic regression, Decision tree, SVM, Random Forest and to evaluate their Accuracy, sensitivity, specificity, precision using different models and compare and collate them to state the best possible model to solve the credit card fraud detection problem.

3. Background

Ability of system to automatically learn and improve from experience without being explicitly programmed is called machine learning and it focuses on the development of computer programs that can access data and use it learn for themselves. And classifier can be stated as an algorithm that is used to implement classification especially in concrete implementation, it also refers to a mathematical function implemented by algorithm that will map input data into category. It is an instance of supervised learning i.e. where training set of correctly identified observations is available.

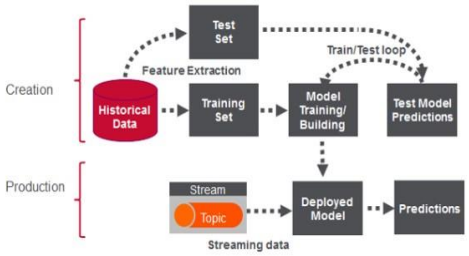


Figure 1: Classifier Steps

Logistic Regression

Logistic Regression is a supervised classification method that returns the probability of binary dependent variable that is predicted from the independent variable of dataset that is logistic regression predict the probability of an outcome which has two values either zero or one, yes or no and false or true. Logistic regression has similarities to linear regression but as in linear regression a straight line is obtained, logistic regression shows a curve. The use of one or several predictors or independent variable is on what prediction is based, logistic regression produces logistic curves which plots the values between zero and one.

Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, binomial logistic models. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors.

$$p = \frac{e^{\alpha+\beta_n X}}{1 + e^{\alpha+\beta_n X}}$$

Above equation represents the logistic regression in mathematical form.

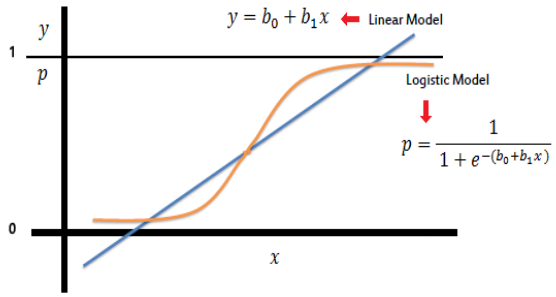


Figure 2: Logistic Curve

This graph shows the difference between linear regression and logistic regression where logistic regression shows a curve but linear regression represents a straight line.

SVM Model (Support Vector Machine)

SVM is a one of the popular machine learning algorithm for regression, classification. It is a supervised learning algorithm that analyses data used for classification and regression. SVM modeling involves two steps, firstly to train a data set and to obtain a model & then, to use this model to predict information of a testing data set. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane where SVM model represents the

training data points as points in space and then mapping is done so that the points which are of different classes are divided by a gap that is as wide as possible. Mapping is done in to the same space for new data points and then predicted on which side of the gap they fall

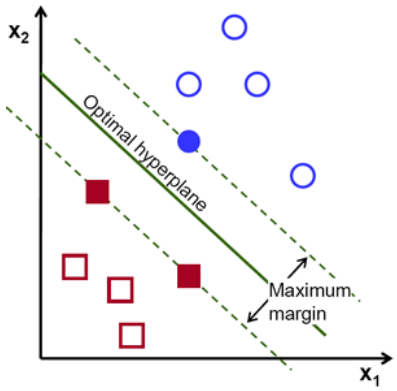


Figure 3: SVM Model Graph

In SVM algorithm, plotting is done as each data item is taken as a point in n-dimensional space where n is number of features, with the value of each feature being the value of a particular coordinate. Then, classification is performed by locating the hyper-plane that separates the two classes very well.

Decision Tree

Decision tree is an algorithm that uses a tree like graph or model of decisions and their possible outcomes to predict the final decision, this algorithm uses conditional control statement. A Decision tree is an algorithm for approaching discrete-valued target functions, in which decision tree is denoted by a learned function. For inductive learning these types of algorithms are very famous and have been successfully applied to abroad range of tasks. We give label to a new transaction that is whether it is legit or fraud for which class label is unknown and then transaction value is tested against the decision tree, and after that from root node to output/class label for that transaction a path is traced.

Decision rules determines the outcome of the content of leaf node. In general rules have the form of 'If condition 1 and condition 2 but not condition 3 then outcome'. Decision tree helps to determine the worst, best and expected values for different scenarios, simplified to understand and interpret and allows addition of new possible scenarios.

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Steps for making a decision tree are that firstly to Calculate the entropy of every

attribute using the dataset in problem then dataset is divided into subsets using the attribute for which gain is maximum or entropy is minimum after that to make a decision tree node containing that attribute and lastly recursion is performed on subsets using remaining attributes to create a decision tree.

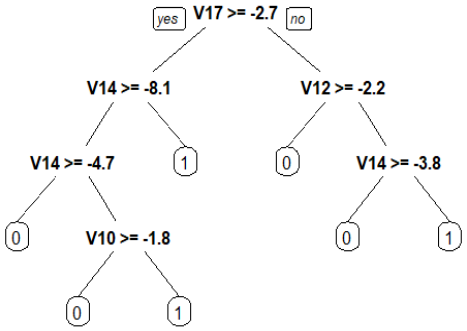
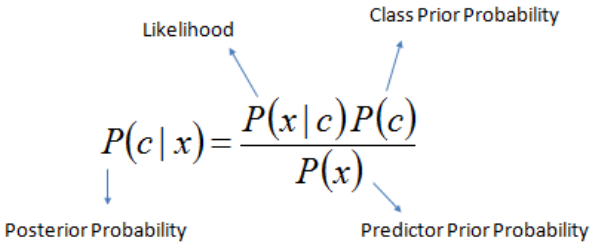


Figure 4: Decision tree

Random Forest

Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of overfitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provides a good estimate of the generalization error and to be resistant to overfitting.



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by Random Forest.

4. Experiments

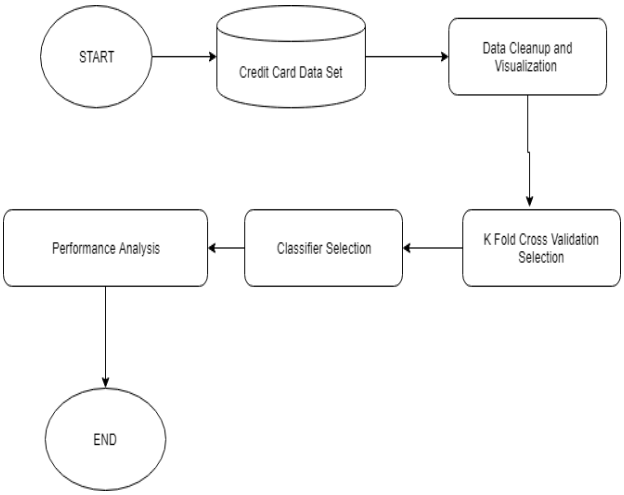


Figure 5: Architecture

First the credit card dataset is taken from the source and cleaning and validation is performed on the dataset which includes removal of redundancy, filling empty spaces in columns, converting necessary variable into factors or classes then data is divided into 2 part, one is training dataset and another one is test data set. Now K fold cross validation is done that is the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data, Models are created for Logistic regression, Decision tree, SVM, Random Forest and then accuracy, sensitivity, specificity, precision are calculated and a comparison is made.

The dataset is sourced from ULB Machine Learning Group. The dataset contains credit card transactions made by European cardholders around September 2013 and occurrence of transactions that happened in two days are presented by this dataset, consisting of 284,786 transactions. The dataset is highly unbalanced and skewed towards the positive class and positive class that is fraud cases make up 0.173% of the transactions data. It contains only numerical (continuous) input variables which are as a result of a Principal Component Analysis (PCA) feature selection transformation resulting to 28 principal components. And total of 30 input features are utilized in this study. Behavioral characteristic of the card is shown by a variable of each profile usage representing the spending habits of the customers along with days of the month, hours of the day, geographical locations, or type of the merchant where the transaction takes place. Afterwards these variables are used to create a model which distinguish fraudulent activities. The details and background information of the features cannot be presented due to confidentiality issues. The time feature stores the seconds that has elapsed between each transaction along with first transaction in the dataset. The 'amount'

feature is the transaction amount. Feature 'class' is the target class for the binary classification and it takes value 1 for positive case (fraud) and 0 for negative case (non fraud).

Four basic metrics are used in evaluating the experiments, namely True positive (TPR), True Negative (TNR), False Positive (FPR) and False Negative (FNR) rates metric respectively.

$$TPR = \frac{TP}{P}$$
$$TNR = \frac{TN}{N}$$
$$FPR = \frac{FP}{N}$$
$$FNR = \frac{FN}{P}$$

where FN , FP ,TP,TN, and are the number of false negative false positive ,true positive and true negative test cases classified while total number of positive and negative class cases under test are represented by P and N. Cases classified rightly as negate are termed with true negative and cases classified as positive which are actually positive are termed with True positive .Cases classified as positive but are negative cases are termed as false positive and cases classified as negative but are truly positive are termed as false negative. The performance of Classifiers is evaluated based on accuracy, precision, specificity and sensitivity.

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

Sensitivity (Recall) gives the accuracy on positive (fraud) cases classification. Specificity gives the accuracy on negative (legitimate) cases classification. Precision gives the accuracy in cases classified as fraud (positive)

In this study, four classifier models based on and logistic regression, SVM, decision tree and random forest are developed. To evaluate these models, 70% of the dataset is used for training while 30% is set aside for validating and testing. Accuracy, sensitivity, specificity, precision are used to evaluate the performance of the four classifiers. The true positive, true negative, false positive and false negative rates of the classifiers in each set of un sampled are shown below in Table 6 and a format of confusion matrix is also illustrated. The accuracy and specificity scores are misleadingly high in the table due to the presence of a large number of true negatives.

5. Results

Table 1: Performance Matrices

Metrics	Classifiers			
	Logistic Regression	SVM	Decision Tree	Random Forest
Accuracy	0.977	0.975	0.955	0.986
Sensitivity	0.97.5	0.973	0.955	0.984
Specificity	0.923	0.912	0.878	0.905
precision	0.996	0.996	0.995	0.997

Table 2: Confusion Matrix Format

Actual/Predicted	Not a fraud	Fraud
Not a Fraud	True Positive	False Positive
Fraud	False Negative	True Negative

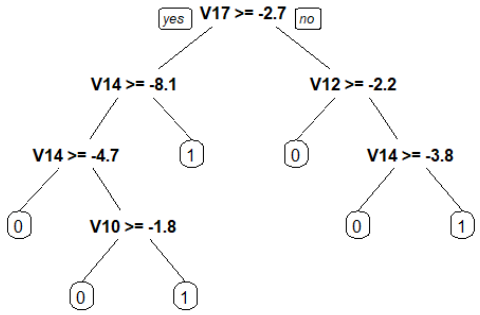


Figure 6: Decison Tree References

6. Conclusion

From the experiments the result that has been concluded is that Logistic regression has a accuracy of 97.7% while SVM shows accuracy of 97.5% and Decision tree shows accuracy of 95.5% but the best results are obtained by Random forest with

a precise accuracy of 98.6%. The results obtained thus conclude that Random forest shows the most precise and high accuracy of 98.6% in problem of credit card fraud detection with dataset provided by ULB machine learning.

The Random forest algorithm will perform better with a larger number of training data, but speed during testing and application will suffer. Application of more pre-processing techniques would also help. The SVM algorithm still suffers from the imbalanced dataset problem and requires more preprocessing to give better results at the results shown by SVM is great but it could have been better if more preprocessing have been done on the data.

Acknowledgement

We sincerely thank the management of SRM Institute of Science and Technology that have provided support and guidance throughout the project.

References

[1] Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.

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

[3] Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).

[4] Phua C., Lee V., Smith, Gayler K.R., A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010).

[5] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.

[6] Carneiro E.M., Dias L.A.V., Da Cunha A.M., Mialaret L.F.S., Cluster analysis and artificial neural networks: A case study in credit card fraud detection, 12th International Conference on Information Technology-New Generations (2015), 122-126.

[7] Hafiz K.T., Aghili S., Zavarsky P., The use of predictive analytics technology to detect credit card fraud in Canada, 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016), 1-6.

[8] Sonepat H.C.E., Bansal M., Survey Paper on Credit Card Fraud Detection, International Journal of Advanced Research in Computer Engineering & Technology 3(3) (2014).

[9] Varre Perantalu K., Bhargav Kiran, Credit card Fraud Detection using Predictive Modeling (2014).

[10] Stolfo S., Fan D.W., Lee W., Prodromidis A., Chan P., Credit card fraud detection using meta-learning: Issues and initial results, AAAI-97 Workshop on Fraud Detection and Risk Management (1997).

[11] Maes S., Tuyls K., Vanschoenwinkel B., Manderick, B., Credit card fraud detection using Bayesian and neural networks,

Proceedings of the 1st international naiso congress on neuro fuzzy technologies (2002), 261-270.

[12] Chan P.K., Stolfo S.J., Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, In KDD (1998), 164-168.

[13] Rousseeuw P.J., Leroy A.M., Robust regression and outlier detection, John wiley & sons (2005).

[14] Wang C.W., Robust automated tumour segmentation on histological and immunohisto chemical tissue images, PloS one 6(2) (2011).

[15] Sait S.Y., Kumar M.S., Murthy H.A. User traffic classification for proxy-server based internet access control,IEEE 6th International Conference on Signal Processing and Communication Systems (ICSPCS) (2012), 1-9.