Credit Card Fraud Detection

Machine Intelligence

BACHELOR OF TECHNOLOGY- V Sem CSE

Department of Computer Science & Engineering

SUBMITTED BY

Name: HARSHIT PRAKASH	SRN: PES2UG20CS137
Name: HEMANTH GA	SRN: PES2UG20CS140
Name: HEMANTH REDDY N.	SRN: PES2UG20CS141

PES UNIVERSITY

(Established under Karnataka Act No. 16 of 2013)

100 Feet Ring Road, BSK III Stage, Bengaluru-560085

PAPER 1

International Conference on Computer, Communication and Electrical Technology – ICCCET2011, 18th & 19th March, 2011 978-1-4244-9394-4/11/$26.00 ©2011 IEEE 152 Analysis on Credit Card Fraud Detection Methods 1 S. Benson Edwin Raj, 2A. Annie Portia 1Assistant Professor (SG), P.G., 2 Scholar Department of CSE Karunya University, Coimbatore

In the flowing paper the author as talked about the importance and the increase in the no.of users using credit cards and also talked about the real life fraudulent transaction . There are number of methods and algorithm for detecting credit card frauds such as artificial neural-network models(ANN) which are based upon artificial intelligence and machine learning approach , distributed data mining systems , sequence alignment algorithm which is based upon the spending profile of the cardholder , intelligent decision engines which is based on artificial intelligence , Meta learning Agents and Fuzzy based systems Web Services-Based Collaborative Scheme for Credit Card Fraud Detection in which participant banks can share the knowledge about fraud patterns in a heterogeneous and distributed environment to enhance their fraud detection capability and reduce financial loss . But the following methods and algorithms are used that are as follows Fusion of Dempster Shafer and Bayesian learning, BLAST-SSAHA Hybridization, Hidden Markov Model, Artificial neural networks and Bayesian Learning approach.

In fusion approach using Dempster–Shafer theory and Bayesian learning is a hybrid approach for credit card fraud detection. This method combines the posterior and prior probability. The FDS as four important components rule based filter which is used to determine the pattren of the input, Dempster–Shafer's theory is used to combine multiple such evidences and an initial belief is computed, The transaction is classified as suspicious or suspicious depending on this initial belief. Once a transaction is found to be suspicious, belief is further strengthened or weakened according to its similarity with fraudulent or genuine transaction history using Bayesian learning. Advantages and application of this approach is high accuracy and high processing Speed, improves detection rate and reduces false alarms and also it is applicable in E-Commerce. But it is highly expensive and its processing Speed is low

In BLAST-SSAHA Hybridization for Credit Card Fraud Detection is an efficient technique and is used for analyzing the spending behavior of the customer. The following method used two stage analyzer that is profile analyzer (PA) and the deviation analyzer (DA). The PA will check the similarity of the current transaction spending to the genuine cardholder and if an unusual transaction is found the transaction is sent to the DA. The further decision is taken from PA and DA. The performance of BLAHFDS is good and it results in high accuracy and is used to detect fraud transactions on the line only. But it does not detect cloning of credit cards.

In Hidden Markov Model if an incoming credit card transaction is not accepted by the trained Hidden Markov Model with sufficiently high probability, it is considered to be fraudulent transactions. The HMM will maintain a log. In each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify whether the transaction is genuine or not. If the FDS confirms the transaction to be malicious, it raises an alarm and the issuing bank declines the transaction. The concerned cardholder may then be contacted and alerted about the possibility that the card is compromised.

In Bayesian and Neural Networks approach is used for automatic credit card detection and for reasoning under uncertainty. The neural network is initially trained with the normal behavior of a cardholder. The suspicious transactions are then propagated backwards through the neural network and classify the suspicious and non-suspicious transactions. Bayesian uses supervised learning and the network are able to process data as needed, without experimentation. Bayesian belief networks are very effective for modeling situations where some information is already known and incoming data is uncertain or partially unavailable. This information or belief is used for pattern identification and data classification. Bayesian and Neural Networks uses supervised algorithms and require high speed processing speed.

PAPER 2

18th International Symposium INFOTEH-JAHORINA, 20-22 March 2019 978-1-5386-7073-6/19/$31.00 ©2019 IEEE Credit Card Fraud Detection - Machine Learning methods Dejan Varmedja, Mirjana Karanovic, Srdjan Sladojevic, Marko Arsenovic, Andras Anderla Faculty of Technical Sciences University of Novi Sad Novi Sad, Serbia

In the following paper the author has used Logistic Regression, Random Forest, Naive Bayes and Multilayer Perceptron and has split the given data set into training and testing data.

There are two types of credit card frauds. One is theft of physical card, and other one is stealing sensitive information from the card, such as card number, CVV code, type of card and other. Models like Gradient Boosting (GB), Support Vector Machines (SVM), Decision Tree (DT), LR and RF had high recall of over 91% in a given dataset. A high precision and recall were achieved inly after balancing the given dataset by oversampling the data. In paper [6], European dataset was also used, and comparison was made between the models based on LR, DT and RF. Among the three models, RF proved to be the best, with accuracy of 95.5%, followed by DT with 94.3% and LR with accuracy of 90%. K-Nearest neighbors (KNN) and outlier detection techniques can also be efficient in fraud detection. They are proven useful in minimizing false alarm rates and increasing fraud detection rate. KNN algorithm also performed well in experiment for paper [9], where the authors tested and compared it with other classical algorithms. In paper 10 a comparison was made between some classical algorithms and deep learning techniques with the accuracy of 80%.

Here the author of the paper uses the oversampling technique rather than the under sampling.

The author talks about the dataset used which contains transactions, occurred in two days, made in September 2013 by European cardholders. Since the dataset contain confidential information the value are changed to PCA. Dataset contains 284,807 transactions where 492 transactions were frauds and the rest were genuine. Considering the numbers, we can see that this dataset is highly imbalanced, where only 0.173% of transactions are labeled as frauds.

In preprocessing carefully choosing appropriate features and removing the less important one can reduce overfitting, improve accuracy and reduce training time. By using Feature selector tool , features

that do not contribute to the cumulative importance of 95% are removed. After the feature selection technique, 27 features were selected for additional experiment.

Since the dataset is not balance it is necessary to perform some kind of balancing, so that model can be efficiently trained. The methods include undersampling the majority class, oversampling the minority class, or combination of those two and the Synthetic Minority Oversampling Technique (SMOTE) is a popular oversampling method that has proven useful when used on imbalanced dataset SMOTE was proposed method to improve random oversampling.

The author of the paper has used numpy, pandas, matplotlib, sklearn and imblearn for coding.

Logistic regression model describes relationship between predictors that can be continuous, binary, and categorical. Based on some predictors we predict whether something will happen or not. We estimate the probability of belonging to each category for a given set of predictors.

Naïve bayes uses a supervised learning algorithms where there is no dependency.

In this following paper Bernoulli distribution is used for detecting fraud transactions.

Random forest can be used in both classification and regression problems. to prevent the model to overfit there should good no.of trees in the forest.

In the following paper the ANN has 4 hidden layers with 50, 30, 30 and 50 units in each hidden layer, respectively and uses relu activation function. In the following observation it has been shown that deeper networks acquire better results than those with smaller number of layers. So the author of the paper starts with a smaller number of layers gradually increasing them in order to get acceptable results. Therefore, the best hyper-parameters were chosen based on exhaustive research. Further increasing the network cause greater computational time and obtained results didn't differ much from the chosen architecture. Weight optimization was accomplished with Adam, stochastic gradient-based optimizer

In the paper the ratio of train and test is 80:20 ratio and the model was updated through multiple epochs, based on tolerance for the optimization (TOL). When the loss or score is not improving by at least TOL for specified consecutive iterations, convergence is considered to be reached and training stops.

Most used metric for this paper is accuracy, recall and precision and this can be calculated by the confusion matrix.

The results of each model in the paper were

LR model

• precision: 58.82%• recall: 91.84%• accuracy: 97.46%.

NB model

• precision: 16.17%, • recall: 82.65%, • accuracy: 99.23%

RF model

• precision: 96.38%, • recall: 81.63%, • accuracy: 99.96%.

MLP model

• precision: 79.21%, • recall: 81.63%, • accuracy: 99.93%

According to the paper the accuracy must be taken “with a grain of salt” – desirably it should be interpreted in combination with some other metrics.

So according to the paper Random Forest algorithm gives the best results i.e. best classifies whether transactions are fraud or not. This was established using different metrics, such as recall, accuracy and precision.

Paper 3

Real-time Credit Card Fraud Detection Using Machine Learning Anuruddha Thennakoon1, Chee Bhagyani2 , Sasitha Premadasa3, Shalitha Mihiranga4, Nuwan Kuruwitaarachchi5 Faculty of Computing Sri Lanka Institute of Information Technology Colombo, Sri Lanka 1anuruddha.thennakoon@gmail.com,2 bhagyani.lochana@my.sliit.lk,3 sasitha.premadasa@my.sliit.lk, 4 shalitha.mihiranga@my.sliit.lk,5 nuwan.ku@sliit.lk

Here the paper address more on real-time credit card fraud detection. For this, paper make use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent.

The dataset used is not disclosed due to confidential disclosure agreement.

The paper explains avoidance of the fraud that is prevention and detection. Prevention avoids any attacks from fraudsters by acting as a layer of protection. Detection happens once the prevention has already failed. Credit card fraud can be classified into several categories. The two types of frauds that can be mainly identified in a set of transactions are Card-not-present (CNP) frauds and Card-present (CP) frauds and the author uses the CNP fraud category.

The paper uses supervised learning.

The given dataset is divided into 4 dataset that are Transactions with Risky Merchant Category Code, Transactions larger than $100, transactions with risky ISO Response code, Transactions with unknown web addresses. And those 4 datasets were used in two different ways by transforming raw data into a numerical form. (Type A) and by preparing raw data categorically without making any transformation (Type B)

To prepare the dataset type A the following method were used data cleaning (finding the missing values), Data Integration, Data Transformation, Data Reduction (by using PCA).

To prepare the dataset type A the following method were used Resampling Techniques, Modelling and testing, Real time Fraud Detection, Fraud Detection System

Here Resampling Techniques is used because there are more genuine transactions than the fraudulent transaction so to overcome this, we conducted under-sampling and over-sampling by reducing the majority occurrences and by raising the minority occurrences respectively. For over-sampling, Synthetic Minority Oversampling Techniques (SMOTE) and for under-sampling, condensed nearest neighbor (CNN) and random under-sampling (RUS) were used.

The following paper uses 10-fold cross-validation.

In Modelling and testing the machine learning algorithms were prioritized by analysis with the help of the literature. They are Support Vector Machine, Naive Bayes, K-Nearest Neighbor and Logistic Regression. Optimal models were selected by filtering them out comparatively against an appropriate performance matrix.

In the paper the real-time fraud detection system consists of three main units API MODULE, FRAUD DETECTION MODELS and DATA WAREHOUSE.

In the paper Data Warehouse has been used for storing live transactions.

Paper 4

Credit Card Fraud Detection Using Machine Learning by Sanobar khan, Sanovar, Suneel Kumar, Mr Hitesh Kumar, Department of Electronics and Communication

This paper centres around four principal fraud events in certifiable transactions. Every fraud is tended to strategy is chosen through an assessment. Significant key territory which we discourse in our venture is constant credit card scam identification, the following detection modules are applied:

1- RANDOM FOREST ALGORITHM
2- LOCAL OUTLIER ALGORITHM

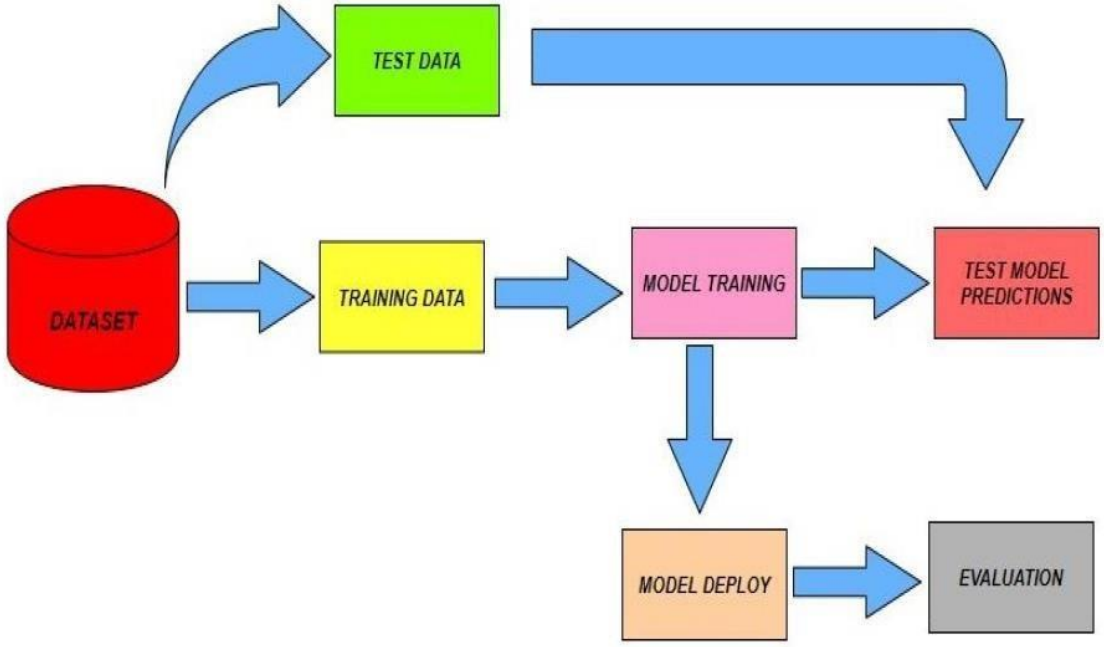
1- RANDOM FOREST ALGORITHM
Random Forest is additionally known for Random Decision Forest (RFA) that will utilized for categorization, Regression with different assignments which is carried out building numerous decision trees. That Random Forest Algorithm depends onto supervised learning along with significant preferred position for that technique is which that tends to be utilized for categorization and Regression. Random Forest Algorithm has more good accuracy if contrasted and any remaining existing frameworks with that's the most generally utilized technique. This approaches the utilization for Random Forest technique in credit card scam identification could show us accuracy of around 90 to 99%.
In credit card scam identification, the Random Forest technique show best accuracy into outcomes.

2-LOCAL OUTLIER ALGORITHM
This is an Unsupervised Outlier identification technique. 'Local Outlier Factor' says abnormal score for every example. It calculates the confined difference of trial data respectively to the adjacent ones. More accurately, placing will be done by k-nearest neighbour, which has displacement use for measures the local data.

Paper 5

"Credit Card Fraud Detection Using Machine Learning" Ruttala Sailusha, R. Ramesh, V. Gnaneswar, G. Ramakoteswara Rao, Department of Information Technology

The main aim of this paper is to classify the transactions that have both the fraud and non-fraud transactions in the dataset using algorithms like that The Random Forest and the Adaboost algorithms. Then these two algorithms are compared to choose the algorithm that best detects the credit card fraud transactions. The process flow for the credit fraud detection problem includes the splitting of the data, model training, model deployment, and the evaluation criteria.



The detailed architecture diagram for the credit card fraud detection system includes many steps from gathering dataset to deploying model and performing analysis based on results. In this model we take the Kaggle credit card fraud dataset and pre-processing are to be done for the dataset. Now to prepare the model we have to split the data into the training data and the testing data. We use the training data to prepare the Random Forest and the Adaboost models. Then we develop both the models. Finally, the accuracy, precision, recall, and F1-score are calculated for both the models. Finally, the comparison of the credit card fraud transactions more accurately.

Paper 6
"Machine Learning for Credit Card Fraud Detection System"
Lakshmi, Selvani, Deepthi Kavila, Department of CSE, Anil Neerukonda Institute of Technology and Sciences

The proposed techniques are used in this paper, for detecting the frauds in credit card system. The comparison is made for different machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions.

1 Logistic Regression:
Logistic Regression is one of the classification algorithms, used to predict a binary value in a given set of independent variables (1 / 0, Yes / No, True / False). To represent binary / categorical values, dummy variables are used. For the purpose of special case in the logistic regression is a linear regression, when the resulting variable is categorical then the log of odds is used for dependent variable and also it predicts the probability of occurrence of an event by fitting data to a logistic function.

2 Decision Tree Algorithm:
Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.

TYPES OF DECISION TREE
1. Categorical Variable Decision Tree: Decision Tree which has categorical target variable then it called as categorical variable decision tree.

2. Continuous Variable Decision Tree: Decision Tree has continuous target variable then it is called as Continuous Variable Decision Tree

3 Random Forest:
Random forest is a tree-based algorithm which involves building several trees and combining with the output to improve generalization ability of the model. This method of combining trees is known as an ensemble method. Ensembling is nothing but a combination of weak learners (individual trees) to produce a strong learner. Random Forest can be used to solve regression and classification problems. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical.

PAPER 7

CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING AND DATA SCIENCE S P Maniraj Assistant Professor (O.G.) Department of Computer Science and Engineering SRM Institute of Science and Technology Aditya Saini, Swarna Deep Sarkar Shadab Ahmed Department of Computer Science and Engineering SRM Institute of Science and Technology International Journal of Engineering Research & Technology (IJERT)

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUS&T at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that aren't genuine. They have taken attributes of customer's behaviour and based on the value of those attributes they've calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to perceive illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium sized online transaction. There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alertfeedback interaction in case of fraudulent transaction. In case of fraudulent transaction, the authorised system would be alerted and a feedback would be sent to deny the ongoing transaction. Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction

It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

PAPER 8

AN EFFICIENT TECHNIQUES FOR FRAUDULENT DETECTION IN CREDIT CARD DATASET: A COMPREHENSIVE STUDY Akanksha Bansal and Hitendra Garg Department of Computer Engineering & Applications, GLA University, Mathura (UP) India

The supervised and unsupervised techniques are used to classify the fraudulent transaction in order to determine the risk in credit card transaction. In the proposed summary of credit card fraud detection, data visualization techniques like histograms and correlations matrix are better techniques for data analysis. Credit card fraud detection is an important application of outlier detection implemented in batch mode but these techniques are not so effective as for huge amount of dataset . Therefore, PCA is used that requires low computation and low memory requirement. In the analysis, using Principal Component Analysis (PCA) and using Anomaly Detection Algorithm (ADA) are used. Principal Component Analysis (PCA) is the common method in detecting the frauds in various financial organizations, insurance companies and government agencies . The occurrence of such transactions are comparatively less in the dataset of credit card transactions therefore, its detection is little bit difficult and tricky . PCA is used to shrink the dimensionality of the dataset maintaining the correlation and variation present in the dataset. PCA transform the original dataset with new reduced dataset. Anomaly Detection Algorithm are used to identify the datasets with unusual patterns in the datasets. These anomalous have prospective of transforming into some problems like errors, structural defects and frauds. The Machine Learning (ML) techniques enhance the speed of classification of suspected transactions. The efficiency of

the method depends on the dataset and classifier used. The proposed summary will be beneficial to the banker, credit card user, and researcher to analyze to prevent credit card frauds. The future scope of this credit card fraud detection is to explore the things in each and every associations and banks to live safe and happily life. The data must be balanced in each place and we are getting the best results

PAPER 9

ADVERSARIAL LEARNING IN CREDIT CARD FRAUD DETECTION

In order to deploy robust systems and create an adaptive model, knowledge of an adversary's most effective strategy is beneficial.

Previous work has utilized a game theoretic model, the adversarial classifier reverse engineering (ACRE) approach and Markov processes to model the interactions between the fraudster and the classifier .

This project adds to current research in this area by extending the game theoretic framework to a real-world data set in fraud detection, implementing the most effective adversarial strategy and retraining the classifier in multiple rounds of a game.

To implement this approach, we employ a simple logistic regression model to classify charges as fraudulent or nonfraudulent, and then play a series of games to imitate the adversary's learning process and preemptively retrain the classifier. The contributions of this paper are the following:

• Introducing an adaptive fraud detection system that utilizes repeated games in the form of a feedforward model and incorporates the synthetic minority oversampling technique (SMOTE) to mitigate class imbalance.

• Utilizing Gaussian Mixture Models (GMMs) to segment the distribution space of continuous attributes as a means to find possible adversarial strategies.

The use of a GMM in determining a best strategy proved an effective way of finding optimal new transactions an adversary is likely to replicate. The use of SMOTE provided a useful tool in our ability to produce synthetic transactions of this best strategy. Overall, these two contributions provided tools able to mimic an adversary's learning and thought processes, giving the credit card company the ability to preemptively react to the changing transaction strategies.

In future research, there are many possible additions to our framework that would provide more information and realism in our models and could possibly improve our results. In order to differentiate the various possible fraud strategies, our GMM could be optimized to produce more regions of possible transaction types. In our SMOTE algorithm, we choose to introduce enough fraud for the next round to have 15% fraudulent transactions, though this number could also be optimized to the percentage of fraud that yields the most effective classifier.

CONCLUSION

The algorithm to be used for the project will be Hidden Markov model. Since the HMM produces the most efficient and accurate output. In Hidden Markov Model if an incoming credit card transaction is not accepted by the trained Hidden Markov Model with sufficiently high probability, it is considered to be fraudulent transactions. The HMM will maintain a log. In each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify whether the transaction is genuine or not. If the FDS confirms the transaction to be malicious, it raises an alarm and the issuing bank declines the transaction.The concerned cardholder may then be contacted and alerted about the possibility that the card is compromised. The model is supposed to have an accuracy of around 80-90%.