RESEARCH ARTICLE OPEN ACCESS

Credit Card fraud analysis and detection Using Machine Learning

K.Banupriya [1], Mr.S.Baskaran [2]
1(Mphil (Research Scholar), Tamil University, Thanjavur, Tamilnadu, India)
2 (Assistant Professor, Department of Computer Science, Tamil University, Thanjavur, Tamilnadu, India)

Abstract:

Credit card fraud is a serious and ever-growing problem with billions of dollars lost every year due to fraudulent transactions. Fraud has always been present and will always be. It is also ever changing, as the technology and usage patterns change over time, which makes CCFD (credit card fraud detection) a particularly hard problem. Furthermore, a classification of mentioned techniques into two main fraud detection approaches, namely, misuses (supervised) and anomaly detection (unsupervised) is presented. Again, a classification of techniques is proposed based on capability to process the numerical and categorical datasets. Data will be processed by machine learning based on Decision tree, SVM, Linear regression. Thus, it gives chart based result which has been prejudiced by data mining.

Keywords — **Credit Card, Fraud Classification, Fraud Detection Techniques, Machine Learning.**

I. INTRODUCTION

Credit card fraud is a wide-ranging term for theft and fraud committed using a credit card as a fraudulent source of funds in a given transaction. Generally, the statistical methods and many data mining algorithms are used to solve this fraud detection problem. Credit card fraud can be divided into two types: inner card fraud and external card fraud. Inner card fraud intends to defraud the cash. Usually, it is the combination of merchants and cardholders, using false transactions to defraud banks cash. External card fraud is mainly embodied at using the stolen, fake or counterfeit credit card to consume, or using cards to get cash in disguised forms, such as buying the expensive, small volume commodities or the commodities that can easily be changed into cash.

The credit card fraud detection using Decision tree, SVM (Support Vector Machine), Linear Regression are automatic credit card fraud detection system by means of machine learning approach. These three machine learning approaches are appropriate for reasoning under uncertainty. **Decision tree** is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node [1]. A **SVM** performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (case) that define the hyperplane are the support vectors [2]. **Linear Regression** performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting [3].

Rest of the paper is described as follows: section 2 describes the related work about existing fraud analysis and detection system, section 3 describes the proposed fraud analysis and detection system, section 4 describes the credit card fraud detection techniques section 5 described the pattern discovery, section 6 shows the Experimental results, and section 7 shows the conclusion.

II. LITERATURE REVIEW

Fraud detection involves monitoring the behaviour of users in order to estimate, detect, or avoid undesirable behaviour. To counter the credit card fraud effectively, it is necessary to understand the technologies involved in detecting credit card frauds and to identify various types of credit card frauds [4].

Bhattacharyya.Siddhartha[5],author evaluated two advanced data mining approaches, namely support vector machines and random forests together with logistic regression to detect credit card fraud and

examines the performance of these techniques with the varying level of data under sampling and these techniques can detect only few fraudulent transaction when it is applied to a real world data set.

Li.Jinjiu.Wei.Wei. [6], online banking fraud detection framework recommended i.e. based on utilizing resources and advanced data mining techniques and algorithms such as contrast pattern mining, neural network and decision forest are implemented and their outcomes are integrated with an overall score measuring the risk of an online transaction being fraudulent or genuine.

Sahin.Yusuf [7], the security mechanism such as CHIP and PIN are developed for credit card system that does not prevent from fraudulent credit card usages over online fraud and the author have developed and implemented a cost sensitive decision tree approach to detect fraudulent transactions and this approach is compared with the traditional classification models on a real world credit card data set.

R.Huang [8], a hybrid model recommended for online fraud detection of Video-on-demand system to improve the current Risk Management Pipeline (RMP) by adding Artificial Immune System (AIS) based fraud detection for logging data in which AIS based model combines two artificial immune system algorithms with behaviour based intrusion detection using Classification and Regression trees (CART),so the proposed approach can help e-commerce better understand the issues and plan the activities involved in a systemic approach to E-fraud.

A.Shen etal (2007)[9] demonstrate the efficiency of classification models to credit card fraud detection problem and the authors proposed the three classification models i.e., decision tree, neural network and logistic regression. Among the three models neural network and logistic regression outperforms than the decision tree.

III. IMPLEMENTATION

The proposed system is used in this paper, for detecting the frauds in credit card system. The comparison are made for different machine learning algorithms such as Decision Trees, SVM, and linear Regression to determine which algorithm gives suits best and can be adapted by credit card merchants for identifying fraud transactions. The main objective of this paper is to identify the different types of credit card frauds involves in physical or virtual cards. Then to review on data analytical techniques that detect credit card frauds and finally to study how we can safeguard the credit card and some precautions to avoid credit card frauds.

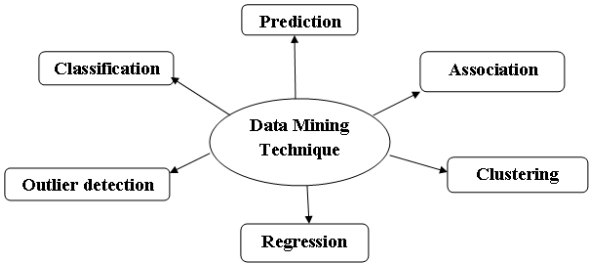
Advantages

- This paper presents a comprehensive investigation on financial fraud detection practices using such data mining methods, with a particular focus on computational intelligence-based techniques.
- Classification of the practices based on key aspects such as detection algorithm used, fraud type investigated, and success rate have been covered.

THE CREDIT CARD FRAUD DETECTION TECHNIQUES:

The credit card fraud detection techniques are classified in two general categories: fraud analysis (misuse detection) and user behaviour analysis (anomaly detection) [10].

Fraud Detection has been usually seen as a data mining problem where the objective is to correctly classify the transactions as legitimate or fraudulent [11].



Classification:

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of object whose class label is unknown.

Prediction:

The predictive attribute of a predictive model can be geometric or categorical. This models

continuous-valued functions. It is also encompasses the identification of distribution trends based on the available data.

Association:

Association rule mining is a procedure which is meant to find frequent patterns, correlations, association, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.

Clustering:

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

Regression:

Regression is a data mining techniques used to predict a range of numeric values (also called continuous values) given a particular dataset.

Outlier analysis:

A data object that does not comply with the general behavior of the data. It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis.

FRAUD DETECTION IS BASED ON TWO APPROACHES:

Supervised:

Supervised learning, no teacher is provided that means no training will be given to the machine [12].

Unsupervised:

Unsupervised learning is the training of machine using information that is neither classified nor labelled and allowing the algorithm to act on that information without guidance [13].

THE PATTERN DISCOVERY:

Credit card is one of the most divisive products among all the available financial tools. Credit card becomes popular mode of payment for both online as well as offline purchase. Besides being convenient, there are credit card frauds which become main threat for the users. Credit card frauds are increasing day by day. The fraudulent transactions are scattered with genuine transactions. Hence the simple pattern matching techniques are often insufficient to detect these frauds accurately. Credit card fraudsters are becoming more sophisticated. So, we need to develop /invent few new techniques to combat and to prevent such fraudulent attack.

The Parameters used for comparison of various Fraud Detection Systems are correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error (MAE), root mean squared error (RMSE), root relative squared error (RRSE), and total number of instances.

SNO	PARAMETERS	DECISION TREE	SVM	LINEAR REGRESSION
1	Correctly classified instances	89%	68%	72%
2	Incorrectly classified instances	21%	31%	30%
3	Kappa statistic	0.6173	0.3977	0.4564
4	Mean absolute error (MAE)	0.0793	0.1919	0.1845
5	Root mean squared error (RMSE)	0.1991	0.2988	0.2988
6	Root relative squared error(RRSE)	51.64%	125.06%	99.90%
7	Total number of instances	366	366	366

Correctly classified instances: Correctly classified instances means the sum of TP and TN.

TP: the true positive rate represents the portion of the fraudulent transactions correctly being classified as fraudulent transactions.

TN: the true negative rate represents the portion of the normal transactions correctly being classified as normal transactions.

Incorrectly classified instances: Incorrectly classified instances means the sum of TP and FN.

FP: the false positive rate indicates the portion of the non-fraudulent transactions wrongly being classified as fraudulent transactions.

FN: the false negative rate indicates the portion of the fraudulent transactions wrongly

being classified as normal transactions.

Kappa statistic: The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance).

Mean absolute error (MAE): MAE measures the average magnetic of the errors in a set of predictions, without considering their direction.

Root Mean Squared Error (RMSE): RMSE is a quadratic scoring rule that also measure the average magnitude of the error.

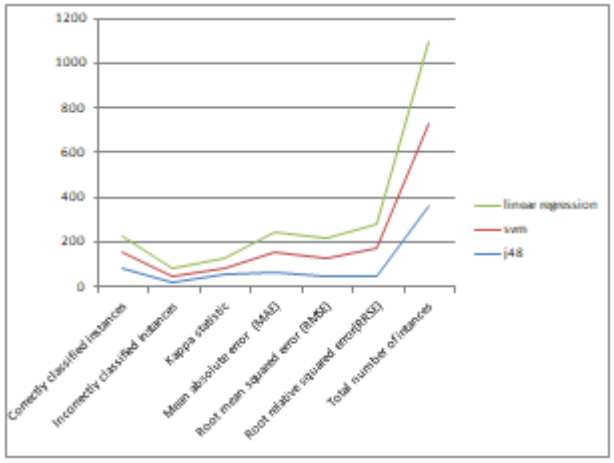
Total number of instances: Total number of instances selection (or dataset reduction, or dataset condensation) is an important data preprocessing step that can be applied in many machine learning (or data mining) tasks.

Root relative squared error (RRSE): RRSE takes the square root of (actual, predicted) divided by (actual, mean (actual)), meaning that it provides the squared error of the predictions relative to a naïve model that predicted the mean for every data point.

EXPERIMENTAL RESULTS:

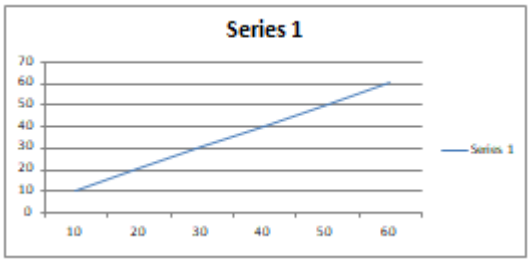
The credit card fraud detection using Decision tree, SVM (Support Vector Machine), Linear Regression are automatic credit card fraud detection system by means of machine learning approach.

J48 Decision vs SVM vs LR



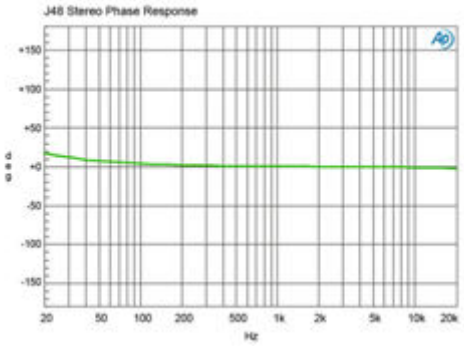
Linear Regression:

Linear regression is a way to model the relationship between two variables. The equation has the form Y=a+bX, where Y is the independent e dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.



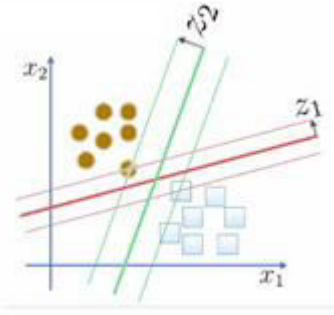
Decision Tree:

Decision Tree is a data representing and mining structure which comprises of a root node, branches and leaf nodes. Each internal node in the tree represents an attribute; each branch holds the outcome of a test, whereas each leaf node denotes a class label. The topmost node in the tree is the root node. And, the work of decision tree algorithm is to recursively partition the given data set of records using Depth-First or Breadth First Approach. Basically this tree structure reduces the complexity of the data sets by partitioning the unknown data sets.



Support Vector Machines:

Support vector machines (SVMs) are statistical learning techniques that can be used in a classification tasks. This technique is based on the supervised learning algorithm. An SVM model is a representation as points in space, and different points are mapped so that the separate categories are divided by a clear gap that is as wide as possible.



IV. CONCLUSIONS

In this paper, Machine learning technique like Linear Regression, Decision Tree and SVM used to detect the fraud in credit card system. since, correctly classified instances, incorrectly classified instances, kappa statistic, mean absolute error (MAE), root mean squared error (RMSE), root relative squared error (RRSE) are used to evaluate

the performance for the proposed system. Some of these techniques have been applied and produce an efficient system which can easily detect and report credit card frauds. The main objective of this paper is to identify the different types of credit card frauds involves in physical or virtual cards. Then to review on data analytical techniques that detect credit card frauds and finally to study how we can safeguard the credit card and some precautions to avoid credit card frauds.

REFERENCES

1. *Fan, W. (2004). Systematic Data Selection to Mine Concept- Drifting Data Streams. Proc. of SIGKDD04, 128-137.*
2. *C. Cortes and V. Vapnik, ―Support-vector networks,‖ Machine Learning, vol. 20, no. 3, pp. 273–297, 1995*
3. *Foster, D. & Stine, R. (2004). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. Journal of American Statistical Association **99**: 303-313.*
4. *Parvinder Singh, Mandeep Singh, "Fraud Detection by Monitoring Customer Behavior and Activities"International Journal of Computer Applications (0975 – 8887) Volume 111 – No 11, February 2015 23*
5. *Bhattacharya.S , Jha.S, Tharakunnel.K and Westland.C.J, "Data mining for credit card fraud", Science Direct, Decision Support System pp 602-613, 2010.*
6. *Li.J, Wei.W, Yuming.O and Chen.J, "Effective detection of sophisticated online banking fraud on extremely imbalanced data", Springer World Wide Web, pp 449-475, 2012.*
7. *Sahin.Y, Bulkan.S and Duman.E, "A cost-sensitive decision tree approach for fraud detection", Science Direct, Expert System with Applications 40 pp-5916-5923, 2013.*
8. *Huang.R, Tawfik.H and Nagar.A.K, "A Novel Hybrid Artificial Immune Inspired Approach for Online Break-in Fraud Detection", International Conference on Computer Science, Science Direct, pp 2733-2742, 2012.*
9. *Sanchez.D, Cerda.L, Serrano.J.M and Vila-.M.A, "Association Rules applied to Credit Card Fraud Detection", Science Direct Expert System with applications 36 pp 3630-3640, 2009.*
10. *Elkan, C. (2001). Magical Thinking in Data Mining: Lessons from CoIL Challenge 2000. Proc. of SIGKDD01, 426-431.*
11. *Linda Delamaire, Hussein Abdou, John Pointon, "Credit card fraud and detection techniques: a review," Banks and Bank Systems, pp. 57-68, 2009.*
12. *Syeda, M., Zhang, Y. & Pan, Y. (2002). Parallel Granular Neural Networks for Fast Credit Card Fraud Detection. Proc. of the 2002 IEEE International Conference on Fuzzy Systems.*
13. *Taniguchi, M., Haft, M., Hollmen, J. & Tresp, V. (1998). Fraud Detection in Communication Networks using Neural and Probabilistic Methods. Proc. of 1998 IEEE International Conference in Acoustics, Speech and Signal Processing, 1241-1244.*