IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 01, August 2019)

A Novel Approach of Credit Card Fraud Detection using Random Forest Classifier

Anuja Yawar Shivangi Chouhan Brijesh Taunk

Abstract – **Nowadays, as internet speed has increased and the prices of mobile have decreased very much in past few years. Also the data prices too are very much affordable to most of the people. This has resulted into the digitization of most of the institutes as it is easy and convenient for the people and also for the authority to maintain the records. So, it resulted in most of the banks and other institutes receiving and transferring money through credit cards. But with the hackers and other cyber criminals around there is always chances of the frauds in the transactions. The possibility of the fraud transaction is very less but it is not negligible and even having one fraud transaction is unacceptable because it is crime and we can't neglect it even if it is very less as it harms both the customer and credibility of the institute. This paper presents an implementation of automatic credit card fraud detection system using random forest classifier on Kaggle dataset. The selection of proper attributes for reducing the training overhead and claiming higher accuracy for the fraud detection using soft computing. Performance evaluation is achieved using confusion matrix plot with accuracy, sensitivity and precision values.**

Keywords – **Artificial Intelligence, Data Mining, KDD, Machine Learning, Random Forest Classifier.**

I. INTRODUCTION

Machine learning is a subfield of artificial intelligence (AI). In general, the goal of machine learning is to understand the structure of data and integrate it into models that can be understood and used by everyone.

Although machine learning is a field of computer science, it differs from traditional IT approaches. Indeed, in the latter, algorithms are sets of explicitly programmed instructions used by computers to calculate or solve problems. Machine learning algorithms allow computers to train on data inputs and use statistical analysis to produce values that fall within a specific range. For this reason, machine learning facilitates the use of computers in building models from sampling data to automate decision-making processes based on the data entered [1].

Any user of the latest technology benefits from machine learning. Facial recognition technology, for example, enables social media platforms to help users tag and share photos of friends. Optical Character Recognition (OCR) technology converts images of text into moving characters. Recommendation engines, powered by machine learning, suggest movies or TV shows to watch according to the preferences of the user. Autonomous cars that will use machine learning to navigate will soon be available to consumers.

Machine learning is an area in continuous development. For this reason, you need to consider some considerations when working with machine learning technologies or analyze the impact of machine learning processes.

Data Mining) is the process by which understandable and useful knowledge - previously unknown - is extracted from databases, in various formats and automatically. Then, Information Exploitation poses two challenges: working with large databases and applying techniques that automatically convert these data into knowledge [1]. Likewise, Data mining is a fundamental element for a broader technique whose objective is to discover knowledge in large databases (Knowledge Discovery in Databases —KDD) [2] [3].

The further development of the use of Information Exploitation in activities related to systems auditing has to do with the detection of intruders in telecommunications networks. Even in the scientific literature there are antecedents linked to the location of fraud using data mining [4].

This text refers to a specific case of fraud associated with credit cards and commonly known as the card cloning, a circumstance that represents a risk for clients attached to a bank.

Humans, having cognitive ability, develop a series of behaviours that can be defined as pattern depending on certain situations. In turn, the moment in which a crime is committed is no exception; a group of psychologists determined that there are patterns of behaviour associated with factors such as location, time of day and temperature. Such information, managed through data mining, allows us to develop a predictive model of ideal situations -

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 01, August 2019)

scenarios - where a crime could happen. For the cited example, three scenarios are identified that are identified with the aforementioned sponsors: bicycle theft, firearm theft and wallet theft [5].
Consequently, the development of this predictive tool generates a positive impact on society since it allows - to the forces of public order - to have faster reaction times and thus avoid being delayed by reaching the scenes of crime. However, it can also generate a negative impact if a citizen is wrongly prejudged due to poor system documentation (falsification of public documents, for example) [6].
The manual and technical review of fraud prevention does not detect some of the most prevalent patterns such as the use of a credit card several times, in multiple locations (physical or digital) and in a short time.
This paper develops a framework for automatic credit card fraud detection using random forest classifier using Kaggle dataset. Rest of paper is organized as follows. Section 2 provides theoretical details about machine learning algorithms. Section 3 presents the proposed methodology followed by the simulation and results in section 4 and finally the conclusion and future aspects are detailed in section 5.

II. Algorithms of Machine Learning

The principle of Machine Learning is to create a model from a dataset, from which performance can be evaluated by a validation method on a subset of this dataset. The larger the dataset used, the more powerful the model will be.
In the field of supervised learning, here are some algorithms:

A. Linear Classifiers

It classifies on the basis of the value of a linear combination of characteristics.

Logical Regression: Logistic regression is a known statistical technique used to model binary results. Use: Exploration and evaluation of the factors that contribute to a result such as looking for factors that influence customers to visit a store several times.

Naïve Bayes Classifier: Allows to classify easily and quickly. Use: Categorize good and bad customers.

Perceptron: It is an algorithm of linear separation techniques.

Support Vector Machine: This is a supervised learning algorithm to solve problems of discrimination and regression based on two key ideas, the maximum margin and transforming the representation space of the input data into a space larger dimensions [7].
Use: SVMs can be used to solve discrimination problems, i.e. decide which class a sample belongs to, or regression, and i.e. predict the numerical value of a variable.

B. Quadratic Classifiers

A quadratic classifier Jump is used to separate two or more classes of objects using a quadratic surface.

C. K-Means Clustering

The k-means algorithm divides observations into K partitions (clusters) in which each observation belongs to the score with the closest average.
Usage: K-Means is useful when you have an idea of how many partitions exist in your space.

D. Boosting

Boosting optimizes the performance of binary classification algorithms. It is a method for converting poor performance prediction rules into a (very) powerful prediction rule. The final prediction comes from a combination (weighted vote) of several predictions.
Use: Prediction of horse races.

E. Decision Tree

This is a classification and regression algorithm used for predictive modeling of discrete and continuous attributes. The algorithm is easy to understand.
Use: identify the characteristics of a customer [8].

F. Random Forest

The decision tree forest algorithm performs learning on multiple decision trees trained on slightly different subsets of data.

G. Neural Networks

The algorithm associates each possible state of the input attribute with each possible state of the predictable attribute, and uses the training data to calculate the probabilities [9].
Use: Prediction of stocks.

H. Bayesian Networks

The algorithm calculates conditional probabilities. Based on the information observed, it calculates the probability of unobserved data. For example, according to the symptoms of a patient, the probabilities of the different pathologies compatible with his symptoms are calculated. We can also calculate the probability of unobserved symptoms, and deduce the most interesting complementary examinations
Use: Risk modelling.

IJDACR
ISSN: 2319-4863

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 01, August 2019)

III. PROPOSED METHODOLOGY

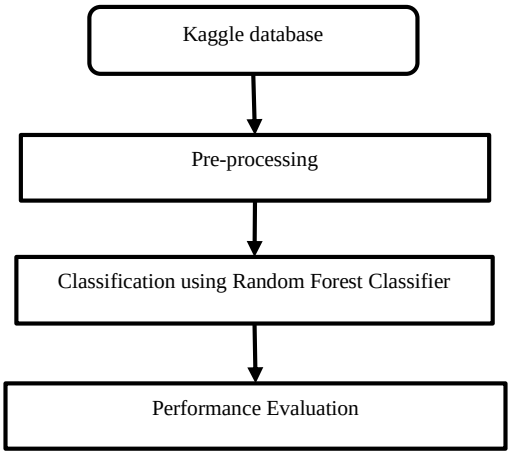

Figure 1: Flow diagram for proposed credit card fraud detection system

A. *Dataset Description*

The Kaggle datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions [10].

B. *Pre-Processing*

- Create a response variable in a categorical form where:
  1= fraudulent activity
  0 = non-fraudulent or normal activity
- Remove Time attribute from categorical class

C. *Classification by Random Forest Classifier*

A random forest is a classifier containing of one group of structured tree predictors $[T(x,\ominus_k), k = 1, \ldots.]$ where the $[\ominus_k]$ are random vectors of identical distributions and where every tree provides a unit poll for the furthermost common class of each entry x.

The main advantage of this structure is that it avoids the danger of over-learning for any method of prediction based on induction. BREIMAN [11] shows that when the number of trees involved in the prediction forest increases, the generalization error rate converges to a limit value, of which an upper bound can be estimated on the basis of the characteristics intrinsic features of the forest.

The classification trees in RF is built by selecting features from random samples to obtain a class label. The RF classifier is formed by a number of base learners and each base learner acts as an independent binary tree which adapts recursive partitioning.

The best feature is selected by Gini index, which is used to build the binary tree. It has the following advantages:

- RF is one of the most accurate classifier in present scenario.
- Overfitting is reduced due overgrowing the trees hence it is ease to handle.
- It accepts a large number of input variables without any deletion of the variables.
- The number of base learners is the only setting parameter to give the highest accuracy.

If the marginal function of a random forest is $T(X,\ominus)$

$$mr(X,Y) = P_{\ominus}(T(X,\ominus) = Y) - \max_{j \neq T} P_{\ominus}(T(X,\ominus) = j) \quad (1)$$

Which represents the confidence level of the ranking established by the trees of this forest on the population (X, Y), measured by the difference of probability between the prediction of the correct class Y and the best class erroneous j ≠ Y, one can define the prediction value of a game of trees {T (x, Θ)} by the mathematical expectation of this function:

$$s = E_{x,y}[mr(X,Y)] \quad (2)$$

The dependency between trees in a forest ρ (⊖,⊖') is measured by the correlation between their gross marginal functions, and it is evaluated for fixed and distinct parameter values ⊖,⊖′. By means of these definitions, an upper limit to the error in generalization (TEG) of any random forest is given by the relation.

$$TEG \leq \bar{\rho}(1 - s^2)/s^2 \quad (3)$$

IV. SIMULATION AND RESULTS

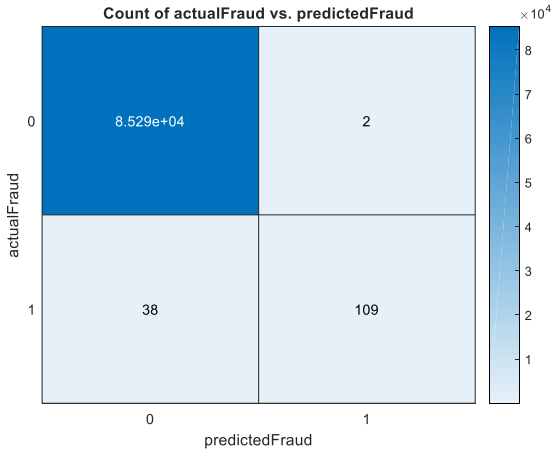

Figure 2: Confusion matrix plot for proposed credit card fraud detection

IJDACR
International Journal Of Digital Application & Contemporary Research

International Journal of Digital Application & Contemporary Research
Website: www.ijdacr.com (Volume 8, Issue 01, August 2019)

Here, TP=85290, TN=109, FP=2 and FN=38

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{85290+109}{85290+109+2+38} = 99.95\%$$

$$Precision = \frac{TP}{TP+FP} = \frac{85290}{85290+2} = 99.99\%$$

$$Sensitivity = \frac{TP}{TP+FN} = \frac{85290}{85290+38} = 99.95\%$$

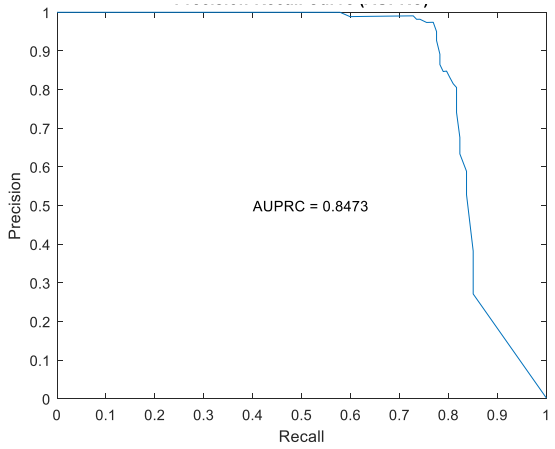

Figure 3: Precision-Recall curve

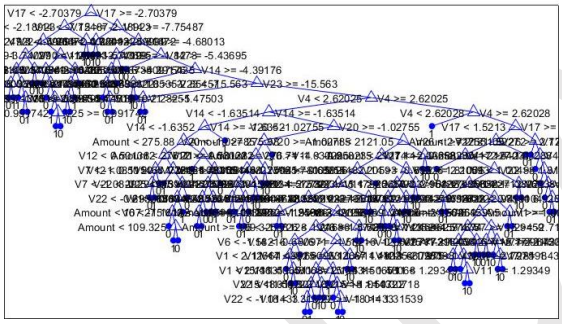

Figure 4: Tree diagram

V. CONCLUSION

This paper has examined the performance of random forest classifier. The Kaggle dataset on credit card transactions is used in this paper. This work achieves maximum accuracy of 99.95%.

Although random forest obtains good results on small set data, there are still some problems such as imbalanced data. Our future work will focus on solving these problems. The algorithm of random forest itself should be improved. For example, the voting mechanism assumes that each of base classifiers has equal weight, but some of them may be more important than others. Therefore, we also try to make some improvement for this algorithm.

REFERENCE

[1] Britos, Paola, Oscar Dieste, and Ramón García-Martínez. "Requirements elicitation in data mining for business intelligence projects." In *IFIP World Computer Congress, TC 8*, pp. 139-150. Springer, Boston, MA, 2008.

[2] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17, no. 3 (1996): 37-37.

[3] Britos, Paola, Hernan Grosser, Dario Rodríguez, and Ramon Garcia-Martinez. "Detecting Unusual Changes of Users Consumption." In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pp. 297-306. Springer, Boston, MA, 2008.

[4] Gunderson, L. F. "Using data mining and judgment analysis to construct a predictive model of crime." In *IEEE International Conference on Systems, Man and Cybernetics*, vol. 7, pp. 5-pp. IEEE, 2002.

[5] Brown, Donald E., and Rosemary B. Oxford. "Data mining time series with applications to crime analysis." In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, vol. 3, pp. 1453-1458. IEEE, 2001.

[6] Mota, Gabriel, Joana Fernandes, and Orlando Belo. "Usage signatures analysis an alternative method for preventing fraud in E-Commerce applications." In *2014 International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 203-208. IEEE, 2014.

[7] Şahin, Yusuf G., and Ekrem Duman. "Detecting credit card fraud by decision trees and support vector machines." (2011).

[8] Seyedhossein, Leila, and Mahmoud Reza Hashemi. "Mining information from credit card time series for timelier fraud detection." In *2010 5th International Symposium on Telecommunications*, pp. 619-624. IEEE, 2010.

[9] Yee, Ong Shu, Saravanan Sagadevan, and Nurul Hashimah Ahamed Hassain Malim. "Credit card fraud detection using machine learning as data mining technique." *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 10, no. 1-4 (2018): 23-27.

[10] Credit Card Fraud Detection Dataset. Online available at: https://www.kaggle.com/mlg-ulb/creditcardfraud

[11] Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.