International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 4 Issue 5, August 2020 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Credit Card Fraud Detection using a Combined Approach of Genetic Algorithm and Random Forest

M. Bhavana Lakshmi Priya, Dr. Jitendra Jaiswal

Department of Computer Science and Engineering,
Jain University, Kanakapura, Bangalore, Karnataka, India

ABSTRACT

Nowadays the companies are growing around the world and a lot of data is also processing daily. This data helps the companies for future business-related purposes for this they will store the data. Is the data is stolen the company will affects it. In this paper, we are discussing credit card fraud detection. Credit card fraud detection is of two types mainly first is through online and second is through the physical card. By stealing the information related to the credit card they can fraud large amounts of money transfer or a large amount of purchase before the cardholder finds out. For detecting the frauds, the companies are using many machine learning techniques for finding transactions that are fraudulent or not. This paper is a combined approach of genetic algorithm and random forest the genetic algorithm is used for feature selection and in the random forest, we used random forest classifiers by splitting the training and testing set. The combination of both gives good results then alone.

KEYWORDS: *Credit Card Fraud Detection, Genetic Algorithm, Random Forest*

How to cite this paper: M. Bhavana Lakshmi Priya | Dr. Jitendra Jaiswal "Credit Card Fraud Detection using a Combined Approach of Genetic Algorithm and Random Forest" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-5, August 2020, pp.230-233, URL: www.ijtsrd.com/papers/ijtsrd31774.pdf

IJTSRD31774

1. INTRODUCTION

The companies are growing around the world and a huge amount of data is processing daily. This data helps the companies for the future business purpose for this they will store the data. If the data is stolen it affects the company. Credit card fraud is the main problem that affects the entire company costumers. Then detecting the fraud, the following transactions of both valid or not valid payments. In this process we are detecting the frauds by using parameters mainly amount, time, location, card details, and transaction, etc., there are major types of frauds. 1) Clone transaction in this fraud will happen when the organization of a person tries for payment for multiple times. 2) Account theft suspicious transactions: It happens with personal information such as social security member or date of birth from the costumers for their criminals. 3) False application fraud: It is often accomplished by account or credit card. It happens with the documents and replacing by supporting the fake application. 4) Credit card skimming: It happens with a credit card by replicating the duplicate. Then copy the device that reads the duplicate information and transaction are made up of illegal electronic transactions or manual credit cards. 5) Account Take over: The criminals will send the emails to cardholders or by sending an illegal message from this they will take the personal information from cardholders. And mainly online passwords This paper is mainly based on credit card fraud detection by using machine learning algorithms mainly genetic programming and random forest and dealing with highly imbalanced dataset. In these we are combining both the algorithms we will get good accuracy results and also improves the performance. And this combined approach named Genetic programming with random forest.

2. OBJECTIVES

2.1. General Objective:

The first objective is to analyse the algorithm itself, discussing the best implementation parameters. After that we will combine both the genetic algorithm with random forest algorithm in this, we are using a random forest classifier.

2.2. Specific Objective:

The first objective we will initialize the population and perform feature selection with the operator's crossover and mutation for good fitness value. Second, we will split the dataset according to fitness values using a random forest classifier for a good result.

3. LITERATURE SURVEY

Satvik vats, Suryakant Dubey, Naveen Kumar Pandey they proposed "Genetic algorithms for credit card fraud detection" in this genetic algorithm used for transaction whether it is fraudulent or not and mainly they took genetic algorithm because it is the best evolutionary algorithm to

find the best solution for any problem. They took their example over boundary value testing in the terms of application domain. Based on the transactions with the optimization technique the genetic programming got the solution that satisfies the minimum criteria. The fitness function helps the fitness function for successive iterations and they adapted anti-fraudulent strategies are predicted. This algorithm does not fir the situation that they decided multi-population to an optimized parameter [1].Ruchi Oberoi, they proposed the "Credit card fraud detection using genetic algorithm" in this paper the genetic algorithm a technique for the optimal solution and also generates the fraudulent transactions in this first they removed the fraudulent transactions are done with time and online payment after the iterations are done finally they got the best solution is obtained. Here the set of interval-valued parameters is optimized in the short of the period of time we can reduce the risks [2]. Ibtissum, Samira Douzi, Bouabid El Quahidi proposed "Using genetic algorithm to improve the classification of imbalanced datasets for credit card" in this they combined k-Means Clustering and genetic algorithm. In this k-means clustering done the sampling and under-sampling and genetic programming is used for fraud transaction there are dealing if highly imbalanced data by generating new minority classes. It is done the sample for each cluster into chromosomes it inherits the individual's characteristics of generation. It aims to get more accuracy in fraudulent transactions. So, the genetic algorithm is used for removing transactions and by the parameters, it will decrease transactions [3].

Sahil Dhankhand, Behrouz Far proposed "Supervised machine learning algorithm for credit card fraud detection: A comparative study". In this paper, they did different classification algorithms with card numbers, account numbers, CVV in these they are dealing with highly imbalanced data. Their data can balance by under-sampling technique 70% used for training data 30% for testing data. The model evaluation is done by using accuracy, precision, specificity, G-mean by the confusion matrix. Random forest and Logistic regression have good performance in the comparative study [4].

M. Suresh Kumar, V. Soundarya, Kavitha, E.S. Keerthika, E. Ashwini proposed "Credit card fraud detection using random forest" in this mainly they detected the fraud transactions. For that they collected from a bank they took amount, transaction, and time for fraud detection and they divided the data into the training set and testing set. The accuracy results around 90% approximately they got the result [5].
Sonali Bakshi "Credit card fraud detection a classification analysis" in this paper they discussed they discussed the arrangement of charge card and challenges looked for visa cardholders and they detected visa card fraudster [6].

Vyom shah, Parin shah, Harish Shetty proposed "A review of credit card fraud detection techniques" in this paper the genetic algorithm gave a good accuracy result when compare with HMM and ANN. In this, no absolute values are used for standardized techniques [7]. Combined approach random forest and genetic algorithm gives better result.

4. METHODOLOGY

In this paper, we used python language for comparing the algorithms with the help of a genetic algorithm operator's crossover, mutation and tournament selection, fitness function, population initialization, and a random forest classifier and getting a good result.

Population initialization: It is divided into two types mainly 1) Heuristic 2) Random. Heuristic will initialize the population with complete random solutions. Initialize the population using heuristic for the problem.

Fitness function: The fitness function will calculate the result repeatedly there for sufficient works.

Crossover: It is applied with high probability and here more than one parents is selected and one more offspring are produced.

Mutation: It Produces small random changed by choosing a single bit randomly and it maintains a small random tweak.

Tournament selection: It is extremely popular it can work on negative fitness values also.

Random forest classifier: It works on no of decision tress for various subsets to improve the accuracy result.

5. PROJECT OUTLINE

This project is structured as the follows:

- we introduced the project idea along with background about the topic, the problem statement, literature review, objectives and the followed methodology.
- (System design), we explain the plan of the project using the system model. In our system model, we have the following three main parts:
  The part which we can initialize the population and getting fitness values.
  The part which the feature selection can be done with operators' crossover, mutation.
  The part which divide the data into training and testing for accurate result.
- (Analysis and Implementation), we give details about the algorithms which based on them we conducted our work. Then using Python language, we determine:
- We Initialize the population and performs fitness function based on operators.
- The split them by using random forest classifier with the operators.
- The implementation details.
- (Results and Discussion), we present and analyze the obtained results.
- (Conclusions and Recommendations), we draw conclusions and recommendations regarding the future scope of this project.

6. BACKGROUND

Fraud is the main problem for the credit card companies. It is growing bigger in our daily life and detecting the scam in time is a big challenge. Primarily in this paper, we are detecting the fraudulent transactions by using the combined approach of genetic algorithm and random forest with a highly imbalanced dataset.

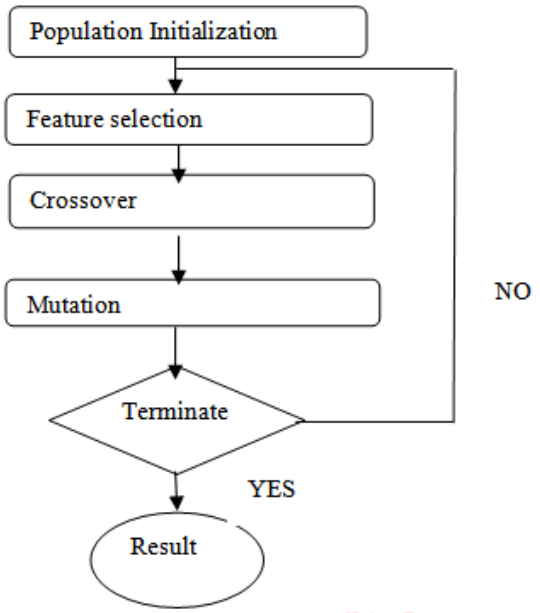
Imbalanced data: It is mainly referring to a classification problem where the observations per class are not equally distributed. We have to declare a large number of observations in one class. The main advantage with this is no

loss of information and the main disadvantage is over fitting (when a function is closely fit to set of data points).

The main reason for this project to detect the fraud which is related to credit card and by using the genetic algorithm with random forest the genetic algorithm will perform feature selecting by initializing the population values and with the help of operators cross over and mutation it inhibits the fitness values and with the values, the random forest classifier will split into training and testing and helps for good accuracy result.

7. SYSTEM DESIGN

The system model of our project which forms the plan for the rest of project and then break down this model into its main parts.



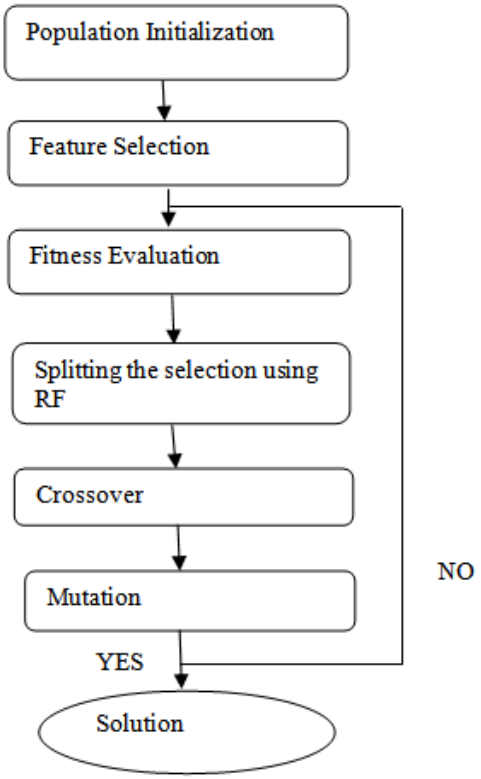
We can describe the previous system model by considering the three main parts of it.

Genetic Algorithm: The Genetic algorithm is based on a methodology inspired by the biological evolution and evolutionary algorithms to find solutions. Genetic algorithm falls in the category a machine learning technique. It optimizes a population. Operators of crossover and mutation in genetic algorithm help to generate better solutions.

Random Forest: The random forests algorithm, in machine learning, can also be thought of as an ensemble method for classification. A dataset with attributes is the input to this algorithm. Random subsets of the given dataset are formed. Then, on each of the random subsets that are created, a decision tree will be formed. The resultant class of any test record is decided by the algorithm, which in this case, uses the majority vote technique. Suppose 'x' is the input in the form of a matrix. On the matrix, there are trees formed randomly. Say there are 'b' number of trees namely tree1, tree2, …., treeb, each of which gives decision k1, k2, …, kb respectively. The majority voting method is applied. Final vote is k class, which will be decided as the class of the test record under consideration. The random forests algorithm works on the method of randomly selecting the subsets. This means that there is no bias when the random forests algorithm is used.

Genetic Algorithm with Random forest: First we initialize the population and we performed feature selection with operators and later random forest classifier will split the training and testing and perform the operators crossover and mutation with the evaluation of fitness function later the graphs are drawn based on the fitness values and which operator got the highest fitness values will help to give the good accuracy result.

8. BLOCK DIAGRAM



9. ANALYSIS AND IMPLEMENTATION

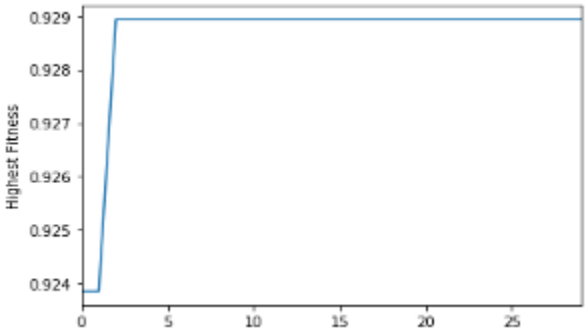
Analysis and evaluation methodology were built, then we will explain how we implemented this methodology and finally we will explain the implementation of fitness function along with the required diagrams.

Feature selection and its evaluation: The feature selection can be done with the initialization of population can be done for evaluating the fitness value with the operator's crossover and mutation and which operator gets the highest fitness value will help for accurate result.

Crossover: In Crossover its selects one parent to one or more off springs with high probability and in these it is replaced with one tree with another tree.

Mutation: Mutation step involves from one generation to another generation of population for the chromosomes and random forest is replaces one with another for the best accuracy.

10. RESULTS



Results of analyzing the graphs: The mutation got the highest fitness value and here the graph we took based on genetic algorithm operator with the help of random forest classifier. With the initialization the population the fitness values can be evaluate by fitness function with these we applied for both crossover and mutation the mutation got the highest fitness value and around 92% accuracy we got when we combined the both genetic algorithm and random forest.

Comparison of accuracy in percentage:

Algorithms	Dataset	Accuracy
Random Forest	Credit Card	90%
Genetic Algorithm	Credit Card	89%
Genetic Algorithm with RF	Credit Card	92%

When random forest algorithm applied alone, we got the accuracy result around 90% and genetic algorithm applied alone we got the accuracy result around 89% when we combine the both algorithms, we got the better accuracy around 92%.

11. SCOPE AND LIMITATION

In the scope of this project, we will analyze the operators using fitness values with both genetic algorithm and a random forest algorithm. Beyond that which operator got the highest fitness value, we will plot the graph with the parameters and we will implement the accurate result. In the limitation of this project, the data set is highly imbalanced so applying random under-sampling on the top if applying cross-validation, the accuracy won't increase because data set already reduced.

12. CONCLUSION AND FUTURE SCOPE

For datasets in Table.1, we compared the performances of random forests and genetic algorithm individually and also combined approach of genetic algorithm and random forest when compare with single accuracy results the combined approach got 92% accuracy. In this we took dataset from Kaggle and I divided the data set into training and testing dataset with the help of genetic algorithm and later with the random forest we performed the operators. Finally, mutation got the highest value with the fitness function, got the good performance. In future I want to implement a software with high security and costumers can pay their money whenever they want with the help of eye recognition and finger print.

13. REFERENCES

[1] Satvik vats, Suryakant Dubey, Naveen Kumar Pandey they proposed "Genetic algorithms for credit card fraud detection".

[2] Ruchi Oberoi, they proposed the "Credit card fraud detection using genetic algorithm".

[3] Ibtissum, Samira Douzi, Bouabid El Quahidi proposed "Using genetic algorithm to improve the classification of imbalanced datasets for credit card".

[4] Sahil Dhankhand, Behrouz Far proposed "Supervised machine learning algorithm for credit card fraud detection: A comparative study".

[5] M. Suresh Kumar, V. Soundarya, Kavitha, E. S. Keerthika, E. Ashwini proposed "Credit card fraud detection using random forest".

[6] Sonali Bakshi "Credit card fraud detection a classification analysis".

[7] Vyom shah, Parin shah, Harish Shetty proposed "A review of credit card fraud detection techniques".