Credit Card Fraud Detection System Using Machine Learning

Neera Chaudhary[1], Deepanshu Srivastava[2], Abhishek Verma[3], Jagriti Varshney[4], Awanish Katiyar[5]
[1,2,3,4,5]*Department of Computer Science and Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad*

Abstract **- Credit card use is not always beneficial for everyone, and it can result in significant financial losses in some cases. The frauds of credit cards are now increasing day by day. As the digitalization of internet purchasing grows, so does the use of credit cards for online transactions. Thus, most of the financial institutions and banks now prefer credit card fraud detection application. There are many types of fraudulent transactions which can happen in various ways with anyone, anywhere. Credit card firms must be able to recognize credit card fraud transactions in order to prevent or identify fraudulent transactions of products that the consumer did not purchase. Data science and machine are now helping to identify these fraud transactions. Fraud transactions are frequently mixed up with valid transactions, and simple recognition approaches that compare both the fraud and normal data are never enough to effectively detect fraud transactions. This study uses Credit Card Fraud Detection to demonstrate the modelling of a knowledge set using machine learning. The Credit Card Fraud Detection Problem entails credit card transaction modelling, which has previously been done with fraud transaction data.**

Index Terms **- Credit Card, Machine learning, Detection, Random Forest.**

I.INTRODUCTION

A Fraudulent transaction from a credit card is basically referred to the use of someone's account intentionally without awareness of its owner. Some steps are required to be taken against these types of frauds activities by evaluating and researching all these fraud transactions to stop similar circumstances in future transactions.

In brief, Frauds in credit card transaction can be easily described as a situation in which a fraudster uses someone's credit card for their personal benefits without the authorization of the owner of the credit card. For preventing these fraudulent activities, there is a requirement to observe and monitor these activities to prevent abnormal behavior which comprises of intrusion, defaulting, and fraud. Data sciences and machine learning are only possible ways through which institutions can monitor and detect all these fraudulent activities. As it seems, it is a serious issue and it is a very challenging problem as it is illustrated by several components such as class disparity [2][3][6]. Generally, the numbers of the legal transactions are more than the fraud transactions. Moreover, the statistical characteristic of the arrangement of transactions modifies regularly over a short time period. On-time execution of detection of credit card fraud faces several complications. Nonetheless, most of the automatic tools scan the wide stream of transaction requests that mainly notify about the legal transaction. The algorithm of machine learning is mainly used to identify all the legal transactions and also inform about the transaction which is doubtful.

At the present time, professionals are now monitoring the reported doubtful transactions. They firstly contact the cardholder to confirm the transaction that whether it was fraudulent or authentic. The investigator then stores the response given by the customer as their feedback in the automated system which supports the professionals for further analysis and enhances the value of detection of fraud over time. Regular updates of detection of fraud are very essential to prevent further fraudulent activities and understand the fraudulent strategies of the criminals.

System Framework

In order to average prior fraudulent conduct, the raw data is being taken and pre-processed, in pre-processing data is tested and trained using different machine learning algorithms. Based on the result of

training the data is classified as fraudulent or non-fraudulent categories.

Anomalous transactions detected by the profile analyser will then be transmitted to the Deviation Analyzer (DA). On the basis of the observations of these two analyzers the final conclusion on the nature of a transaction is made. We offer a new technique to combine two sequence alignment in order to achieve online response time for PA and DA.

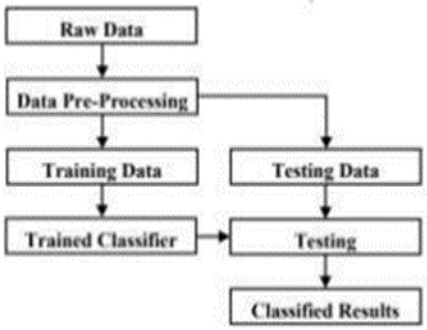

Fig.1 System Framework

Fig.1 demonstrates the process involved in developing the model. The diagram represents the key steps involved in the development of the proposed model. The sequence of operations like data processing, data cleaning and feature extraction takes place and in the end the classification is performed.

We addressed further in this paper several models of detection of fraud, innovative fraud detection techniques and the general conclusion that we concluded from this research following the examination of all strategies.

II. LITERATURE SURVEY

Fraud transactions are basically referred to an act of illegal activities for their personal benefits and financial profit. This act is intentional and considered a crime.

As per the explanation of G. Singh et al. Dempster-Shafer is an evidence theory, it combines all possible outcomes of the problem. Hence it is used to solve problems where there may be a chance that different evidence will lead to some different result. Dempster-Shafer theory and Bayesian Learning, a technique used in credit card fraud detection combines evidence from both current and past behavior. Dempster Shafer's theory used to combine multiple such shreds of evidence and an initial belief is computed. The transaction is accessed as normal, abnormal, or suspicious counting on this first 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 the significance of the categorization of models to credit card fraud detection issue and the writers introduced the 3 different models such as decision tree, neural network, and logistic regression. Out of these 3 models, logistic regression and neural network are far better than a decision tree [7].

As per the theory proposed by F. N. Ogwueleka et al. the chances of the theory for making choices under certainty. After reviewing of Bayesian theory, Naïve Bayes classifier and l-nearest neighbor classifier is mainly imposed to the dataset for the system of a credit card.[6]

As cited by Amian Kandu et al. A Hidden-Markov Model is a double embedded stochastic process used to model much more stochastic processes as compared to the traditional Markov model. An HMM is initially trained with the normal behavior of a cardholder. If an incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent [3].

In research by S.Benson et al. Hidden Markov model, each transaction is submitted to FDS. FDS receives the card details and the purchase value to verify whether it is a genuine transaction or not. If the transaction is confirmed as malicious or fraudulent, an alarm is raised and the bank declines the transaction [2].

The study about the detection of credit card fraud by Maja Puh et al. mainly used seven classification methods which took a major role. In this theory, they have basically comprised the decision tree and SVMs to reduce the risk of the banks. They also recommended artificial neural networks and Logistic regression classification models are far better and help to enhance the performance in detecting credit card frauds. Here the distribution of training data sets becomes more partial and the importance of all models reduced in catching the fraud transactions of credit cards.[4]

III. MODES FOR FRAUD DETECTION

Algorithms which can be used to form the model for frauds detection in transaction of credit cards are as follows:

A. Random Forest Classifier

It is a very essential classifier as it can be used to categories the information into one of many parts. This model can be attained by using the Random Forest class which is an essential part of the classification module. There is a requirement of the information in a significant format for better evaluation and performance. It also depends on several factors such as data, numClasses, categorical Features Info, num Trees, feature Subset Strategy, impurity, max Depth, max Bins, seed [9]

B. Decision Tree

The root node and each node essentially represent a "test" on a trait of an example in the dataset, the results of each test are conveyed by the resultant branches, and the node that does not have a branch is known as the leaf node and it also symbolizes the class labels. It is primarily determined by three elements, namely Occurrences – the collection of instances for which a class label has already been determined.[8] The class label attribute is said to insert is called Target Attribute. The list of predictor attributes is called Attributes List.

C. Support Vector Machine

It is another classification algorithm that categories the information or data into one category on the basis of the training data set. Support vector Machine significantly forms a model to determine the more possible margins [2]. SVM forms a separating hyperplane by converting the data into greater dimensions where the data is separable by using the kernel trick.

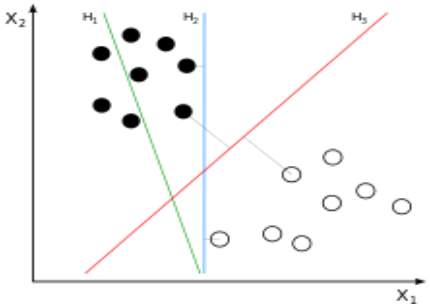

Fig.2. SVM Model

H1 does not separate the classes H2 does, but only with a small margin. H3 separates them with the maximal margin

IV. NOVEL TECHNIQUE TO DETECT CREDIT CARD FRAUD

A. INTRODUCTION

BLAH-FDS algorithm is the improved form comprised of BLAST and SSAHA algorithm. These two algorithms are pretty much proficient sequence aligning algorithms in detecting credit card frauds. In the sequence alignment of the BLAH-FDS algorithm, there are two stages where a profile analyzer obtains the correspondence between the transactions that are incoming in sequence with all the past and the sequence of genuine transactions made within the past [1]. The abnormal transactions detected by the profile analyzer are then passed into a deviation analyzer for checking with the past fraudulent transaction's behavior if present. Thus, based on these two analyzers a conclusion is drawn and the final decision is taken. The performance of this mechanism in detecting MasterCard frauds is sweet and its accuracy is high. Also processing speed is fast but the problem using this credit card fraud detection approach is that it cannot detect duplicate transactions or cloned credit card frauds.

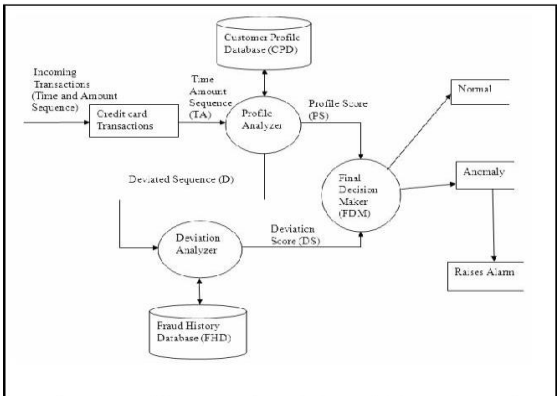

Fig.3. Structure of Model (BLAST-SSAHA)

B. DEFINITION AND NOTATION

BLAH is a two-stage algorithm. A clustered k-tuple table is created in the first stage which is used in the second stage to find the database similarity regions [10]. k-tuple: Let S¼ <s1;s2;s3:::::s4>be a sequence of length n. Then, any consecutive k elements of this sequence form a k-tuple. Two k-tuples are called

overlapping if they share some elements between them. The total number of overlapping k-tuples in S is kþ1. [4]. Tuple-weight: Every distinct k-tuple is assigned a unique integer value which is called Tuple-weight.

Tuple-weight	Sequence-index	Sequence-offset
AB_1	1	0
AB_1	1	3
AC_2	1	6
AC_2	2	2
AC_2	2	4
BC_5	1	1
BC_5	1	4
CA_6	1	2
CA_6	1	5
CA_6	2	1
CA_6	2	3
CC_8	2	0
CC_8	2	5

Table 1.

Tuple-offset: The position in a sequence where a k-tuple starts is called tuple-offset. There are n-k+1 distinct Tuple-offsets in a sequence of length n [4]. Sequence-index: If a database has n number of sequences, then the sequence-index of the ith sequence in the database is i. [5]. Sequence-base: Number of distinct elements present in the sequences of the database is called the Sequence-base.

C. QUERY SEQUENCE SEARCH METHOD

In the query-sequence search method, a query sequence is aligned with the existing sequences in D using BLAST. The k-tuple table is used here to choose some database regions on which the alignment is performed. KT, thus, is useful in reducing the search space for the alignment process. The query sequence is broken into overlapping k-tuples and its Tuple-weight is evaluated. A list of Sequence-index and Sequence-offset is obtained from KT for each k-tuple in query sequence. The sequence-indexes are ordered according to the number of distinct k-tuples present in that sequence. Let us consider a query sequence < ABCACB > which gives five overlapping 2-tuples < AB >,< BC >,< CA >,< AC >,and < CB > .Table 2 shows thepositionsofthese2- tuples in D. As< BC > exists in S1 at offsets 1 and 4, Sequence-index column of the Table 2 contains {1, 1} and Sequence-offset column contains {1, 4} for < BC >. The information shown in Table 2 is generated from the k-tuple table. Next, the Sequence-indexes are arranged according to the number of distinct query k-tuples present in the database sequence. The ordered Sequence-indexes along with the positions of k-tuples for the above example are shown in Table 1. If there are many distinct query k-tuples in database sequence, it leads to good alignment score. For a given pair of query and database sequence, there can be different possible local alignments having alignment score above a given threshold. BLAST can identify these alignment regions by extending each small hit segment in both directions. The BLAH algorithm may be stopped once high alignment score is achieved. Since we keep overlapping k-tuple information in BLAH, the searching sensitivity of this algorithm does not defer from BLAST [1]. However, some additional database space is required for storing the k-tuple information. A profile database with p sequences of average length q requires p(q-k+1) number of database entries for k-tuple information. As an example, Table 1 contains 2-tuple information for the sequences S1 (of length 8) and S2 (of length 7). It is seen that there are 13 entries in the table since in this example, p = 2; q= 7:5, and k = 2.

Query k-Tuples and Their Occurrences in D

Tuple-weight	Sequence-index	Sequence-offset
AB_1	{1,1}	{0,3}
BC_5	{1,1}	{1,4}
CA_6	{1,1,2,2}	{2,5,1,3}
AC_2	{1,2,2}	{6,2,4}
CB_7	{}	{}

Table 2.

V. CONCLUSION

An efficient credit card fraud detection system is an utmost important for any credit card issuing bank. Credit card fraud detection has drawn quiet a lot of attention of the research community and a number of techniques have been proposed to counter credit card fraud. The Fuzzy Darwinian fraud detection system improved the system's accuracy. The neural network based CARD WATCH shows good accuracy in fraud detection processing speed which is exceptionally high but is limited to one network per customer [1]. The fraud detection rate in Hidden Markov Model is very less as compared to other methods. The hybridized algorithm named BLAH-FDS identifies and detects fraudulent transactions using the sequence alignment tool. The processing speed of BLAST-SSAHA is fast enough to enable on-line detection of

credit card fraud. BLAH-FDS can be effectively used to counter frauds in other domains such as telecommunication and banking fraud detection.

REFERENCES

[1]. Sunil Bhatia, Rashmi Bajaj, Santosh Hazari, "Analysis of Credit Card Fraud Detection Techniques", Index Copernicus Value (2013)
[2]. S.Benson Edwin Raj, Annie Portia, "Analysis on credit card Fraud detection Methods",2011.
[3]. Amian Kandu, Suvasini Panigrahi, Shamik Sural, Arun K.Majumdar, "BLAST-SSAHA Hybridization for Credit Card Detection", 2009.
[4]. Maja Puh, Ljilijana Brkic, "Detecting credit card Fraud Using Selected Machine Learning Algorithms", MIPRO 2019
[5]. B.Meena, I.S.L.Sarwani, S.V.S.S.Lakshmi," Web Service mining and its techniques in Web Mining" IJAEGT,Volume 2,Issue 1 , Page No.385-389.
[6]. F. N. Ogwueleka, "Data Mining Application in Credit Card Fraud Detection System", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2011.
[7]. G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "A Machine Learning Approach for Detection of Fraud based on SVM", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2012, ISSN ISSN: 2277-1581.
[8]. "Survey Paper on Credit Card Fraud Detection by Suman", Research Scholar, GJUS&T Hisar HCE,Sonepatpublished by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014
[9]. "Research on Credit Card Fraud Detection Model Based on Distance Sum –by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence
[10]. SVSS & Kavila (2018) "Machine Learning For Credit Card Fraud Detection System" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 24 (2018) pp. 16819-16824