International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Artificial Intelligence based Credit Card Fraud Identification using Fusion Method

D.Uma Devi, Gnanaprakasam Thangavel, P. Anbhazhagan

***Abstract*: Increase of online transactions has given a greater scope for increasing of credit card frauds. In this work we develop a general framework with Artificial Intelligence based Hadoop. Also that fuses multiple detection algorithms to improve accuracy, reliability. Further to support large amount of transactions storage. The workflow satisfies the design ideas of current credit card fraud identification systems. The verification process for all the transactions is implemented. If incoming transaction that passed through trained model with low probability then it is rejected.**

***Index Terms*: Credit Card, Artificial Intelligence, Hadoop, Model Fusion, Fraud Detection.**

I. INTRODUCTION

Use of credit cards has increased in recent times due to rapid development of e-commerce and internet. Figure 1.1 shows the projected growth of global cards from 2012 to 2019 that was estimated by Statista, a statistics portal (https://www.statista.com/statistics/283572/projected-growth-in-global-credit-cards/). But unfortunately this also leads to increase in frauds as little important card information is enough for an unauthorized person to make a fraudulent transaction using the credit card. The online transactions are deceived by identification of the card details (e.g., Card ID, secure code); generally the types of deception are classified into:

i) Synthetic identity fraud is the use of probable but fictitious identities. These are trouble free to generate but harder to apply successfully.

ii) Real identity fraud is to illicit use of innocent people's complete identity details. These can be harder to obtain, but easier to successfully apply.

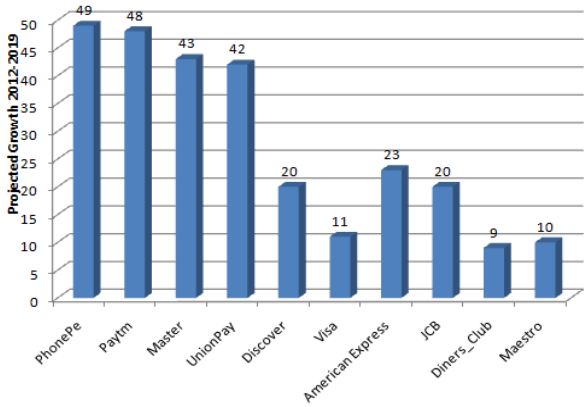

Figure 1.1 Projected growth of global cards from 2012-19

In reality, identity crime can be devoted to a mix of both synthetic identity fraud and real identity fraud details. Identity crime has become famous because there is so much real identity data available such as Web and unsecured mailboxes. It has also become easy for hackers to bury their accurate identities which are enabling them to make more frauds. In this paper we propose a Credit Card Fraud Detection model to overcome this problem by fusing multiple algorithms and using big data technology Hadoop.

II. LITERATURE SURVEY

Many new algorithms are developed by researchers for fraud detection to improve accuracy. Popular supervised algorithms include neural networks, logistic regression models, decision trees etc. Ghosh and Reilly proposed a feed forward back propagation neural network algorithm. The neural network was trained on examples of fraud due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud and NRI (non-received issue) fraud. The network detected significantly more fraud accounts (an order of magnitude more) with significantly fewer false positives (reduced by a factor of 20) over rule-based fraud detection procedures [2].Raghavendra Patidar, Lokesh Sharma used neural network along with genetic algorithm. Genetic algorithm was used for making the decision about the network topology, number of hidden layers, and number of nodes that will be used in the design of neural network [3].

Several fusion models are also proposed. Montek Singh, Ashraf Zakee proposed fusion of Hidden Markov Model and Naive Bayesian. The transaction amount outliers are calculated using Hidden Markov Model which were combined using Dempster Shafer Adder. If the value obtained is high then the transaction is considered fraudulent else Bayesian Learning methodology was used [6].

Manuscript published on November 30, 2019.
* Correspondence Author
Dr. D. Uma Devi*, Associate Professor, Department of IT, Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam, India.
Dr. Gnanaprakasam T, Associate Professor, Department of CSE, Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam, India..
Dr. Anbhazhagan, Assistant Professor, Department of IT, Gayatri Vidya Parishad College of Engineering (Autonomous), Visakhapatnam, India.

You Dai, Jin Yan, Xiaoxin Tang, Han Zhao and Minyi Guo proposed fusion of supervised and unsupervised algorithms with a new framework which consists of four layers: distributed storage layer, batch training layer, key-value sharing layer and streaming detection layer. With the four layers, we are able to support massive trading data storage, fast detection model training, quick model data sharing and real-time online fraud detection, respectively and implemented with latest big data technologies like Hadoop, Spark, Storm, HBase, etc[1].

V. Bhusari, S. Patil proposed a HMM based system that initially studied spending profile of the card holder and followed by checking an incoming transaction against spending behaviour of the card holder [5].

K. Rama Kalyani, D. Uma Devi proposed the use of Genetic Algorithm that is based on customer behaviour [4].The algorithm is an optimization technique and evolutionary search based on the principles of genetic and natural selection, heuristic used to solve high complexity computational problems.

Aman Gulati, Prakash Dubey, Md. Fuzail C, Jasmine Norman and Mangayarkarasi R proposed a methodology which facilitates the detection of fraudulent exchanges while they are being processed by means of Behaviour and Locational Analysis which considers a cardholder's way of managing money and spending pattern. A deviation from such a pattern will then lead to the system classifying it as suspicious transaction and will then be handled accordingly [7].

The proposed workflow in this paper is a fusion approach of supervised algorithms namely Logistic Regression, Support Vector Machines, Bayesian Classifier and Decision Trees including Big Data technology Hadoop.

III. PROPOSED METHODOLOGY

Figure 3.1 shows the proposed methodology of this work,

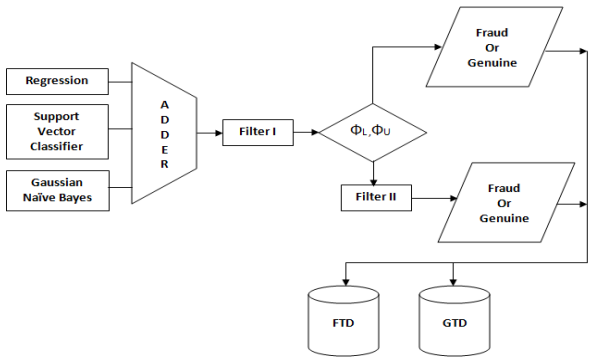

Figure 3.1 Proposed Method of Data Flow

Given a transaction T, it will be processed through the following components.

A. Regression

It is used to describe the data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. It uses categorical variables as dependent variable using a log function explaining the probability of success or failure. If X is an independent variable and a, b are parameters of the model then log function is given as

$$P = \frac{eA+Bx}{1+eA+Bx}$$

B. Support Vector Classifier

In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes. In case of multiple hyper planes it selects the hyper-plane which classifies the classes accurately prior to maximizing margin.

C. Gaussian Naive Bayesian Classifier:

Naive Bayesian classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. If the input variables are real-valued, a Gaussian distribution is assumed. This model is easy to build and useful for large datasets. Probability distribution function is given as:

$$\text{PDF}\ (X, \text{Mean}, \text{Std}_{\text{Dev}}) = \left(\frac{1}{Sqrt\ (2*\pi)} * Std_{Dev}\right) * \frac{e-(X-Mean^2)}{2*Std_{Dev}2}$$

Where,

X = Transaction

$$Mean\,(X) = \frac{1}{n} * Sum\ (X)$$

$$\text{Std}_{\text{Dev}}(X) = Sqrt(\frac{1}{n} * Sum\ (Xi - Mean\,(X)^2)$$

D. Adder

At this stage the fraud scores are merged to generate a merged result *f(t)*

$$f(t) = f_1(t) \oplus f_2(t) \oplus f_3(t)$$

E. ΦL and ΦU

The threshold value is to determine whether transaction is genuine or fraud or suspicious. θU is the Upper bound of fraud transactions and θL is the lower bound of genuine transactions.

When f(t)<ΦU the transaction is fraudulent.

When f(t)> ΦL the transaction is genuine.

When ΦU< f(t) <ΦL the transaction is suspicious.

F. Filter 1*:*

At filter F1 we used three supervised algorithms Logistic Regression (LR), Gaussian Naive Bayesian (GNB) and Support Vector Classifier (SVC). Each algorithm learns patterns from historical data and fraud scores are calculated for an incoming transaction based on the pattern. Fraud score fi(t) is the probability of whether transaction is genuine or not.

G. Filter 2:

Further decision that whether the transaction is genuine or fraudulent is made. Decision Tree Classifier (DTC) and Gini Index values are used in this filter operation.

a. Decision Tree Classifier:

Decision Tree is a non-parametric method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The tree is constructed by recursive partitioning based on the Gini Index.

b. Gini index:

Gini index says that if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure and is calculated as sum of squares of probabilities for success and failure. Node split will take on the feature that has higher Gini index.

IV. EXPERIMENTAL SETUP

A. Transaction Dataset

For experimental purpose the dataset is taken from (https://www.kaggle.com/mlg-ulb/creditcardfraud/data) which has 284807 transactions out of which 492 are fraudulent i.e., a highly unbalanced one. Each transaction record has 30 payment attributes and a class attribute where value 0 corresponds to Genuine and 1 corresponds to Fraud. A total of 12 attributes are removed from dataset as they are proven unimportant by correlation graphs and amount attribute is scaled to [-1,1]. Then, 25% of the dataset is taken as test data in random fashion.

Table 4.1 Training and Testing Analysis

Dataset/Class	Class 0	Class 1
Training Set	213606	372
Testing Set	71201	120

B. Performance Analysis

Execution time taken for training the model on single node cluster of hardware configuration: Core i5 processor, 4GB RAM is:

Table 4.2 Performance Analysis

Algorithm	Time taken to training set (in ms)
LR	6.506594
SVC	0.191343
GNB	0.487519
DTC	9.047001

Implementing the above method in various algorithms, we are getting the following performance.

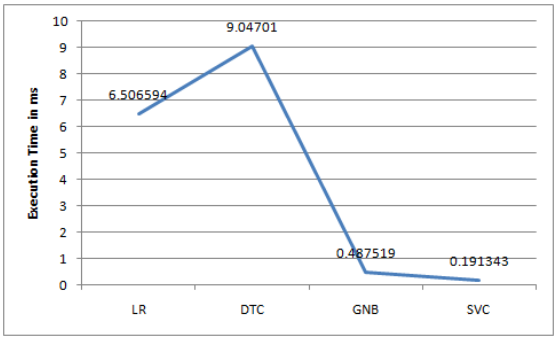

Figure 4.1 Performance Analysis

Figure 4.1 illustrate the execution times of different algorithms, in which SVC taken less time for the computation. Among all the algorithms SVC provided solution with less time.

V. CONCLUSION

Fraud detection has become quite a challenge in recent times. Our framework aims at solving this problem by fusion of detection algorithms and implementation of Artificial Intelligence based big data approach to provide scalable and reliable system. However, the work has lots of things to do e.g., optimisation of the model, testing with larger amounts of real transaction data on better hardware configured system with multiple nodes. Further the future researchers can extend this work for different kinds of datasets like image, audio and video. For text data, SVC provided good result. Different algorithms may be used for different forms of data in future.

REFERENCES

1. You Dai, Jin Yan, Xiaoxin Tang, Han Zhao and Minyi Guo, 2016, "Online Credit Card Fraud Detection: A Hybrid Framework with Big Data Technologies", IEEE International Conference , 23-26-Aug-2016.
2. S. Ghosh and D. L. Reilly, 2011, "Credit card fraud detection with a neural network," International Journal of Computer Applications, Volume:1, Issue:1, pp: 28-32.
3. Raghavendra Patidar, Lokesh Sharma, 2012, "Credit Card Fraud Detection Using Neural Network", International Journal of Soft Computing and Engineering, Volume:1, Issue:1, pp: 32-38.
4. K.Rama Kalyani, D.UmaDevi, 2012, "Fraud Detection of Credit Card Payment System by Genetic Algorithm", International Journal of Scientific & Engineering Research, Volume 3, Issue 7, pp: 1-6.
5. V. Bhusari, S. Patil , 2012, "Application of Hidden Markov Model in Credit Card Fraud Detection", International Journal of Distributed and Parallel Systems, Volume 2, Issue 6, pp: 203-211.
6. Montek Singh, Ashraf Zakee, "Credit Card Fraud Detection Using Hidden Markov Model, Dempster–Shafer Theory and Bayesian Learning a Better Approach to Credit Fraud Detection", International Journal for Research in Applied Science and Engineering Technology, Volume 6, Issue 7, pp: 112-118
7. Aman Gulati, Prakash Dubey, Md. Fuzail C, Jasmine Norman and Mangayarkarasi R, "Credit card fraud detection using neural network and geo-location", Materials Science and Engineering, Volume 263, Issue 4, pp: 39-42.

AUTHORS PROFILE

Dr. D. Uma Devi currently working as Associate Professor in the department of Information Technology, Gayatri Vidya Parishad College of Engineering (A). She have 18 years of experience in Teaching and Research. She have published papers in Scopus indexed journals. Her research interests are Safety Critical Systems, Artificial Intelligence, Machine Learning and Deep Learning.

Dr. Gnanaprakasam Thangavel currently working as Associate Professor in CSE at Gayatri Vidya Parishad College of Engineering (Autonomous), Vizag. He have 10+ years of experience in Teaching and Research. He have published more than 10 research papers in SCI, and Scopus indexed journals. His research interests are Cloud Security, Software defined Networks and Artificial Intelligence,

Dr. P. Anbhazhagan currently working as Assistant Professor in the department of Information Technology, Gayatri Vidya Parishad College of Engineering (A). Having 8+ years of teaching experience in Computer Science and Information Technology Departments, guided the PG and UG students in Computer Networks, Wireless Sensor Networks, Image Processing and Internet of Things. His Research areas includes Wireless sensor networks, Image processing and Internet of Things. He authored and coauthored more than 5 international articles and presented 3 papers in International conferences.