Credit Card Fraud Detection Technique using Hybrid Approach: An Amalgamation of Self Organizing Maps and Neural Networks

Harsh Harwani[1], Jenil Jain[1], Chinmay Jadhav[1], Manasi Hodavdekar[2]

[1]Undergraduate Research Scholar, Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai-50, Maharashtra, India
[2]Undergraduate Research Scholar, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai-53, Maharashtra, India

---

Abstract - *Fraud is an extensive term that refers to acts intended to swindle someone. Millions of people each year fall victim to it. There are several ways to commit fraud, as criminals keep on finding new ways to gain wealth by cheating someone. The most common varieties of fraud are committed through media, including mail, phone, and the Internet. Credit card fraud falls under the category of Internet fraud. Credit card frauds take place regularly and as a result, lead to huge financial losses. The main reason behind the increasing Credit card frauds is the surge in online transactions which thereby lead to the hijacking of personal details. Thus, a powerful fraud detection system is the need of the hour. The system needed must learn from past committed frauds and should be proficient in identifying frauds in the future with impeccable accuracy. Over the years, various techniques are used for fraud detection system viz. Support Vector Machine (SVM), K-nearest Neighbour (KNN), Fuzzy Logic, Decision Trees, and many more. All these techniques have yielded decent results but to improve the accuracy even further, a hybrid learning approach is needed for detecting frauds. In this paper, the hybrid learning approach i.e. a two-step approach is implemented.*

***Key Words*: credit card fraud, fraud detection system, Support Vector Machine, K-nearest Neighbour, Fuzzy Logic, Decision Trees, hybrid learning approach.**

1. INTRODUCTION

The use of credit cards has been increasing drastically with the progression of state-of-art technology and worldwide communication. Transactions completed with credit cards seem to have become more in demand for the introduction of online shopping and banking. But this increase in credit card users has paved a smooth way for credit card fraudsters. Credit card fraud can be classified into various categories:

1. Counterfeit credit cards-

To create fake cards criminals, use the most recent technology to "skim" information contained on magnetic strips of cards and to pass security measures like holograms.

2. Lost or stolen cards-

Fraudsters use the misplaced cards or stolen cards to carry out online transactions as using an ATM requires a PIN.

3. No-Card frauds-

No-Card frauds occur when cardholders give credit card information such as credit card number, PIN, or CVV on the phone to strangers or fraudsters acting as bank employees and also deceptive Internet sites that sell non-existent goods.

4. Non-Receipt fraud-

It occurs when a customer has applied for a credit card and the card is lost or stolen during the process of being mailed. This fraud is also called Never Received Issue fraud.

5. Identity Theft fraud-

Identity Theft Fraud refers to fraud when someone applies for a credit card using someone else's identity and information.

The study proposed in this paper aims at addressing these frauds through the hybrid approach. Machine learning and Deep learning are the generation's solution which replaces such methodologies and can easily work on large datasets which is not possible for humans. During the last few years, many supervised and unsupervised algorithms have been used. In this study, the Hybrid Deep Learning approach is used. Firstly, Unsupervised learning approach is used to find potential frauds using Self Organizing Maps [SOM], and then the supervised learning approach is implemented using Artificial Neural Network [ANN] with output from Self Organizing Maps as the target variable of the network.

The very nature of this study allows for multiple algorithms to be integrated as modules and their results can be combined to increase the accuracy of the final result.

The data which is being used in this study is Credit Approval Dataset from the UCI repository [1]. The dataset is chosen due to its contents i.e. it contains about 15 features divided into 3 classes (integer, real and categorical) which would surely help in training the model to achieve high accuracy. This dataset was very interesting because there was a good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of value.

This paper aims at eliminating the four fraud patterns via the hybrid approach. The rest of the paper is as follows Section 2 presents the dataset description. Section 3 presents the Literature Review. The Hybrid Approach is explained in Section 3, followed by the Conclusion and Future Scope in sections 4 and 5 respectively.

2. DATASET DESCRIPTION

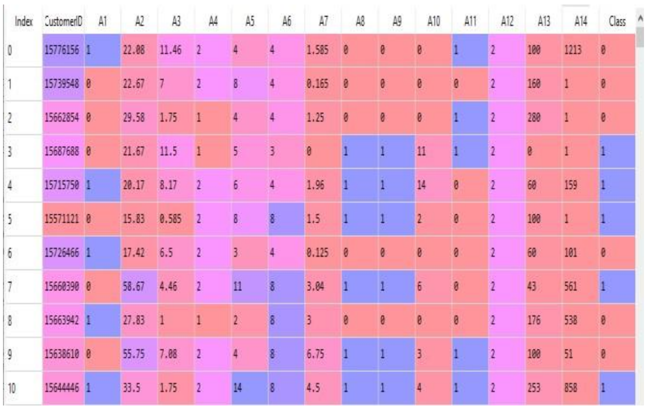
All the Attribute names have been changed and encoded to meaningless names/symbols to protect the confidentiality of the data. The dataset is made of multiple and a good mix of attributes such as continuous, nominal with small number of values, and nominal with larger numbers of values [1].

Incorrect or inconsistent data can create a vast number of problems which affects the result and leads to drawing of incorrect conclusions and predictions which results in the system giving inefficient efficiency. The dataset included missing values, inconsistencies due to inaccurate entry, database corruption etc.

All necessary steps to minimize inconsistency have been taken into consideration and all the missing values have been removed.

The 15 features of the data set are merely divided into three classes: -

- Integer
- Real
- Categorical



Index	CustomerID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	Class
0	15776156	1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	0
1	15739548	0	22.67	7	2	8	4	0.165	0	0	0	0	2	160	1	0
2	15662854	0	29.58	1.75	1	4	4	1.25	0	0	0	1	2	280	1	0
3	15687688	0	21.67	11.5	1	5	3	0	1	1	11	1	2	0	1	1
4	15715750	1	20.17	8.17	2	6	4	1.96	1	1	14	0	2	60	159	1
5	15571121	0	15.83	0.585	2	8	8	1.5	1	1	2	0	2	100	1	1
6	15726466	1	17.42	6.5	2	3	4	0.125	0	0	0	0	2	60	101	0
7	15660390	0	58.67	4.46	2	11	8	3.04	1	1	6	0	2	43	561	1
8	15663942	1	27.83	1	1	2	8	3	0	0	0	0	2	176	538	0
9	15638610	0	55.75	7.08	2	4	8	6.75	1	1	3	1	2	100	51	0
10	15644446	1	33.5	1.75	2	14	8	4.5	1	1	4	1	2	253	858	1

Fig -1: Dataset

3. LITERATURE REVIEW

J. Esmaily, R.Moradinezhad [2]. Proposed a hybrid of Artificial Neural Network and Decision Trees, in 2015. One of the reasons to use this model was because it promises reliability by giving very low false detection rate. Their model consists of a two-phase approach, wherein the first phase was the classification results of Decision Trees and Multilayer perceptron. This first layer was used to generate a new dataset which in turn was fed into Multilayer perceptron in the second layer to classify the data. In 2011, Siddhartha Bhattacharyya, Sanjeev Jha, Kurian Tharakunnel and J. Christopher Westland [3] conducted a comprehensive comparative study on Support Vector Machine (SVM) and Random Forest along with Logistical Retrogression. This model is very accurate by providing very low detection rates. In 2011, Raghavendra Patidar and Lokesh Sharma [4] proposed the Artificial Neural Network and Genetic Algorithms hybrid. They concluded with experiments to show that Random Forest methodology is most accurate, followed by Logistic Regression and Support Vector Machines. They utilized neural nets to classify transactions & genetic algorithms so that solution is optimized & the system is not trained. In 2015, Tanmay Kumar and Suvasini Panigrahi [5] in this paper, they proposed a novel approach credit card detection in which the fraud detection is done in three phases. The first phase does the initial user authentication and verification of card details. If the check is successfully cleared, then the transaction is passed to the next phase where fuzzy c-means clustering cleared algorithm is applied to find out the normal usage patterns of credit card users based on their past activity. In another paper published by Wen-Fang YU & Na Wang [6] proposed the distance-based method. This method judges whether it is outlier or not according to the nearest neighbors of data objects. They only showed the highest accuracy of about 89.4 percent but did not talk about FP & FN. Ayushi Agrawal and others [7] proposed testing a transaction, wherein they used the Hidden Markov Model to maintain the record of previous transactions, Behavior based technique for grouping of datasets and lastly genetic algorithm for optimization i.e. calculating the threshold value. Sam Maes [8] proposed detecting frauds in credit card using two machine learning techniques namely Bayesian Networks and Artificial Neural Network. The paper discussed that how Bayesian Networks after a short training gave good results and their speed was enhanced by the use of ANN. Y. Sahin and E. Duman [9] proposed fraud detection in credit card using a combination of Support Vector Machines and Decision Trees. Geoffrey F.Miller, Peter M.Todd and Sailesh Hegde [10] have elaborated the concept of designing of Neural Networks using Genetic Algorithms. It aims to free the network design process from the constraints of human biases. They built a system which would have applications in biological, neurological and psychological modelling as well as the engineering and design applications using automated network design.

Ekrem Duman and M. Hamdi Ozcelik [11] proposed a system to credit each transaction a certain score and based on that score the transaction was judged, and to implement this they combined Neural Networks with Scatter Search. Alireza Pouramirarsalani1, Majid Khalilian, Alireza Nikravanshalmani [12] proposed a new method of fraud detection which used a hybrid of feature selection and genetic algorithm. They observed the salient features of the transactions and used the same while detecting any unusual feature and flagging it to be the fraud one. Pooja Chougule and others [13] in their paper proposed simple K-means and Simple Genetic Algorithm for fraud detection. They showed that how k-means algorithm grouped the transactions based on the distinct attribute values and genetic algorithm. This was used for optimization since with the increase in size of the input k-means algorithm produced outliers. S.Fashoto, O.Adeleye and J.Wandera [14] have used a hybrid of K-means clustering with Multilayer Perceptron (MLP) and the Hidden Markov Model (HMM) in their paper. They have used K-means clustering in order to group together the suspected fraudulent transactions into a similar cluster. The output of this stage is used to train the HMM and the MLP which then classify the incoming transactions. M.R. Harati Nik, M. Akrami, S. Khadivi and M. Shajari [15] in their paper have proposed a fusion on Fuzzy expert system and Fogg behavioral analysis thus naming it the Fuzzy hybrid model. The Fogg behavioral model describes the merchant behavior in two dimensions: motivation and ability to make a fraud. The fraud tendency weight is then calculated for each merchant followed by the degree of suspicion for the incoming transactions. Krishna K. Tripathi and Mahesh A. Pavaskar [16] have done a comparative study of different techniques in their paper and one of the techniques they have worked upon is a fusion of Dempster-Shafer theory and Bayesian learning which combines the evidences or datasets from past as well as the current behavior. The rule-based filter transaction history and Bayesian learner are the 4 stages of this system via which we decide the suspicious and unsuspicious transactions altogether. In the first component the extent to which the incoming transaction has deviated is determined so as to get the suspicion level.

4. HYBRID APPROACH

In our model, we've used a hybrid of SOM and ANN.

The dataset which is employed was unlabeled, therefore unsupervised algorithm namely Self Organizing Map is used in order to seek out the outliers within the data, further improving precision using Artificial Neural Network.

The main objective of SOMs is to rework a posh high dimensional discrete input space into an easier low-dimensional discrete output space by preserving the topology within the data but not the particular distances. It is an unsupervised learning algorithm which uses simple heuristic method capable of discovering hidden non-linear structure in high dimensional data. SOMs are more advantageous to use than other clustering algorithms because they are doing not make assumptions regarding the distributions of variables nor do they require independence among variables they're easier to implement and are ready to solve non-linear problems of high complexity. They effectively deal with noisy and missing data, very small dimensional and samples of unlimited size.

In order to perform unsupervised learning, SOMs apply a competitive learning rule where the output neurons compete among themselves for the chance to represent distinct patterns within the input space.

SOM mapping steps starts from initializing the load vectors. From there a sample vector is chosen randomly and the map of weight vectors is searched to seek out which weight best represents that sample. Each weight vector has neighboring weights that are on the brink of it. The load that is chosen is rewarded by having the ability to become more like that randomly selected sample vector. The neighbors of that weight also are rewarded by having the ability to become more just like the chosen sample vector. This enables the map to grow and form different shapes. Most generally, they form square/rectangular/hexagonal/L shapes in 2D feature space.

Algorithm:

1. The load of every node is initialized.
2. A vector is chosen at random from the set of training data.
3. Every node is examined to calculate the weight and compared to the input vector. The winning

node is commonly referred as the simplest Best Matching Unit (BMU).

4. Then the neighboring nodes of the BMU is calculated. The number of neighbors decreases over time.

5. The winning weight is rewarded with becoming more like the sample vector. The neighbors also become more just like the sample vector. A node that is closer to the BMU, undergoes a higher alteration in its weight than a node that is farther. A neighbor that is farther away from the BMU, the less it learns.

6. Repeat step 2 for N iterations. Best Matching Unit is a technique which calculates the distance from each weight to the sample vector, by running through all weight vectors. The load with the shortest distance is the winner. There are numerous ways to work out the distance, however, the foremost commanly used method is the *Euclidean Distance* which is being proposed in the paper.

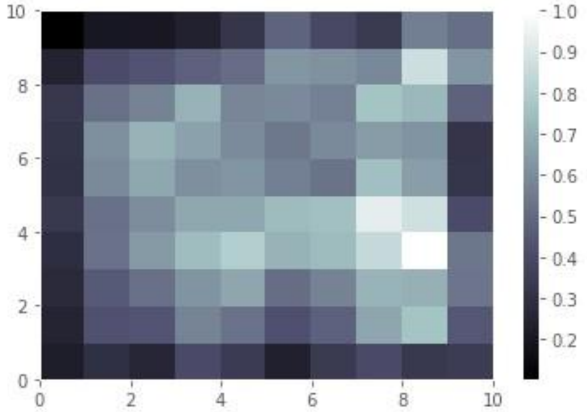


Fig -2: Heatmap

Inference:

If the typical distance is high, then the encompassing weights are very different, and a light color is assigned to the location of the weight. If the average distance is low, a darker color is assigned. The above heatmap shows that the concentration of different clusters of species are more predominant in three zones. First figure tells us only about where the density of species is bigger (darker regions) or less (lighter regions). The second visualization tells us how they are specifically clustered.

Artificial Neural Network:

Further we created a Neural Network with three layers. The output generated by the Self Organizing Map is taken as a Target (customer being fraud or not) and the fifteen attributes as features. First layer with 15 input nodes as we have 15 attributes in our dataset and apply Relu (Rectified Linear Unit) activation function to the nodes. The second layer has 2 nodes with Relu as activation function and third layer with one node tells the probability of the customer being fraud. Sigmoid activation function is applied which fits best for output layer.

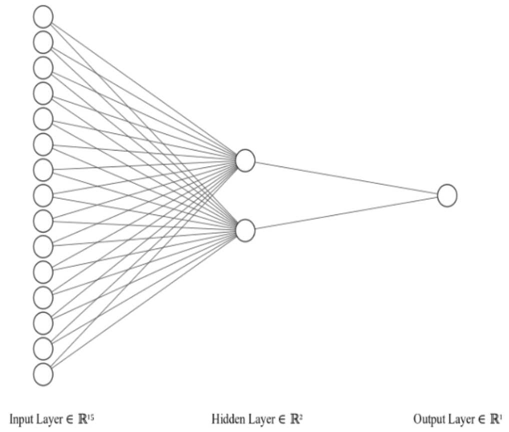


Fig -3: Artificial Neural Network

5. CONCLUSIONS

Out of the plethora of fraud detection techniques available today, most of them detect the fraud after it has been committed. This creates a need of real time system which needs to be in place which will not only be able to detect these frauds taking place but also have the ability to catch these frauds in real time. One of the reasons of this drawback is that out of all the transactions that take place a very small number of frauds are only fraudulent, and the rest are legitimate transactions. Hence, one can say that lack of fraudulent data is a major reason for this drawback.

There are quite a few drawbacks with the techniques vastly used today, some of which are that they do not give the same result when applied to a different dataset i.e. in different environment the technique's performance will vary. One of the ways to overcome these drawbacks is to use a Hybrid Approach i.e. merge two or more technique's together to give better, accurate and sustainable result.

J. Esmaily and R. Moradinezhad [2] have proposed a hybrid of Decision Tree and Neural Network; R. Patidar and L. Sharma have proposed a hybrid of Neural Network and Genetic Algorithm [4]; T. Kumar and S. Panigrahi [5] have proposed a hybrid of Fuzzy Clustering and Neural Network Similarly In our system we have proposed a hybrid of combining SOM and ANN technique together to cancel out the limitations of using a single technique and to get enhanced results. One of the major reasons to combine SOM with ANN was to enhance the result. With this model, we achieved better accuracy precision and cost compared to the using SOM or ANN alone.

An important aspect to keep in mind while developing a good hybrid model is to always pair an expensive technique with takes long time to run but gives efficient results with an optimizing technique which will help in reducing down the cost of the entire system.

6. FUTURE SCOPE

From the working of the model, it is clear that Artificial Neural Network improves the accuracy in this scenario. Artificial Neural Networks on the other hand can be easily over-trained and also, they are very expensive to train. This can be overcome by creating a neural network with some optimization technique, thus reducing the expense. Some of the optimization techniques that could be used with Neural Network are Genetic Algorithm, Artificial Immune System, Case Based Reasoning and any other similar technique. Genetic Algorithm helps by selecting the optimized weight of the edges in neural network. Case Based Reasoning first tries to predict the outcome on the basis of a direct match with the user's profile and Artificial Immune System reduces the cost by eliminating the weights that cause the maximum error.

REFERENCES

[1] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository Credit Approval Dataset. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Credit+Approval

[2] Esmaily, J., Moradinezhad, R., & Ghasemi, J. (2015). Intrusion Detection System based on Multilayer Perceptron Nueral networks and Decision Tree. 2015 7th Conference on Information and Knowledge Technology. Urmia,Iran.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, C. (2011). Data Mining for Credit Card Fraud: A Comparitive Study. Elsevire, 50, 602-613.

[4] Patidar, R., & Sharma, L. (2011). Credit Card Fraud Detection Using Neural Network. International Journal of Soft Computing and Engineering, 32-38.

[5] kumar, T., & Panigrahi, S. (2015). Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering and Neural Netwrok. Second International Conference on Advances in Computing and Communication Engineering. Dehradun,India: IEEE.

[6] Yu, W.-F., & Wang, N. (2009). Research on Credit Caard Fraud Detection Model Based on Distance Sum. International Joint Conference on Artifical Intelligence. Hainan Island, China: IEEE.

[7] Agarwal, A., Kumar, S., & Kumar, A. (2015). Credit Card Fraud Detection: A Case Study. 2nd International Conference on Computing for Sustainable Global Development . Delhi, India : IEEE.

[8] Maes, S., Puyls, K., Vanschoenwinkel, B., & Manderick, B. (2002). Credit Card Fraud Detection Using Bayesian and Neural Network.

[9] Razoogi, T., Khurana, P., Raahemifar, K., & Abhari, A. (2016). Credit Card Fraud Detection using Fuzzy logic and Neural Network. Proceedings of The 19th Communication and Networking Symposium, (pp. 1-5).

[10] Y, S., & E, D. (2011). Detecting Credit Card Fraud by Decision Trees and Support Vector Machines . Proceedings of the International MultiConference of Engineering and Computer Scientists. Hong Kong.

[11] Miller, G., Todd, P., & Hegde, S. (1989). Desiging Neural Networks using Genetic Algorithms. Proceeding of 3rd International Conference on Genetic Algorithms, (pp. 379-384). Fairfax, Virginia.

[12] Duman, E., & Ozcelik, H. (2011). Detection Credit Card Fraud by Genetic Algorithm and Scatter Search. Expert Systems with Applications: An International Journal, 13.57-13.63.

[13] Chougule, P., Thakare, A., Kale, P., Gole, M., & Nanekar, P. (2015). Genetic K-Means Algorithm for Credit Card Fraud Detection. International Journal of Computer Science and Information Technologies.

[14] Fashoto, S., Owolabi, O., Adeleye, O., & Wandera, J. (2016). Hybrid Methods For Credit Card Fraud Detection. British Journal Of Applied Science and Tenchonlogy .

[15] Reza, M., Akrami, M., Khadibi, S., & Shajari, M. (2012). FUZZGY: A Hybrid Model for Credit Card Fraud Detection. 6th International Symposium on Telecommunications. Tehran, Iran: IEEE.

[16] Tripathi, K., & Pavaskar, M. (2012). Survey On Credit Card Fraud Detection Methods. International Jounal Of Emerging Technology and Advanced Engineering.