Available online at http://icsejournal.com/index.php/JCSE
Journal of Computer Science an Engineering (JCSE)
e-ISSN 2721-0251
Vol. 2, No. 2, August 2021, pp. 112-119

A One Step Further Approach to Fraud Detection

Debjyoti Bagchi[1*], Abhishek Mukherjee[2], Sarannak Pal[3]

[1]Dept. of Computer Science and Engineering, Calcutta Institute of Engineering and Management, Kolkata, India
[2]Dept. of Computer Science and Engineering, Calcutta Institute of Engineering and Management, Kolkata, India
[3]Dept. of Computer Science and Engineering, Calcutta Institute of Engineering and Management, Kolkata, India
* corresponding author

ARTICLE INFO

Article History:
Received August 16, 2020
Revised September 21, 2020
Accepted September 26, 2020

Keywords:
Machine Learning,
Unsupervised learning,
Supervised Learning,
Neural Network,
Data mining

Correspondence:
Telephon: +91 8336849046
E-mail: sarannakpal@gmail.com

ABSTRACT

This paper will discuss about the different approaches to fraud detection such as Data mining, machine learning and artificial intelligence and statistical data analysis. Then we list some of the technical and troublesome challenges to modern fraud detection techniques. A comparison study of these techniques is also done according to the metrics like precision, False Alarm Rate (FAR), Accuracy, Cost, True Positive Rate (TPR) against different categories of frauds such as internal bank fraud, credit card fraud, loan fraud. Finally, we discuss the disadvantages of the existing fraud detection systems and we attempt to recommend a specific technique or algorithm for detecting a specific type of fraud with their advantages and disadvantages.

1. Introduction

When someone defrauds a person of his/her money or other assets from you through deception or illegal action which hurts the financial well-being is known as financial fraud. According to [1] financial frauds in India are increasing rapidly. The number of frauds doubled in the financial year 2019-20 than the previous year worth nearly Rs 1.85 lakh crore. In 2019-20, the maximum count of fraud cases involving Rs 1 lakh or more accelerated by 28% in volume and 159% in value [2]. Hence it is important to develop new fraud detection systems using latest technologies as machine learning, deep learning, pattern recognition, etc. to counter the new kinds of frauds and decrease the fraud rate which also helps in the growth of the economy.

Different approaches to fraud detection, data mining methods [3] discover relationships and rules in data to discover interesting fraud patterns. This is done by classifying, clustering and finally segmenting the data. Machine learning algorithms [4] are suitable for learning fraud patterns from data and identify them in future. Machine learning algorithms are of 2 types (i) Supervised Learning - For this kind of learning curve the data is tagged as either fraudulent or non-fraudulent. e.g., Convolutional Neural Networks[5], Logistic Regression[6], SVM[7], Random Forest[8], (ii) Unsupervised Learning - In this type the machine itself tries to find patterns in the input data(unlabeled). e.g., Apriori[9], K-means[10], FP Growth[11], Statistical techniques like probability distribution models, regression analysis, data matching, etc... are used to detect fraudulent activities. Data Matching[12] - An organization can check into the records of the departments where the company is losing money to uncover suspected fraudulent behavior by matching records that tally with each other within a database. Regression Analysis [13] - Frauds can be predicted by identifying the relationships between variables based on the patterns of fraud

http://dx.doi.org/10.36596/jcse.v2i2.250

variables in previous fraud data. Probability Distributions [14] - Various corporate fraud models and probability distributions are mapped, either in basis of distinct parameters or probability distributions. Clustering Analysis [15] - Cluster analysis is used in conjunction with outlier analysis to put an extra layer of scrutiny for suspicious activity, Challenges in Fraud Detection: Real world dataset for fraud detection is very rare, Fraudsters' erratic behavior, Highly skewed data, Huge data that require scalable algorithms, When new events are recognized, real-time analysis is particularly desirable for updating models.

There are several algorithms and techniques for detecting frauds out there but there is no particular algorithm or default choice for detecting a particular kind of fraud. A person or an organization trying to implement a fraud detection system for a particular type might get confused about which algorithm to select. In this paper we will try to make a classification of different kinds of frauds. Then compare the different algorithms and techniques for each kind of fraud. Finally, we try to suggest a particular technique or algorithm for detecting a particular kind of fraud.

We know very well that all kinds of frauds or fraudulent transactions follow a similar type of pattern. Like fraud detection Based on bagging ensemble classifier [16] , Fraud Detection System Using Machine Learning Techniques [17], Fraud detection using Bayesian and neural networks [18], detection using deep learning based on auto-encoder and restricted boltzmann machine [19], fraud detection using machine learning models and collating machine learning models [20], Fraud Detection Using Meta-Learning [21], fraud detection using self-organizing maps [22].

Bagging Ensemble Classifier Approach, Leo Breiman proposed the tagging classifier in 1994, and it is an ensemble technique. It has the ability to control classification and regression techniques. Its purpose is to improve the stability and accuracy of machine learning algorithms used in classification and regression. It creates a final forecast by merging classifications from randomly generated training sets.

Machine Learning Techniques, In this technique two different approaches are taken, which is Decision Tree and Logistic Regression. On This technique we get the following results which are Confusion Matrix, Accuracy (Decision Tree: 99.92%), Kappa value, Active Class, Balanced Accuracy Value. This strategy has been shown to be effective in lowering the amount of false positives and minimizing the number of fraudulent transactions. In terms of application domain, Machine Learning techniques are a new addition to the literature.

Bayesian and Neural Networks, A Bayesian network is a directed acyclic graph with such a well-defined collection of random variables as its nodes. There's a whole finite number of mutually exclusive states for every variable. Neural networks come in a number of different forms, Like Input Layer, Hidden Layer, output Layer. It's a common misconception that neural networks are a quick, simple, and dependable way to get good results in a variety of fraud detections. In fact, the most difficult part of using neural networks is deciding on a decent collection of pre-processing as well as a reasonable trade-off among the many parameters that must be determined.

Auto Encoder and Restricted Boltzmann machine, A restricted Boltzmann machine (RBM) and an auto encoder that can reassemble common transactions to detect deviations in usual patterns. The auto-encoder (AE) deep learning algorithm presented here is an unsupervised learning technique that uses backpropagation to make inputs and outputs equal. The Restricted Boltzmann technique has two layers. 1. Input Layer 2. Hidden Layer. And using this method we get mean squared error, root mean squared error and area under curve as a result.

http://dx.doi.org/10.36596/jcse.v2i2.250

Machine Learning Models and Collating Machine Learning Models, in this Approach three different kinds of models are taken. Which are SVM Model (Support Vector Machine), Logistic Regression and Random Forest. With the help of these models, we get results which are 97.5%, 97.7% and 98.6% respectively. After getting the result it was noticed that the Random Forest algorithm is giving us the highest and most precise result which is 98.6% accurate. It has been found that Random Forest Algorithm will give us better performance if we use larger training data sets. But the other two algorithms mainly suffer from the imbalanced data sets and they require more pre-processing.

Meta-Learning, This Meta Learning technique is used to learn models of fraudulent transactions. In this algorithm it is observed that despite the fact that the original data had a skewed distribution, the artificially more balanced training data proved to be a stronger classifier. This approach is used to combine different classifiers and maintain and also improves the performance of the best classifier. Using this Meta-Learning we get 50% / 50% distribution of fraud / non-fraud training data, and that is why this will generate the maximum TPR (true positive rate) and minimum FPR (false positive rate).

Self-Organizing Maps, This Self-Organizing Maps approach is used to create a model of typical human behavior in order to analyze transaction deviation and thus find anomalies in transactions. The SOM main principles are used to analyze the transaction. SOM is a neural network with a feed-forward topology and an unsupervised training model algorithm that creates output neurons based on the topological structure of the input data using a self-organizing process.

2. Method

For this time, we take each and every algorithm which is listed above, that works on the datasets i) Credit Card Fraud Detection Anonymized credit card transactions labeled as fraudulent or genuine [29] ii) Synthetic Financial Datasets For Fraud Detection Synthetic datasets generated by the PaySim mobile money simulatore [30] iii) Fraud detection bank dataset 20K records binary 20k records of customer transactions with 112 features [31] and do a comparative analysis among all of them. For this Test We Use a type of data set which satisfies all the requirements for each algorithm and have enough pre-processing so that we can test our Algorithms. Based on the result that is produced by this test we can get which algorithm is the best. Now There Might Be Two Qn's, (i) Why Do We Do this Test? (ii) What do we get from this test ?.

Answer for Q1, we have all these algorithms for so many years, but up to this point we don't know properly which algorithm is the best and when to use that in which circumstances. So, For This Answer we perform a Comparative Analysis test to get our answer. Because up to this point, we don't have any algorithm which is best for all the frauds and transactions. Answer for Q2, After Performing these tests we get a very good and understandable result, so based on the result we can determine which algorithm we should use in which circumstances. The methodology is shown below.

For Comparison of these many techniques, we use the true positive, false positive, true negative, false negative values that are generated by our algorithm and use this for our analysis to get a better idea by comparing performance difference which is observed each time by using these algorithms, during our test. (i) True Positive (TP) this is the number of transactions that were false and were likewise delegated deceitful by the algorithm (ii) False Positive (FP) this is the number of transactions that were real in any case, were wrongly named deceitful exchanges (iii) True Negative (TN) this is the number of transactions that were genuine and were likewise named authentic (iv) False Negative (FN) this is the number of transactions that were false yet, were wrongly named authentic exchanges by the algorithm

The Various Metrics That we used for this analysis, (i) False Alarm Rate (FAR) this is a rate measure of all occurrences classified as fraudulent and how many were misclassified.

http://dx.doi.org/10.36596/jcse.v2i2.250

FAR = FP/FP+TN (1)

Accuracy, this is the negligible part of transactions that were effectively categorized. It is one of the most remarkable and commonly used metrics for these kinds of operations.

Accuracy = (TN + TP) / (TP + FP + FN + TN) (2)

Precision (P) -> Precision otherwise called detection rate also known as hit rate, this is the number of exchanges either real or deceitful that were accurately classified.

p = TP / TP + FP (3)

Cost (c), it tells us the effective cost for our system.

c = 100 x FN + 10 x (FP + TP)

True Positive Rate (TPR) -> It measures the negligible part of strange records (the records that have greatest possibilities of being deceptive) effectively arranged by the algorithm

TPR = TP / TP + FN (4)

True Negative Rate (TNR), it measures the small portion of ordinary records (the records that have least possibilities of being deceitful) accurately categorised by the algorithm.

TNR = TN / TN + FP (5)

3. Results and Discussion

After using these metrics and the above-mentioned algorithms we get the following table. The following table is completely generated based on our methodology. We've done a comparison of all of the strategies outlined in the previous section based on our 3 different datasets. Sample Calculation, For Load Fraud and [16], TN = 4825, TP = 276, FP = 40, FN = 105

$$\text{Accuracy} = \frac{4825+276}{4825+276+40+105} \text{ x } 100\% = 97.24\%$$

$$\text{Precision} = \frac{276}{276+40} \text{ x } 100\% = 87.34\%$$

$$\text{False Alarm Rate} = \frac{40}{40+4825} \text{ x } 100\% = 0.82\%$$

Similarly, to the previous calculation, we compute all of the various results for various algorithms with their respective datasets.

Table 1: Analysis on Loan Fraud

Fraud	Algorithms	Metrics	Results	Average Results
	Bagging Ensemble Classifier Approach [16]	Accuracy	97.24%	60.85%
		Precision	87.34%	
		False Alarm Rate	0.82%	
	Machine Learning Techniques [17]	Accuracy	98.19%	93.66%
		Precision	84.96%	
		False Alarm Rate	0.81%	
	Bayesian and Neural Networks [18]	Accuracy	93.90%	68.81%
		Precision	51.22%	
		False Alarm Rate	2.44%	

http://dx.doi.org/10.36596/jcse.v2i2.250

Loan Fraud	Auto Encoder and Restricted Boltzmann machine [19]	Accuracy	95.33%	78.39%
		Precision	71.54%	
		False Alarm Rate	1.43%	
	Machine Learning Models and Collating Machine Learning Models [20]	Accuracy	99.14%	96.94%
		Precision	92.86%	
		False Alarm Rate	0.40%	
	Meta-Learning [21]	Accuracy	94.85%	60.54%
		Precision	64.94%	
		False Alarm Rate	1.94%	
	Self-Organizing Maps [22]	Accuracy	93.90%	38.84%
		Precision	62.50%	
		False Alarm Rate	2.48%	

Table 2: Analysis on Credit Card Fraud

Fraud	Algorithms	Metrics	Results	Average Results
Credit Card Fraud	Bagging Ensemble Classifier Approach [16]	Accuracy	97.27%	93.85%
		Precision	71.31%	
		False Alarm Rate	0.69%	
	Machine Learning Techniques [17]	Accuracy	97.64%	99.66%
		Precision	78.05%	
		False Alarm Rate	0.53%	
	Bayesian and Neural Networks [18]	Accuracy	96.32%	97.28%
		Precision	71.43%	
		False Alarm Rate	1.58%	
	Auto Encoder and Restricted Boltzmann machine [19]	Accuracy	95.58%	83.11%
		Precision	57.24%	
		False Alarm Rate	1.30%	
	Machine Learning Models and Collating Machine Learning Models[20]	Accuracy	99.31%	99.69%
		Precision	86.61%	
		False Alarm Rate	0.33%	
	Meta-Learning [21]	Accuracy	95.02%	69.62%
		Precision	50.30%	
		False Alarm Rate	1.69%	
	Self-Organizing Maps [22]	Accuracy	93.98%	52.62%
		Precision	59.66%	
		False Alarm Rate	1.45%	

Table 3: Analysis on Internal Frauds

Fraud	Algorithms	Metrics	Results	Average Results
Internal Frauds	Bagging Ensemble Classifier Approach [16]	Accuracy	94.55%	80.79%
		Precision	79.84%	
		False Alarm Rate	0.53%	
	Machine Learning Techniques [17]	Accuracy	96.57%	97.16%
		Precision	51.74%	
		False Alarm Rate	1.64%	
	Bayesian and Neural Networks [18]	Accuracy	94.11%	86.75%
		Precision	54.26%	
		False Alarm Rate	1.75%	
		Accuracy	93.80%	
		Precision	45.45%	

http://dx.doi.org/10.36596/jcse.v2i2.250

Auto Encoder and Restricted Boltzmann machine[19]	False Alarm Rate	3.15%	86.19%
Machine Learning Models and Collating Machine Learning Models [20]	Accuracy	95.88%	
	Precision	59.66%	
	False Alarm Rate	1.42%	97.95%
Meta-Learning [21]	Accuracy	94.47%	
	Precision	65.12%	
	False Alarm Rate	1.53%	81.50%
Self-Organizing Maps [22]	Accuracy	92.83%	
	Precision	51.82%	
	False Alarm Rate	2.18%	64.27%

4. Conclusion

Nowadays when the whole world is converting itself to a digital world, these kinds of frauds are the major threats to our system. In spite of the fact that there are a few fraud detection strategies accessible today, but yet none can identify all frauds totally when they are really happening, they for the most part recognize it after the fraud has been done completely. This is a very major problem, now that's also happening because an exceptionally small amount of exchanges from the complete transaction are really deceitful in nature. As a result an innovation which can distinguish the suspicious exchanges at the point when it is occurring, as a result, it is possible that it will be halted, at that point of time and also with an excessively minimum cost. Therefore, the major task is, build a system which can detect these banking frauds with a very precious, accurate manner and also with a very fast detection technique, which not only detect frauds which are happening over internet like many internal frauds or phishing attacks but also various banking fraud by alert through an alarm when a fraudulent transaction or initiates.

Apart from this there are also some drawbacks of this kind of system. The significant disadvantage of these methods is they're not guaranteed to produce similar results in every conditions. They provide better outcomes with a specific kind of dataset and poor or unacceptable outcomes with other kinds. The Techniques which are discussed above Like the Machine Learning techniques which include Support Vector Machine (SVM) gives us excellent results with a very small dataset. But that also comes with a drawback that it is not scalable with a humongous number of datasets. There is also some technique which includes Neural Network which produces a very good number of fraud detection rates and also gives us very high accuracy rates. But this technique also has a drawback, that is the model and dataset which are required for this algorithm is very expensive to train. There is an algorithm which includes the Bayesian Network approach which is also capable of detecting frauds with a very high detection rate and also with great accuracy. But this algorithm also suffers due to the expensive cost for training. There are some techniques like the Random Forest algorithm which generate decision trees inside and also give us a very good result with very efficient value for all our metrics (Accuracy, Precision and False Alarm rate) that are used in our analysis. But that also comes with a drawback, like the dataset should be sampled and also need a very good amount of pre-processing. There are some techniques Like Meta learning and self-organizing maps which only gives us good results on raw and unsampled data.

Now as we talk about both the positive outcomes and the drawbacks for each and every approach, let's talk about some solutions. An efficient solution for these holes by making a fused approach of different methods which are now utilized in banking fraud detection to drop out their impediments and get upgraded performance measures. The way to create a decent hybrid model is to combine a costly method which takes long to prepare yet gives profoundly exact and exact outcomes with an advancement method to bring down the expense of the algorithm and make the system learn rapidly. The selection of the

http://dx.doi.org/10.36596/jcse.v2i2.250

techniques for the hybrid solution will rely upon the applications and environment of the fraudulent transaction identification system.

Based on the comparative analysis we get a clear idea about fraud detection by some renowned algorithms. But that is not enough to detect frauds in real life scenarios. There are also some drawbacks in our system (i) Failure of the algorithm to adjust viably to evolving environment and to new fake procedures also, certified changes made in purchase propensities for a user. (ii) Inaccessibility of a solitary amazing technique that can perform reliably in every condition and also outflank any algorithm (iii) Inaccessibility of complete information for all the banking frauds because they are a personal property and neither banks nor clients can reveal their data along these lines prompting us inappropriately and inefficiently to train our algorithm. (iv) We have an absence of good and proficient metrics that cannot just depict the accuracy of all the algorithm, however can give a superior near perfect result among various methodologies

Now according to our analysis on Banking Frauds using all the described algorithms it is clear that the machine learning and neural network approach is the best among all of them. However, there are some downsides of Artificial Neural Networks and machine learning which are very costly to prepare and train and can become overtrained easily. To limit their cost, it is required to make a fused solution of neural network and machine learning techniques with some advancement and optimizing procedure.

Enhancement strategies that could be effectively matched with Neural network technique and machine learning technique for Organization of a Genetic Algorithm [23], Artificial Immune System [24], Case Based Reasoning [25] and some other comparative enhancement procedures. There are some algorithms like Genetic Algorithm [26], Artificial Immune System [27] and Case Based Reasoning [28] which help us to maximize the efficiency for a hybrid model approach. These techniques are used for choosing the upgraded weight of the edges in a neural network and lessens the expense by wiping out the weight that causes the maximum error and first attempts to foresee the result on the premise of an immediate match with the client's profile respectively.

References

[1] RBI Annual Report 2019-20: Bank Frauds More Than Double [Online]. Available: https://www.bloombergquint.com/business/rbi-annual-report-2019-20-bank-frauds-more-than-double

[2] In pandemic year, bank frauds down 25% by value: RBI annual report [Online]. Available: https://www.business-standard.com/article/finance/frauds-reported-at-banks-financial-institutions-decreased-in-2020-21-rbi-121052700748_1.html

[3] Phua, Clifton, et al. "A comprehensive survey of data mining-based fraud detection research." arXiv preprint arXiv: 1009.6119 (2010).

[4] J. O. Awoyemi, A. O. Adetunmbi and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," 2017 International Conference on Computing Networking and Informatics (ICCNI), 2017, pp. 1-9, doi: 10.1109/ICCNI.2017.8123782.

[5] Fu, Kang, et al. "Credit card fraud detection using convolutional neural networks." International conference on neural information processing. Springer, Cham, 2016.

[6] Maranzato, Rafael, et al. "Fraud detection in reputation systems in e-markets using logistic regression." Proceedings of the 2010 ACM symposium on applied computing. 2010.

[7] Gyamfi, Nana Kwame, and Jamal-Deen Abdulai. "Bank fraud detection using support vector machine." 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2018.

http://dx.doi.org/10.36596/jcse.v2i2.250

[8] Xuan, Shiyang, et al. "Random forest for credit card fraud detection." 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018.
[9] Tripathi, Diwakar, Bhawana Nigam, and Damodar Reddy Edla. "A novel web fraud detection technique using association rule mining." Procedia computer science 115 (2017): 274-281.
[10] Min, Xing, and Rongheng Lin. "K-means algorithm: fraud detection based on signaling data." 2018 IEEE World Congress on Services (SERVICES). IEEE, 2018.
[11] CHOUDHARY, VIPIN KUMAR, and ER DIVYA. "Credit Card Fraud Detection using Frequent Pattern Mining using FP-Modified Tree and Apriori Growth." (2017).
[12] Islam, Asadul Khandoker, et al. "Fraud detection in ERP systems using scenario matching." IFIP International Information Security Conference. Springer, Berlin, Heidelberg, 2010.
[13] Mercer, Lindsay CJ. "Fraud detection via regression analysis." Computers & Security 9.4 (1990): 331-338.
[14] Taniguchi, Michiaki, et al. "Fraud detection in communication networks using neural and probabilistic methods." Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181). Vol. 2. IEEE, 1998.
[15] Carneiro, Emanuel Mineda, et al. "Cluster analysis and artificial neural networks: A case study in credit card fraud detection." 2015 12th International Conference on Information Technology-New Generations. IEEE, 2015.
[16] Masoumeh Zareapoor, Pourya Shamsolmoali. Application of Credit Card Fraud Detection: Based on Bagging Ensemble Classifier, Procedia Computer Science, Volume 48, 2015, Pages 679-685, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.201.
[17] Bethapudi, Dr Prakash & Murthy, G. & Ashok, P. & Prithvi, B. & Kira, S.. (2018). ATM Card Fraud Detection System using Machine Learning Techniques. International Journal for Research in Applied Science and Engineering Technology. 5. 10.22214/ijraset.2018.4836.
[18] Maes, Sam & Tuyls, Karl & Vanschoenwinkel, Bram & Manderick, Bernard. (2002). Credit Card Fraud Detection Using Bayesian and Neural Networks.
[19] Pumsirirat, Apapan & Yan, Liu. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. International Journal of Advanced Computer Science and Applications. 9. 10.14569/IJACSA.2018.090103.
[20] Navanshu Khare and Saad Yunus Sait, "Credit Card Fraud Detection using Machine Learning Models and Collating Machine Learning models", International Journal of Pure and Applied Mathematics, vol. 118, no. 20, pp. 825-838, 2018.
[21] Stolfo, Salvatore & Fan, David & Lee, Wenke & Prodromidis, Andreas & Chan, Philip. (1998). Credit Card Fraud Detection Using Meta-Learning: Issues and Initial Results.
[22] Zaslavsky, Vladimir & Strizhak, Anna. (2006). Credit Card Fraud Detection Using Self-Organizing Maps. Information & Security: An International Journal. 18. 10.11610/isij.1803.
[23] Genetic Algorithm - an overview | ScienceDirect Topics [Online]. Available: https://www.sciencedirect.com/topics/engineering/genetic-algorithm
[24] Artificial Immune Systems - an overview | ScienceDirect Topics [Online]. Available: https://www.sciencedirect.com/topics/computer-science/artificial-immune-systems
[25] What is Case-Based Reasoning (CBR)? Definition from WhatIs.com [Online]. Available: https://searchenterpriseai.techtarget.com/definition/case-based-reasoning-CBR
[26] “Raghavendra Patidar and Lokesh Sharma,” International Journal of soft computing and engineering, vol. 1, no. NCAI2011, 2011.
[27] J. I. T. L. N. de Castro, “Artificial immune systems as a novel soft computing paradigm,” Journal of Soft Computing, vol. 7, p. 526–544, 2003.
[28] A. S. Wheeler R, “ Multiple algorithms for fraud detection. Knowledge-Based Systems,” no. S0950-7051(00)00050-2, 2000

http://dx.doi.org/10.36596/jcse.v2i2.250