

Advanced Big Data Analysis Approach for Credit Card Fraud Detection

Honey Jain¹, Rohan Benhal² and Tanmayee Parbat³

^{1, 3}B.E IT, Maeer's MIT, Pune

Email: tanmayee.parnat01@gmail.com

²BBA IT, Maeer's MIT, Pune

Abstract—Fraud and default payments are two major anomalies in credit card transactions. Experts have been hard at work developing solutions, and one idea is to use data mining techniques. Credit card data, on the other side, can be tough for researchersto work with. This is related to the data qualities listed below: we have an unequal class distribution, and and redundant class samples. Both of these characteristics lead to low detection rates for minor anomalies in the data. Furthermore, defects in general learning algorithms contribute to the difficulty in categorising anomalies because the algorithms favour the majority class samples m. Each neural network configuration's findings are displayed and explained. To proposed advance big data classification technique based on fusion machine learning technique. The best methoddetects 98.81 percent of total fraud cases while minimising alarms by 30.32 percent.

Index Terms— Credit Card,Fraud,Big Data, Error rate.

I. INTRODUCTION

In this research paper to described a number of credit card fraud difficulties. However, one of the most pressing issues is the issue of class disparity. This problem is defined as a data distribution that is very uneven and severely skewed. To put it another way, the proportion of fraudulent or criminal activity is significantly lower than the proportion of legitimate and authentic activities. Depicts a problem with class imbalance that we discovered in one of our experiments. An imbalance based on two characteristics of our dataset. A class imbalanceexacerbates the problem of finding the characteristics of fraudulent conduct and the extraction of fraud patterns. Mainly because of one class, almost all of the categorization algorithm's optimization phases (concerning accuracy) strive to correctly classify Minority observations, such as observations of fraud, are the most important tocorrectly characterize. If fraud tendencies are not discovered in the classification system, illicit transactions are treated as genuine, resulting in significant financial loss for individuals and businesses. Numerous studies have been conducted to investigate the topic of imbalance classification.However, the issue of class disparity has remained unresolved. The goal of this work is to highlight flaws in existing class imbalance methodologies, notably machine learning methods. To address the imbalance classification problem, we conducted an experimental study using state-of-the-art technology already in use in applications. Our goal is to identify the difficulties that must be addressed in order to develop a highly efficient solution to the problem of class imbalance.

A. Fraud Detection Systems

Although the study does not focus on FDS, it provides a small history on the issue to understand which strategies have shown more efficient for providing the initial line of defence against fraud before decisions of human analysts are made. Recent study has identified multiple techniques employing new or upgraded methods, which improve the state of the art in numerous respects. In terms of categorization accuracy, the researchers determined that unsupervised learning is the most promising technique.

II. RELATED WORK

In this section, we highlight work on detecting credit card fraud. Furthermore, we place a great emphasis on research to address the problem of class imbalance in fraud detection.

S. Makki, et al[1] Identified their flaws and reported the results we acquired using the data set of a credit card theft. Unbalanced classification algorithms are, according to this article, unsuccessful, especially when the data is substantially skewed. According to this research, the present practices produce a number of false alarms, which are costly for financial organizations. This could lead to misunderstandings and an increase in fraud cases.

H. Shahriar, et al[2] In this project, Google's CoLab was used to construct a study case-based portable laboratory experience for an ML cybersecurity software, allowing students with hands-on laboratories to handle onerous installs and configurations from anywhere and at any time. This method, students can focus on learning fundamental ideas and receive important experience by resolving problems.

J. Wang, et al[3] The framework was built using an ensemble of sample methodologies, feature engineering approaches and a range of machine learning techniques.

M. Shao et al[4] A model based on frequency domain features that are integrated with a generative adversarial network (GAN) to boost the minority class is discussed here.

M. Erfani et al[5] We employ a new pre-processing and subsampling stage to identify fraud, followed by a profound vector data description of support. We provide a trend analysis using the area of the ROC- AUC (ROC-APC) and the Average Precision (AP) as metrics in order to evaluate the size of the training, testing and model data sets. Finally, the SVM and Random Forest techniques are the cutting-edge binary classifier in many instances, according to our findings.

Prusti, D., et al[6] A fraud detection system based on the use of the graph database concept has been proposed in this study. The characteristics found in the Graph utilizing the Neo4j tool are retrieved to many different transaction database attributes.

III. PROPOSED METHODOLOGY

Fraud is a significant concern for companies and individuals alike. Many fraudulent transactions cost billions of dollars annually. The most common and costly sort of fraud is credit card fraud, which has led to a widespread world worry. It is difficult to detect credit card fraud. Several problems make it difficult to identify fraud detection solutions, one of which is class imbalance. A number of remedies have been presented to address this issue[7]. These options were investigated in this research. We analysed the results of eight machine-learning approaches for detecting credit card fraud and identified their flaws. More specifically, we examined the approaches to imbalanced categorization and looked at how effective they are in extreme cases of imbalance. According to the three performance indicators studied, the because the other models' assumptions and criteria were difficult to meet. When the imbalance is extreme, the procedures commonly employed to many imbalance problems can have unfavourable repercussions[8], such as generating a large number of false positives, according to our research[9]. Despite the fact that these strategies improve the performance of the classifier, many fraud cases continue to be undetected. In addition, our research revealed that using only one performance measure to assess uneven learning is deceptive. During the study, we ran into a few issues that need to be looked at further. The initial step is to identify[10] a structure that is optimal. On the other side, the proposed fusion machine-learning algorithm for training phase's time commitment is tough to manage. Furthermore, identifying several research are in the works for the subsequent round of this experimental investigation. To establish a trade-off between sensitivity[11] and accuracy, we proposed fusion model for the class imbalance problem. We will create a Big Data-driven ecosystem and put our approach to the test with vast amounts of data. Our study was obviously size of the data, as we were unable to run it on significant volumes of data. As a result, we intend to create a scalable ecosystem based on Big Data technologies.

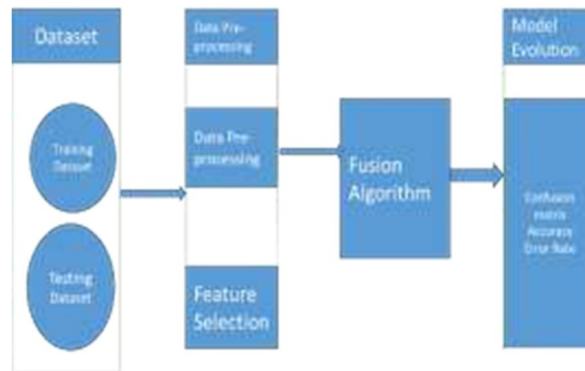


Figure 1: Anomaly detection in big data

Anomaly detection: A useful tool for detecting fraud, suspicious actions, network intrusion, and other unusual events that may be relevant but difficult to detect. The detection of anomalies is important because it converts data into widely usable information and transforms useful insights across a wide range of industries. For example, the use of data machine must generate cancer treatment plans. Finding anomalies is crucial before suggesting a patient's radiation dose. Even if anomaly is detected in part, the accuracy of the data is essential because inefficient examination of the data discovering aberrant behaviour of the data. The majority of applications were important for hundreds or millions of features of high-dimensional data sets. Since anomalies are rare and typically occur in fractional views of subsets of dimensions or subspaces, current methodologies in high-dimensional space to discover anomalies are problematic. We also show that an anomaly identification method based on the concept of closeness deteriorate qualitatively with time and necessitates a technological reinterpretation

However, when data dimensionality increases, standard techniques fail because they rely on tactics that assume low data dimensionality. Furthermore, for data sets with high dimensionality, only a proportion of data points may be useful. Anomaly detection may be used in both online and offline modes to solve issues with high dimensionality. Anomalies in historical data sets are found in an offline mode known as "batch processing." This is related to big data's "volume" characteristic. In contrast, when abnormalities are being discovered in live mode, fresh data points, known as "data streams," are constantly introduced. This is related to big data's "velocity" characteristic. In many domains, such as deep learning and data mining, a number of existing studies and studies emphasise the challenge of high dimensionality. The literature regularly addresses the "size" aspect of the volume and the "speed" component of the velocity; however, the "dimension" element is often overlooked. This gap has been dubbed "large dimensionality" as well as to comprehend the underlying difficulty impeding the performance and accuracy of the available solutions.

A. Fusion Learning

The dynamic state of the nodes responds to inputs in FL. FL[11] are built up of basic and highly linked nodes that respond to inputs via the dynamic state of the nodes. As illustrated in, it is built up of layers, which include an input layer, a concealed layer, and an output layer. These are the layers.

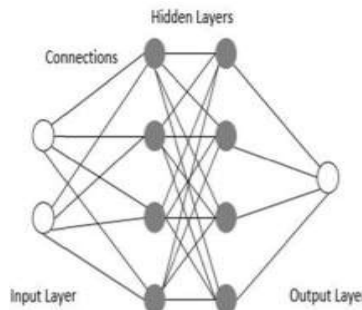


Figure 1: The Fusion Learning with an input layer, hidden layers and output layer is shown

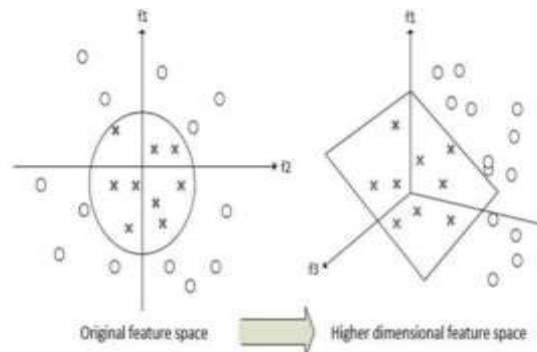


Figure 2: To identify a better decision boundary, SVM turns an original dimensionality into a higher dimensional feature space

are made up of linked 'nodes' that have activation functions attached with them. Learning by examples are two types of learning rules that adjust the weights of the connections based on the input patterns. When dealing with imbalanced data sets, Fusion Learning, like other single classifiers[12], are biased towards majority classifications. Minority class samples will be undetectable by Fusion Learning due to overwhelming samples of majority classes. This framework is created to utilize features from a pre-trained deep learning model. Deep FFNN models acquire useful insights or features when trained on image collections. The pre-trained models are previously trained extensively[13] on vast data collections such as created card datasets. The training weights acquired by such a model can be readily transferred to a new model. This is generally called transfer learning. This is a promising method to create a cost-effective solution for new domains.

Usually, a pre-trained DL model is made up of several layers that compute various features when the input Credit Card Fraud Detection is processed through them. This is as we concluded in the last chapter, learned features acquired by a DL model are more suitable for Credit Card Fraud Detection retrieval than the regular features. Here in this research we are exploring the possibility of transferring the learned features acquired by a pre-trained deep DL[14] model. We created a methodology to position a pre-trained DL[15] model to generate features and use them for operation as depicted methodology above depicted uses a SVM. Here, features learned from the model are extracted by tapping a fully connected (FC2) layer as depicted[16]. This is the output of the second and penultimate FC layer of the model. A similar implementation can be done using different pre-trained models, for instance, which made creation and training of extremely deep DL model a reality. Each stage consists of a convolution block and an identity block.

IV. RESULTS ANALYSIS

The results are analyzed in terms of Accuracy, Precision, Recall and F-Score. We compared three different classifiers namely Naive Bayes, SVM, FFNN with Fusion model. Fusion model performed well in classifying the Credit Card Fraud Detection dataset by yielding maximum accuracy of 97.33%, followed by Naive Bayes, SVM, FFNN with Fusion model 90.5%, 75.65% with 81.33% accuracies.

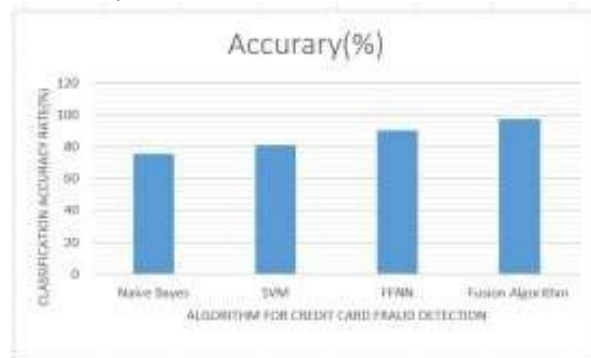


Figure 3: Credit Card Fraud Detection using Deep learning



Figure 4: Error rate Credit Card Fraud Detection using Deep learning

Comparing Classification results on both image Data and Real Time Data, it was found that Real time data was classified with high accuracy by applying Fusion model as it has less features compared to image data. Fusion model achieved highest accuracy among the rest of the deep learning algorithms and Fusion deep learning algorithm on real time data.

V. CONCLUSION AND FUTURE WORK

Credit card fraud detection is critical for increased credit card usage. Given the increasing difficulty of identifying credit card fraud and the enormous and continuous financial losses faced by financial organisations, it is critical to develop more effective ways for identifying fraudulent credit card transactions. The experimental findings show that the suggested methodology beat all existing machine learning methods, achieving the best accuracy. The findings show that the proposed Fusion algorithm outperforms other classification algorithms. The results also highlight the relevance and utility of an efficient parameter optimization technique for increasing the predictive performance of the suggested strategy.

REFERENCE

- [1] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M. Hacid and H. Zeineddine, "An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection," in *IEEE Access*, vol. 7, pp. 93010-93022, 2019, doi: 10.1109/ACCESS.2019.2927266.
- [2] H. Shahriar, M. Whitman, D. C. Lo, F. Wu, C. Thomas and A. Cuzzocrea, "Experiential Learning: Case Study-Based Portable Hands-on Regression Labware for Cyber Fraud Prediction," 2019 IEEE International Conference on Big Data (BigData), 2019, pp. 5814-5818, doi:10.1109/BigData47090.2019.9005713.
- [3] J. Wang, R. Martins de Moraes and A. Bari, "A Predictive Analytics Framework to Anomaly Detection," 2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), 2020, pp. 104-108, doi: 10.1109/BigDataService49289.2020.00023.
- [4] M. Shao, N. Gu and X. Zhang, "Credit Card Transactions Data Adversarial Augmentation in the Frequency Domain," 2020 5th IEEE International Conference on Big Data Analytics (ICBDA), 2020, pp. 238-245, doi:10.1109/ICBDA49040.2020.9101344.
- [5] M. Erfani, F. Shoeleh and A. A. Ghorbani, "Financial Fraud Detection using Deep Support Vector Data Description," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2274-2282, doi: 10.1109/BigData50022.2020.9378256.
- [6] Prusti, D., Das, D. & Rath, S.K. Credit Card Fraud Detection Technique by Applying Graph Database Model . *Arab J Sci Eng* (2021). <https://doi.org/10.1007/s13369-021-05682-9>
- [7] Hassanien, A.E., Darwish, A. & Abdelghafar, S. Machine learning in telemetry data mining of space mission: basics, challenging and future directions. *Artif Intell Rev* **53**, 3201–3230 (2020). <https://doi.org/10.1007/s10462-019-09760-1>
- [8] Dong, H., Yang, F. & Wang, X. Multi-label charge predictions leveraging label co-occurrence in imbalanced data scenario. *Soft Comput* **24**, 17821–17846 (2020). <https://doi.org/10.1007/s00500-020-05029-w>
- [9] Talal, M., Zaidan, A.A., Zaidan, B.B. *et al.* Comprehensive review and analysis of anti-malware apps for smartphones. *Telecommun Syst* **72**, 285–337 (2019). <https://doi.org/10.1007/s11235-019-00575-7>.
- [10] A. Singh Rajawat and S. Jain, "Fusion Deep Learning Based on Back Propagation Neural Network for Personalization," 2nd International Conference on Data, Engineering and Applications (IDEA), 2020, pp. 1-7, doi: 10.1109/IDEA49133.2020.9170693.

- [11] Rawat R., Rajawat A.S., Mahor V., Shaw R.N., Ghosh A. (2021) Surveillance Robot in Cyber Intelligence for Vulnerability Detection. In: Bianchini M., Simic M., Ghosh A., Shaw R.N. (eds) *Machine Learning for Robotics Applications*. Studies in Computational Intelligence, vol 960. Springer, Singapore. https://doi.org/10.1007/978-981-16-0598-7_9
- [12] Somasundaram, A., Reddy, S. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Comput & Applic* **31**, 3–14 (2019). <https://doi.org/10.1007/s00521-018-3633-8>.
- [13] Rajawat A.S., Rawat R., Mahor V., Shaw R.N., Ghosh A. (2021) Suspicious Big Text Data Analysis for Prediction—On Darkweb User Activity Using Computational Intelligence Model. In: Mekhilef S., Favorskaya M., Pandey R.K., Shaw R.N. (eds) *Innovations in Electrical and Electronic Engineering*. Lecture Notes in Electrical Engineering, vol 756. Springer, Singapore. https://doi.org/10.1007/978-981-16-0749-3_58
- [14] Somasundaram, A., Reddy, S. Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance. *Neural Comput & Applic* **31**, 3–14 (2019). <https://doi.org/10.1007/s00521-018-3633-8>
- [15] Wei, W., Li, J., Cao, L. *et al.* Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web* **16**, 449–475 (2013). <https://doi.org/10.1007/s11280-012-0178-0>.
- [16] Mishra, K.N., Pandey, S.C. Fraud Prediction in Smart Societies Using Logistic Regression and k-fold Machine Learning Techniques. *Wireless Pers Commun* (2021). <https://doi.org/10.1007/s11277-021-08283-9>.
- [17] Bhatore, S., Mohan, L. & Reddy, Y.R. Machine learning techniques for credit risk evaluation: a systematic literature review. *J BANK FINANC TECHNOL* **4**, 111–138 (2020). <https://doi.org/10.1007/s42786-020-00020-3>