Technoarete Transactions on Advances in Computer Applications (TTACA)
Volume 4, Issue 1, March 2025
e-ISSN: 2583-3472

Modified Isolation Forest Algorithm for Credit Card Fraud Detection

Krizzia Ydel M. Merino [1*], Ma. Alexandra M. Ong [2], Raymund M. Dioses [3], Vivien A. Agustin [4], Ariel Antwaun Rolando C. Sison [5]

[1, 2] Student, Department of Computer Science, College of Information Systems and Technology Management, University of the City of Manila, Manila, Philippines
[3, 4, 5] Professor, Department of Computer Science, College of Information Systems and Technology Management, University of the City of Manila, Manila, Philippines
*Corresponding Author Email: ydelmerino@gmail.com

Abstract

The Isolation Forest algorithm is an isolation-based method used for detection of anomaly. The algorithm has a problem with swamping which is the misclassification of the normal data points as anomalies. The said problem of Isolation Forest reduces its accuracy and effectiveness. The Modified Isolation Forest used an undersampling method called Near Miss method to address the problem of the Isolation Forest regarding false positives or swamping. The algorithm results in misclassification if a large imbalance dataset is used. Hence, incorporating Near Miss undersampling method to obtain a balanced dataset helps reduce the false positives and improves the overall performance of the algorithm. A dataset containing transactions of European cardholders is used in this study, which has 492 fraudulent transactions among 284, 807 transactions. Both the original algorithm and the modified algorithm are tested for anomaly detection using the same dataset. The original algorithm resulted in 158 True Positive (TP), 235619 True Negative (TN), 48696 False Positive (FP), and 334 False Negative (FN). While, the modified algorithm resulted in 395 True Positive (TP), 391 True Negative (TN), 101 False Positive (FP), and 97 False Negative (FN). The modified algorithm of Isolation Forest results in a significantly better performance compared to the original algorithm. With an accuracy rate of 0.79878 or 79.88%, a precision of 0.79637 or 79.64%, a recall of 0.80285 or 80.29% and an f1-score of 0.79960 or 0.80, the Modified Isolation Forest algorithm addressed the issue of false positives or swamping.

Keywords

Anomaly, Fraud Detection, Isolation Forest, NearMiss, Tree.

INTRODUCTION

Isolation Forest is an algorithm proposed and authored by Liu et. al. in year 2008 which is commonly used for detecting anomalies. Anomalies, also referred to as outliers, refers to data that has a great difference from normal data in a dataset [1]. It is important to detect anomalies as it may cause security risks and other significant problems regarding information and data [2]. Detecting anomalies is the process of determining data patterns that are significantly different from the norm [3]. Various sectors, such as finance and medical, applied anomaly detection. Some examples of its application include fraud detection in credit cards which may suggest theft. There are various techniques used for detecting anomalies. The Local Outlier Factor and K-Nearest Neighbor are some of the approaches used for detection tasks.

The Isolation Forest is a powerful and effective method with low execution time and memory requirements [1]. However, the algorithm faces the challenges due to the problems of swamping or the false positive [1] [4] [5]. Swamping or false positives happens when normal data are considered or identified as anomalies. The said problems may reduce the accuracy and effectiveness of the Isolation Forest algorithm.

In summary, the existing Isolation Forest is designed with the simple principle that (1) the data that are closer to the root node of the tree, are more likely to be an anomaly as it does not conform with the normal patterns, and in contrast, (2) the points that are at the end of the tree are likely to be normal points, as it conforms with the normal patterns in an entire dataset. So, as a data set is given to the algorithm, it builds an ensemble of isolation trees and considers those instances with short average paths as anomalies. The variables needed for this algorithm are the number of trees to be built as well as the required sub sampling size [6].

In this study, the goal is to modify the existing algorithm in order to address the mentioned problem. This paper implements the NearMiss undersampling method to reduce the number of data. By utilizing the said method, this paper aims to improve the ability and overall performance of Isolation forest in detecting anomalies. This paper also aims to benefit the services that rely on anomaly detection methods to keep the overall efficiency of their services.

METHODS AND METHODOLOGY

The Figure 1 shows the framework which is used as a guide for this study. The framework includes modification before the implementation of the Isolation Forest. The modification is represented by the red outline. Data acquisition is the start of the process where the researchers gathered data relevant to the problem. During the next stage, reducing the number of data happened through the use of an undersampling method, specifically the NearMiss undersampling method. Then, the training phase of Isolation

Technoarete Transactions on Advances in Computer Applications (TTACA)
Volume 4, Issue 1, March 2025
e-ISSN: 2583-3472

Forest took place which generates the isolation trees. Next steps include determining the path length and the decision thresholding where data is classified as anomaly or not.

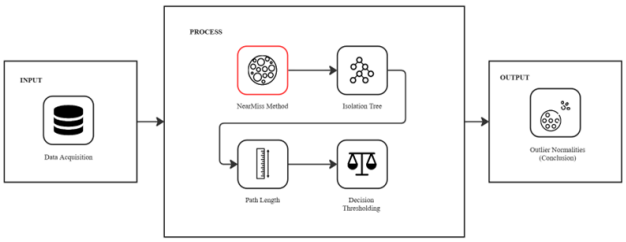

Figure 1. Conceptual Framework

Existing Algorithm of Isolation Forest

The original algorithm is an algorithm used for detection of anomaly designed in year 2008 by Liu et. al. [7]. It splits the data continuously resulting in the isolation of data from the trees which are identified as the anomalies [1]. There are two phases or stages in Isolation Forest, the training stage, and the scoring phase. In the training stage, a subset of data is randomly chosen to build a tree. The chosen subset will be the root node of the tree which will be split into two nodes. The internal nodes will be split into two nodes, and this process will continue until it reaches the maximum tree depth or when the data is completely isolated [2]. In the scoring stage, the calculation of the score of each instance or item in the X dataset takes place. The computed score represents the similarity between the item and the rest of the data which will be used to classify the data as anomaly or normal.

Decision Thresholding

As stated previously, path length of each item (x) is calculated during the scoring phase. The path length is determined by the total count of nodes that x traversed from the root to the node that it belongs to, represented as h(x). Once the item has passed through each isolation tree in the forest, the algorithm will calculate the average path length of the item represented as E(h(x)) [2].

Isolation Forest algorithm calculates the score of each item which is equal to 2 raised to the quotient of the average path length of the item indicated as E(h(x)) and the average path length of an unsuccessful search indicated as C(n), as seen in equation (2) [1] [2].

$$C(n) = 2H(n-1) - \left(\frac{2(n-1)}{n}\right) \quad (1)$$

$$s(x,n) = 2^{-\frac{E(h(x))}{C(n)}} \quad (2)$$

Once the score of the items is calculated, it will be classified as normal data or anomaly using the threshold given: (1) The dataset does not have an anomaly, if all instances have a score of approximately 0.05. (2) An instance is a normal data, if its score is close to 0.1. (3) An instance is an anomaly data, if it has a score lower than 0.05. Hence, it implies that an item is an anomaly if it has a short average path length.

Design

This part of the paper discusses the design of the modified isolation forest algorithm with regard to the techniques that will be used by the proponents to address its issues. The study used the Near Miss undersampling method as a way to reduce a dataset by selecting samples from the majority class based on the nearness or distance of majority and minority class samples [8] [9]. This technique prevents the likeliness of the occurrence when normal instances are misclassified as an anomaly in larger sample sizes, therefore contributing to the modification of the algorithm.

Testing

The proponents include two simulations, one for the simulation of the original algorithm and another for the simulation of the proposed algorithm. Credit card fraud dataset is used to check the existence of the said problems and to test the effectiveness of the objective.

The credit card fraud dataset used for this study is collected from Kaggle which is a result of the collaboration between Worldline and the Machine Learning Group. It includes credit card transactions of the European cardholders in the month of September year 2013. A two-day worth of transactions has been recorded in the dataset with a total of 284, 807 transactions. The highly imbalanced number between normal and fraudulent transactions is important as it affects the accuracy and effectiveness of the Isolation Forest resulting in various problems such as swamping.

The proponents used an under-sampling method, the NearMiss method, to reduce the data in the dataset aiming to lessen the partitions required to identify the anomalies. Figure 2 displays the scatter plot of the original dataset with a total of 284, 807 transactions on the left with Time on the y axis and Amount on the x axis. And Figure 3 is the scatter plot of the dataset after implementing the NearMiss undersampling method on the right with the same dataset.

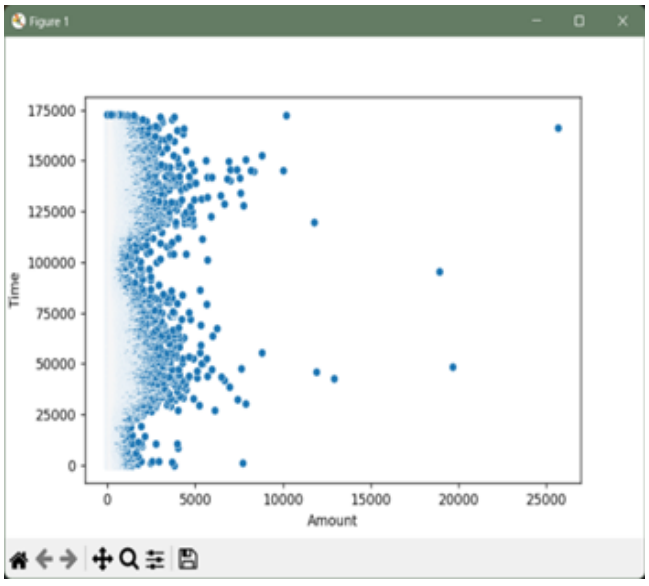

Figure 2. Scatter Plots of the Original Dataset

Technoarete Transactions on Advances in Computer Applications (TTACA)
Volume 4, Issue 1, March 2025
e-ISSN: 2583-3472

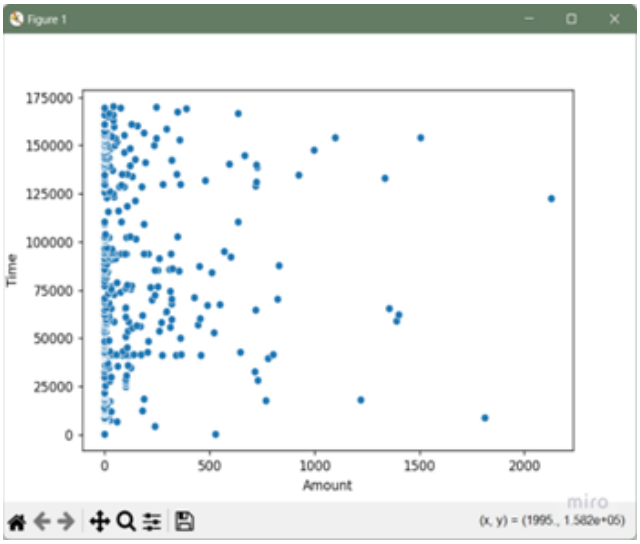

Figure 3. Scatter Plots of the Dataset after NearMiss

The data obtained after utilizing NearMiss under-sampling method is then used in Isolation Forest algorithm to identify the anomalies. After the training phase of the algorithm where the features time and amount are used, scoring phase takes place where the anomaly score will be determined using the calculated average path length of each item. Once calculated, the decision threshold happens where a data or item with anomaly score near 0.05 will be considered as a normal data, represented by 1. On the other hand, instances with an anomaly score lower than 0.05 will be considered as anomaly data, represented by -1.

RESULTS

The performance of the original algorithm and the modified algorithm used in the credit card dataset anomaly detection is displayed in Table 1. The performance of the algorithms is measured through the accuracy, precision, recall, and f1-score. The modified algorithm has a 79.88% accuracy, precision of 79.64%, 80.29% recall, and f1-score of 0.80. On the other hand, the original IF algorithm has an 82.78% accuracy, precision of 0.32%, 32.11% recall, and f1-score of 0.01.

Table 1. The Evaluation of the Original and Modified Isolation Forest

Algorithm	Measure	Score
Modified Isolation Forest	Accuracy	0.79878
	Precision	0.79637
	Recall	0.80285
	F1-Score	0.79960
Isolation Forest	Accuracy	0.82784
	Precision	0.00323
	Recall	0.32114
	F1-Score	0.00640

The table shows the results of the original algorithm and the modified algorithm in various measurements.

The Confusion Matrix of the original IF algorithm and the modified Isolation Forest which shows the number of TP, TN, FP, FN is in Table 2 and Table 3.

Table 2. The Confusion Matrix of the Original Isolation Forest

	Positive	Negative
Positive	158 TRUE POSITIVE	334 FALSE NEGATIVE
Negative	48696 FALSE POSITIVE	235619 TRUE NEGATIVE

The original Isolation Forest has resulted to 158 TP, 235619 TN, 48696 FP, and 334 FN.

Table 3. The Confusion Matrix of the Modified Isolation Forest

	Positive	Negative
Positive	395 TRUE POSITIVE	97 FALSE NEGATIVE
Negative	101 FALSE POSITIVE	391 TRUE NEGATIVE

The modified Isolation Forest has resulted to 395 TP, 391 TN, 101 FP, and 97 FN.

DISCUSSION

Accuracy

The accuracy of the algorithm represents the correctly classified data instances. Accuracy is computed based on the sum of TP and TN divided by the total number of samples. The accuracy of the original Isolation Forest algorithm resulted to 82.78%, and it is higher compared to the accuracy of the modified IF algorithm which resulted in 79.88%. However, basing solely from the accuracy may not completely represent the performance of the algorithm especially for the imbalanced dataset.

Precision

The precision of the algorithm measures the reliability of a classifier in identifying True Positives (TP) from all positive predictions [1]. Ideally, precision should be 1, meaning there are no false positives (FP = 0). As the False Positive FP increases, precision decreases, indicating a poorer performance. In terms of this evaluation, the modified Isolation Forest resulted to a precision of 79.64%, with 395 True Positives (TP) and 101 False Positives (FP), significantly outperforming the original Isolation Forest, which has a precision of only 0.32% due to a high number of false positives specifically 48, 696.

Recall

The Recall of the algorithm measures the ability of the classifier to predict the true positives. It is computed by dividing the TP by the sum of TP and FN [1] [10]. The

e-ISSN: 2583-3472

original Isolation Forest algorithm resulted to a recall of 32.11% significantly lower than the result of the modified Isolation Forest algorithm with a recall of 80.29%. This shows that the modified IF algorithm can predict the true positives in the European cardholders' transactions dataset better than the original algorithm.

F1 Score

The f1-score of the algorithm measures the predictive performance of the classifier since it considers both the precision, and the recall [10]. It is computed by getting the product and the sum of the precision and the recall, and then dividing the two together, and lastly by multiplying it by two [1]. The original Isolation Forest algorithm has an f1-score of 0.01 which is significantly lower than the f1-score of the modified Isolation Forest algorithm which is 0.80.

CONCLUSION

The Modified Isolation Forest algorithm scores 79.88% in accuracy, 79.64% in precision, 80.29% in recall, and 0.80 in f1- score, which has an average performance rate of 0.7994 or 79.94% for European cardholders' transactions dataset used in this study. Meanwhile, the original Isolation Forest algorithm scores 82.78% in accuracy, 0.32% in precision, 32.11% in recall, and 0.01 in f1-score, which results in an average performance rate of 0.28965 or 28.97% for the said dataset. The modified algorithm resulted in fewer false positives compared to the original algorithm indicating an improved management of false positive issues.

Comparing the accuracy rate of the algorithms, the original Isolation Forest has a higher accuracy rate of 82.78% than the modified algorithm with an accuracy of 79.88%. However, for precision, and recall, the Modified Isolation Forest outperformed the original algorithm indicating a better ability and reliability in detecting anomalies. With a higher f1-score, the modified algorithm shows a better predictive performance. In overall, the Modified Isolation Forest has the best performance for anomaly detection in credit card fraud.

Acknowledgement

We thank Krizzia Ydel Merino and Ma. Alexandra Ong for their contributions to this paper. And, special thanks to Sir Raymund M. Dioses, Ma'am Vivien A. Agustin, and Sir Ariel Antwaun Rolando C. Sison for their guidance and assistance. As well as DOST-SEI for financial support.

Funding Statement

Financial support was provided by DOST-SEI under scholarship program RA 7687. The funders had no participation in the preparation and content of the manuscript, in deciding for publication, or in data analysis and collection.

Data Availability

The dataset used in this study is collected and available from Kaggle named Credit Card.

Conflict of Interest

The authors state that there is no conflict of interest.

Miscellaneous

There are no other figures, tables, supplements, appendices, or annexes placed after the reference.

REFERENCES

[1] Chabchoub, Y., Togbe, M. U., Boly, A., Chiky, R. An in-depth study and improvement of Isolation Forest. IEEE Access. 2022; 10, 10219-10237.

[2] Tokovarov, M., Karczmarek, P. A probabilistic generalization of isolation forest. Information Sciences. 2022; 584, 433–449.

[3] Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access. 2021; 9, 78658–78700.

[4] Liu, T., Zhou, Z., Yang, L. Layered isolation forest: A multi-level subspace algorithm for improving isolation forest. Neurocomputing. 2024; Volume 581, 127525.

[5] Hariri, S., Kind, M. C., Brunner, R. J. Extended isolation forest. IEEE Transactions on Knowledge and Data Engineering. 2021; 33(4), 1479–1489.

[6] Meriem, C., Amel, B., Mohamed, T., S., Karem, S., G., & Mohamed I., L. Fuzzy Isolation Forest for Anomaly Detection. 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems. 2022.

[7] Liu, F. T., Ting, K. M., & Zhou, Z. Isolation Forest. Eight IEEE International Conference on Data Mining, Pisa, Italy. 2008; pp.413–422.

[8] Mamun, A. A., Enan, A., Indah, D. A., Mwakalonge, J., Comert, G., & Chowdhury, M. Crash Severity Risk Modeling Strategies under Data Imbalance. arXiv preprint. 2024; arXiv:2412.02094.

[9] Tahfim, S. A., & Chen, Y. Comparison of Cluster-Based sampling approaches for imbalanced data of crashes involving large trucks. Information. 2024; 15(3), 145.

[10] Elnour, M., Meskin, N., Khan, K. M., Jain, R. A Dual-Isolation-Forests-Based Attack Detection Framework for Industrial Control Systems. IEEE Access. 2020; 8, 36639–36651.