International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Hybrid Machine Learning Approach for Data Anomaly Detection in Credit Card Transactions

Rashmi R. More[1], Mrs. Dipalee Divakar Rane[2]

[1]PG Student D.Y. Patil college of engineering, Akurdi, Pune, Maharashtra, India
Email Id: more.rashmi2912@gmail.com
[2]Assistant Professor D.Y. Patil college of engineering, Akurdi, Pune, Maharashtra, India
Email Id: ddrane@dypcoeakurdi.ac.in

Abstract - Data anomaly detection is crucial for maintaining the integrity, security, and efficiency of systems, helping to detect and respond to abnormal events promptly. In this paper, a hybrid approach of a combination of two unsupervised machine learning algorithms, Isolation Forest (IF) and One Class Support Vector Machine (OCSVM) with one supervised machine learning algorithm Random Forest Classifier as an ensemble method is used in credit card transactions system for data anomaly detection. For this, credit card fraud detection dataset from Kaggle is used for anomaly detection. As credit card transactions data is huge, varying and unlabeled, unsupervised algorithms like IF and OCSVM become suitable for it. The ability of OCSVM of defining of what constitutes the anomaly, power of isolation forest to detect outliers efficiently and use of random forest classifier as an ensemble method of decision trees, complementing both IF and OCSVM are associated to increase accuracy of a system. The result of this paper shows that this hybrid approach provides more accuracy than individual algorithm.

Keywords: Anomaly detection, Isolation Forest, One Class SVM, Random forest classifier, Hybrid approach

I. INTRODUCTION

Today credit card transactions are the salient part of the economy. Increased use of credit card results in increased number of frauds during transactions and consequently these frauds affect on consumers, financial institutes and overall economy system. Consumers suffer from financial losses, credit score impact and loss of trust while financial institutes suffer monetary losses, increased security costs and reputation damage. Overall, these frauds make bad effects on financial stability, consumer behavior, market dynamics, adoption of digital transactions, global commerce, international trading and economic relation. The solution for all these problems is fraud detection. In this paper, we are going to focus on anomaly detection which is a crucial component of fraud detection in credit card transactions. Anomaly detection involves identifying patterns in data which do not adhere to the predefined way of behaving. In the context of credit card transactions, it means that transactions which deviates the way from the spending habits of the consumers. Anomaly detection system finds the different transactions which may or may not be fraudulent. It analyzes specific pattern of the previous data to decide the baseline for normal behavior of the transactions and checks whether the processed data is within baseline or out of the baseline called as outliers. If data is out of the baseline, it is considered as anomalous data of the transactions. It is the process of detecting outliers. It can be extreme values or unusual patterns. Extreme values are the transactions with amount that are significantly lower or higher that typical transactions while unusual patterns contain sudden spikes or drops in transaction frequency deviated from normal behavior. In this paper, we are using machine learning models to detect data anomalies based on patterns in the data.

II. RELATED WORK

Here we present review on existing literature about anomaly detection in credit card transactions.

Prerna singh and et.[1] proposed isolation trees and LOF as classifiers in credit card transactions.

Dr. V. Ceronmani and et.[2] proposed the anomaly detection techniques in credit card fraud detection by using Isolation Forest and Local Outlier Factor. Through these algorithms, data is visualized to detect frauds in transactions.

Swaroop K. and et.[3] proposed the technique of finding fraud events in credit card fraud transaction process by using isolation forest and Local Outlier Factor to calculate number of fraud transactions. They calculated accuracy and number of frauds at a time.

Arun Kumar Rai and et.[4] proposed NN based approach which gives better accuracy than existing approaches as IF, LOF,KNN and AE

Ravi Ghevariya and et.[5] focused on first all the information about credit card transactions i.e involved parts in it, how communication goes and then proposed approach which is based on Isolation forest and LOF. It gave comparison between accuracy of IF and LOF.

Esenogho, Ebenezer, Ibomoiye Domor Mienye [7] proposed neural network ensemble method used in feature engineering for improving accuracy of credit card fraud detection system.

III. PROPOSED APPROACH:

The existing approaches have used supervised, unsupervised and semi supervised machine learning techniques to detect anomalies in credit card transaction process. The use of a single algorithm leads to low accuracy, higher risk of overfitting or underfitting, lack of robustness and scalability challenges. To address these issues, hybrid approach that combines multiple machine learning algorithms leverages qualities of these algorithms to improve accuracy of the anomaly detection system. The hybrid approach is more robust, comprehensive, reliable and adaptable than single algorithm approach. In the proposed approach, the combination of Isolation Forest(IF) and One Class Support Vector Machine(OCSVM) algorithms is used with Random Forest Classifier(RFC) as an ensemble method. IF and OCSVM are unsupervised algorithms while RFC is a supervised algorithm. As IF and OCSVM are unsupervised, these are beneficial in effectively identifying outliers without requiring labeled data. Fig. shows the system architecture of the proposed hybrid method. Basically the system architecture is designed to show how to preprocess the data, to reduce dimensionality for deriving predictions from multiple algorithms as IF, OCSVM and RF classifier and combining them to achieve robust anomaly detection system. Each step ensures that data is preprocessed properly, features are accurately standardized,

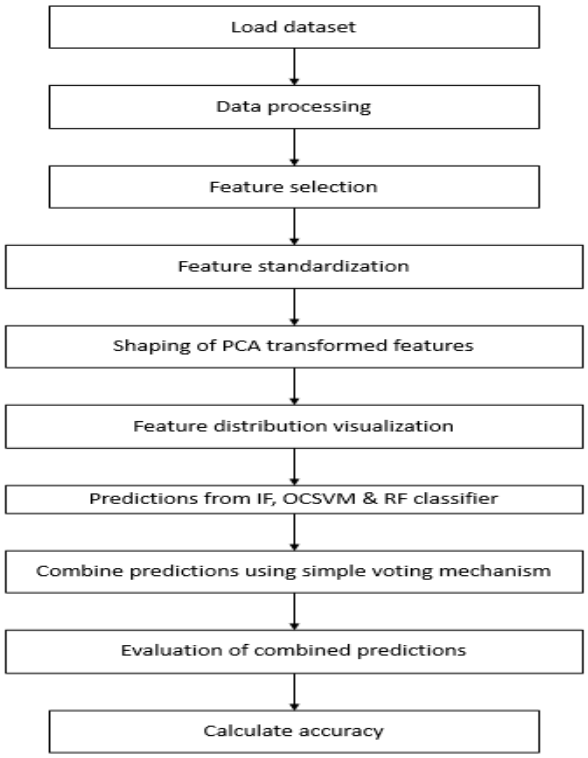

Fig 1. System architecture of Hybrid approach

models are fully trained and predictions are derived and combined properly to achieve accurate results.

Dataset loading: The credit card dataset from Kaggle for data anomaly detection is loaded for anomaly detection in credit card transactions.

Credit card Dataset Preprocessing: The raw information about credit card transactions data is cleaned and preprocessed to handle missing values, outliers, and inconsistencies.

Feature selection: The relevant features are selected from the dataset which are used for model training.

Feature standardization: Features are standardized for unit variance and zero mean

Shaping of PCA transformed features: The Principal Component Analysis (PCA) is applied to the standardized features for dimensionality reduction and shaping the data.

Feature distribution visualization: This step is used to visualize the transformed features through box plot graph.

Predictions from IF, OCSVM and random forest classifier: The data is trained and tested on three algorithms Isolation Forest, One Class SVM and Random Forest classifier.

Combined predictions using simple voting mechanisms: The predictions from algorithms are combined using simple voting mechanism. The Voting mechanisms is one of the ensemble voting mechanism. It is also called as majority voting. In this mechanism, each classifier is of equal importance and contributes single vote. The final prediction is the most frequent vote.[23]

Evaluation of combine prediction: The evaluation of combined prediction is done using confusion matrix, precision, recall, F1 score and accuracy.

Calculate accuracy: This step calculates overall accuracy based on combined prediction.

IV. ALGORITHM FOR DATA ANOMALY DETECTION

Algorithm: Hybrid approach for anomaly detection with Isolation Forest and One-Class SVM

Input: selected features from a dataset

Output: Accuracy of a hybrid algorithm for data anomaly detection

1.BEGIN

2.dataset <- pd.read_csv("creditcard.csv")

3. dataset <- dataset.dropna()label_encoder <- LabelEncoder()

4.dataset['Label'] <- label_encoder.fit_transform(dataset[' Label'])

5. features <- ['list', 'of', 'selected', 'features']

6. X <- dataset[features]

7. y <- dataset['Label']

8.scaler <- StandardScaler()

9.X_scaled <- scaler.fit_transform(X)

10.boxplot(X_scaled)

11. isolation_forest <- IsolationForest(contamination=0.05)

12. isolation_forest.fit(X_scaled

13. isolation_forest_preds<- isolation_forest.predict(Scaled)

14. ocsvm <- OneClassSVM(kernel='rbf', gamma='auto')

15. ocsvm.fit(X_scaled)

16. ocsvm_preds <- ocsvm.predict(X_scaled)
17.combined_preds <- (isolation_forest_preds == -1) OR (ocsvm_preds == -1)

18.combined_preds <- 1 if combined_preds else 0
true_labels <- dataset['Label'].values

19. confusion_matrix <- confusion_matrix(true_labels, combined_preds)

20.classification_report <- classification_report(true_labels, combined_preds)

21. accuracy_score <-accuracy_score(true_labels, combined_preds)

PRINT confusion_matrix

PRINT classification_report

PRINT accuracy_score

END

V. ALGORITHMS

- Isolation Forest (IF):

Isolation Forest detects anomalies by isolating outliers based on their different characteristics. It examines the structure and distribution of transaction data. It randomly selects the split value between maximum and minimum values of that selected features. This process is repeated until all the instances are isolated. Instances with short average path lengths are considered as anomalies as they have distinct values requiring in fewer splits.

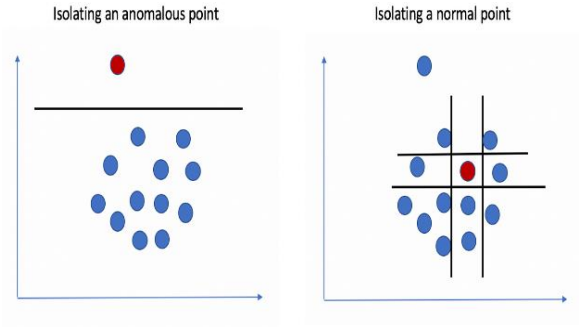

Fig. 2. Working of IF

- One Class Support Vector Machine(OCSVM)

OCSVM is another important algorithm which complements IF by its distinct working technique. It separates data into two regions: one containing majority of the data(normal transactions) while other containing anomalies(fraudulent transactions). It makes boundary around normal data points. The transactions which are outside the boundaries are considered as anomalies.

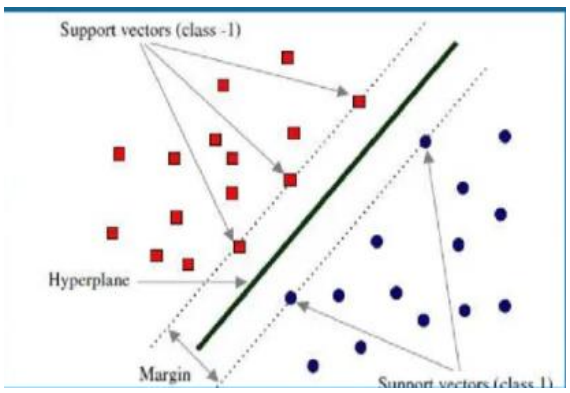

Fig. 3 Working of OCSVM

- Random Forest Classifier

RFC is considered as ensemble algorithm which constructs multiple decision trees during training and outputs the prediction of the individual tree. RF creates a forest of many decision trees. Each tree is trained on a random subset of data. Each tree is unrelated and RF can handle high dimensional data.[10] For classification purpose, RF uses majority voting scheme where each tree votes for a class and class with the most votes is considered as final prediction. For regression purpose, it averages the predictions of the individual trees.

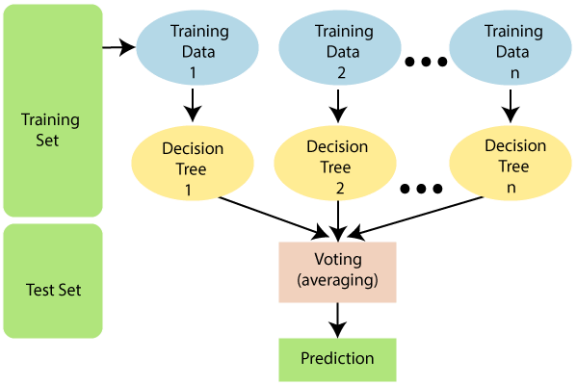

Fig. 4 Working of RFC

VI. Performance Evaluation

The proposed work is implemented in python and performed on credit card dataset from Kaggle[12] The performance is evaluated by confusion matrix and classification report with accuracy. In this section, we are going evaluated system step by step:

```
Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10',
       'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20',
       'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount',
       'Class'],
      dtype='object')
```

Fig. 5 Features of dataset

The fig. 5 shows all the features of the dataset creditcard from Kaggle and then these features are reduced and standardized

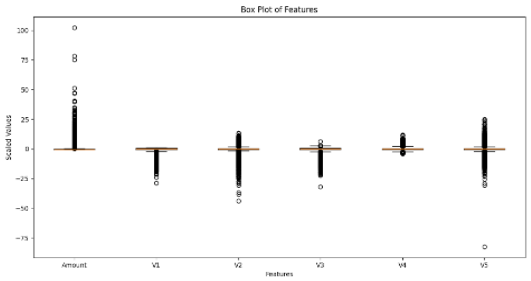

Fig 6. Box plot of features vs its scaled values of creditcard dataset

In the fig. 6 we plotted box plot of features vs its scaled values. It provides us the visual representation of features after dimensionality reduction using PCA.

```
Confusion Matrix:
[[280567   3748]
 [   240    252]]

Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99    284315
           1       0.06      0.51      0.11       492

    accuracy                           0.99    284807
   macro avg       0.53      0.75      0.55    284807
weighted avg       1.00      0.99      0.99    284807


Accuracy: 98.60%
```

Fig. 7 confusion matrix and classification report

Fig shows 7 shows performance evaluation parameters confusion matrix and classification report. The classification report shows good accuracy with good precision, recall value and f1-score for class 0. Also shows moderate recall value for class 1.

VII. CONCLUSION

The hybrid model of combination of Isolation Forest and One Class SVM with Random Forest classifier performs exceptionally well in detecting data anomalies in the credit card dataset, achieving good precision, recall, F1-score, and accuracy. As compared to the model containing single algorithm, this model performs well. If we consider credit card transactions data, this is highly imbalanced, so it is effective to use unsupervised algorithms IF and OCSVM with supervised approach RF classifier to enhance accuracy of the system.

VIII. FUTURE SCOPE

As the hybrid approach gives good accuracy but there is scope to improve precision and f1-score for class 1.

IX. REFERENCES

[1] Prerna Singh, Khyati Singla, Prince Piyush, Bharti Chugh, "Anomaly Detection Classifiers for Detecting Credit Card Fraudulent Transactions", 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2024.

[2] Dr.V. Ceronmani Sharmila,Kiran Kumar R ,"Credit Card Fraud Detection Using Anomaly Techniques",IEEE 2023.

[3] Swaroop K, Amruta D, Sanath J, "Credit Card Fraud Detection Using Machine Learning",International Journal of Engineering Research &Technology (IJERT),NCRACES Conference Proceedings 2019

[4] Arun Kumar Rai, Rajendra Kumar Dwivedi"Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme",International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)

[5] Ravi Ghevariya, Rahul Desai, Mohammed Husain Bohara, Dr. Dweepna Garg,"Credit Card Fraud Detection Using Local Outlier Factor & Isolation Forest Algorithms: A Complete Analysis",Fifth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2021)

[6] Chugh, Bharti, and Nitin Malik, ”Machine Learning Classifiers for Detecting Credit Card Fraudulent Transactions”, Information and Communication Technology for Competitive Strategies (ICTCS 2021), June 2022

[7] Esenogho, Ebenezer, Ibomoiye Domor Mienye, Theo G. Swart, Kehinde Aruleba, and George Obaido, ”A neural network ensemble with feature engineering for improved credit card fraud detection”, IEEE Access, vol. 10,, January 2022.

[8] Kittidachanan, Kittikun, Watha Minsan, Donlapark Pornnopparath, and Phimphaka Taninpong, ”Anomaly detection based on GS-OCSVM classification”, 2020 12th International Conference on Knowledge and Smart Technology (KST), IEEE, April 2020.

[9] S. N. Kalid, K. H NG, G. K Tong, K. C Khore., “A Mult iple Classifiers System for Anomaly Detect ion in Credit Card Data With Unbalanced and Overlapped Classes”, IEEE Access (2020), Vol. 8.

[10] Liang Zhang, Lingyun Liu, "Data Anomaly Detection Based on Isolation Forest Algorithm", IEEE International Conference on Computation, Big-Data and Engineering (ICCBE) 2022

[11] Jiangtao Ma, Yaqiong Qiao, Guangwu Hu , Yongzhong Huang, Arun Kumar Sangaiah , "De-Anonymizing Social Networks With Random Forest Classifier", IEEE 2017

[12] Dataset for credit card fraud. (2020). https://www.kaggle.com/mlgulb/creditcardfraud