A Study of Credit Card Fraud Detection using Adaptive Pattern Matching with Optimize Itemset

Mukesh Kumar Mandal
PG Scholar
Dept. of CSE, MITS, Bhopal

Abstract- Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. This work focuses on four main fraud occasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the frauds and we illustrate the evaluation with an appropriate performance measure. Another major key area that we address in our project is real-time credit card fraud detection. For this, we take the use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent. We also assess a novel strategy that effectively addresses the skewed distribution of data. The data used in our experiments come from a financial institution according to a confidential disclosure agreement. As the developed machine learning models APM-OI (Adaptive Pattern Matching with Optimize Itemset) present an average level of accuracy, we hope to focus on improving the prediction levels to acquire a better prediction. Precision improve upto 21.32% during characterization process, hence maximum fraud detection may be relevant with train data. Recall improves upto 14.1% during arrangement process, hence maximum relevant train data to be classified as fraud detection. Accuracy improves upto 23.4%, hence high inspecting unpredictable examine. F1-Score improves upto 18.05%, Uncertainty of characterization becomes reduce.

Keywords: credit card frauds, fraud detection system, fraud detection, confidential disclosure agreement, real-time credit card fraud detection, skewed distribution.

I. INTRODUCTION

In commonsense application, numerous datasets are imbalanced, i.e., a few classes have substantially more occurrences than others. Imbalanced learning is regular as a rule like data sifting and misrepresentation location. Datasets irregularity must be thought about in classifier structuring, generally the classifier may will in general be overpowered by the larger part class and to disregard the minority class. Re-testing method is a successful way to deal with lopsidedness learning. Numerous re-examining strategies are utilized to diminish or dispense with the degree of datasets lopsidedness, for example, over-inspecting the minority class, under-testing the greater part class and the blend of the two techniques.

Yet, it demonstrated that under-inspecting can conceivably evacuate certain significant examples and lose some helpful data, and over-testing may prompt over fitting. Over-inspecting strategies additionally experience the ill effects of commotion and exceptions. Bolster Vector Machine (SVM) has been broadly utilized in numerous application territories of AI. Be that as it may, standard SVM is never again appropriate to unevenness class particularly when the datasets are incredibly imbalanced. A successful way to deal with improve the presentation of SVM utilized in imbalanced datasets is to inclination the classifier so it gives more consideration to minority cases. This should be possible by setting distinctive misclassifying punishment.

II. BACKGROUND

Ruttala Sailusha et. al, Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the car was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. This project aims to focus mainly on machine learning algorithms. The algorithms used are random forest algorithm and the Ada boost algorithm. The results of the two algorithms are based on accuracy, precision, recall and F1-score. The ROC curve is plotted based on the confusion matrix. The Random Forest and the Adaboost algorithms are compared and the algorithm that has the greatest accuracy, precision, recall, and F1-score is considered as the best algorithm that is used to detect the fraud. [1]

Anuruddh Thennakoon et. al, Credit card fraud events take place frequently and then result in huge financial losses. The number of online transactions has grown in large quantities and online credit card transactions hold a huge share of these transactions. Therefore, banks and financial institutions offer credit card fraud detection applications much value and demand. Fraudulent transactions can occur in various ways and can be put into different categories. This paper focuses on four main fraud occasions in real-world transactions. Each fraud is addressed using a series of machine learning models and the best method is selected via an evaluation. This evaluation provides a comprehensive guide to selecting an optimal algorithm with respect to the type of the frauds and we illustrate the evaluation with an appropriate performance measure. Another major key area that we address in our project is real-time credit card fraud detection. For this, we take the use of predictive analytics done by the implemented machine learning models and an API module to decide if a particular transaction is genuine or fraudulent. We also assess a novel strategy that effectively addresses the skewed distribution of data. The data used in our experiments come from a financial institution according to a confidential disclosure agreement. [2]

J. Gao et. al, With the rapid development of big data and machine learning technologies, many fields have begun to use related algorithms and methods. Classification algorithms have been widely used in the fields of financial risk identification, fault diagnosis, medical diagnosis, etc. However, the datasets are often unbalanced in these cases and the original methods fail to classify instances correctly. Many methods such as over-sampling, under-sampling and ensemble methods were raised to improve the classifier’s performance, but which one to choose for a certain dataset still remains a problem. Therefore, this paper aims at a experimental conclusion on which kind of method can perform best on unbalanced classification problems generally. In detail, we evaluated the performances of 13 kinds of methods for unbalanced classification on several unbalanced datasets which have different amounts of instances and different ratios of positive instances, and finally came to a conclusion. [3]

Victor et. al, In imbalanced classification tasks, the training datasets may show class overlapping and classes of low density. In these scenarios, the predictions for the minority class are impaired. Although assessing the imbalance level of a training set is straightforward, it is hard to measure other aspects that may affect the predictive performance of classification algorithms in imbalanced tasks. This paper presents a set of measures designed to understand the difficulty of imbalanced classification tasks by regarding on each class individually. They are adapted from popular data complexity measures for classification problems, which are shown to perform poorly in imbalanced scenarios. Experiments on synthetic datasets with different levels of imbalance, class overlapping and density of the classes show that the proposed adaptations can better explain the difficulty of imbalanced classification tasks. [4]

Alex et. al, This paper presents Fraud-BNC, a customized Bayesian Network Classifier (BNC) algorithm for a real credit card fraud detection problem. The task of creating Fraud-BNC was automatically performed by a Hyper-Heuristic Evolutionary Algorithm (HHEA), which organizes the knowledge about the BNC algorithms into a taxonomy and searches for the best combination of these components for a given dataset. Fraud-BNC was automatically generated using a dataset from PagSeguro, the most popular Brazilian online payment service, and tested together with two strategies for dealing with cost-sensitive classification. Results obtained were compared to seven other algorithms, and analyzed considering the data classification problem and the economic efficiency of the method. Fraud-BNC presented itself as the best algorithm to provide a good trade-off between both perspectives, improving the current company's economic efficiency in up to 72.64%. [5]

III. PROBLEM IDENTIFICATION

The basic objections of my hypothesis work are according to the accompanying:

1. Unrelated information are arrange for explicit dataset, hence minimum fraud detection may be relevant with train data.
2. Inconsistency exists during arrangement process, hence minimum relevant train data to be classified as fraud detection.
3. Due to low inspecting unpredictable examining rate create, hence obtain accuracy is low.
4. Uncertainty of characterization, hence obtain F1-measure becomes down.

IV. PROPOSED METHODOLOGY

The proposed methodology Adaptive Pattern Matching with Optimize Itemset (APM-OI) is as follows. The pseudo code of training algorithm is given in Algorithm 1.

Algorithm 1: Training Phase of Proposed Method (APM-OI)

```
Input: Customer Transactions Database D, Support S
Output: Legal Pattern Database LPD, Fraud Pattern Database FPD
Begin
Group the transactions of each customer together.
Let there are —n‖ groups corresponds to —n‖ customers f or i = 1 to n do
Separate each group Gi into two different groups LGi and FGi of legal and fraud transactions. Let there are —m‖ legal and —k‖ fraud transactions
FIS = Apriori (LGi, S, m); //Set of frequent itemset LP = max (FIS); //Large Frequent Itemset
LPD(i) = LP;
FIS = Apriori (FGi, S, k); //Set of frequent itemset FP = max (FIS); //Large Frequent Itemset
FPD(i) = FP;
End for
Return LPD & FPD;
End
```

The pseudo code of training algorithm is given in Algorithm 2.

Algorithm 2: Testing Phase of Proposed Method (APM-OI)

```
Input: Legal Pattern Database LPD, Fraud Pattern Database FPD, Incoming Transaction T, Number of customers' —n‖, Number of attributes —k‖, matching percentage —mp‖
Output: 0 (if legal) or 1 (if fraud)
Assumption
1. First attribute of each record in pattern databases and incoming transaction is Customer ID
2. If an attribute is missing in the frequent itemset (ie, this attribute has different values in each transaction and thus it is not contributing to the pattern) then we considered it as invalid.
Begin
lc = 0; //legal attribute match count fc = 0; //fraud attribute match count
for i =1 to n do
if (LPD(i, 1)=T(1)) then //First attribute
for j =2 to k do
if (LPD(i, j) is valid and LPD(i, j)= T(j)) then
lc = lc + 1;
endif endfor
endif endfor
for i =1 to n do
if (FPD(i, 1)= T(1)) then for j =2 to k do
if (FPD(i,j) is valid and FPD(i,j)= T(j)) then
fc = fc + 1;
endif endfor
endif endfor
if (fc = 0) then //no fraud pattern
if ((lc/no. of valid attributes in legal pattern) ≥ mp) then return (0); //legal transaction
else return (1); //fraud transaction
endif
elseif (lc = 0) then //no legal pattern
if ((fc/no. of valid attributes in fraud pattern) ≥ mp) then return (1); //fraud transaction
else return (0); //legal transaction
endif
```

```
elseif (lc > 0 && fc > 0) then //both legal and fraud patterns are available
if (fc ≥ lc) then return (1); //fraud transaction
else return (0); //legal Transaction
endif endif
End
```

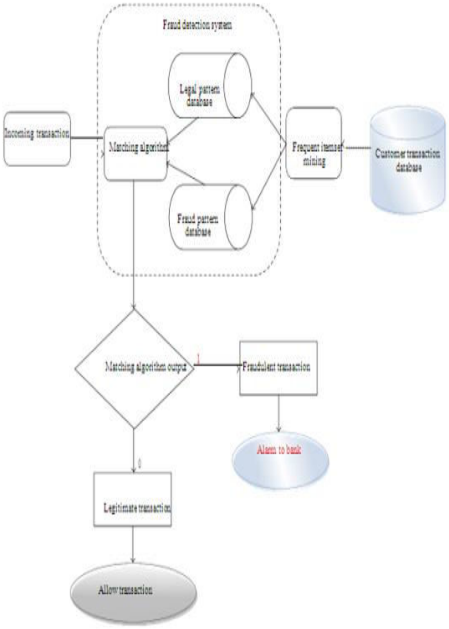

Figure 1: Outline of proposed work

VI. RESULTS AND ANALYSIS

The performance of the proposed scheme is evaluated in terms of various metrics such as precision, recall, accuracy and F1-Score. These metrics play very important role during performance evaluation. The proposed scheme requires high precision, recall, accuracy and F1-Score as compared to existing schemes Random Forest[1].

Table 1: Confusion Matrix as per number of transactions

Number of Transactions	TP	TN	FP	FN
1500	896	297	198	109
3000	1742	312	487	459
4500	2746	916	712	918
6000	4088	1642	1114	656
7500	4878	1214	1724	1184
9000	4878	1214	1724	1184
10500	10500	1098	1644	896

Table 2: Analysis of precision

Number of Transactions	Random Forest [1]	APM-OI (Proposed)
1500	0.67	0.82
3000	0.72	0.78
4500	0.56	0.73
6000	0.59	0.73
7500	0.71	0.78
9000	0.67	0.74
10500	0.65	0.81

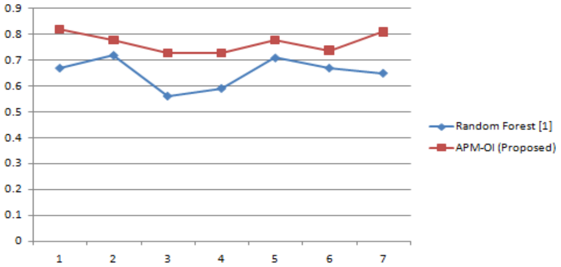

Figure 2: Graphical analysis of precision

The proposed method APM-OI (Adaptive Pattern Matching - Optimize Itemset) performs outstanding result in case of precision. When specify 1500 transactions then precision of APM-OI is 0.82 instead of 0.67. Similarly for 7500 transactions, precision of APM-OI is 0.78 instead of 0.71.

Table 3: Analysis of Recall

Number of Transactions	Random Forest [1]	APM-OI (Proposed)
1500	0.78	0.89
3000	0.67	0.79
4500	0.53	0.68
6000	0.62	0.71
7500	0.72	0.86
9000	0.69	0.8
10500	0.71	0.88

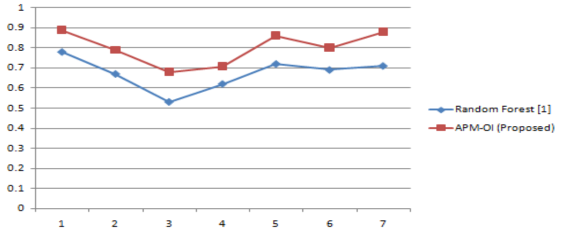

Figure 3: Graphical analysis of recall

The proposed method APM-OI (Adaptive Pattern Matching - Optimize Itemset) performs outstanding result in case of recall. When specify 1500 transactions then recall of APM-OI is 0.89 instead of 0.78. Similarly for 7500 transactions, precision of APM-OI is 0.86 instead of 0.72.

Table 4: Analysis of Accuracy

Number of Transactions	Random Forest [1]	APM-OI (Proposed)
1500	0.64	0.79
3000	0.58	0.68
4500	0.51	0.64
6000	0.52	0.64
7500	0.61	0.76
9000	0.62	0.68
10500	0.63	0.76

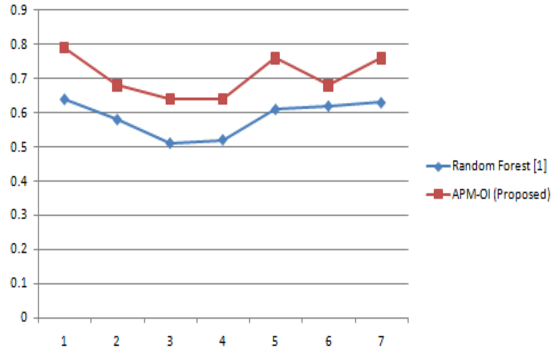

Figure 4: Graphical analysis of accuracy

The proposed method APM-OI (Adaptive Pattern Matching - Optimize Itemset) performs outstanding result in case of accuracy. When specify 1500 transactions then accuracy of APM-OI is 7ms instead of 9ms. Similarly for 7500 transactions, precision of APM-OI is 21ms instead of 23ms.

Table 4: Analysis of F1-Score

Number of Transactions	Random Forest [1]	APM-OI (Proposed)
1500	0.72	0.85
3000	0.66	0.79
4500	0.62	0.71
6000	0.64	0.72
7500	0.7	0.82
9000	0.65	0.77
10500	0.71	0.84

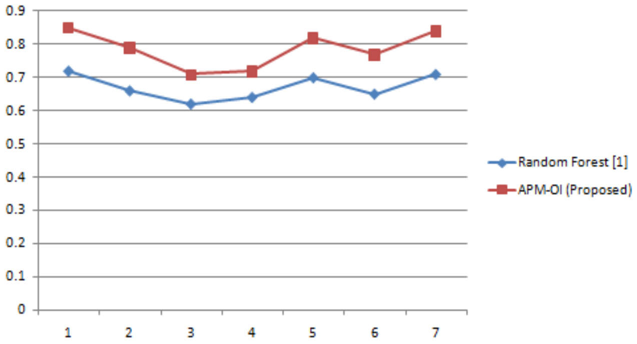

Figure 4: Graphical analysis of F1-Score

The proposed method APM-OI (Adaptive Pattern Matching - Optimize Itemset) performs outstanding result in case of F1-Score. When specify 1500 transactions then F1-Score of APM-OI is 0.85 instead of 0.72. Similarly for 7500 transactions, precision of APM-OI is 0.82 instead of 0.7.

VII. CONCLUSION

Credit card fraud detection has been a keen area of research for the researchers for years and will be an intriguing area of research in the coming future. This happens majorly due to continuous change of patterns in frauds. As the developed machine learning models APM-OI (Adaptive Pattern Matching with Optimize Itemset) present an average level of accuracy, we hope to focus on improving the prediction levels to acquire a better prediction.

1. Precision improve upto 21.32% during characterization process, hence maximum fraud detection may be relevant with train data.
2. Recall improves upto 14.1% during arrangement process, hence maximum relevant train data to be classified as fraud detection.
3. Accuracy improves upto 23.4%, hence high inspecting unpredictable examine.
4. F1-Score improves upto 18.05%, Uncertainty of characterization becomes reduce.

As per analysis, number of observation has been taken on multiple dataset and appreciable of finding where achieved.

VIII. FUTURE SCOPE

The future work of this dissertation task is as per the following:

Also, the future extensions aim to focus on location-based frauds. One thing worth investigating in the future is whether the strategies related to cost-sensitive classification could be added to the components given to the hyper-heuristic.

REFERENCES

[1] Ruttala Sailusha, V. Gnaneswar, R. Ramesh, G. Ramakoteswara Rao, "Credit Card Fraud Detection Using Machine Learning", IEEE International Conference on Intelligent Computing and Control Systems, 2020.

[2] Anuruddh Thennakoon, Chee Bhagyani, Sasith Premadasa, Shalith Mihiranga, Nuwan Kuruwitaarachchi,"Real-time Credit Card Fraud Detection Using Machine Learning", IEEE Transaction on Machine Learning, 2019.

[3] J. Gao, L. Gong, J. Y. Wang, Z. C. Mo, "Study on Unbalanced Binary Classification with Unknown Misclassification Costs", IEEE Transaction on Machine Learning, 2019.

[4] Victor H. Barella, Lu´ıs P. F. Garcia, Marcilio P. de Souto, Ana C. Lorena, Andr´e de Carvalho, "Data Complexity Measures for Imbalanced Classification Tasks", IEEE Transaction on Machine Learning, 2018.

[5] Alex G.C. de Sá, Adriano C.M. Pereira, Gisele L. Pappa, "A customized classification algorithm for credit card fraud detection", Springer Journal of Engineering Applications of Artificial Intelligence, 2018.

[6] Qi Dong, Shaogang Gong, Xiatian Zhu, "Class Rectification Hard Mining for Imbalanced Deep Learning", Springer Journal of Artificial Intelligence, 2017.

[7] Josey Mathew, Chee Khiang Pang, Ming Luo and Weng Hoe Leong, "Classification of Imbalanced Data by Oversampling in Kernel Space of Support Vector Machines", IEEE Transactions On Neural Networks And Learning Systems, 2017.

[8] Guillaume Lematre, Fernando Nogueira, Christos K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning", Journal of Machine Learning Research, 2017.

[9] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, Gong Bing, "Learning from class-imbalanced data: Review of methods and applications", Elsevier Journal of Expert Systems With Applications, 2017.

[10] Alberto Fernández, Sara del Río, Nitesh V. Chawla, Francisco Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges", Arebian Journal of Complex Intelligent System, 2017.