Unbalanced feature selection for credit card fraud detection with KNN classification with Feature Enhancement

Nikita Sawhney
Department of Information Technology
Chandigarh Engineering College Landran
Mohali, India

Dr. Bikrampal Kaur
Department of Information Technology
Chandigarh Engineering College Landran
Mohali, India

***Abstract*—The credit card fraud detection is very important task for the financial institutions offering the credit cards and other forms of debt to its customers. The fraudulent behaviour classification is known to significantly reduce the losses of the financial institutions by predicting the loan defaults as early as possible. The prediction of the loan defaulters creates the possibility of stopping the lending of debt to such customers for loss minimization. In this paper, the KNN and SVM based classifiers are deployed for the purpose of credit card fraud classification, where the performance of both of the classifiers is analyzed on the basis of various performance parameters. The comparison of the performance has been analyzed in the form of recall, precision, F1 measure and accuracy. The credit card fraud classification based upon KNN is found more accurate in comparison with SVM classification, where KNN has been recorded with approx 99.95% against 99.94% for the SVM on an average.**

Keywords—KNN classification, SVM classification, loan fraud, predictive analysis.

I. INTRODUCTION

Secure credit services of banks and development of E-business a reliable fraud detection system is essential to support safe credit card usage, Fraud detection based on analyzing existing purchase data of cardholder (current spending behavior) is a promising way for reducing the rate of credit card frauds. Fraud detection systems come into scenario when the fraudsters exceed the fraud prevention systems and start fraudulent transactions. Along with the developments in the Information Technology and improvements in the communication channels, fraud is spreading all over the world with results of large amount of fraudulent loss. Anderson (2007) has identified and described the different types of fraud. Credit card frauds can be proceed in many different ways such as simple theft, counterfeit cards, Never Received Issue (NRI), application fraud and online/Electronic fraud (where the card holder is not present). Credit card fraud detection is dreadfully difficult, but also common problem for solution. As there is limited amount of data with the transactions being confided, for example, transaction amount, merchant category code (MCC), acquirer number and date and time, address of the merchant. Various techniques in Knowledge Discovery, such as decision tree, neural network and case based reasoning have broadly been used for forming several fraud detection systems/ models. These techniques usually need adequate number of normal transactions and fraud transactions for learning fraud patterns. However, the ratio of fraudulent transactions to its normal transactions is low extremely, for an individual bank.

II. LITERATURE SURVEY

Kulkarni, Pallavi et. al. [1] has worked on the unbalanced financial data for the credit card fraud detection using the regression model. Traditionally, machine learning area has been developing algorithms that have certain assumptions on underlying distribution of data, such as data should have predetermined and fixed distribution.

Bahnsen, Alejandro Correa et. al. [2] has worked towards the feature engineering in order to improve the feature descriptors for the purpose of credit card fraud investigations. In this paper the authors have expanded the transaction aggregations strategy, and proposed to create a new set of features based on analyzing the periodic behavior of the time of a transaction using the von Mises distribution.

Dal Pozzolo, Andrea et. al. [3] has developed the practitioner perspective method for the purpose of credit card fraud investigation. In this paper, the authors have analyzed the threats on cloud users' activity logs considering the collusion between cloud users, providers, and investigators.

Halvaiee, Neda Soltani et. al. [4] has worked on the unique credit card fraud detection model using the artificial immune systems. The authors have proposed the system which detects fraud in credit card transaction processing using a decision tree with combination of Luan's algorithm and Hunt's algorithm. Luhn's algorithm is used to validate the card number. Address matching check does not guarantee whether a transaction is fraud or genuine. But if the two addresses match, the transaction can be classified as genuine with a high probability.

Van Vlasselaer et. al. [5] has worked on the APATE model, which utilizes the network-based approach for the credit card fraud detection. This paper proposes APATE, a novel approach to detect fraudulent credit card transactions conducted in online stores.

Prakash, A. et. al. [6] has proposed the multiple semi-hidden markov model for credit card fraud detection. The main intent of this research is automating the use of Multiple Semi-Hidden Markov Model, by liberating customers from the necessity of statistical knowledge.

A. Literature Table

Index	Authors	Problem Addressed	Technologies Used	Algorithm Model
1	Kulkarni, Pallavi et. al. [1]	Financial fraud detection using imbalanced data	Application of artificial intelligent fraud detection model using machine learning methods	Variable distribution normalization, outlier marking, semi-fragile features
2	Dal Pozzolo, Andrea et. al. [3]	User activity log thread detection	Time series analytics applied to analyze the user activity logs and threat consideration	Moving averages, Log access attempt based cross evaluation.
3	Halvaiee, Neda Soltani Soltani et. al. [4]	Credit card fraud detection with transactional analysis	Artificial immune systems (AIS) applied to detect the credit card frauds	Geo-location based reliability, Feature line up with decision tree
4	Van Vlasselaer et. al. [5]	Credit card fraud detection using APATE model	Recency, Frquency & Monetary based evaluation to detect credit card frauds	Intrinsic credit card features with transaction history and spending patterns
5	Prakash A et. al. [6]	Automation of semi-hidden markov models (HMM)	Multiple instance based semi-hidden markov model to assess the hidden parameters of data by combining multiple parameters or variables	Hidden Markov Models (HMM)

III. EXPERIMENTAL DESIGN

In order to eliminate the problems in the existing model, the proposed model will be designed with the unbalanced metric normalization methods, where the combination of averages and floating averages are be utilized to create the state-of-art system in order to minimize the feature unbalance in the feature matrix.

In addition to averaging factor based feature description, the flexible and robust feature scaling practices can be utilized, which may vary from column to column in the given data according to its volatility and overall variance, to precise the features in order to create the high accuracy based credit card fraud detection model. The model with best feature selection with probabilistic classification for the purpose of credit card fraud detection would be deployed with certain improvements or enhancements during the proposed model implementation. The best feature selection method will incorporate the selection of the features on the basis of their compatibility, which can be measured with the column or feature variance. The probabilistic classification algorithm involves the probability based matching between the training and testing data, which is decided with the maximum likeliness or similarity between the entries of test and train data. The following algorithm describes the overall workflow of the proposed model:

Read the source healthcare data, and Extract the features from the numerical (quantitative) or categorical (qualitative) data source. Feature descriptor will be the set of selective features, and will describe smaller details than the original feature matrix.

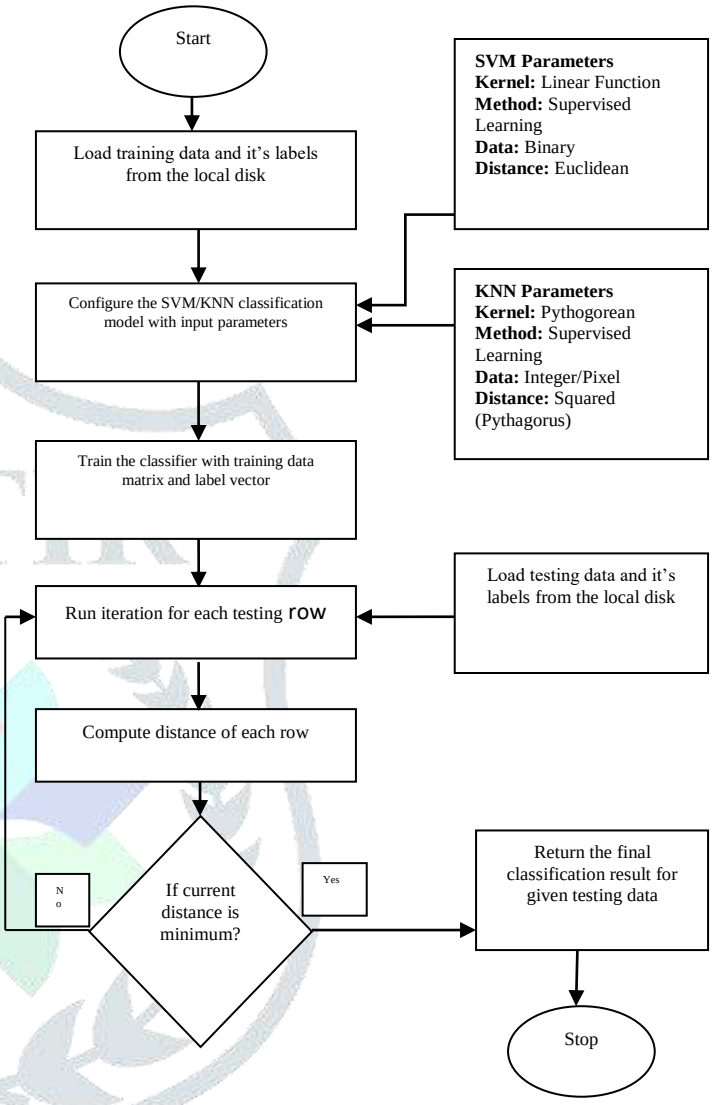

Figure 1: Classification model for credit card fraud detection

Algorithm 1: Supervised Classification Algorithm

1. Perform pre-processing step to validate the feature descriptor set and arrange all of the feature descriptors in the single feature sets as the training set.
2. Prepare the group data by adding the group IDs corresponding with all of the samples or feature descriptors in the training set.
3. Run Linear training on the feature descriptor training set and return the weight and bias information for all feature descriptors in the training set.
4. Run Supervised classification classifier by submitting the KNN or SVM Equation to create feature data, group data and the testing feature descriptor vector.
5. Return the matching classification information.
6. Evaluate the classification information and return the decision logic.

Testing: Feature vector of training set is fed to learnt model to assign a class label to given test samples in the form of a unclassified set.

IV. RESULT ANALYSIS

The results of the proposed model are analyzed on the basis of various accuracy based parameters, which includes the performance parameters of precision, recall, F1 measure and accuracy. The following table shows the results obtained from the KNN classifier, where The KNN classifier has been analyzed for the various performance parameters in the given scenario under the 10 rotations. The results are obtained for the statistical type 1 and type 2 errors, which have been further used to compute the various types of the accuracy, precision, recall and f1-measure parameters. The results of KNN are obtained in the form of various performance parameters as per shown in the following tables. The statistical analysis includes the parameters of accuracy, precision, recall and f1 error.

Round Index	Precision	Recall	F1 Measure	Accuracy
1	77.04918	87.03704	81.73913	99.94733
2	76.1194	94.44444	84.29752	99.95235
3	82.35294	91.80328	86.82171	99.95736
4	78.125	92.59259	84.74576	99.95486
5	80.59701	94.73684	87.09677	99.95987
6	69.23077	91.52542	78.83212	99.92727
7	80.30303	92.98246	86.17886	99.95736
8	78.46154	89.47368	83.60656	99.94984
9	77.14286	93.10345	84.375	99.94984
10	75.5814	87.83784	81.25	99.92476

Table 1: Statistical parameter based analysis of KNN

The higher recall for KNN has been recorded at 94.73%, whereas the higher precision has been recorded at approx. 82% for the KNN model. The highest accuracy in all 10 rounds is recorded at 99.957%, wheras the higher F1 measure has been recorded approx 87.10% The accuracy is higher recommended for the real-time deployment of the KNN classification based model, however the SVM is also a close candidate. The results of SVM are described in the following table in the form of similar performance parameters.

Round Index	Precision	Recall	F1 Measure	Accuracy
1	85.2459	77.61194	81.25	99.93981
2	76.1194	87.93103	81.6	99.94232
3	85.29412	85.29412	85.29412	99.94984
4	76.5625	89.09091	82.35294	99.94733
5	83.58209	81.15942	82.35294	99.93981
6	74.35897	87.87879	80.55556	99.92978
7	87.87879	87.87879	87.87879	99.95987
8	83.07692	81.81818	82.44275	99.94232
9	80	81.15942	80.57554	99.93229
10	75.5814	84.41558	79.7546	99.91724

Table 2: Statistical parameter based analysis of SVM

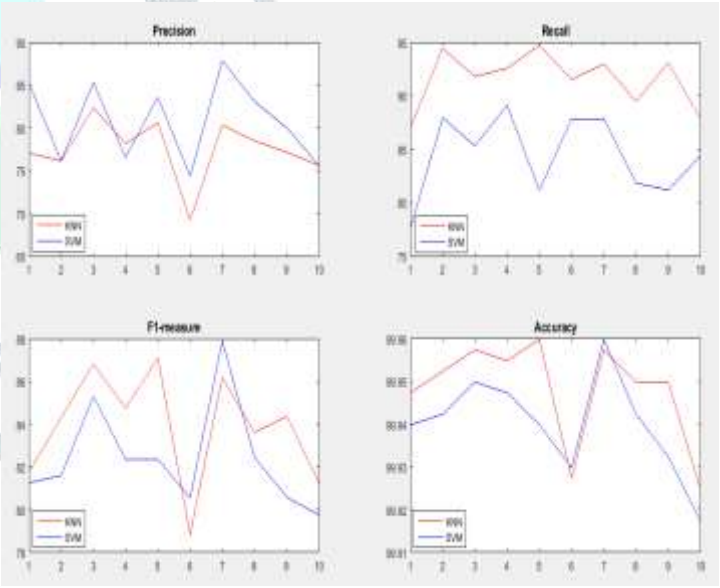

Figure 2: Performance based comparison of KNN and SVM

The higher precision for SVM has been recorded at 87.88%, whereas the higher recall has been recorded at approx. 89% for the KNN model. The highest F1 measure in all 10 rounds is recorded at 87.88%, wheras the highest accuracy is recorded approx 99.959%. The highest accuracy of SVM (99.959%) equal the highest accuracy of KNN, however the average accuracy of KNN (approx. 99.95%) is slightly higher than SVM (approx. 99.94%), which shows its better performance in comparison.

IV. CONCLUSION

In this paper, the credit card fraud detection models are prepared using the KNN and SVM based classifiers. The KNN and SVM based credit card fraud classification models are tested over the dataset containing nearly 300,000 user data, out of which nearly 200,000 signatures are used for the training of the classifier, and remaining approx 88,000 signatures are used to test the classifiers. The average accuracy has been recorded at 99.95% for the KNN and 99.94% for SVM classifier. The SVM has been found higher only on the basis of recall, where it has been recorded with 80.77% against the 77.50% of KNN. The comparison between the KNN and SVM proves the higher efficiency of KNN model for the purpose of credit card fraud classification.

V. REFERENCES

[1] Kulkarni, Pallavi, and Roshani Ade. "Logistic Regression Learning Model for Handling Concept Drift with Unbalanced Data in Credit Card Fraud Detection System." In *Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 681-689. Springer India, 2016

[2] Bahnsen, Alejandro Correa, Djamila Aouada, Aleksandar Stojanovic, and Björn Ottersten. "Feature engineering strategies for credit card fraud detection." *Expert Systems With Applications*. 51 *pp*. 134-142 (2016)

[3] Dal Pozzolo, Andrea, Olivier Caelen, Yann-Ael Le Borgne, Serge Waterschoot, and Gianluca Bontempi. "Learned lessons in credit card fraud detection from a practitioner perspective." *Expert systems with applications* 41, no. 10 *pp*. 4915-4928 (2014)

[4] Halvaiee, Neda Soltani, and Mohammad Kazem Akbari. "A novel model for credit card fraud detection using Artificial Immune Systems." *Applied Soft Computing* 24 *pp*. 40-49 (2014).

[5] Van Vlasselaer, Véronique, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. "APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions." *Decision Support Systems* 75 *pp.* 38-48 (2015).

[6] Prakash, A., and C. Chandrasekar. "An optimized multiple semi-hidden markov model for credit card fraud detection." *Indian Journal of Science and Technology* 8, no. 2 *pp*.165-171 (2015).

[7] Bahnsen, Alejandro Correa, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. "Improving credit card fraud detection with calibrated probabilities." In *Proceedings of the 2014 SIAM International Conference on Data Mining*,. Society for Industrial and Applied Mathematics *pp*.677-685 (2014)

[8] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." *Procedia Computer Science* 48 *pp*.679-685 (2015).

[9] Seeja, K. R., and Masoumeh Zareapoor. "FraudMiner: a novel credit card fraud detection model based on frequent itemset mining." *The Scientific World Journal* 2014 (2014).