Principal Component Analysis (PCA) based Hyperplane Reallocation with KNN Classification

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 systems are in the development all across the globe to centralize the health records of all citizens in order to facilitate the flexible provision to all health facilities to the historical health records whenever required. These systems are known to integrate the data of all of the possible diseases together on one platform. These systems can be also used to predict the chances of appearance of disease on the basis of the healthcare analysis done periodically specifically in the developed nations. In this thesis, the work has been performed on the credit card fraud detection. The credit card fraud detection methods are known to claim a number of lives every year across the globe. Hence it becomes important to implement such system, which can predict the chances of committing the fraud in person specific. KNN has been recorded with higher accuracy (approx. 99.95%) than SVM (99.94%) and Logistic Regression (99.92%).**

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

Aktepe, Adnan et. al. [1] has worked on the customer satisfaction and loyalty analysis with classification algorithms and structural equation modeling. Businesses can maintain their effectiveness as long as they have satisfied and loyal customers. Customer relationship management provides significant advantages for companies especially in gaining competitiveness. In order to reach these objectives primarily companies need to identify and analyze their customers.

Gaiardelli, Paolo et. al. [2] has worked towards the classification model for product-service offerings. In this paper, the authors have developed a comprehensive model for classifying traditional and green Product-Service offerings, thus combining business and green offerings in a single model.

Prakash, A. et. al. [18] 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. The number of states and also its model parameters are decided by the Cuckoo Search algorithm.

Bahnsen, Alejandro Correa et. al. [19] has worked on the improvement of the credit card fraud detection models with the calibrated probabilities. In this paper two different methods for calibrating probabilities are evaluated and analyzed in the context of credit card fraud detection, with the objective of finding the model that minimizes the real losses due to fraud.

Seeja, K. R. et. al. [20] has worked upon the frequent itemset mining based credit card fraud investigation. This paper proposes an intelligent credit card fraud detection model for detecting fraud from highly imbalanced and anonymous credit card transaction datasets. The class imbalance problem is handled by finding legal as well as fraud transaction patterns for each customer by using frequent itemset mining.

III. EXPERIMENTAL DESIGN

The credit card frauds are on the rise with the rise in the number of credit card holders. The credit card frauds takes place with the non-awareness of the holders or IT based expertise of the hackers, which as a result accounts for the loss of heavier amounts every year. To commit the credit card fraud, the hackers utilize the various methods from the social engineering, point of sales hacking, phishing attacks, etc to steal the information related to the credit cards. The credit card fraud detection model in existing scheme is utilizing the unbalanced metrics and practices to balance them in order to produce the decision, whether the credit card fraud has taken place or not. The feature scaling has been found missing from the existing model design, which account for the dilution of the data column dominance. The feature scaling methods can further improve the overall accuracy. The mean

absolute error (MAE) has been recorded with the value of 74.83%, which is considerably very high and must be curbed or neutralized in order to achieve the higher accuracy. The presence of the statistical type 1 and 2 errors (i.e. False Positive & False Negative) in the results has been indicated by the relative absolute error (RAE), which has been recorded at 0.406 and have resulted the lower accuracy based results of 96.6243%. The input data acquisition is done on the historical data of the credit card for last 1 to 10 years. The historical data of credit card spending contains the readings of the spending based listings in the day to day interval analysis, which contains the starting and closing value of each day accounted in the historical data. During the implementation, the proposed model would be designed using the ensemble of the regression models with squared distance based classification for the purpose of historical data processing, which is generally utilized for the long term prediction.

IV. RESULT ANALYSIS

The three classification algorithms, which are tested for the performance of credit card fraud detection model, are tested side by side for their performance in order to find the best candidate for the fraud classification modeling. In the following table, KNN, Logistic Regression and SVM are compared side by side on the basis of various performance parameters, out of which the higher accuracy has been recorded with the KNN classifier. The KNN classifier is found to be higher than SVM and Logistic Regression for Recall and F1 Measure based parameters, however, the precision is higher in SVM. Overall analysis shows the most successful classification results from the KNN classifier, which has been found stronger on the basis of more parameters than any other classifier.

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.

Table 1:Average Accuracy based analysis of the classification models

Classifier Name	Precision	Recall	F1 Measure	Accuracy
K-Nearest Neighbor (KNN)	77.49631	91.5537	83.89434294	99.94809
Maximum Entropy (Logistic Regression)	62.97191	87.23197	72.85516	99.91874
Support Vector Regression (SVM)	80.77001	84.42382	82.40572	99.94006

Table 2: Error based analysis of the classification models

Data Bag No.	Existing Model		Proposed Model	
	Standard Error (epsilon)	Normalized Error (beta)	Standard Error (epsilon)	Normalized Error (beta)
1	0.2594	0.9782	0.052667218	0.182609
2	0.2452	1.0427	0.047651293	0.157025
3	0.2224	1.1335	0.042635367	0.131783
4	0.2657	0.9652	0.04514333	0.152542
5	0.2837	0.9905	0.040127405	0.129032
6	0.2169	1.1668	0.072730921	0.211679
7	0.2744	0.9448	0.042635367	0.138211
8	0.2743	0.9544	0.050159256	0.163934
9	0.202	1.0165	0.050159256	0.15625
10	0.2436	1.0611	0.075238883	0.1875
Average	0.24876	1.02537	0.05191483	0.1610565

The error based analysis has been conducted between the existing model and KNN classification model for the detection of the credit card frauds. The proposed model based on KNN has been reportedly found with significantly lower error in comparison to the existing model. The standard error for the existing model on an average has been computed at 0.25%, which is nearly 0.20% higher than the proposed model's 0.05% error overall. Also, the normalized error in the case of proposed model has been recorded at 0.16%, which is more than 0.85% lower than existing model. This shows the robustness of the proposed model in maintaining the higher accuracy of the proposed model.

Table 3: Average Error and Accuracy based analysis of the classification models

	Mean Absolute Error (MAE)	Relative Absolute Error (RAE)	Accuracy
Existing	74.83%	0.406	96.62%
Proposed	64.71%	0.051	99.95%

The proposed model based upon KNN classification has been recorded with nearly 3% higher accuracy than the existing model. The significant increase in the accuracy has been recorded due to the use of hyperplane shift in the proposed model, which eventually shifts the fraud patterns by 150% and creates the definite hyperplane. This phenomenon is known to reduce the overall false negative rate, which eventually increases the overall accuracy. Also, the proposed model has been found with lower MAE and RAE errors than the existing model based upon Logistic Regression.

IV. CONCLUSION

The proposed model based upon KNN has been evaluated against the existing models to test its compatibility to the real-time credit card fraud detection models. The values of F1-measure (83.89%) and recall (91.55%) are significantly higher than SVM and Logistic Regression. This proves the high acceptability and adaptability of the KNN based credit card fraud detection model for the final proposed credit card fraud classification. The KNN has been compared to the existing model on the basis of the standard error and normalized error, where the KNN has outperformed the existing model by nearly 0.20%. The KNN has been recorded with 0.05% standard error against the 0.25% for existing model. The KNN based credit card classification has also been found better than existing model on the basis of normalized error, where KNN has been recorded with 0.16% against the 1.03% for existing model. This shows the robustness of the KNN

based proposed model in maintaining the higher accuracy of the credit card fraud classification. In the future, the deep learning model can be deployed over the significantly engineered features in order to enhance the features, which carries the higher probability of improving accuracy. The proposed model can be also improved by infusing the multiple classification algorithms.

V. REFERENCES

[1] A. Aktepe, S. Ersöz, and B. Toklu, "Customer satisfaction and loyalty analysis with classification algorithms and Structural Equation Modeling," *Computers & Industrial Engineering*, vol. 86, pp. 95–106, 2015.

[2] Gaiardelli, Paolo, Barbara Resta, Veronica Martinez, Roberto Pinto, and Pavel Albores. "A classification model for product-service offerings." *Journal of cleaner production* 66 pp: 507-519, 2014

[3] Lu, Ning, Hua Lin, Jie Lu, and Guangquan Zhang. "A customer churn prediction model in telecom industry using boosting." *Industrial Informatics, IEEE Transactions on* 10, no. 2, pp.1659-1665, 2014

[4] K. Coussement and D. V. Poel, "Churn Prediction in Subscription Services: An Application of Support Vector Machines while Comparing Two Parameter-Selection Techniques," *Expert Systems with Applications*, Vol. 34, no 1, pp. 313-327, 2008.

[5] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research*, Vol. 218, no 1, pp. 211-229,2012.

[6] W. J. Reinartz and V. Kumar, "The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration," *Journal of Marketing*, Vol. 67, no1, pp. 77-99,2013.

[7] P. Datta, B. Masand, D. R. Mani, and B. Li, "Automated Cellular Modeling and Prediction on a Large Scale," *Artificial Intelligence Review*, Vol. 14, no 6, pp. 485-502, 2000.

[8] D. Popović and B. D. Bašić, "Churn Prediction Model in Retail Banking Using Fuzzy C-Means Algorithm," *Informatica*, Vol. 33, no2, pp. 235-239,2009..

[9] C.-P. Wei and I. T. Chiu, "Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach," *Expert Systems with Applications*, Vol. 23, no2, pp. 103-112,2002.

[10] M. Owczarczuk, "Churn Models for Prepaid Customers in the Cellular Telecommunication Industry Using Large Data Marts," *Expert Systems with Applications*, Vol. 37, no 6, pp. 4710-4712,2010.

[11] J. Burez and D. V. Poel, "Handling Class Imbalance in Customer Churn Prediction," *Expert Systems with Applications*, Vol. 36, no 3, pp. 4626-4636,2009

[12] N. Kim, K.-H. Jung, Y. S. Kim, and J. Lee, "Uniformly Subsampled Ensemble (USE) for Churn Management: Theory and Implementation," *Expert Systems with Applications*, Vol. 39, no 15, pp. 11839-11845,2012

[13] 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*, Springer India pp. 681-689., 2016.

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

[15] 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

[16] 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.

[17] 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

[18] 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 : 165-17,2015

[19] 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.

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

.