INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 3)
Available online at: www.ijariit.com

Selective feature processing with k-Nearest Neighbor classification to predict credit card frauds

Simranjeet Kaur
brarajit18@gmail.com
Punjabi University, Patiala, Punjab

Sikander Singh Cheema
ersimran721@gmail.com
Punjabi University, Patiala, Punjab

ABSTRACT

The predictive analytics are being used in many applications across the globe ranging from financial risk to avalanche studies. In this paper, a new approach is designed to predict the credit card frauds. This approach utilizes the imbalanced feature correction methodology, which eventually reduces the levitation of the features towards one class. The proposed model is designed to filter the credit card data by analyzing the multiple factors to analyze and predict the fraudulent transactions. The proposed model utilizes the maximum-minimum scaling method to scale the quantitative variables on a 0-1 scale, after handling the missing values with column mean value. The SVM and KNN based classification method are used to predict the patterns for the credit card frauds. The experimental results have proved the proposed model based on SVM classification as the most efficient algorithm for the purpose of fraudulent pattern prediction. The SVM has been recorded with 99.94% (mean) of accuracy, which is slightly lower than KNN's 99.95% (mean). Also, KNN outperformed SVM on the basis of recall with (approx 91%) and F1 measure (approx 84%) against approx. 84.50% (recall) and approx 82.50% (F1 measure).

Keywords: *Predictive Analysis, Imputation, Feature Scaling, Imbalanced Features.*

1. INTRODUCTION

The financial risk assessment models are used to determine the losses in the business, which must be curbed in the near future. Afterward, when a certain database of various frauds is prepared, the predictive models help the organizations to predict the losses in advance by analyzing the user behavior. In this paper, the financial risk analysis model is designed to predict the losses from the credit card business. In this paper, the imbalanced features are used for predictive for a fact check of the target data on the transaction database, which eventually describes the characteristics of the credit card users by analyzing its transaction history. The data mining methods are used to determine the significance of the various features, which are processed by the means of imputation, scaling, denoising, outlier elimination and other practices.

The data mining approaches are used to discover the matching patterns in the larger volumes of data, which is almost impossible to discover manually. A number of software techniques are used to design the solutions for predictive analytics, which are used to store and analyze the transactional data. The database systems are being improved with every version and every past decade. There are several new techniques offered in the past years, which are nowadays being used to work with larger data volumes. Each of the analytical programs in the data mining uses the three-step model as per described in the following:

- Collection of data and building a database
- Management of the transactional data in the online transactional processing (OLTP) database
- Preparing a fixed or dynamic interval based online analytical processing (OLAP) database for analytical studies

In the case of credit card businesses, the customer transactional data is procured in the target databases, which is further used to determine the customer behavior. The customer behavior helps the organizations to match the certain patterns in the fraudulent transactions, which are concretely used to determine the future frauds to curb the financial losses. The credit card data is very sensitive data and must be processed with precision. A small change in the analytical model may produce the misguided results, which is clearly taken under observation in the proposed model.

2. LITERATURE REVIEW

Kulkarni Pallavi et. al. [1] has worked towards the realization of the credit card analytics models on the unbalanced transactional database and applied the regression model to determine the irregularities in the target data. The authors have used a certain set of

© 2018, www.IJARIIT.com All Rights Reserved

rules to correct the feature distribution, either by fixed lineage method, or deterministic dynamic method to repair the target feature. Bahnsen, Alejandro Correa et. al. [2] has developed the model to process the features in the way, where it is capable of processing the features according to the need of machine learning algorithms. The authors have used their experience on a variety of databases to determine the most effective set of practices to create a process the features under the automatic feature engineering method. Dal Pozzolo, Andrea et. al. [3] has worked on the vision of financial analyst to create a model for determination of the credit card frauds. The security threats over the credit card OLAP and OLTP databases are also studied in this paper. Prakash A. et. al. [6] worked towards the detection of credit card frauds by employing the semi-HMMs (hidden markov models) to determine the hidden patterns in the data by analyzing the multiple combinations of the input data, which is credit card data in this case to prepare the processing knowledge. Seeja, K. R. et. al. [9] worked towards the processing of financial transactional data of credit card organizations by the means of itemset mining. The itemset mining methods analyze the input data in the unique patterns to determine the anonymous and hidden patterns in the given data.

3. EXPERIMENTAL DESIGN

The existing model is observed with multiple problems in the classification of credit card frauds, which are evident from the errors and overall accuracy based parameters. The proposed model is designed to eliminate the classification problems associated with the existing credit card fraud detection model. The classification problems are primarily caused by the imbalanced features in the datasets. The imbalanced features are predominantly not normally distributed features, which is an indicator showing their leveraging behavior towards a particular category. These features must be processed properly to align their distribution in order to improve the classification decision on target feature.

The feature imputation is one the best method to scale the features, which eventually transform the features to a certain range (eg. -1 to 1, or 0 to 1). This process is known as feature scaling, and its eventually known to reduce the feature bias up to certain extent varying from feature to feature.

After imputation, the logarithmic computation of the feature leverages the features around the central limit and reduces the imbalanced nature of the target feature. Also, the central limit theorem application reduces the feature bias by reducing the overall standard deviation of feature distribution.

Further, the application of supervised classification model on the processed feature data to create the classification model for a credit card is performed. The supervised classifier is used to create the data model, which is applied to the credit card data to discover the anomalies (possible fraud transactions) in the given data.

Algorithm 1: Supervised Classification Algorithm

1. Perform the acquisition of the credit card data from the online transactional database
2. Then, perform the preprocessing on the acquired features in order to leverage their quality by eliminating the outliers and less significant features.
3. Acquire the pre-assigned labels from the transactional data, which labels the data as fraud or normal according to its actual behavior.
4. Initialize the supervised classification model for predictive analysis based upon KNN classifier with a number of neighbors and metric.
5. Split the acquired data and labels into testing and training with the permutation-based method
6. Train the KNN based predictive analyzer model with training data and training labels
7. Test the performance of proposed KNN based credit card detection model with testing data
8. Obtain the observation of the classifier on testing data and match them against the original observations
9. Evaluate the performance of KNN based classifier in the form of multiple performance indicators
10. Return the performance indicator observation

4. RESULT ANALYSIS

The detection of the credit card frauds becomes very important nowadays to minimize the operative losses by detecting the frauds in future. The credit card transactions are firstly used to determine the unique patterns in the historical data for the already occurred frauds and non-frauds. The data prepared from the historical transactional data is used with the machine learning models to determine the credit card frauds in advance, even before their occurrence in most of the times or as soon as possible after its occurrence to nab the fraudulent people. These models are typically known as predictive analytical models and implemented with the machine learning algorithms, such as Logistic Regression, k-NN (k-nearest neighbor), deep learning, random forest, decision tree and SVM (support vector machine). In this paper, the logistic regression aka Maximum entropy (MaxEnt) is used along with kNN and SVM classifiers, the results of which are shown in the following table.

© 2018, www.IJARIIT.com All Rights Reserved

Table 1: Comparison of the machine learning algorithms based upon KPIs (key performance indicators)

Supervised Classification Algorithm	Precision	Recall	F1 Measure	Accuracy
k-NN	77.5	**91.6**	**83.9**	**99.95**
MaxEnt	62.9	87.2	72.9	99.92
SVM	**80.8**	84.4	82.4	99.94

Table 2: Comparison of the machine learning algorithms based upon errors as the KPIs

Index	Analysis of Existing Model		Analysis of Proposed Model	
	Normalized Error	Standard Error	Normalized Error	Standard Error
1	0.98	0.26	0.183	0.053
2	1.04	0.25	0.157	0.048
3	1.13	0.22	0.132	0.043
4	0.97	0.27	0.153	0.045
5	0.99	0.29	0.129	0.040
6	1.17	0.22	0.212	0.073
7	0.94	0.28	0.138	0.043
8	0.95	0.28	0.164	0.050
9	1.01	0.20	0.156	0.050
10	1.06	0.24	0.188	0.075
Average Error	1.03	0.25	0.161	0.051

The KNN classification model is observed with the lowest error and highest accuracy after the synthesis of table 1 and 2. The KNN model is recorded for various performance indicators, which includes its best outputs involving f1-measure (approx 84%), recall (91.6%) and over prediction accuracy at approx 99.95%. In the only analytical exception, the precision value based result of SVM classifier is higher than kNN, where the different of nearly 3% is reported. The precision value works on the basis of false-positive cases, which can be again verified to discover the customer's loyalty and record. On the other hand, recall value is most important, because it works with false negative value, which reflects the chances of a fraud being ignored. This case becomes most important in the credit card scenario, hence, the KNN is preferred over SVM in this paper.

The analysis of errors also proves the domination of k-NN in comparison with existing models, where k-NN is recorded with a normalized error of 0.161% against the whopping 1.03% of existing model. The standardized error 0.051% error is recorded for kNN against 0.25% of existing model. This describes the efficiency of the proposed credit card fraud detection model based upon k-NN classification.

Table 3: Comparison of machine learning algorithms based upon average values of KPIs

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

5. CONCLUSION

The proposed model is analyzed on the basis of various key performance indicators (KPIs), which involves the statistical parameters of precision, recall, overall accuracy, and F1-measure. The kNN observations of the performance indicators are 77.5% (precision), 91.6% (recall), 99.95% (overall accuracy) and 83.9% (F1-measure). The k-NN is observed the best performer among all of the parameters except the precision, which is a least important factor among all four KPIs. The 80.8% of precision is observed for SVM, which is slightly higher than kNN's 77.5%. The standard error of existing model is observed with a value of 0.25% against proposed model's 0.051%. Similarly, the normalized error of proposed model (0.16%) is certainly improved than existing model (approx 1%). This shows the efficiency of kNN classification with imbalanced feature correction. In the future, the credit card fraud detection model can be improved further by using the deep learning with convolutional features up to multiple levels.

© 2018, www.IJARIIT.com All Rights Reserved

6. 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 (2016): 134-142.
[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 (2014): 4915-4928.
[4] Halvaiee, Neda Soltani, and Mohammad Kazem Akbari. "A novel model for credit card fraud detection using Artificial Immune Systems." Applied Soft Computing 24 (2014): 40-49.
[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 (2015): 38-48.
[6] Prakash, A., and C. Chandrasekar. "An optimized multiple semi-hidden markov models for credit card fraud detection." Indian Journal of Science and Technology 8, no. 2 (2015): 165-171.
[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, pp. 677-685. Society for Industrial and Applied Mathematics, 2014.
[8] Zareapoor, Masoumeh, and Pourya Shamsolmoali. "Application of credit card fraud detection: Based on bagging ensemble classifier." Procedia Computer Science 48 (2015): 679-685.
[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).

© 2018, www.IJARIIT.com All Rights Reserved