Credit Card Fraud Detection using Non-Overlapped Risk based Bagging Ensemble (NRBE)

Conference Paper · December 2018

DOI: 10.1109/ICCIC.2017.8524418

2 authors:

Akila Somasundaram
National Institute of Technology Tiruchirappalli
7 PUBLICATIONS **12** CITATIONS
SEE PROFILE

U. Srinivasulu Reddy
National Institute of Technology Tiruchirappalli
21 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Credit Card Fraud Detection using Non-Overlapped Risk based Bagging Ensemble (NRBE)

S. Akila
Research Scholar, Department of Computer Applications,
National Institute of Technology, Trichy-620015
Email: akila29@gmail.com

U. Srinivasulu Reddy
Assistant Professor, Department of Computer Applications,
National Institute of Technology, Trichy-620015
Email: usreddy@nitt.edu

***Abstract*—Fraud due to credit card misuse costs consumers several billions of dollars annually. This is due to the huge usage levels and inability of the systems to automatically detect the anomalies. This paper analyzes the implicit nature of data with noise and imbalance and proposes a Non-overlapped Risk based Bagged Ensemble model (NRBE) to handle imbalance and noise contained in the credit card transactions. The bagging model has been enhanced in terms of a novel bag creation model and an effective risk based base learner. Non-overlapped bag creation generates training subsets to handle data imbalance and the risk based Naïve Bayes eliminates the issues arising due to noise. Experiments were conducted and comparisons were performed with existing state-of-the-art fraud detection models, which indicates that NRBE exhibits improved performances of 5% in terms of BCR and BER, 50% in terms of Recall and 2X to 2.5X times reduced cost.**

Keywords— Ensemble Modelling; Bagging; Credit card fraud detection; Risk based modelling

I. INTRODUCTION

Credit card transaction data tends to be large in number due to high usage levels of cashless transactions. Major factors aiding customers to use such models is its ease-of-use. However, these methods are highly vulnerable due to the usage of electronic information rather than actual information. Analysis of every transaction is required to solve this issue. However, this is not feasible for a human being, hence an automated fraud detection system is required to verify the authenticity of the transactions. These verifications are to be performed in real time, before the actual cash withdrawal occurs.

Several implicit factors are constituents of credit card transactions. They are noise, imbalance in data and concept drift. Imbalanced dataset is one in which one of the classes dominates the other classes by a huge margin [1]. The class that dominates is referred to as the majority class, while the others are referred to as the minority classes. When trained with imbalanced data, the prediction model gets biased. This is due to the large number of majority classes and very small number of minority classes. Such biased training leads to reduced performance levels during prediction. Data is considered to be noisy if an instance of one class is surrounded by the instances of other existing classes [2]. Concept drift is one where the data exhibits variations over time due to the changing behavioral patterns of users and is inherent in several domains dealing with customer transactions. Data elimination is the proposed solution to handle imbalance and noise. In the domain of credit card fraud detection, since the data is generated by a customer, it cannot be eliminated as such irrespective of its state, instead they are to be handled appropriately by the learning algorithm. Credit Card Fraud detection is actually a business problem that requires solutions which aid in making business decisions. Usefulness of business decisions are proportional to the cost in applying the decisions. Hence the predictions provided by a credit card fraud detection model must be able to provide cost effective solutions. Hence cost should also be considered as a metric during predictions.

This paper presents a Non-overlapped Risk based Bagged Ensemble (NRBE) to be operated upon data with high imbalance levels. The NRBE model was observed to effectively handle noisy entries and the data imbalance contained in the transaction data. Comparisons with existing state-of-the-art models exhibits the high performing nature of the NRBE model.

II. RELATED WORK

Credit card fraud detection has been researched since the beginning of digital transactions. However, it is still pondered due to the ever-changing customer behavior. Some of the more recent contributions to the domain of credit card fraud detection are discussed in this section.

Current orientation of researchers is towards cost-based modelling, due to the increased requirements for business oriented modelling. A metaheuristic based credit card fraud detection system was proposed by Gadi et al. [3]. This model uses Artificial Immune System (AIRS) as the base predictor and parameter fine-tuning is performed using Genetic Algorithm (GA). This is a cost-based model with its focus towards business goals. An extension of AIRS was presented by Halvaiee et al. [4]. The Artificial Immune System based Fraud Detection Model (AFDM) incorporates multiple contributions to the AIRS in terms of memory cell regeneration, an updating and scoring system for prediction and a modified distance function. Ghobadi et al. proposed an extension AFDM in [5]. This model presents a Cost Sensitive Neural Network (CSNN) that identifies the best network topology to be used for fraud detection. This model also incorporates cost effectiveness as one of the major factors for prediction. Fraud detection models proposed by Ottersten, Bahnsen, Aouada and Stojanovic are also relied towards developing cost sensitive classifiers. Contributions by Bahnsen includes Cost Sensitive Logistic Regression models [6], Cost Sensitive Decision Trees [7], Bayes Minimum Risk based fraud detection [8] and calibrated property based fraud detection [9].

978-1-5090-6621-6/17/$31.00 ©2017 IEEE

Feature creation is one of the major pre-processing functionalities used to enhance the predictability of models. Feature creation is performed by creating new attributes from the existing attributes in such a way that it enhances the effectiveness of the model. A pattern generation model proposed by Bahnsen et al. in [10] is used to generate additional spending patterns of customers to build their complete profiles. Pattern identification was performed using behaviour analysis using the von Mises distribution. Another feature creation based model utilizing a sliding window mechanism for generating features was proposed by Vlasselaer et al. [11]. Genetic Algorithm (GA) has also been used as a base classifier for fraud detection. A credit limiting technique that uses GA for identification of maximum credit allocation to a customer was proposed by Patel et al. in [12]. Other similar models include, a Genetic Algorithm (GA) and Scatter-Search (SS) incorporated model for fraud detection by Duman et al. in [13] and a GA based evolutionary model by Ramakalyani et al. [14]. Although these models are cost based, analysis on imbalance and noise are not performed.

Ensembles for solving complex problems has gained prominence. This is due to the ability of ensembles to handle huge amounts of data and to perform analysis from varied perspectives. A bagging based ensemble model proposed by Zareapoor et al. [15] uses decision trees as the base learners for performing predictions. Lin et al. [16] proposed a Random Forest based credit card fraud detection model. A Minimal Learning Machine based ensemble model (MLM) incorporating Neareset Neighbor was presented by Mesquita et al. [17]. This model utilizes Cubic Equation and it can support for classification and regression based predictions. An ensemble model that also performs data imputation in its pre-processing phase was presented by Conroy et al. [18]. AdaBoost is used as the base classifier and it also utilizes several low dimensional classifiers for the process of imputation. Zhang et al. [19] proposed a weighted model based on differential evolution algorithm for fraud detection. Although several such models exist in literature, not many are inclined towards cost based analysis, further most models do not concentrate on the implicit data imbalance and noise contained in credit card fraud detection. The proposed model concentrates on handling these issues effectively.

III. Non-Overlapped Risk based Bagging Ensemble (NRBE)

Fraud detection in credit card transactions has become highly complicated due to imbalance and noise intrinsically contained in the data. This work focusses on a non-overlapped risk based bagged ensemble (NRBE) for fraud detection in credit card transactions. The NRBE model is composed of three major phases; creation of Non-Overlapped Bags, base learner creation and the combiner phase. Conventional bagging models creates overlapping subsets of data by selecting 60% of training data (randomly) for each base learner. This process is performed to make sure that each base learner obtains data with different imbalance levels. This scheme operates well for data with low imbalance levels, however, data with very high imbalance levels might even ignore the minority classes during random selection. This leads to the base learner getting trained only on a single target class. In order to avoid this, the proposed model introduces non-overlapped bagging model that creates bags such that all bags contain samples of all representative classes. This aids in effective handling of data with huge imbalance levels. The major gain in using bagging models is that they can effectively handle data imbalance [20]. The base learners are formulated using Modified Risk based Naïve Bayes. Studies conducted by the authors [21] shows the efficiency of using Naïve Bayes on noisy data. The combiner aggregates the results using voting to provide the final predictions.

A. Ensemble Learning

Ensemble learning is an emerging paradigm that uses multiple learning models to perform predictions. It is based on the analysis that using multiple models on subsets of data and finally combining for the final prediction improve results to a large extent, in comparison to a single model. Bagging, boosting, stacking and bucket of models are some of the ensemble modelling techniques. According to literature boosting and bagging are the mostly used ensembles for predictions. Bagging generates several training models by providing each model with overlapped training data subsets for modelling. Every base learner creates an independent fully trained prediction model. All the models are made to predict the test data, where multiple predictions are obtained. Result aggregation is performed by the combiner, which produces a single prediction result. The proposed NRBE extends the regular bagging model and incorporates a non-overlapped bag creation method, modified Naïve Bayes as the training method and a voting combiner for prediction.

B. Non-Overlapped Bag Creation

Bag or training data-subset creation is the initial phase of bagging models. Usual bagging models selects 60% of the random data for bags. The proposed non-overlapped bag creation model modifies this approach by creating non-overlapping bags. However, creating completely non-overlapping bags would lead to data loss. Hence this model divides the minority and majority classes into two sections. Every bag is provided with a part of the majority class data and all of the minority class data. The majority class instance selection is done in such a way that overlaps between bags are eliminated.

Consider the training data $T= \{(x_1,c_1),(x_2,c_2),....(x_m, c_m)\}$ containing m instances. $x_1,x_2...x_m$ represents the attributes of the training data and $c_1,c_2...c_m$ represents their corresponding target class. The current work performs fraud detection, hence it considers the targets to be either fraud or legitimate, hence $c_x \in \{0,1\}$. Training data selection for the Non-Overlapped Bagged Ensemble model is obtained by combining the minority class instances and a part of the majority class instances. This is given by

$$T' = Min \cup Maj' \qquad \forall 1 \leq i \leq n \tag{1}$$

Where *Min* is the minority class instances and *Maj'* is obtained by equally splitting the majority class instances for all the n bags.

C. Risk based Base Learner

Bag creation is followed by application of the training data on each base learner. The proposed NRBE model uses Naïve

Bayes as the base learner. It is conventional to incorporate a weak learners as the base learners for bagging models. However, it was identified that Naïve Bayes exhibits highly effective results when used as a base learner for bagging.

Naïve Bayes performs predictions based on conditional probability. Let the attributes corresponding to the training instances be $At_1, At_2 \ldots At_m$, their corresponding values be, $av_1, av_2 \ldots av_m$ and C be the class to be predicted, where $C \in \{0,1\}$. According to Bayes' rule, this can be represented as

$$\frac{P(At_1=av_1 \Lambda \ldots \Lambda At_k=av_k|C=c)}{P(At_1=av_1 \Lambda \ldots \Lambda At_k=av_k)} P(C=c) \quad (2)$$

The base probability $P(C=c)$ and example probability $P(At_1 = av_1 \Lambda \ldots \Lambda At_k = av_k)$ are obtained from the training data and they do not change. Hence predictions depend on the numerator, which differs between training samples. Then, the probability of the instance j belonging to class x is given by

$$p_{jx} = P(C = x|At_1 = av_1 \Lambda \ldots \Lambda At_k = av_k) = \left(\prod_{j=1}^{k} p_{jx}\right) b_x/z \quad (3)$$

where z is the normalizing constant.

Risk associated with predicting a fraudulent transaction is given by

$$R_1 = C_{admin}(FP * P_{j0} + TP * P_{j1}) \quad (4)$$

where C_{admin} refers to the administrative cost spent for contacting the customer.

Risk associated with predicting a legitimate transaction is given by

$$R_0 = FN * C_{Amt} * P_{j1} \quad (5)$$

where C_{Amt} is the amount for which the transaction was performed. Prediction carrying the lowest risk value is selected as the final prediction for instance i of base learner l which is given by

$$Pred_{lj} = \begin{cases} 1 & if\ R_1 < R_0 \\ 0 & Otherwise \end{cases} \quad (6)$$

D. Voting Combiner

Voting [22] is the modus-operandi of bagged ensembles. Voting operates by selecting the prediction that has been provided by majority of the base learners in the bagging model. The maximum voted prediction for each instance is considered as the final prediction. The proposed NRBE model considers equal weights for all the base learners, hence it is not possible for a single base learner to influence the results in any manner. Final predictions for an instance x is given by

$$P(x) = argmax_{c \in C} \sum_{i=1}^{n} c = m_i(x) \quad (7)$$

Where x is the instance to be predicted, $C \in \{0, 1\}$, n is the number of bags/models and m_i is the prediction given by i^{th} model for instance x.

IV. Results and Discussion

Apache Spark is used as the base environment and the proposed NRBE model is implemented in PySpark. Brazilian Bank data was used to verify the efficiency of the proposed NRBE model. The Brazilian Bank dataset consists of 0.3 million instances and exhibits an imbalance ratio of 25.7. Comparisons were performed with AIRS [3] and CSNN [5] to validate the efficiency of the proposed model.

A. Experimental Analysis on Brazilian Bank Data

ROC curves representing TPR and FPR levels of AIRS, CSNN and NRBE is presented in figure 1. ROC is plotted with FPR levels in x-axis and TPR in y-axis. An ideal classifier exhibits low False Positive levels and high True Positive levels. From figure 1 it could be observed that the proposed NRBE classifier exhibits high TPR levels and moderate FPR levels. However, the curve of NRBE exhibits higher area under the curve, exhibiting its dominance over the other state-of-the-art models.

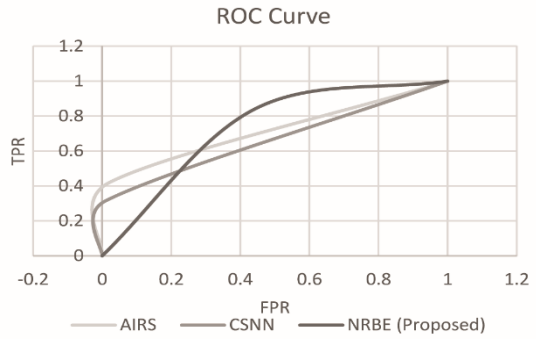

Fig 1. Receiver Operating Characteristic Curve - Comparison

Area-Under-the-Curve(AUC) signifies the area occupied by an ROC curve beginning from its contact with the origin (0,0) to its contact with the point (1,1). The model exhibiting highest area (approaching 1) is considered to be the best predictor. NRBE and the AIRS models exhibits the maximum AUC value (figure 2), proceeded by CSNN.

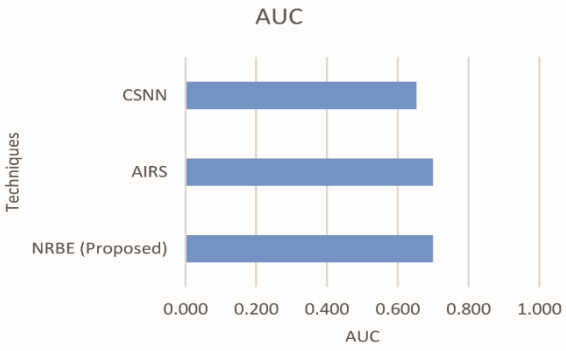

Fig 2. Comparison of Area Under the Curve (AUC)

Fraud Detection Rate (FDR) refers to the level of frauds identified by the models. A comparison of the FDR values is presented in figure 3. The proposed NRBE model is observed to exhibit very high FDR levels of 84.2% compared to other existing state-of-the-art models. NRBE model exhibits elevated fraud detection levels at the rate of 40% to 50% higher than AIRS and CSNN.

Additional performance measures such as Recall, Balanced Correction Rate (BCR), Balanced Error Rate (BER) and Cost were used for analysis and the results are presented in Table 1.

The best performances are shown in bold. The proposed NRBE model was found to exhibit similar results as that of AIRS in terms of Balanced Correction Rate and Balanced Error Rate, and is found to outperform all the other models in terms of recall and cost, making it the best performing model for credit card fraud detection.

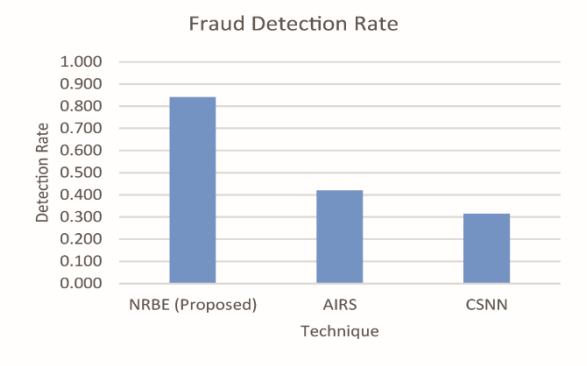

Fig 3. Comparison of Fraud Detection Rate

TABLE I. PERFORMANCE MEASURES

Detection Models	BCR	BER	Recall	Cost
NRBE	**0.70**	**0.30**	**0.84**	**9595**
AIRS	**0.70**	**0.30**	0.42	19216
CSNN	0.65	0.35	0.31	23082

V. CONCLUSION

Huge losses associated with frauds in credit card transactions makes fraud detection in this domain a major need for the highly interconnected world. Efficiency in predictions are usually hindered by the intrinsic properties such as noise and imbalance associated with the transaction data. This paper presents a Non-Overlapped Risk based Bagging Ensemble (NRBE) for detecting frauds in credit card transactions. The NRBE model imposes two major modifications to the conventional bagging approach. The bag creation section is enhanced to generate non-overlapped bags and the weak base learner is replaced with Risk based Naïve Bayes, to handle noise. Efficiency of the proposed NRBE model can be observed from the experimental results. It was observed that the proposed NRBE model exhibits effective and robust fraud detection levels at low cost, thus making the model appropriate for business decision making.

VI. ACKNOWLEDGMENT

The authors would like to thank DEITY for the financial support extended under Visvesvaraya Ph.D. scheme (NITT/RO/DEITY-Ph.D. Cont. grant/2015-16). The authors would like to acknowledge the infrastructure support provided by HPC Lab and Machine Learning & Big Data Analytics Lab, Dept of Computer Applications, NITT. The authors would also like to thank Dr. Manoel Fernando Gadi, Univ. of Sao Paulo, Brazil and Dr. Neda Solatani, Amirkabir University of Technology, Iran for providing the Brazilian Bank Dataset.

REFERENCES

[1] Tomašev, Nenad, and Dunja Mladenić,. “Class imbalance and the curse of minority hubs.” Knowledge-Based Systems, vol.53, 2013, pp.157-172.

[2] Napierała, K., Stefanowski, J., and Wilk, S.” Learning from imbalanced data in presence of noisy and borderline examples.” In Rough sets and current trends in computing, 2010, pp. 58-167.

[3] Gadi, Manoel Fernando Alonso, Xidi Wang, and Alair Pereira do Lago. "Credit card fraud detection with artificial immune system." International Conference on Artificial Immune Systems. Springer Berlin Heidelberg, 2008, pp. 119-131.

[4] Halvaiee, N.S; and Akbari, M.K. “A novel model for credit card fraud detection using Artificial Immune Systems”. Applied Soft Computing, Vol. 24, 2014, pp. 40-49.

[5] Ghobadi, Fahimeh, and Mohsen Rohani. "Cost sensitive modeling of credit card fraud using neural network strategy." Signal Processing and Intelligent Systems, International Conference of. IEEE, 2016, pp. 1-5.

[6] Bahnsen, A. C., Aouada, D., and Ottersten, B. “Example-dependent cost-sensitive logistic regression for credit scoring.” In Machine Learning and Applications,13th International Conference, 2014, pp. 263-269.

[7] Bahnsen, A. C., Aouada, D., and Ottersten, B. “Example-dependent cost-sensitive decision trees.” Expert Systems with Applications, Vol. 42(19), 2015, pp. 6609-6619.

[8] Bahnsen, A. C., Stojanovic, A., Aouada, D., and Ottersten, B. “Cost sensitive credit card fraud detection using Bayes minimum risk.” In Machine Learning and Applications (ICMLA), 12th International Conference, Vol. 1, 2013, pp. 333-338.

[9] Bahnsen, A. C., Stojanovic, A., Aouada, D., and Ottersten, B. “Improving credit card fraud detection with calibrated probabilities.” In Proceedings of SIAM International Conference on Data Mining, 2014, pp. 677-685.

[10] Bahnsen, A. C., Aouada, D., Stojanovic, A., and Ottersten, B. “Feature engineering strategies for credit card fraud detection.” Expert Systems with Applications, Vol. 51, 2016, pp. 134-142.

[11] Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., and Baesens, B. “APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions.” Decision Support Systems, Vol. 75, 2015, pp. 38-48.

[12] Patel, Rinky D., & Dheeraj Kumar Singh., “Credit card fraud detection & prevention of fraud using genetic algorithm”. *International Journal of Soft Computing and Engineering*, Vol. 2 (6), 2013, pp. 292–294.

[13] Duman, E., & Ozcelik, M. H, “Detecting credit card fraud by genetic algorithm and scatter search”. *Expert Systems with Applications*, Vol. *38*(10), 2011, pp. 13057-13063.

[14] RamaKalyani, K., & UmaDevi, D., “Fraud detection of credit card payment system by genetic algorithm”. *International Journal of Scientific & Engineering Research*, Vol. *3*(7), 2012, pp. 1-6.

[15] Zareapoor, M., & Shamsolmoali, P., “Application of credit card fraud detection: Based on bagging ensemble classifier”. *Procedia Computer Science*, Vol. 48, 2015, pp. 679-685.

[16] Lin, L., Wang, F., Xie, X., & Zhong, S.,. “Random forests-based extreme learning machine ensemble for multi-regime time series prediction”. *Expert Systems with Applications,* Vol. *83*,2017, pp 164-176.

[17] Mesquita, D. P., Gomes, J. P., & Junior, A. H. S, “ Ensemble of Efficient Minimal Learning Machines for Classification and Regression”. *Neural Processing Letters*, 2017, pp 1-16.

[18] Conroy, B., Eshelman, L., Potes, C., & Xu-Wilson,. “A dynamic ensemble approach to robust classification in the presence of missing data”. *Machine Learning*, 2016, Vol. *102*(3), pp 443-463.

[19] Zhang, Y., Liu, B., Cai, J., & Zhang, S, “Ensemble weighted extreme learning machine for imbalanced data classification based on differential evolution”. *Neural Computing and Applications*, 2016, 1-9.

[20] S. Akila, and U. Srinivasulu Reddy, “Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data”, Proceedings of ICRECT, 2016, pp. 28-34.

[21] S. Akila and Srinivasulu Reddy U, “Modelling a Stable Classifier for Handling Large Scale Data with Noise and Imbalance”, IEEE International Conference on Computational Intelligence in Data Science, 2017.

[22] Bauer, Eric, and Ron Kohavi. "An empirical comparison of voting classification algorithms: Bag-ging, boosting, and variants." Machine learning, Vol. 36, 1999, pp. 105-139.