IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-12, Dec 2018 E-ISSN: 2347-2693

Credit Card Fraud Detection using Feature Augmentation based Boosted Ensemble (FABE)

V. Sobanadevi[1*], G. Ravi[2]

[1]Department of Computer Science, Jamal Mohamed College, Trichy, India
[2]Department of Computer Science, Jamal Mohamed College, Trichy, India

[*]*Corresponding Author: ammu.soba1987@gmail.com, ravi_govindaraman@yahoo.com*

Available online at: www.ijcseonline.org

Accepted: 20/Dec/2018, Published: 31/Dec/2018

***Abstract*—** Fraud detection in credit card transactions have become mandatory for the financial services industry due to the huge levels of automations observed in the industry. This work presents a Feature Augmentation based Boosted Ensemble (FABE) for credit card fraud detection on huge data. The proposed model integrates two major components; feature augmentation and ensemble creation. Feature augmentation phase performs feature reduction, feature transformation and feature engineering. Feature reduction aids in effective elimination of unnecessary features, while feature transformation and feature engineering aids in creation of new features that can aid in better predictions. The ensemble creation phase models a boosted ensemble using Decision Trees. Multiple training data bags are created, and multiple base learners are created. The learner with highest weight and lowest error levels is iteratively modelled and used as the final learner. Experiments were performed and comparisons with existing models in literature exhibit the high-performance levels of the proposed FABE model.

***Keywords*—** Credit card fraud detection; Ensemble model; Feature Augmentation; Feature Reduction; Feature Engineering; Boosting

I. INTRODUCTION

Big data has become a vital part of the current scenario due to the increase adoption levels of online based systems. Such huge increase in the amount of data poses several unique challenges [1] in fields that require analysis of such huge data. Most real-time data processing applications are currently facing this issue, hence processing Big Data has become a domain in itself. Credit Card Fraud Detection (CCFD) is one of the major use case and has a huge requirement for Big Data processing models that operate in real-time [2]. The huge volume and velocity at which the transactions are generated poses a huge challenge for the prediction models. Significance is to be provided for all transactions irrespective of the transaction value. Both high and low valued transactions are to be provided with the same significance, as fraudsters usually check with the low valued transactions before moving to high valued transactions [3]. Several other issues form major components of credit card transaction data. They are data imbalance and noise [4]. Further, the credit card transactions, being dictated by human behavior, are also affected by concept drift [5].

The problem of anomaly detection in credit card transaction is usually modelled as a supervised learning problem. Machine learning has been the go-to solution for such decision-making problems. The process of classification identifies rules that partition data into sub-sections that involves decisions, which in turn divides them and finally completes by predicting any one of the existing classes [6]. The credit card fraud detection problem can be modelled in several ways even in the supervised context. The two broad categories are models specifically build for the data [7] and models extending generic models [8,9] to perform classification. However, in the current context such simple extensions were found to be insufficient. Currently the data is more complex, as human interaction is the only process that can generate a credit card transaction and human interactions tend to be complex. The generated transactions tend to get more complex, as more customers are involved. Hence simple and single machine learning models are not sufficient to perform this process. It becomes mandatory to move towards techniques that utilizes multiple or more complex models like ensembles or deep learning systems.

This work proposes an ensemble model that can effectively aid in credit card fraud detection on huge transaction data. The proposed model is composed of two major components. The feature augmentation component is used to effectively reduce unnecessary features and include all the necessary features. The ensemble modelling component creates a boosted ensemble for the detection process. Experiments show improvements in performances both in terms of predicting anomalies and normal data.

This paper is structured as follows: Section 2 presents the related literature for credit card fraud detection, section 3 presents the FABE model, section 4 presents the results and compares it with existing model and section 5 presents conclusion and further research directions.

II. RELATED WORKS

Anomaly detection in credit card transactions has witnessed a huge change in the processing methodology in recent years due to the increase in the amount of data being processed. This section discusses some of the most recent works in this domain.

A subsampling based anomaly detection model, specifically designed for anomaly detection in Big data was proposed by Vaughan [10]. This is a stochastic stage-wise operating model that uses data clustering and elimination as a part of the operating component. An ensemble based model called DeepBalance was proposed by Xenopoulos [11]. DeepBalance is an ensemble model, integrated with bootstrapping and feature selection mechanisms. The model is a deep belief network, constructed with multiple Restricted Boltzmann Machines (RBM) in its architecture. The deep belief networks are used as the base learners due to their ability to handle complex data models. However, the computational complexity of the model is very high, as DeepBalance is a secondary ensemble created using a network model. A Bayesian Network Classifier (BNC) model developed for customized fraud detection in credit card transaction data was proposed by Alex et al. [12]. A hyper heuristic evolutionary algorithm was used to create the classifier model. This process is automatic, hence it aids in effective fine-tuning of the model in correspondence with the dataset being used for the classification process. Another major advantage of this model is that it also aids in cost-sensitive predictions. Ensemble modelling has become a standard method for handling large and complex data. Ensemble based models that effectively handle imbalance, noise and concept drift in credit card transactions include Risk Induced Bayesian Inference Bagging (RIBIB) by Akila et al. [13]. This method creates a homogeneous ensemble and proposes a new bagging methodology to effectively detect frauds in credit card transactions. Extensions of this model includes a parallelized model that has been specifically created to handle data imbalance and concept drift [14] and a risk based model (NRBE) [15].

Active learning is another major area that witnesses a large number of works in the domain of credit card fraud detection. The initial approach to active learning was analyzed by Fan et al. [16]. This method also analyzed the presence of concept drift in the credit card domain and also views active learning as one of the solutions for concept drift. A model that analyses active learning methods that uses streaming to provide real-time results was provided by Carcillo et al. [17]. This work highlights the existence of exploration and exploitation as the major trade-offs that aids in effective learning. Another active learning model specifically designed for large scale data was proposed by Pichara et al. [18]. This method was observed to exhibit good computational efficiencies, enabling effective predictions. Other domains where anomaly detection models can be used include outlier detection [24] and insurance fraud detection [25].

III. FEATURE AUGMENTATION BASED BOOSTED ENSEMBLE (FABE)

Credit card fraud detection is a process that requires a detailed analysis of the input data and to effectively detect frauds in the transactions. Credit card transactions are usually machine generated data, hence they are composed of several features identifying the customer, POS terminal, location and the type of transaction. This sometimes leads to several features that might not be entirely useful for the prediction process. Further, the data generated for the transactions is triggered by a customer, hence human cognition and behavior plays a vital role in determining the type of the transaction being generated. This work proposes a Feature Augmentation based Boosted Ensemble (FABE) that aims to perform appropriate feature reduction and engineering and also provides an ensemble model, that can effectively handle the intrinsic complexities associated with the human generated data.

A. Feature Augmentation

Feature Augmentation is the process of analyzing the data and identifying components that can be effectively utilized for best results when applied on a machine learning algorithm. Feature augmentation also deals with eliminating and transforming existing data to create a high-quality training data for the machine learning algorithm [19]. The major advantage of feature augmentation is that it avoids the model from learning irrelevant patterns, which in turn avoids biased results. This work proposes an initial phase that performs feature augmentation in terms of dimensionality reduction and feature engineering.

B. Dimensionality Reduction

Dimensionality reduction is the process of eliminating features from the data. The major challenge in this phase is to identify the features that are to be eliminated. Credit card data is machine generated and hence contains all the information required to identify or pin-point the particular transaction. Not all features contained in the data tends to be useful during the prediction process. Some features might just be identifiers, while others might represent static content. Such data are not required and can be eliminated during the prediction process. Such features exhibit zero influence in correspondence with the prediction class [20]. However, the

major issues is due to the features that exhibits negative influence over the prediction classes. Such features must be eliminated to avoid prediction bias and to obtain improved results [21]. This work uses a correlation-based feature selection model, that identifies the correlation corresponding to each feature. The features exhibiting low correlation levels are eliminated for effective results.

C. Feature Engineering

Feature engineering refers to the process of transforming existing features into more usable ones and also the process of creating additional features from existing features [22]. The major advantage of feature engineering is to aid the prediction model to create accurate decision rules that aids in effective predictions. Machine learning models usually require numerical features to perform predictions. Categorical features that contain high correlation levels with the class variable are transformed to numerical data to enable effective prediction. Categorical data is generally of two types; data that can be ordered and data that cannot be ordered. Categorical content that can be ordered can be numerically encoded by assigning increasing values corresponding to its order level. Features that cannot be ordered is transformed using one-hot-encoding method. One-hot encoding is the process of converting categorical variables into numeric formats. The values of categorical variables are converted to features and are encoded according to their occurrence in the data instance. Although this leads to an increase in the number of features in the data, it has been mathematically proven that this model results in improved prediction efficiency, if applied appropriately.

D. Boosted Ensemble Creation

The previous phase results in modifying and preparing the data for the machine learning model. This section details the construction of the ensemble model for credit card fraud detection. This work uses a boosting ensemble as a learning model.

Boosting is a machine learning ensemble that creates a strong learner using multiple weak learners so as to primarily reduce the bias and variance. Weak learner is a classifier model that exhibits only slight correlation with the prediction variable and performs better than random guessing. This work utilizes a Decision Tree classifier model as the base learner.

Let *DT(x)* refer to the decision tree algorithm (weak learner) used as the base learner for the boosting process. Let *y'* be the predictions from the decision tree classifier, it is given by

$$y' = DT(x)$$

As a weak learner is used for prediction, the process tends to contain errors in prediction. Let *e* be the errors in prediction. The errors can be obtained by

$$e = y' - y$$

Where *y* and *y'* are the actual results and the predicted results respectively. The next phase is to integrate the error into the prediction process, such that the errors are eliminated during the prediction process. This is given by

$$y'' = DT(x) + e$$

However, this does not result in complete elimination of errors, instead, this process might result in a different set of errors. Hence the model is made to train and predict on the same training and testing dataset. The new errors obtained are given by

$$e' = y - y''$$

The next iteration integrates this error component to perform the next level training. This process of error identification and error integration is iteratively performed until the error level reaches an acceptable limit.

The proposed bagging model modifies this single iterative process such that multiple classifier models are created instead of a single model. This results in improved performances and huge reduction in bias and variance.

A 10-fold cross validation dataset is created using the actual training data. Each data group is boosted independently to obtain the trained model. Every iteration results in incorporating weights for the classifier model. This is given by

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x))$$

Where f_m is the m^{th} classifier model and θ_m is the weight assigned for the classifier. Every classifier is assigned a weight according to its prediction efficiency. The best classifier with highest weight and lowest error levels is used for the final prediction process.

IV. RESULTS AND DISCUSSION

The experiments were performed on BankSim [23] dataset and the proposed model is implemented using Python. BankSim is an agent-based simulator developed by Lopez et al. It is a simulator model that generates bank payment transactions and is based on the transactional data obtained

from a bank in Spain. BankSim is a benchmark data that has been calibrated extensively to match the real time distributions of the transaction data. The structure of BankSim is shown in table 1.

Table 1: BankSim Data Description

Feature	Type	Unique Values
step	Numerical (Count Parameter)	-
customer	Text	4112
age	Categorical	8
gender	Categorical	4
zipcodeOri	Categorical	1
merchant	Categorical	50
zipMerchant	Categorical	1
category	Categorical	15
amount	Numerical	-
fraud	Numerical	2

Comparisons were performed with the Big Data based Fraud Detection model proposed by Vaughan et al. [10]. Performances were measured in terms of True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR) and False Negative Rate (FNR). A comparison of the true prediction levels of the models is shown in figure 1. True prediction levels correspond to the TPR and TNR values. TPR represents the prediction level of the model in terms of the anomalous class and TNR represents the prediction level of the model in terms of the normal class. It could be observed that the proposed FABE model exhibits effective prediction both in terms of TPR and TNR levels exhibiting the high prediction efficiency of the model.

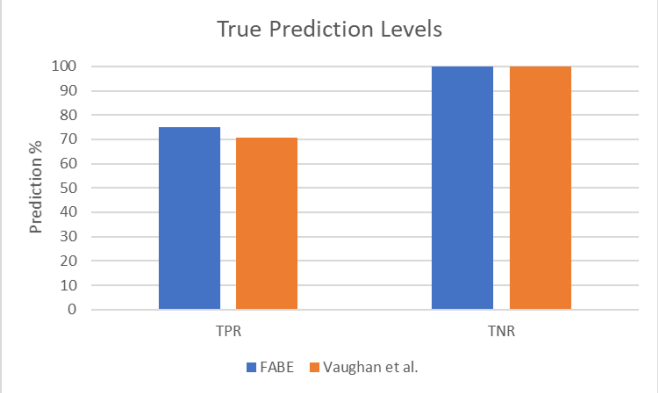

Figure 1. Comparison of True Prediction Levels

A comparison of the false prediction levels of the models is shown in figure 2. False prediction levels correspond to the FPR and FNR values. FPR corresponds to false alarm and FNR corresponds to predicting anomalous transactions as normal transactions. Lower FPR and FNR values exhibit effective models. It could be observed that the proposed FABE model exhibits equal FPR levels and a reduction of 4% in the FNR levels.

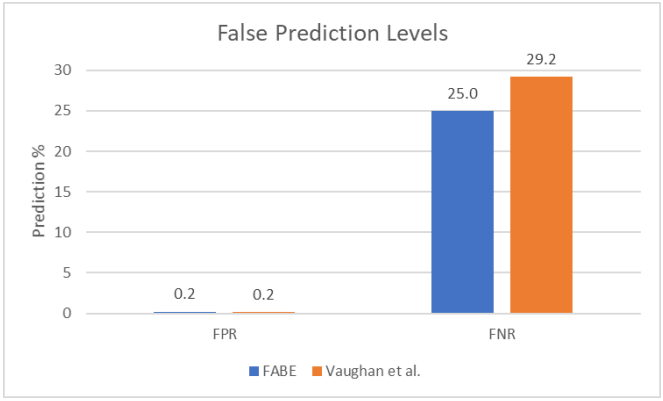

Figure 2. Comparison of False Prediction Levels

A comparison of the ROC curves is shown in figure 3. ROC is plotted by FPR in the x-axis and TPR in y-axis. The models are expected to exhibit low FPR values and high TPR levels. The model representing curve that dominates the ROC space is considered to be the best model. It could be observed that the proposed FABE model dominates the ROC space, exhibiting good prediction efficiency levels.

A tabulated view of the performance metrics is shown in table 2. Accuracy, F-Measure and AUC are aggregated measures, while others correspond to prediction efficiency of a single class. Values for the metrics range between 0 and 1. It could be observed that the FABE model exhibits effective prediction levels and also very good aggregated prediction levels.

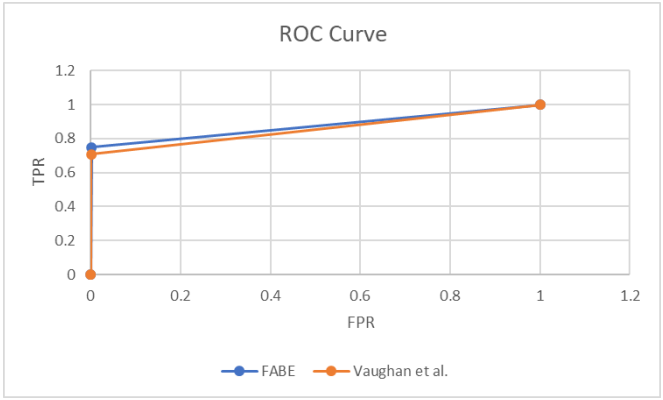

Figure 3. Comparison of ROC Values

Table 2: Performance Metrics

Metrics	Values
FPR	0.002
TPR	0.75
TNR	0.998

FNR	0.25
Recall	0.75
Precision	0.782
Accuracy	0.99
F-Measure	0.88
AUC	0.87

V. CONCLUSION

Fast and accurate detection of credit card frauds is a mandatory component due to the huge usage levels and high loss levels associated with the domain. This paper presents a fast and effective model, FABE for detecting frauds. The model is composed of a feature augmentation phase, that is used to perform dimensionality reduction, hence aids in reduction in the data size, resulting in faster processing. Feature engineering is also incorporated into the model to create and transform features that can enable more accurate predictions. The boosted ensemble model is created to ensure reduced errors in the prediction process. Experiments were performed on benchmark and real-time data, BankSim. Comparison results indicate improved performances by FABE. The major downside of this model is that, although the model exhibits comparatively high performance, the TPR levels still show scope for improvements. Future works will deal with improving to model to achieve better true prediction levels.

REFERENCES

[1] J. Fan, F. Han, H. Liu. "*Challenges of Big Data Analysis*", National science review, Vol.**1**, pp.**293-314, 2014**

[2] R. J.Bolton, D.J. Hand, "*Statistical fraud detection: A review",*. Statistical Science, Vol. **17**, Issue **3**, pp. **235–249, 2002**

[3] A. D. Pozzolo, O. Caelen, Y. Le Borgne, S. Waterschoot, G. Bontempi, "*Learned lessons in credit card fraud detection from a practitioner perspective*", Expert Systems with Applications, Vol. **41**, Issue **10**, pp. **4915– 4928**, **2014**

[4] A. Somasundaram, U.S. Reddy. "Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data", Proceedin,gs of ICRECT 16, pp. **28-34**, **2016**

[5] A. Somasundaram, U.S. Reddy. "Modelling a stable classifier for handling large scale data with noise and imbalance", In Proceedings of International Conference on Computational Intelligence in Data Science (ICCIDS-17), pp. **1-6**, **2017**.

[6] R. Akbani, S. Kwek, N. Japkowicz, "*Applying support vector machines to imbalanced datasets*," Machine Learning: ECML 2004, pp. **39–50, 2004.**

[7] N. S. Halvaiee, M. K. Akbari, "*A novel model for credit card fraud detection using artificial immune systems*", Appl. Soft Comput. Vol. **24**, pp. **40–49**, **2014**

[8] A.B. Hens, M.K. Tiwari, "*Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method*", Expert Syst. Appl. Vol. **39**, Issue **8**, pp. **6774–6781**, 2012

[9] A. Somaundaram., U.S. Reddy, "Risk based bagged ensemble (RBE) for credit card fraud detection." In Proceedings of International Conference on Inventive Computing and Informatics (ICICI-17), pp. **670-674**, **2017**

[10] G. Vaughan, "*Efficient big data model selection with applications to fraud detection*", International Journal of Forecasting, **2018**

[11] P. Xenopoulos, "Introducing DeepBalance: Random deep belief network ensembles to address class imbalance", arXiv preprint arXiv:1709.10056, **2017**

[12] A. G. de Sá, A. C. Pereira, G.L. Pappa, "*A customized classification algorithm for credit card fraud detection*", Engineering Applications of Artificial Intelligence, Vol. **72**, pp. **21-29**, **2018**

[13] A.Somasundaram, and U.S. Reddy, "*Cost Sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for Credit Card Fraud Detection*", Journal of Computational Science, **2018**

[14] A. Somasundaram, U. S. Reddy, "*Parallel and incremental credit card fraud detection model to handle concept drift and data imbalance*", Neural Computing and Applications, pp. **1-12**, **2018**

[15] A. Somasundaram., U.S. Reddy, "*Credit Card Fraud Detection Using Non-Overlapped Risk Based Bagging Ensemble (NRBE)*" In Proceedings of IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. **1-4, 2017**

[16] W. Fan, Y. A. Huang, H. Wang,P. S. Yu, "*Active mining of data streams*", In Proceedings of the 2004 SIAM International Conference on Data Mining, pp. **457–461**. **SIAM** , **2004**

[17] F. Carcillo, Y. L. Borgne, O. Caelen,G. Bontempi, "*Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization*", International Journal of Data Science and Analytics, pp.**1-16**, **2018**.

[18] K. Pichara, A. Soto, A. Araneda, "*Detection of anomalies in large datasets using an active learning scheme based on dirichlet distributions*", In Proceedings of Ibero-American Conference on Artificial Intelligence, pp. **163–172**, **2008**

[19] J. Fan, Y. Feng, J. Jiang, X. Tong, "*Feature Augmentation via Nonparametrics and Selection (FANS) in high-dimensional classification*". Journal of the American Statistical Association, Vol. **111**, Issue. **513**, pp.**275-287**, **2016**

[20] S. T. Roweis, L.K. Saul, "*Nonlinear dimensionality reduction by locally linear embedding*", Science, Vol.**290**, Issue.**5500**, pp.**2323-2326**, **2000**

[21] J. B. Tenenbaum, V.D. Silva,J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction", Science, Vol. **290**, Issue. **5500,** pp.**2319-2323, 2000**

[22] C. R. Turner, A. Fuggetta, L. Lavazza, A. L. Wolf, "*A conceptual basis for feature engineering*", Journal of Systems and Software, Vol. **49**, Issue.**1**, pp.**3-15,** ***1999***

[23] L. Lopez, E. Alonso, A. Stefan, *"Banksim: A bank payments simulator for fraud detection research",* In proceedings of 26th European Modeling and Simulation Symposium, EMSS 2014, pp. **144–152**, **France**, **2014**

[24] V. Jain, "*Outlier Detection Based on Clustering Over Sensed Data Using Hadoop*", International Journal of Scientific Research in Computer Science and Engineering, Vol.**1**, Issue.**2**, pp.**45-50, 2013**

[25] Namrata Ghuse, Pranali Pawar, Amol Potgantwar, "*An Improved Approch For Fraud Detection In Health Insurance Using Data Mining Techniques*", International Journal of Scientific Research in Network Security and Communication, Vol.**5**, Issue.**3**, pp.**27-33**, **2017**

Authors Profile

Mrs. V. Sobanadevi is a research scholar working in the field of big data analytics. She has done her M.Phil at Jamal Mohamed college in the area for network security and has done her M.sc(IT) in SRM college Chennai. Her research area is based on Machine Intelligence based Zero Event Anomaly Detection in Big Data using Human Guided Interactive Visualization.

Dr. G. Ravi is working as associate professor and Head in Department of computer science, Jamal Mohamed College (Autonomous), Tiruchirapalli, Tamil Nadu, India. He has more than 31 years of experience in teaching field. His areas of interest include Artificial Intelligence, Network Management, Big Data Analytics and Cloud computing. His current area of research is wireless sensor Network.

