Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/27/20230146

Research on improvements of fraud detection system: basing on improved machine learning algorithms

Zhiding Zhang

School of Economics, Shanghai University, Shanghai, China

Corresponding author: zhangzhiding@shu.edu.cn

Abstract. Nowadays, commercial fraud behaviors commonly occur in many industries. However, due to obstacles like concept drift, imbalanced dataset and uneven distribution of fraud entries, Fraud Detection System (FDS) fails to identify such behaviors. Among the problems mentioned above, most research focus on dealing with skewed dataset. This paper first presents common application scenarios of FDS which consist of credit card fraud, insurance fraud and supply chain fraud. Then, this study introduces five representative methods in dealing with problems mentioned above, which are K Nearest Neighbors-Synthetic Minority Oversampling Technique-Long Short-term Memory Networks (kNN-SMOTE-LSTM), Generative Adversarial Nets-AdaBoost-Decision tree (GAN-AdaBoost-DT), Wasserstein GAN-Kernel Density Estimation-Gradient Boosting DT (WGAN-KDE-GBDT), Time-LSTM (TLSTM) and Adaptive Synthetic Sampling-Sequential Forward Selection-Random Forest (ADASYN-SFS-RF). KNN-SMOTE-LSTM adopts KNN as an identifying classifier so as to only retain true samples. GAN-AdaBoost-DT generates new samples without referring to real transactions. WGAN-KDE-GBDT uses Wasserstein Distance as distance measurement instead, and thus improves training speed and guarantees successful generation. TLSTM tires to consider the weights of different time intervals and measures the similarity between the simulated behavior and the genuine behavior. ADASYN-SFS-RF employs SFS algorithm, basing on RF, to only reserve optimal subsets of features. Finally, result metrics prove that those improved algorithms do improve the overall performance of FDS, even if with limitations at some indicators.

Keywords: fraud detection system, imbalanced dataset, machine learning

1. Introduction

Commercial fraud generally consists of three parts, namely concealing real situations, notifying false information and inducing customers to act according to their false judgements. Nowadays, fraud behaviors involve many aspects, credit card fraud, insurance fraud, supply chain fraud, etc. Fraud Detection System (FDS) is one way to identify such behaviors in advance through machine learning algorithm by studying the characteristics of fraudulent dataset [1]. Nevertheless, the performance of FDS is obstructed due to obstacles like concept drift, supports real time detection, skewed distribution, large amount of data, etc. [2]. Therefore, how to solve these issues deserves more attention.

Solving the issue related to skewed distribution problem is important in the task of FDS. One possible solution to deal with imbalanced dataset is to propose a new oversampling method based on the K-Means algorithm and the genetic algorithm. More specifically, this method is conducted by first using

DOI: 10.54254/2755-2721/27/20230146

the K-Means to split the minority class, which are fraudulent entries, and then generates new samples within each cluster based on genetic algorithms [3]. Unlike oversampling and undersampling methods which could cause information loss or overfitting, since new samples are generated from every cluster, they are more representative. Another way is to refine Synthetic Minority Oversampling Technique (SMOTE) method. For instance, Ju et al. employ K-Nearest Neighbor (kNN) classifier to filter noise from samples generated by SMOTE so as to improve accuracy [4]. And Lin et al. combine Adaptive Synthetic Sampling Approach (ADASYN) and Optimization of Decreasing Reduction (ODR) with Support Vector Machine (SVM) to cope with SMOTE limitation [5].

Despite the recent progress in dealing with skewed distribution, choosing appropriate detecting algorithm is also challenging. Among the papers related to this issue, the most popular ones are Decision Tree (DT), Long Short-term Memory Networks (LSTM), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and SVM. For instance, Mo et al. improve the performance of the DT based AdaBoost model [6]. Lin et al. come up with ODR-ADASYN-SVM model to predict extreme finance risk [5]. Also, Cao et al. try TLSTM to predict patients' medical behavior due to the uneven distribution of treatment time [7]. Chen et al. uses XGBoost in identifying fraudulent transactions and compare it with RF and GBDT [8]. This paper will present the most representative paper of each promising algorithms and hence compare the attributes of different approaches by evaluating the performance of classifiers.

The remainder of this paper is organized as follows. Section 2 introduces the application in different industrial areas. Then in section 3, modified oversampling methods and potential ML algorithms will be further elaborated. What comes next are the results of models and their discussions. Section 5 summarizes this paper and presents conclusion from the algorithms discussed here.

2. Application

Normally, FDS consists two probable methods: an automatic tool and a manual evaluation. The automatic tool is based on a specific algorithm and hence detects fraudulent behaviors through processing large stream of transactions. Whereas manual evaluation lies on expert's professional experience, like previous situations when such fraudulent behaviors occurred, so it is quite subjective [9]. To obtain a more versatile detecting method, this paper only investigates on the former method. To be more specify, FDS is basically to analyze the characteristic of the incoming new transactions and employ a classifier to identify whether this entry belongs to the fraudulent class. FDS is valuable in many aspects, and this paper aims to aims to elaborate on the application of FDS in light of the following aspects.

2.1. Credit card detection

Credit card fraud detection is the most common problem in fraud criminal. In most cases, such behaviors involve obtaining goods without paying the merchant, transferring money without permission, etc. [3]. Those behaviors have caused severe damage to both the customers and the merchants. Since whether a transaction is labeled "fraudulent" is a binary problem, FDS is applicable in detecting such transactions. For instance, Ju et al. combine kNN with SMOTE in generating new entries and employ LSTM to classify fraudulent transactions [4]. And Benchaji et al. use genetic algorithm to improve classification of imbalanced credit card dataset [3]. Mo et al. employ Generative Adversarial Nets (GAN) to produce minority fraudulent transaction samples and use DT as the base classifier [6]. Carneiro et al. try to keep high-cardinality attributes on credit card detection [10].

2.2. Insurance fraud

Insurance fraud behavior covers many areas, mainly around the medical insurance, like fabricating healthcare card for medical treatment, overprescribing and reselling in violation of regulations, etc. Just like credit card transactions, this dataset is also highly skewed. Wu et al. propose a new fraud detection method for imbalanced healthcare dataset called Wasserstein Generative Adversarial Network-Kernel Density Estimation (WGAN-KDE), and then since the medical features are specialized, they use Auto-

DOI: 10.54254/2755-2721/27/20230146

Encoder to simplify those features [11]. Since treatment time distribution is uneven and dataset sample is imbalanced, Cao et al. employ Time-Aware LSTM (TLSTM) to judge the fraud [7].

2.3. Supply chain fraud

Supply chain fraud often means fraudulent orders or transactions in a supply chain, like unchanging order status, excessively long shipping time, etc. Wan hybridizes the XGBoost and RF and measure it using DataCo smart supply chain datasets, which performs better than other algorithms [12]. Besides, Wang and Zhi use ADASYN to produce new data from DataCo Global dataset, and then employ Sequential Forward Selection (SFS) to retain only the optimal subsets of features, including training set and testing set [13].

3. Methods

3.1. Generalized fraud detecting system

*3.1.1. Raw data prepossessing.*This work includes sifting meaningless variables, disposing missing values, generating new features and improving imbalanced distribution, etc. After finishing those work, then dividing the employed dataset into two parts, namely training set and testing set. Effective and appropriate prepossession could improve the quality of dataset, and thus guarantee the performance of the following algorithms.

3.1.2. Model training and cross-validation. In addition to the model the researcher investigates on, other possible machine learning algorithms are also trained as comparison when evaluating the performance of that base classifier, and those additional models will be shown in the result part. Also, most research employ cross-validation. For instance, 10-fold cross validation means dividing the dataset into 10 mutually exclusive subsets, and then selecting 9 subsets in turn as the training sets and the remaining subset as the testing set. The final testing result is the average of 10 separate results [13].

3.1.3. Testing results evaluation. In order to measure the performance, different evaluation metrics should be introduced. In most situations, evaluation of classification algorithms is commonly based on accuracy, precision and recall. In the following content, the most recent methods related to FDS issues and their improvements as well as contributions are introduced.

3.2. Machine learning models

3.2.1. kNN-SMOTE-LSTM. According to Ju et al., kNN-SMOTE-LSTM credit card fraud detection model is a Long Short Term Memory (LSTM) model based on improving SMOTE technology, and kNN identifying classifier could only keep reliable samples generated by SMOTE [4]. Because of the imbalanced dataset, they adopt SMOTE method to produce new samples. Nevertheless, those new entries are generated basing on distance measurements, so some of them are noise data. kNN identifying classifier could only retain true samples, which would improve the accuracy of the following base LSTM classifier. Also, since the distribution of dataset in historical transactions varies and new fraud transactions could occur, LSTM is more suitable in identifying such behaviors.

3.2.2. GAN-AdaBoost-DT. Mo et al argue that there are mainly two ways in dealing with imbalanced dataset: the first one focuses on dataset, like undersample and oversampling; the other one pays attention to ensemble learning for integrated classifier could avoid bias caused by a single classifier when classifying the unbalanced dataset [6]. They use GAN method to generate minority samples. This model includes Generator model and Discrimination model. When the discrimination model cannot distinguish between generated samples and original datasets, this new dataset resembles reality. This method does not have to be based on real transactions and thus could avoid overfitting. After this, they employ

DOI: 10.54254/2755-2721/27/20230146

AdaBoost with DT as the base classifier because ensemble method could improve the classification performance of a single weak classifier when dealing with uneven distribution.

3.2.3. WGAN-KDE-GBDT. Chai et al. demonstrate that although GAN performs better than SMOTE when treating imbalanced dataset, this model is easy to crash, unable to converge or tends to overfit during training process [14]. WGAN uses Wasserstein Distance as distance measurement instead and turn it into an optimization problem, and it shows that this change improves training speed and guarantees successful generation. According to Wu et al., the random noise data in WGAN does not consider the distribution of data in reality, so they use KDE to change the constitution of noise data in WGAN [11]. After getting the simulation dataset, they employ Gradient Boosting Decision Tree (GBDT) as the base classifier.

3.2.4. TLSTM. Some research concerning medical treatment prediction are based on Recurrent Neural Network (RNN) [15]. However, according to Cao et al., this algorithm cannot be applied to a long-term time series [7]. In other words, the performance of RNN classifier drops when this sequence is extended too long. This situation is quite normal because it is natural for patients to take a long interval between medical treatments. They use TLSTM instead, considering the weights of different time intervals. After getting the predicted medical behavior, they calculate the similarity between the simulated behavior and the genuine behavior. If they are not alike, the behavior in reality presumes to be fraudulent.

3.2.5. ADASYN-SFS-RF. In order to solve imbalanced distribution problem, Wang et al. use ADASYN as a formular to determine the number needed to be generated for the minority class [13]. As for the possible noise data accompanying ADASYN method, they employ Sequential Forward Selection (SFS) algorithm, basing on RF, to only reserve optimal subsets of features. And then they use this subset to train the RF base classifier. In the end, they use Local Interpretable Model-agnostic Explanations (LIME) to identify important features responsible for fraud detection.

3.3. Evaluation metrics

Evaluation of classifier are commonly based on accuracy, precision and recall. Also, some statistical measurements are also applicable, like F1 score and Area Under Curve (AUC). The corresponding equation can be found as follows.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative} \quad (1)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (2)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3)$$

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \quad (4)$$

DOI: 10.54254/2755-2721/27/20230146

4. Results and Discussion

4.1. kNN-SMOTE-LSTM

Table 1. The performances of LSTM under different oversampling methods [4].

Model	F1	AUC
LSTM	**0.8344**	0.9116
ADASYN+LSTM	0.0705	0.9359
SMOTE+LSTM	0.1604	0.9394
BorderlineSMOTE+LSTM	0.2257	0.9391
svmSMOTE+LSTM	0.2446	**0.9396**
SMOTEENN+LSTM	0.1624	0.9395
SMOTETomek+LSTM	0.1604	0.9394
kNN+SMOTE+LSTM	0.8280	0.9247

First of all, when choosing the appropriate base classifier, Ju et al. select LSTM rather than Gaussian Naive Bayes, Logistic Regression, knn, SVM, etc, since it outperforms others when it comes to precision, recall, F-score and AUC due to its noise immunity [4]. Hence, the evaluation basically revolves around how LSTM performs under different oversampling circumstances. Table 1 shows that if the base classifier combines with different oversampling methods, the corresponding result shows poor performance with low F score (at most 0.24). Whereas kNN+SMOTE+LSTM successfully integrates the data generator, the identifying classifier and the base classifier, demonstrating excellent classification performance.

4.2. GAN-AdaBoost-DT

Table 2. AUC and accuracy for different sampling methods [6].

Model	AUC	Accuracy
DT	0.615	0.725
SMOTE-DT	0.614	0.722
RUS-DT	0.621	0.619
ADASYN-DT	0.613	0.720
GAN-DT	**0.655**	**0.755**

Table 3. AUC and accuracy of different classification models [6]

Model	AUC	Accuracy
LR	0.500	0.725
SVM	0.536	0.554
AdaBoost	0.639	0.725
RUSBoost	0.665	0.816
SMOTEBoost	0.658	0.816
MSMOTEBoost	0.662	0.813
GAN-Adaboost-DT	**0.701**	**0.853**

To begin with, Mo et al. try to compare GAN with other sampling methods, including SMOTE, Random Undersampling Method (RUS) and ADASYN. Table 2 shows that GAN outperforms other methods in terms of AUC and accuracy with DT as the base classifier [6]. Then, they contrast GAN-Adaboost-DT with other classification models and successfully prove that the minority class samples generated by GAN in every iteration of adaboost improve the overall performance of classifier.

DOI: 10.54254/2755-2721/27/20230146

4.3. WGAN-KDE-GBDT

Table 4. Different method results on health insurance dataset [11].

Sampling method	Base classifier	Recall	Precision	F1	Accuracy	AUC
RUS	LR	0.809	0.393	0.529	0.862	0.931
SMOTE		**0.866**	0.465	0.605	0.892	**0.955**
WGAN		0.484	0.960	0.644	0.949	0.939
WGAN-KDE		0.509	**0.978**	**0.669**	**0.952**	0.931
RUS	AdaBoost	**0.934**	0.636	0.757	0.943	0.985
SMOTE		0.917	0.691	0.788	0.953	0.985
WGAN		0.931	0.931	0.819	0.969	0.980
WGAN-KDE		0.743	**0.948**	**0.833**	**0.971**	**0.985**
RUS	GBDT	**0.990**	0.835	0.906	0.980	0.979
SMOTE		0.987	0.824	0.898	0.979	0.984
WGAN		0.974	0.842	0.903	0.980	0.975
WGAN-KDE		0.967	**0.937**	**0.951**	**0.991**	**0.988**

Wu et al. use five indicators to assess the overall performance of different methods. In table 4, WGAN-KDE performs better than other sampling methods with LR, AdaBoost and GBDT as base classifiers [11]. In most cases, WGAN-KDE has deficiency in recall comparing to RUS and SMOTE. This is because both the oversampling method and the undersampling method keep minority class samples by duplicating existing minority data or discarding majority class samples so as to maintain a balance in data distribution. Since base classifiers are sensitive to minority data, those sampling methods have relatively high recall rate, whereas perform poorly on precision, F1, accuracy and AUC. On the contrary, WGAN-KDE keeps a better balance of five indicators.

4.4. TLSTM

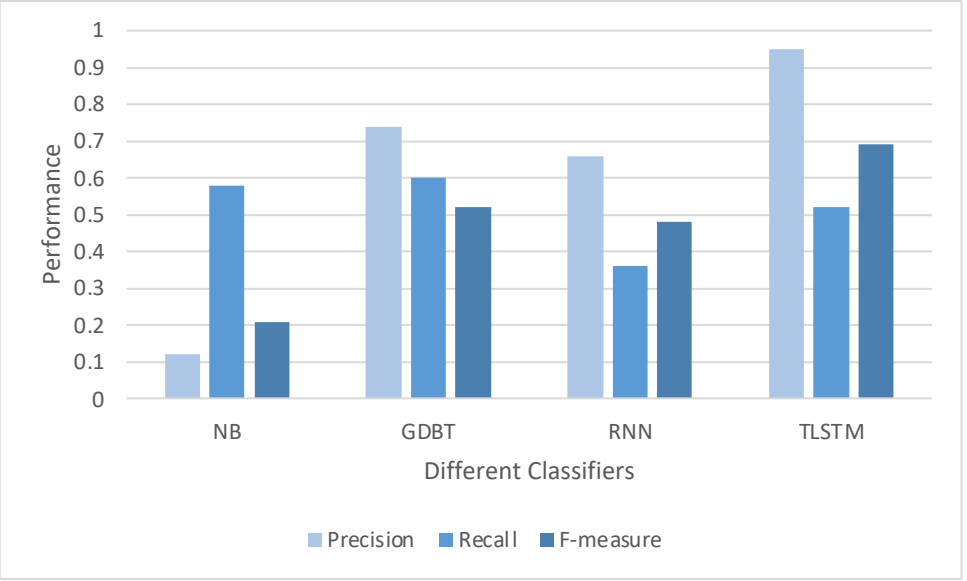

Figure 1. Performances comparison [7].

Cao et al. compares TLSTM with other models like Naïve Bayes (NB), GDBT, and RNN [7]. Figure 1 shows that TLSTM performs better in terms of precision and F-measure, except for recall. Unlike the

DOI: 10.54254/2755-2721/27/20230146

improved sampling methods mentioned above, Cao et al try to solve the uneven distribution of dataset and the imbalanced dataset problem by adding time intervals into LSTM model. Through taking weights into consideration, the modified LSTM model improves the ability in fraud detection.

4.5. ADASYN-SFS-RF

Table 5. Comparision of detecting results of oversampling methods [13]

Model	Accuracy	Recall	F1	AUC
RF	0.9915	0.7961	0.8171	0.9171
SMOTE+RF	0.9812	0.5473	0.6954	0.9675
SvmSMOTE+RF	0.9814	0.5480	0.7037	0.9821
BorderlineSMOTE+RF	0.9811	0.5481	0.6886	0.9543
ADASYN+RF	0.9812	0.5469	0.6948	0.9671
ADASYN+SFS+RF	0.9948	0.8220	0.8945	0.9881

At first, wang et al. test different base classifiers before applying sampling methods, including LR, DT, Back Propagation Neural Networks (BP), kNN, SVM and RF [13]. The result shows that RF has more generalization performance with low false positive rate, which is more suitable for supply chain transaction fraud detection. After this, when evaluating the performance of unbalanced oversampling algorithm, they employ RF as the base classifier. It turns out that SFS successfully eliminate the noise data generated by ADASYN and further clarify the classification boundaries.

5. Conclusion

This paper first presents main application situations for FDS and then introduces corresponding cutting-edge methodologies in improving the performance of FDS classifiers, like kNN-SMOTE-LSTM in credit card detection, TLSTM in predicting medical treatment and ADASYN-SFS-RF for detecting fraudulent transaction in supply chain. In addition, this paper shows the performance of each algorithm when comparing to other methods, proving the merits of each modification. Nevertheless, this paper fails to elaborate on more detailed analysis of corresponding mechanisms, which should be improved in the future.

References

[1] Abdallah A, Maarof M A, Zainal A. Fraud detection system: A survey. Journal of Network and Computer Applications, 2016, 68: 90-113.

[2] P. Saravanan, V. Subramaniyaswamy, N. Sivaramakrishnan, M. Arun Prakash, T. Arunkumar. Data mining approach for subscription-fraud detection in telecommunication sector. Contemporary Engineering Sciences, Vol. 7, 2014, no. 11, 515-522

[3] Benchaji, I., Douzi, S., El Ouahidi, B. Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for Credit Card Fraud Detection. In: Khoukhi, F., Bahaj, M., Ezziyyani, M. (eds) Smart Data and Computational Intelligence. AIT2S. Lecture Notes in Networks and Systems, 2018, vol 66. Springer, Cham.

[4] Chunhua Ju, Guanyu Chen, Fuguang Bao. Consumer Finance Risk Detection Model Based on kNN-SMOTE-LSTM: Taking Credit Card Fraud Detection as An Example. Chinese Journal of Systems Science, 2021, 41(02):481-498.

[5] Yu Lin, Xun Huang, Weide Chun, Dengshi Huang. Early Warning Research on Extreme Financial Risks Based on ODR-ADASYN-SVM. Journal of Management Sciences in China, 2016, 19(05):87-101.

[6] Zan Mo, Yanrong Gai, Guanlong Fan. Credit Card Fraud Classification Based on GAN-AdaBoost-DT Imbalance Classification Algorithm. Journal of Computer Applications, 2019, 39(02):618-622.

DOI: 10.54254/2755-2721/27/20230146

[7] Luhui Cao, Fenglin Qin, Zhongmin Yan. TLSTM Based Medicare Fraud Detection. Computer Engineering and Applications, 2020, 56(21):237-241.

[8] Rongrong Chen, Guohua Zhan, Zhihua Li. Credit Card Transaction Fraud Prediction Based on XGBoost Algorithm Model. Application Research of Computers, 2020, 37(S1):111-112+115.

[9] Johannes Jurgovsky, Michael Granitzer, Konstantin Ziegler, Sylvie Calabretto, Pierre-Edouard Portier, Liyun He-Guelton, Olivier Caelen, Sequence classification for credit-card fraud detection, Expert Systems with Applications, Volume 100, 2018, Pages 234-245

[10] Carneiro, E.M.; Forster, C.H.Q.; Mialaret, L.F.S.; Dias, L.A.V.; da Cunha, A.M. High-Cardinality Categorical Attributes and Credit Card Fraud Detection. Mathematics, 2022, 10, 3808.

[11] Wenlong Wu, Xi Zhou, Yi Wang, Baoquan Wang. WKAG: A Fraud Detection Method for Unbalanced Health Care Data. Computer Engineering and Applications, 2021, 57(09):247-254.

[12] F. Wan, XGBoost Based Supply Chain Fraud Detection Model, 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 355-358.

[13] Wanmin Wang, Luping Zhi. Generalization Performance Improvement and Interpretability of Fraud Detection Model Based on ADASYN-SFS-R.F. Application Research of Computers, 2023, 1-11.

[14] Mengting Chai,Yuanping Zhu. Research and Application Progress of Generative Adversarial Networks. Computer Engineering, 2019, 45(09):222-234..

[15] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17). Association for Computing Machinery, 2017, 1903–1911.