Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 221 (2023) 316–321

Procedia Computer Science

www.elsevier.com/locate/procedia

Information Technology and Quantitative Management (ITQM 2023)

Feature Selection Based on Two-stage Resampling Technique for Imbalanced Dataset

Dan Zhao[a,b], Zhenyi Shen[a,b], Shuangxue Zhao [a,b,*]

[a]*College of Computer Science and Technology, Zhejiang University, No.38 Zheda Road, Xihu Distric, Hangzhou 310027, China*
[b]*Bank of Hangzhou Co., Ltd., No.46 Qingchun Road, Gongshu Distric, Hangzhou 310002, China*

Abstract

The conventional feature selection methods fail to work on imbalanced datasets due to the overfitting issue caused by the extremely imbalanced positive and negative samples. In addition, the large number of negative samples makes the feature selection operation inefficient. To address these issues, an automatic feature selection method for addressing feature selection issues on extremely imbalanced datasets is proposed. An undersampling operation is performed to generate a small balanced dataset at first. Then, the feature importance scores are estimated based on the occurrence of the feature columns in the feature combinations that have a higher score than the base score. Finally, a repeated feature removal operation and classification model retrain are performed on the new dataset generated with the oversampling technique until the removal of the unimportant feature column does not bring any score improvement. In the proposed approach, the oversampling operation can fully utilize the dataset to train the classification model properly and the importance score is easily interpreted as the frequency of each feature column in the satisfied feature combinations. The experiment shows the superior performance of the proposed feature selection method.

Peer-review under responsibility of the scientific committee of the Tenth International Conference on Information Technology and Quantitative Management

Keywords: Feature selection; Imbalanced dataset; Undersampling; Oversampling

1. Introduction

With the increase in credit card fraud transactions, building an explainable and effective anti-fraud model has become an important research direction in the financial field. Traditional detection methods use the expert experience to formulate corresponding rules which are time-consuming, labor-intensive, and not enough to deal with various types of fraud. In recent years, using machine learning for fraud detection has become an inevitable trend with the development of computer technology. Hidden Markov chains [1] and self-organizing networks [2] are often used to construct individual behavior models. However, due to the lack of transaction data of most users, such models are difficult to establish an accurate model and describe new transaction behaviors. K-nearest neighbors (KNN)[3], decision trees [4], random forests [5], and artificial neural networks (ANN)[6][7] are applied to model crowd behaviors. Nevertheless, since the fraud data is a typical class-imbalanced dataset, these classical machine learning algorithms cannot guarantee their performance [8]. So we need to pre-process the data including

*Corresponding author. Tel.: +86-0571-8515-7830.
E-mail address: zhaosx@zju.edu.cn.

1877-0509
Peer-review under responsibility of the scientific committee of the Tenth International Conference on Information Technology and Quantitative Management
10.1016/j.procs.2023.07.043

oversampling, undersampling, and feature selection. In the literature [9] and [10], the original datasets were preprocessed using oversampling and undersampling techniques, respectively. However, due to the large amount of credit card fraud data, oversampling technology used on the train data will make the training efficiency lower. At the same time, the small amount fraudulent data will lead to the loss of some important information when using undersampling technique. Therefore, some research [11][12] combine the two sampling techniques and achieve better results. But it is obviously not enough that only apply sampling techniques to imbalanced data. We also need to select more useful and discriminative features to improve the efficiency and interpretability of the fraud detection model.

The fundamental task of the credit card fraud detection problem is to obtain better feature differentiation to ensure the performance and stability of the model. In fact, the large number of redundant features in the original data increases the calculation amount of the model and reduces the accuracy [7][13], which leads to the necessity of feature selection. Feature selection techniques mainly include feature selection based on importance weights [14][15] and recursive feature elimination (RFE) [16][17]. The former carries out feature selection and classification problem simultaneously, and thus is computationally slow when the classification algorithm is complex. And RFE is considered a wrapper method, but it has the problem of overfitting and time consuming, especially when using big data. Besides, both methods when used on imbalanced data tend to select features that are important only for the majority class and ignore important features specific to the minority class, thus increasing the classification error, as shown in the experimental results in Section 3.

Based on the above description, there are two issues for establishing a credit card detection model on an imbanlanced dataset. The first is to screen out discriminative features to make the model more interpretable, and the second is to improve the efficiency. Therefore, this paper proposes a new feature selection to perform feature selection on undersampling train data, and then test the detection performance of the selected features on the oversampling validation data. The main contributions of this paper are summarized below:

- The proposed method makes full use of the minority samples in the imbalanced data for feature selection through a two-stage resampling technique, which avoids overfitting and improves classification accuracy.
- The undersampling stage and the selection of high-dimensional features reduce the computational complexity.
- The good interpretation benefits from our preservation of the original important features, as opposed to the feature extraction methods that project the original features into new directions.
- We propose a new robust and stable feature selection method based on cross-validation.

2. Feature selection via estimating the feature importance

There are three challenges to be addressed in this paper:

- Feature selection on the imbalanced dataset
  Conventional feature selection methods fail to work on imbalanced dataset due to the overfitting issue caused by the extremely imbalanced positive and negative samples.
- Effective feature importance evaluation method
  The features importance should be estimated properly on the imbalanced dataset such that important features are kept and unimportant features can be removed.
- The interpretability of the feature selection method
  The feature selection process should have high interpretability such that the selected results are trustable.

2.1. Evaluating the feature importance via undersampling

To select the effective features on the extremely imbalanced dataset and ensure the high running efficiency of the proposed method, an undersampling is performed. On the resampled dataset, the distribution of positive and negative samples are balanced and the sample size is much smaller than that of the original dataset.

Assume we have N_{pos} and N_{neg} samples, where $N_{neg} \gg N_{pos}$. To normalize the distribution of positive and negative samples, all N_{pos} fraud samples are selected and N_{pos} non-fraud samples are selected among the N_{neg} non-fraud samples. Therefore, the sample size of the newly generated dataset is $2N_{pos}$.

To evaluate the importance of each feature column, a large number of binary mask vectors are generated based on a predefined Bernoulli distribution. Each element in the binary mask vector $\boldsymbol{k}$ is generated based on the distribution defined below:

$$\begin{aligned} P(X=1) &= p \\ P(X=0) &= 1-p \end{aligned} \tag{1}$$

Assume that we have N binary mask vectors $\boldsymbol{k}_i \in \mathbb{R}^L, 1 \le i \le N$, initialized based on Eq. (1), where L is the total number of features. Each mask vector is designed to work as a vector to filter the feature columns. For a given $\boldsymbol{k}_i$, the j-th element e_j works as below:

$$e_j = \begin{cases} 1, & \text{select the } j\text{-th column} \\ 0, & \text{remove the } j\text{-th column} \end{cases} \tag{2}$$

Therefore, each mask vector is associated with a unique combination of feature sets. To get the score of each feature combination encoded in a mask vector, a logistic regression with 5-fold cross-validation is applied. Then, the averaged score is used as the evaluated score for the mask vector. Assume that the $f_d(\boldsymbol{k_i})$ represents the AUC score computed at the d-th fold for mask vector $\boldsymbol{k_i}$, the evaluated scores with 5 are formulated as:

$$[f_1(\boldsymbol{k_i}), f_2(\boldsymbol{k_i}), f_3(\boldsymbol{k_i}), f_4(\boldsymbol{k_i}), f_5(\boldsymbol{k_i})] \tag{3}$$

and the averaged score for mask vector $\boldsymbol{k_i}$, denoted as $f(\boldsymbol{k}_i)$, is formulated as:

$$score(\boldsymbol{k}_i) = \frac{1}{D}\sum_{i=1}^{D}(f_d(\boldsymbol{k_i})) \tag{4}$$

where D represents the number of folds in the evaluation process.

2.2. Remove the undesired features based on the cumulative occurrence

To remove the undesired features, a base score computed with the all features is computed as:

$$score_{base} = score(\boldsymbol{k}),\ \boldsymbol{k} = [1, 1, ..., 1, 1] \in \mathbb{R}^L \tag{5}$$

Any mask vector $\boldsymbol{k}_i, 1 \le i \le N$, with score $score(\boldsymbol{k}_i) > score_{base}$ are selected into a set. Then, the occurrence vector $\boldsymbol{o}$ that records the occurrence of each feature column in the feature combination that has a higher score than the $score_{base}$ can be computed as:

$$\boldsymbol{o} = \sum_{i}^{N} \boldsymbol{k}_i I_{\{x|x>score_{base}\}}(score(\boldsymbol{k}_i)) \tag{6}$$

where $I_{\{x|x>score_{base}\}}(\cdot)$ is an indicator function that outputs 1 when its input is greater than $score_{base}$ and outputs 0 otherwise. Each element in the occurrence vector $\boldsymbol{o}$ can be interpreted as the recorded frequency of feature columns in the feature combinations that has a higher score than the benchmark. The order of feature importance is obtained when these elements in the vector $\boldsymbol{o}$ are ranked in ascending order.

Once the feature importance order is computed, one can remove the unimportant feature column one by one and then calculate the averaged AUC score with the cross-validation operation. To improve the classification performance, the Synthetic Minority Oversampling TEchnique (SMOTE) is applied to generate a balanced dataset that fully includes information in the fraud and non-fraud samples. An unimportant feature column is removed if the averaged score is higher than the base score and then the base score is updated by the newly computed averaged score. The feature elimination process is repeated until the feature removal operation does not bring any increase in AUC values. In summary, the proposed two-stage resampling technique for imbalanced dataset is summarized in Algorithm 1.

Algorithm 1 The two-stage resampling technique for imbalanced dataset

Undersample the dataset
Evaluate the feature importance by Eq. (4) individually
Rank the feature importance with ascending order
Oversample the dataset with the SMOTE
for `Select the most unimportant feature from the ranked feature set` **do**
    Remove the selected feature and generate the new feature set with the left ones
    Evaluate the model's performance on the evaluation dataset with the new feature set
    Break the loop once the model's performance on the evaluation dataset is not improved
end for
Output the selected feature set

3. Experiemnts

In the experiment, the proposed feature selection method is applied to the credit card fraud dataset. This dataset has 284807 transaction records and only 492 records have frauds. Therefore, this dataset is extremely imbalanced with the probability of having a fraud transaction record equal to 0.172%. The probability p in the Bernoulli distribution is equal to 0.85, meaning there is an 85% probability of being selected independently for each feature. The initial number of binary mask vectors used in feature selection is 300. The data samples are shuffled, and 80% samples are used in training and the left 20% samples are used in testing.

3.1. The importance of the feature selection

In this section, the proposed method is utilized before the application of the classic logistic regression model and the random forest model. Table 1 summarizes the model performance on the test dataset before and after the proposed feature selection. The KS value, AUC values are used as two metrics to evaluate the overall performance of the model for both fraud and non-fraud samples. The fraud and non-fraud F1 score are used to evaluate the model's performance for the fraud and non-fraud cases, respectively.

Table 1. The summarized performance before and after the feature selection

Methods	KS value	AUC value	Fraud F1 score	No Fraud F1 score
Logistic regression	0.856	0.927	0.2058	0.9942
Logistic regression with feature selection	0.867	0.933	0.2248	0.9948
Random forest	0.663	0.831	0.7975	0.9997
Random forest with feature selection	0.673	0.836	0.8000	0.9997

In view of all four metrics used, the feature selection method applied before training the logistic regression model and random forest model can have higher metric scores than applying the model directly without the proposed feature selection method. This shows the effectiveness of the proposed method, as the feature columns that do not have a positive contribution to the prediction have been removed.

3.2. The efficiency of the proposed feature selection process

The running efficiency of the proposed method can be ensured as the feature selection process is performed in the under-sampling stage. The proposed feature selection method is compared with the feature selection from the model weight and feature selection via recursive feature elimination. Logistic regression is used as the prediction model after the listed feature selection methods.

Table 2. The running efficiency of the proposed method

Methods	Time spent for each iteration (Second)	Total time spent
The proposed feature selection method	0.015	1.71
Feature selection from model weight	2.15	2.15
Feature ranking with recursive feature elimination	2.471	7.43

Table 2 shows that the proposed feature selection method has the smallest total running time and the running time spent on each iteration. This is because the proposed feature selection is applied to the undersampled dataset, where the samples are greatly smaller than the original dataset. Since all fraud samples are kept on the undersampled dataset, our feature selection process can keep features that are sensitive to the fraud samples and ensure the recall rate at a high level.

3.3. The achieved performance visualized by the ROC curve

Fig. 1 visualizes the achieved ROC curve for the feature selection with logistic regression. It is seen that the plotted ROC curve achieved is much higher than the ROC score of 50%, which is regarded as a benchmark. The final roc score obtained is 0.927, showing the effectiveness of the feature selection combined with the logistic regression in solving the prediction tasks on the imbalanced dataset.

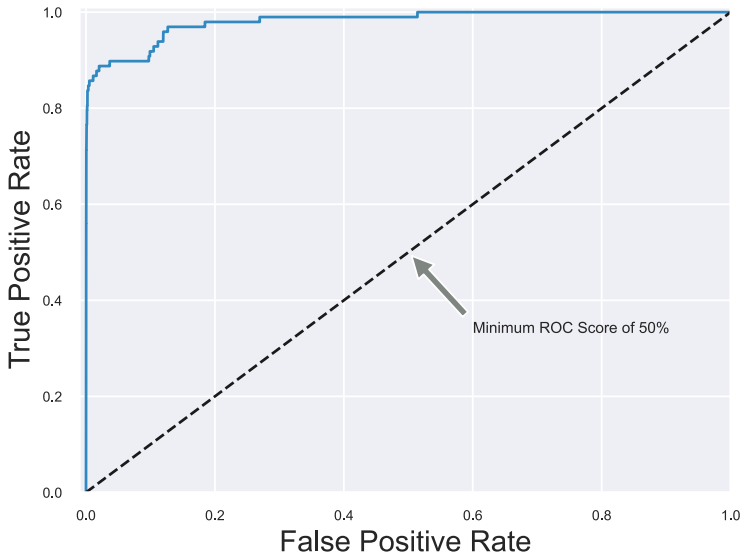

Fig. 1. ROC curve of feature selection with logistic regression

4. Conclusion

In conclusion, an automatic feature selection method has been proposed to address the issues of overfitting and inefficiency on extremely imbalanced datasets. This method involves an undersampling operation to generate a small balanced dataset, estimation of feature importance scores based on the occurrence of feature columns in high-scoring feature combinations, and a repeated feature removal and classification model retrain operation until no further score improvement is achieved. The oversampling operation allows for full utilization of the dataset to properly train the classification model and the importance score is easily interpreted. Experimental results demonstrate the superior performance of this proposed method.

References

[1] Y. Lucas, P.-E. Portier, L. Laporte, L. He-Guelton, O. Caelen, M. Granitzer, S. Calabretto, Towards automated feature engineering for credit card fraud detection using multi-perspective hmms, Future Generation Computer Systems 102 (2020) 393–402.
[2] D. Olszewski, Fraud detection using self-organizing map visualizing the user profiles, Knowledge-Based Systems 70 (2014) 324–334.
[3] S. Nami, M. Shajari, Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors, Expert Systems with Applications 110 (2018) 381–392.

[4] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Systems with Applications 40 (15) (2013) 5916–5923.
[5] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, C. Jiang, Random forest for credit card fraud detection, in: 2018 IEEE 15th international conference on networking, sensing and control (ICNSC), IEEE, 2018, pp. 1–6.
[6] C. Mishra, D. L. Gupta, R. Singh, Credit card fraud identification using artificial neural networks, International Journal of Computer Systems 4 (07) (2017) 151–159.
[7] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, H. Zeineddine, An experimental study with imbalanced classification approaches for credit card fraud detection, IEEE Access 7 (2019) 93010–93022.
[8] Z. Zojaji, R. E. Atani, A. H. Monadjemi, et al., A survey of credit card fraud detection techniques: data and technique oriented perspective, arXiv preprint arXiv:1611.06439.
[9] D. Varmedja, M. Karanovic, S. Sladojevic, M. Arsenovic, A. Anderla, Credit card fraud detection-machine learning methods, in: 2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH), IEEE, 2019, pp. 1–5.
[10] F. Zhang, G. Liu, Z. Li, C. Yan, C. Jiang, Gmm-based undersampling and its application for credit card fraud detection, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.
[11] H. Shamsudin, U. K. Yusof, A. Jayalakshmi, M. N. A. Khalid, Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset, in: 2020 IEEE 16th International Conference on Control & Automation (ICCA), IEEE, 2020, pp. 803–808.
[12] J. Ahammad, N. Hossain, M. S. Alam, Credit card fraud detection using data pre-processing on imbalanced data-both oversampling and undersampling, in: Proceedings of the International Conference on Computing Advancements, 2020, pp. 1–4.
[13] W. N. Robinson, A. Aria, Sequential fraud detection for prepaid cards using hidden markov model divergence, Expert Systems with Applications 91 (2018) 235–251.
[14] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.
[15] Y. Liu, H. Zhao, Variable importance-weighted random forests, Quantitative Biology 5 (2017) 338–351.
[16] C. V. Priscilla, D. P. Prabha, A two-phase feature selection technique using mutual information and xgb-rfe for credit card fraud detection, Int. J. Adv. Technol. Eng. Explor 8 (2021) 1656–1668.
[17] N. Rtayli, N. Enneya, Enhanced credit card fraud detection based on svm-recursive feature elimination and hyper-parameters optimization, Journal of Information Security and Applications 55 (2020) 102596.