Towards novel financial risk prediction method with machine learning model

Kaiyu Xiong *

Economics, University of South Florida, Tampa, United States

* Corresponding Author Email: Xiongkaiyu123@gmail.com

Abstract. Among the multiple risks faced by enterprises, financial risk is particularly prominent in the big data environment. Serious data imbalance has become a major challenge in the analysis of corporate financial risks. Aiming at the sample imbalance problem in enterprise competitive intelligence analysis, this paper proposes an enterprise risk identification method oriented to unbalanced samples, taking credit risk prediction of financial enterprises as a starting point. The method utilizes intelligent analysis means such as feature selection, unbalanced sample balance processing and integrated learning in the field of artificial intelligence, aiming to provide a solution to the problem of enterprise risk identification in enterprise competitive intelligence under the big data environment.

Keywords: Financial risk, machine learning, big data, risk prediction, data processing.

1. Introduction

With the advent of the big data era, the competitive environment of enterprises has changed dramatically. This change constantly brings threats as well as opportunities to enterprises. In enterprises, the number of performing customers is much larger than the number of defaulting customers, and this data category imbalance problem leads to the trained model being more likely to incorrectly categorize risky customers as performing customers, which increases the difficulty of predicting the risk of credit default in financial enterprises. In order to solve the problem of enterprise risk identification in enterprise competitive intelligence under the big data environment, this paper proposes an enterprise risk identification method oriented to unbalanced samples for the unbalanced sample problem in the analysis of enterprise competitive intelligence, taking the prediction of credit risk of financial enterprises as an entry point.

1.1. Intelligence Information Analysis methods for Unbalanced Sample Problems

The unbalanced sample problem has been an existential challenge in the application scenario of intelligence information analysis. Literature [1] proposed an opinion mining method based on unbalanced data analysis by using a method based on sentiment knowledge and machine learning, based on the linguistic characteristics of unbalanced e-commerce opinion mining data; literature [2] used convolutional neural network as a classification method of fusion classifier to improve the citation accuracy of the middle graph classification method under unbalanced dataset. In the field of medical intelligence information analysis, the unbalanced sample problem is more common. Literature [3] addresses the sample imbalance phenomenon in heart failure medical data and uses the under-sampled data to train a locally sensitive discriminant matrix-type classifier and achieves good prediction results; literature [4] proposes a weight-based oversampling method and combines it with two integrated learning models, Bagging and Boosting, to solve the imbalanced samples under the scenarios of identifying the hepatotoxicity of herbal medicines Problem.

1.2. Machine Learning Model Research of Credit Risk Prediction

In credit risk prediction, data in realistic scenarios are usually characterized by large dimensionality, containing a large amount of irrelevant information and information redundancy, which leads scholars to actively explore and apply feature selection methods. In text feature selection, scholars have adopted the binary fireworks algorithm [5] and the feature screening method based on

information gain [6]. In the field of credit risk prediction, some scholars have also tried various feature selection methods, for example, one article [7] proposed a feature selection method based on Relief evaluation and SVM cross-validation, and another article [8] used Random Forest for feature extraction and feature importance assessment before training credit assessment models. However, although related studies have dealt with feature selection, most of the methods have not been improved for the sample imbalance problem of credit data, which may lead to the fact that some features that are effective for classification are not successfully filtered out. Methods for constructing credit risk prediction models for financial enterprises in the present day are mainly divided into two categories: traditional methods, represented by logistic regression and discriminant analysis; and machine learning methods, which usually outperform traditional methods due to their lack of strict assumptions on data distribution. In specific applications, scholars use SVM algorithm to predict credit risk and try to combine SVM with other algorithms to improve the prediction effect [9]; on the other hand, LightGBM algorithm based on integrated learning is applied to credit risk assessment, and compared with machine learning algorithms such as plain Bayes, decision tree, random forest, XGBoost, etc., and the results show that when using AUC as the model evaluation index, LightGBM algorithm scores the highest [10]; In addition, there are researchers who use Boost algorithm algorithm to construct P2P default prediction model and compare it with LightGBM and XGBoost algorithms to comprehensively analyze the performance advantages of Boost algorithm algorithm [11].

1.3. Unbalanced sample processing method

However, if the unbalanced sample problem is not dealt with and the data are used directly for model training, the classifier obtained may tend to judge the data to be predicted as the majority class. Therefore, in order to obtain a more reasonable prediction model, some scholars will use the sampling method to deal with unbalanced data before constructing a credit risk prediction model. Sampling methods usually have two forms: oversampling and under-sampling. The oversampling method increases the minority class samples through some strategy to balance the number of minority class samples with the number of majority class samples. The simplest oversampling method is to randomly select the minority class sample data for replication until the number of minority class samples equals the number of majority class samples. However, this approach may lead to a large number of identical samples in the minority class in the training set, which greatly increases the risk of model overfitting. Therefore, an improved oversampling algorithm, SMOTE [12], has been proposed, which synthesizes new minority class samples by randomly performing linear interpolation between the minority class samples and their nearby samples of the same class, which reduces the overfitting problem to some extent. Another improved oversampling algorithm is Borderline-SMOTE, which focuses more on synthesizing new samples using samples near the boundary of the minority class samples, thus making the newly synthesized samples more conducive to model training [13]. There is also an oversampling method, KMeans SMOTE, based on k-Means and SMOTE, which has been tested on 71 datasets and shown to have performance advantages over other oversampling algorithms [14]. Unlike oversampling, under-sampling balances the number of minority class samples with the number of majority class samples by reducing the number of majority class samples, thus avoiding the noise associated with synthetic samples. The simplest method of under-sampling is random under-sampling, which randomly draws the same number of samples from the majority class as the number of minority classes. However, this method may result in some key information of the majority class samples being missing from the training data. In order to fully utilize the data of the majority class, some scholars have proposed the EasyEnsemble algorithm based on the idea of integrated learning [15], whose original version uses AdaBoost as the base classifier, and whose sampling idea is still an important reference value in the current construction of predictive models, despite the fact that it does not have the performance advantage of the gradient boosting algorithm at present. In general, in the current research on credit risk prediction of financial enterprises, scholars generally discuss the issues related to feature selection and model prediction under unbalanced data separately, lacking a set of

comprehensive solutions, and seldom explore the application effect of multiple unbalanced data sampling methods in credit risk prediction in a comprehensive way. To solve with this problem, we adopt the Boost algorithm classifier with excellent performance and combines the improved AUCRF and EasyEnsemble algorithms to construct a combined optimization algorithm, which provides a new algorithmic solution for science and technology intelligence analysis scenarios in which there are high-dimensional features and unbalanced samples. This method is expected to achieve better prediction results in credit risk prediction and provide a valuable reference for related research.

2. Corporate Financial Risk Prediction Methodology

In this paper, our proposed algorithm consists of three key steps. First, we use AUC-based model evaluation metrics and adopt Random Forest algorithm for feature selection. Then, based on the sampling idea of EasyEnsemble algorithm, we combine Bagging and under-sampling to obtain several balanced training sets. Finally, we trained and integrated Boost algorithm classifier on these balanced training sets.

2.1. Design of Feature selection algorithm

Random forest is an integrated learning algorithm which uses decision tree as a base model. Each decision tree h(x) is generated according to the following rules: 1. Assuming that the number of samples in the total dataset is N and the number of samples in the training set is n, n samples are randomly and with a putback from N as the training set for this decision tree. 2. Assuming that the number of features in the data is M, m features are randomly selected from M for the training of the decision tree.

Where Y denotes the set of category labels in the dataset and P (-) is the indicator function. In order to improve this algorithm. First, we calculate the importance of the features based on the amount of change in the Gini index of the random forest at the time of node splitting. Then, we set the threshold variable and gradually increase the threshold. After each increase of the threshold, we select the feature combinations whose feature importance is greater than the threshold for model training, and calculate and record the AUC of the model at this time.Finally, we compare the AUCs of all the models, and find out the feature combination corresponding to the model with the largest AUC as the optimal feature combination.

2.2. Sampling Strategy Based on the Combination of Bagging and Undersampling to Construct a Balanced Training Set

We adopt a sampling strategy based on the combination of Bagging and under-sampling, inspired by the EasyEnsemble algorithm. Specifically, for a given training dataset, we assume that the number of minority class samples is M and the number of majority class samples is N. Then, we perform iterative random under-sampling with put-back on the majority class samples N to obtain T majority class subsets N_1, N_2, N_3..., N_t, where the number of samples in each subset is equal to M. Next, we combine the obtained T majority class subsets are merged with the minority class samples respectively, so that we obtain T sample-balanced training sets with a sample size of 2M. With this sampling approach, we retain as much information as possible from the majority class samples while avoiding the noise associated with synthetic samples. The advantage of this sampling strategy is that it provides sample equalization while maintaining the distribution of the data, thus helping us to train the predictive model more accurately. By using a balanced training set, we are able to better capture the features of the minority class samples and effectively address the problem of degraded prediction due to sample imbalance in credit risk prediction.

2.3. Training and integrating Boost algorithm base classifiers

Boost algorithm is a strong classifier belonging to the Boosting algorithm, which is structured using an additive model and is combined by linear summation of a series of weak classifiers. It

measures the gap between the predicted and true values, and the penalty term, which is used to prevent the model from overfitting.

Compared to other members of the Boosting algorithm family, Boost algorithm not only has excellent performance, but also handles categorical features efficiently. Its method of handling categorical features is called "Target-based Statistics", i.e., it replaces the elements of categorical features by computed values. The specific approach is to introduce a priori optimization in the Greedy TS algorithm, and then according to the principle of sorting the training samples are randomly sorted and numbered, and in each calculation, only the value of the category label of the sorting number is smaller than that of the sample is calculated. In addition, Boost algorithm uses the greedy method to combine features in different categories to create new features.

In the algorithm of this paper, the Boost algorithm base classifier is trained using the balanced training set of T samples obtained in step 2, and the Boost algorithm classifier is integrated, and the final integration result, H(x). In this equation, Y denotes the set of category labels in the dataset and P (-) is the indicator function.

Through this series of steps, the algorithm in this paper realizes the prediction of credit risk of financial enterprises and achieves better prediction results by utilizing the powerful performance of Boost algorithm algorithm and the advantage of processing category-type features.

3. Experiment and result analysis

3.1. Data sources

The experiments in this paper use a customer default dataset provided by a domestic financial institution as the training and testing data for the model. The observation period of this dataset is the whole year of 2018 and the performance period is the whole year of 2019. After removing missing values and outliers, there are 124,880 samples, which contain 12,984 positive samples, accounting for 10.40% of the total samples. The dataset contains 50 customer performance and attribute variables during the observation period, and one labeling variable used to label whether a customer defaulted during the performance period. According to the five-level classification criteria for loans, all states except "normal" are considered as default.

3.2. Data feature preprocessing

In the data preprocessing stage, for numerical variables, this paper adopts the standardization method. This is done by subtracting the value of each numerical variable from its mean μ, and then dividing it by its standard deviation σ, thus making the data obey a standard normal distribution with mean 0 and variance 1.

For categorical variables, the Boost algorithm itself can handle this on its own, but to facilitate comparisons with other algorithms, this paper encodes the individual categories of dichotomous and categorical variables as numerical values. For example, the mortgage status of the customer's loan product is converted to 1 for "mortgage" and 0 for "non-mortgage"; the marital status of the customer is converted to 1 for "unmarried" and 1 for "married". "Married" as 2, "Widowed" as 3, and "Divorced" as 4. By pre-processing the data in this way, we transform the dataset into a form suitable for training machine learning models, allowing different types of variables to be used in the training of machine learning models, so that different types of variables can be used in the training of machine learning models, so that they can be used in the training of machine learning models. This allows different types of variables to be efficiently input into the model, preparing it for subsequent model training and performance evaluation.

3.3. Model Evaluation Metrics

In the unbalanced classification problem, although the number of majority samples is much larger than the number of minority samples, we are often more concerned about the classification of the minority samples when making predictions.

3.4. Algorithm Feature Selection Results

The ROC curve is a commonly used performance evaluation tool that represents the relationship between the true positive rate (TPR) and the false positive rate (FPR) as a curve. Where TPR represents the probability of correctly categorizing a positive example, while FPR represents the probability of misclassifying a negative example as a positive example. In financial corporate credit risk prediction, when positive samples belong to the minority class and negative samples belong to the majority class, the ROC curve describes the trend of misclassification of the majority class in the process of the model's ability to continuously search for the minority class. In order to facilitate the comparison of the performance of different models, the area under the ROC curve, i.e., AUC (Area under the Curve) value, is generally used to measure the classification performance of the model. In this paper, AUC is also used as a model evaluation metric.

3.5. Experimental Results of Prediction Algorithms with Multiple Sampling Methods

Feature selection is to select the features that contribute most to the prediction effect of the model from the original large number of features. However, with the further increase of the threshold, the features that are helpful for improving the predictive performance of the model are gradually screened out, which leads to a decrease in the predictive ability of the model. According to the feature combination corresponding to the highest point of the feature selection curve, 17 features are selected as the optimal feature combination in this paper.

Further analyzing the feature selection results, it is found that the feature combination contains the customer's historical credit record, basic situation, and basic information of the product. Historical credit record is a crucial reference element for judging the credit risk of financial enterprises. Customers with default records are still more likely to default during the observation period and the performance period. In addition, among customers with poor historical credit records, especially those with long overdue days and a high number of overdue payments, they should receive more attention. Basic customer profiles, such as high educational qualifications, adequate asset balances, and close business dealings with financial firms, usually mean that such customers are less likely to default. Factors such as the balance of the credit product, the interest rate, and the type of security can also have an impact on the risk of a financial firm's credit business.

3.6. Experimental results of prediction algorithm under multiple classification algorithms

In the experiment, this paper also compares the effect of combining different unbalanced data processing methods with Boost algorithm. In addition, the overall performance of the oversampling algorithm is not as good as the under-sampling algorithm, and the credit data of financial enterprises have a certain degree of complexity, and the use of the oversampling algorithm to synthesize the samples may introduce noise, which affects the training effect of the model. Among the under-sampling algorithms, the sampling strategy in this paper has a better classification effect than the simple under-sampling algorithm, and for that the sampling method in this paper can more fully utilize the information in the credit data of financial enterprises through the integrated way to improve the prediction performance of the model.

To verify the effectiveness of the proposed risk prediction method, the authors use the optimal feature selection results that have been obtained and use AUC as the model evaluation index to compare the Boost algorithm with several machine learning algorithms that are more commonly used at present in the experiments. At the same time, the SMOTE algorithm is also added as a reference to observe the effect enhancement brought by EasyEnsemble compared to ordinary sampling methods.

It can be found that the 3 integrated learning algorithms such as AdaBoost, LightGBM and Boost algorithm outperform the 3 machine learning algorithms with single learner in all the experiments. Especially, among the 3 integrated learning algorithms, Boost algorithm performs the best. Among all the experimental results, the combined optimization algorithm proposed in this paper achieves the highest AUC value, and compared with other algorithms, this method has better prediction effect in credit risk prediction of financial enterprises under unbalanced samples.

4. Conclusion

In summary, this paper proposes a risk prediction method for the sample imbalance problem in the analysis of enterprise competitive intelligence, taking the credit risk prediction of financial enterprises as an entry point. The method carries out feature selection through the improved Random Forest, and uses the strategy of Bagging and under-sampling combination to construct multiple balanced training sets, and uses Boost algorithm as the base learner for training and integration. Experiments have proved that the feature combinations screened by the algorithm in this paper have strong interpretability and are informative for high-dimensional feature screening in unbalanced samples. Meanwhile, the method proposed in this paper performs optimally among multiple classification algorithms and effectively solves the problem of credit risk prediction of financial enterprises.

In the future, we'll continue to optimize the machine learning algorithm to reduce financial enterprise credit risk under the data category imbalance scenario, which can perform better accuracy.

References

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., pp. 785-794, Aug. 2016.

[2] A. Dorogush, V. Ershov and A. Gulin, "Boost algorithm: Gradient boosting with categorical features support", Proc. Workshop ML Syst. Neural Inf. Process. Syst. (NIPS), pp. 1-7, 2017.

[3] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree", Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS), pp. 3146-3154, 2017.

[4] H. B. He and E. A. Garcia, "Learning from imbalanced data", IEEE Transactions on Knowledge and Data Engineering, no. 9, pp. 1263-1284, 2009.

[5] Y. M. Sun, A. K. C. Wong and M. S. Kamel, "Classification of imbalanced data: a review", International Journal of Pattern Recognition and Artificial Intelligence, vol. 23, no. 04, pp. 687-719, 2009.

[6] X. Y. Liu, J. Wu and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning", IEEE Transactions on Systems Man and Cybernetics- Part B, vol. 39, no. 2, pp. 539-550, 2009.

[7] P. Ray and A. Chakrabarti, "A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis", Appl. Comput. Informatics, 2019.

[8] J. Wang, L.-C. Yu, K. R. Lai and X. Zhang, "Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model", Proc. 54th Annu. Meet. Assoc. Comput. Linguist. (Volume 2 Short Pap., vol. 2, pp. 225-230, 2016.

[9] Feng Xia, "Label Oriented Hierarchical Attention Neural Network for Short Text Classification", Academic Journal of Engineering and Technology Science, pp. 5-8, 2022.

[10] J. Kiefer and K. Dorer, "Double Deep Reinforcement Learning," 2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Tomar, Portugal, 2023, pp. 17-22, doi: 10.1109/ICARSC58346.2023.10129640.

[11] Junhong He and Ke Ma. 2021. Enterprise Financial Risk Management and Control. In 2021 2nd Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC2021). Association for Computing Machinery, New York, NY, USA, 393–396. https://doi.org/10.1145/3452446.3452547

[12] Olga Arkadeva and Natalia Berezina. 2021. Digitalization in state financial risk management. In Proceedings of the 2nd International Scientific Conference on Innovations in Digital Economy (SPBPU IDE '20). Association for Computing Machinery, New York, NY, USA, Article 5, 1–7. https://doi.org/10.1145/3444465.3444491

[13] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal Multi-Task Financial Risk Forecasting. In Proceedings of the 28th ACM International Conference on Multimedia (MM '20). Association for Computing Machinery, New York, NY, USA, 456–465. https://doi.org/10.1145/3394171.3413752

[14] Xianping Yuan and Yue Zhang. 2021. Analysis of Bank Loan Risk Management Based on BP Neural Network. In 2021 4th International Conference on Information Systems and Computer Aided Education

(ICISCAE 2021). Association for Computing Machinery, New York, NY, USA, 2457–2461. https://doi.org/10.1145/3482632.3487450

[15] V. T. and J. L. 2019. The Blend of Credit Scoring Model for Individual in the Dmaic Process for Reducing Non-Performing Loan Risk. In Proceedings of the 2019 International Conference on Management Science and Industrial Engineering (MSIE '19). Association for Computing Machinery, New York, NY, USA, 195–202. https://doi.org/10.1145/3335550.3335583