Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5251228, 8 pages
https://doi.org/10.1155/2022/5251228

Research Article

Credit Evaluation of SMEs Based on GBDT-CNN-LR Hybrid Integrated Model

Lei Zhang[1,2] **and Qiankun Song**[2]

[1]*School of Economic and Management, Chongqing Jiaotong University, Chongqing 400074, China*
[2]*School of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Lei Zhang; zhangleicqjtu@163.com

Received 30 December 2021; Revised 19 January 2022; Accepted 21 January 2022; Published 11 February 2022

Academic Editor: Yingjie Wang

Under the background of the increasing demand for credit evaluation and risk prediction, the establishment of an effective credit evaluation model for small- and medium-sized enterprises has become a research hotspot. Based on previous studies, this paper proposes a two-layer feature extraction method based on Gradient Boosting Decision Tree (GBDT) and Convolutional Neural Network (CNN). First, based on the original features, GBDT is used to combine and automatically screen them, the missing values in the feature are processed, and the transformed high-dimensional sparse features are obtained. Then, CNN is used to extract features further, and finally, the logistic regression (LR) model is used to predict. In the simulation experiment, this paper takes a dataset of 14,366 small- and medium-sized enterprise credit evaluations as the analysis samples to verify the results. The results show that the GBDT-CNN-LR model has the best performance. The model also shows good generalization ability and stability in the reliability test.

1. Introduction

For the credit financing of small- and medium-sized enterprises, on the one hand, due to their small scale, high operating, and capital flow risks, financing channels and financing limits will be restricted; on the other hand, the high debt repayment risk and fraudulent behavior of small- and medium-sized enterprises will bring a huge risk of capital loss to the banking industry. How to address the problems of financing difficulties and high credit risks for small- and medium-sized enterprises caused by the asymmetry of information between the two parties to establish a high-precision credit evaluation and prediction model has become the focus of current research.

The SME credit evaluation based on artificial intelligence algorithms has high accuracy and fast speed, which are more often used in the bank credit evaluation business. At the same time, the requirements for the accuracy of the evaluation algorithm are also increasing. Scholars have done extensive research on machine learning algorithms for SME credit classification prediction, including statistical methods, single machine learning algorithms, integrated learning algorithms, and multimodel hybrid integrated learning algorithms [1–4]. Compared with credit evaluation methods based on machine learning algorithms, traditional statistical methods often require more complicated feature engineering in the early stage, which is not only inefficient, but the accuracy of the model is largely affected by the early feature engineering work. The data mining models of machine learning algorithms mainly include artificial neural networks [5–8], support vector machines [9–11], and decision trees [12, 13]. Huang et al. [14] compared the classification accuracy and applicability of several common neural network models. The empirical results show that the probabilistic neural network (PNN) has the lowest classification error rate. Uddin et al. [15] applied the random forest (RF) method to the robust modeling of credit default prediction, which has been proven as an efficient classifier than others. Wang et al. [16] selected appropriate indicators and used an improved SVM model for analysis to be able to

detect the credit risk of SMEs. Luo et al. [17] used a deep learning network and applied a deep belief network with Restricted Boltzmann Machines to credit scoring, which has higher accuracy than that of traditional logistic regression methods. Zhong et al. [18] compared the machine learning training effects of BP, ELM, I-ELM, and SVM, and the results showed that the effects of ELM and BP neural networks are better.

The characteristics of missing values, high dimensionality, and redundancy in the credit evaluation of small- and medium-sized enterprises make it difficult to find the optimal evaluation feature integration of the evaluation classifier, which is also a key factor that leads to the low accuracy of the current evaluation classification. In order to further enhance the evaluation effect, algorithm research based on hybrid integrated machine learning has been innovated and improved for the existing problems so that the integrated model is better than the original model in various evaluation indicators of the predicted results. The RS-PSO-SVM model [19] solves the problem of nonlinear modeling and multicollinearity, which has high accuracy and efficiency. It uses the PSO algorithm to optimize the SVM model parameters and to assess and classify corporate credit risks. Sun et al. [20] combined SMOTE and Bagging to propose the DTE-SBD model, which can not only dispose of the class imbalance problem of enterprise credit evaluation but also increase the diversity of base classifiers for DT ensemble. Ma [21] put forward a hybrid integrated method RS-Boosting based on boosting and random subspace sampling to predict corporate credit risk and verified the effectiveness and feasibility of the method through empirical comparisons. Arora and Kaur [22] used the Bolasso algorithm to select consistent and relevant features from the feature library and applied the generated candidate features to different classification algorithms such as the random forest. The results showed that the BS-RF algorithm has a good performance in the classification accuracy of credit evaluation.

The credit evaluation of SMEs has complex features and high redundancy, and the evaluation data often contain a lot of missing values. Therefore, when using machine learning methods for corporate credit evaluation, high requirements are often placed on the processing of missing data in the early stage, and good feature engineering is also required. However, most of the above models simply remove the redundant features in the metadata and put their subsets into one or several base models for training. However, they do not compare and verify the results of the selected subsets based on different base models. In addition, when the number of feature indicators in the dataset changes, the original model will no longer be applicable.

Aiming at the shortcomings of existing research, this paper proposes a hybrid ensemble model using the GBDT-CNN method for feature extraction to evaluate corporate credit. The model uses the GBDT-CNN method to extract the original data features, which can effectively deal with the missing values of the samples while reducing the difficulty of feature engineering, thereby reducing the assumption of the data missing mechanism and the dependence on the data distribution model, which also has better robustness to abnormal situations in the original data.

2. Enterprise Credit Evaluation Techniques and Procedures

2.1. GBDT Model. Gradient Boosting Decision Tree, based on the idea of Boosting and CART algorithm, is an iterative decision tree algorithm. Except that the first decision tree is generated using the original predictive index, the goal in each iteration is to minimize the loss function of the current learner, that is, to make the loss function always drop along its gradient. Through continuous iteration, it makes the final residual error close to 0. Then by adding up the results of all trees, we can get the final prediction results [23].

The credit risk identification of SMEs is an obvious binary classification problem, which predicts risks through a series of basic corporate information, stocks, capital, investment, income, and other indicators. Let y denote the credit behavior of the enterprise, $y = 1$ denote dishonesty behavior, and $y = 0$ denote nondishonesty behavior. $x = \{x^1, x^2, \ldots, x^K\}$ is a k-dimensional variable composed of a series of basic information of the enterprise. For a training dataset $T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ containing N samples, the GDBT modeling process is as follows:

$$f_0(x) = \arg_c \min \sum_{i=1}^{N} L(y_i, c), \tag{1}$$

where $f_0(x)$ is the initial decision tree with only one root node, y_i is the i-th training data, c is the constant that minimizes the loss function $f(x)$, and $L(y_i, c)$ is the loss function.

In the GBDT model, different loss functions can be used for binary classification problems, but log-likelihood is generally used:

$$L(y, f(x)) = \log(1 + \exp(-yf(x))), \tag{2}$$

where f is the binary classification model to be solved.

Let the number of iterations be $m = 1, 2, \ldots, M$, and then the negative gradient of the i-th training sample is

$$r_{mi} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)} = \frac{y_i}{1 + \exp(yf(x_i))}. \tag{3}$$

According to all samples and their negative gradient directions (x_i, r_{mi}) $(i = 1, 2, \ldots, N)$, a decision tree T_m composed of J leaf nodes is obtained. The j-th leaf node area is R_{mj} $(j = 1, 2, \ldots, J)$, and the best fit value of each leaf node is

$$c_{mj} = \arg\min \sum_{c_{x_i} \in R_{mj}} \log(1 + \exp(y_i f_{m-1}(x_i) + c)). \tag{4}$$

The learners obtained in this round are

$$f_m(x) = f_{m-1}(x_i) + \sum_{i=1}^{N} \sum_{j=1}^{J} c_{mj} I \quad x_i \in R_{mj}, \tag{5}$$

where I is the indicative function of the i-th training sample in the j-th leaf node region and

$$I = \begin{cases} 1, & X_i \in R_{mj}, \\ 0, & X_i \notin R_{mj}. \end{cases} \tag{6}$$

After M rounds of iteration, the final decision model is

$$f(x) = f_M(x) = c + \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I \quad x \in R_{mj}. \tag{7}$$

According to the number of times the variable is selected as the split variable in the regression tree during the iteration process and the degree of improvement of the model during the split process, the importance of each variable can be obtained as

$$\begin{aligned} R_k^2 &= \frac{1}{M} \sum_{m=1}^{M} R_k^2(T_m), \\ R_k^2(T_m) &= \sum_{j=1}^{J} E_j^2 I_j\left(x^k\right), \end{aligned} \tag{8}$$

where T_m is the decision tree trained in the m-th iteration, $I_j(x^k)$ is the k-th variable x^k , which is selected as the indicator function of the j-th leaf node split variable in the decision tree T_m, E_j^2 denotes the improvement of the prediction result when the variable x^k is used as the leaf separate variable, and R_k^2 represents the importance value of the variable x^k in the decision tree.

2.2. CNN. Convolutional Neural Network (CNN) consists of one or more convolutional layers and a fully connected layer, which also includes associated weights layers and pooling layers. CNN's features such as local connection, weight sharing, and pooling processing can effectively reduce network complexity and decrease the number of training parameters. To some extent, they make the model have a certain degree of invariance to translation, distortion, and scaling. While maintaining strong robustness and fault tolerance, it is also easy to train and optimize the network structure [2, 7, 24].

Here, this paper will map the combined feature and feature classification automatically (searched by GBDT) to higher dimensions through the CNN to truly reflect the distribution of the data.

2.3. Logistic Regression. Logistic regression is used for classification problems. The decision boundary can be expressed as $w_1 x_1 + w_2 x_2 + b = 0$, assuming that a certain sample point satisfies the condition $h_w(x) = w_1 x_1 + w_2 x_2 + b > 0$. Then, the category is judged as 1. For the binary classification problem, the given dataset is as follows:

$$\begin{aligned} D = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N), \quad x_i \subseteq R^n, \\ y_i \in 0, 1, \quad i = 1, 2, \ldots, N. \end{aligned} \tag{9}$$

Because the value of $w_T x + b$ is continuous, it is used to fit the conditional probability $p(Y = 1|x)$. However, for $w \neq 0$, the value of $w_T x + b$ is R, and the probability of nonconformity ranges from 0 to 1, so we use a generalized linear model. The unit step function is as follows:

$$p(Y = 1 \mid x) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \quad z = w^T x + b. \\ 1, & z > 0, \end{cases} \tag{10}$$

The step function is not differentiable, and the log probability function is a commonly used substitute function:

$$y = \frac{1}{1 + e^{-(w^T x + b)}}. \tag{11}$$

Then, there are

$$\ln(\text{odds}) = \ln \frac{y}{1 - y}. \tag{12}$$

Regarding y as a class posterior probability estimation,

$$\begin{aligned} w^T x + b &= \ln \frac{P(Y = 1 \mid x)}{1 - P(Y = 1 \mid x)}, \\ P(Y = 1|x) &= \frac{1}{1 + e^{-(w^T x + b)}}. \end{aligned} \tag{13}$$

The output $Y = 1$ log odds is a model represented by a linear function of the input x, that is, a logistic regression model. The closer the value of $w_T x + b$ is to positive infinity, the closer the value of $P(Y = 1|x)$ probability is to 1. Therefore, logistic regression first fits the decision boundary and then establishes the probability link between this boundary and the classification, which gives the probability in the dichotomous case.

3. Enterprise Credit Evaluation Model GBDT-CNN-LR

The samples used by SMEs for credit evaluation often contain a large amount of missing data. The use of machine learning and other methods for credit evaluation has high requirements for the processing of missing data in the early stage. In addition, features of SMEs' credit evaluation have the characteristics of large number, complexity, and high redundancy. Traditional machine learning methods must be based on good feature engineering in the early stage. Therefore, finding the optimal evaluation feature set of the evaluation classifier is the key to improving the accuracy of the evaluation classification. Most of the existing missing value processing methods use certain approaches to fill in data artificially. It is necessary to assume that the dataset obeys a certain distribution model. However, in practical applications, the feature missing data are often intertwined. If the assumptions and models are unreasonable, they will affect the follow-up learning effect of the classifier.

According to the analysis above, if a method adopted can make full use of the information contained in the known dataset, there is no need for the bank and other financial

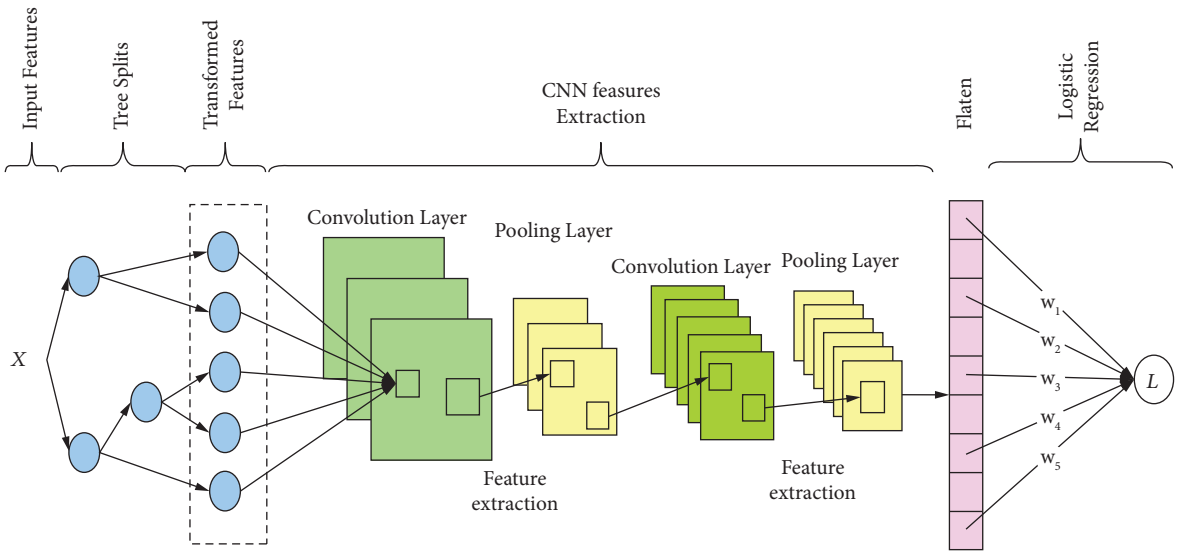

Figure 1: Frame diagram of GBDT-CNN-LR model.

institutions to process the missing data before they classify SMEs' credit, thereby reducing the assumption of the data missing mechanism and the dependence on the data distribution model. Thus, it improves the quality of the evaluation feature set in the evaluation classifier, thereby enhancing the classification accuracy. Therefore, this research mainly focuses on how to simplify the preliminary feature processing for enterprise credit data as much as possible so as to achieve the highest possible discrimination accuracy while realizing feature extraction and feature combination.

This problem can be considered from two aspects: first, compared with human feature engineering, whether the method adopted can reflect the information covered by the original data features so as to ensure the correct rate of subsequent classification of untrustworthy companies; second, whether the adopted method can better adapt to and deal with the outliers and missing values in the data, including whether it is sensitive to the data and whether it can still maintain high accuracy even in the case of massive data distribution.

Therefore, this paper proposes a method based on GBDT-CNN to extract the features of the original data. First of all, it is based on the idea of Boosting. In the GBDT feature generation part, except for the first decision tree generated by the original predictor, the goal of each subsequent iteration will minimize the loss function of the current learner; that is, the loss function always descends along its gradient, and the final residual error tends to 0 through continuous iterations. Finally, the prediction result can be obtained by combining the results of all the trees through a specific aggregation function. Different from the traditional model, this paper uses GBDT as a tool to automatically combine and filter the features of the original data, discover distinguishable features, and generate new feature combinations, thereby retaining the information contained by the original data. In addition, when the loss function is properly selected, GBDT has strong robustness to abnormal conditions in the original dataset and is not sensitive to hyperparameters. It can achieve good prediction accuracy without long-time parameter adjustments. Considering that the original dataset has two types, continuous and discrete values, GBDT can also handle them flexibly without preceding operations, which simplifies the complexity of early feature engineering.

In the GBDT model section, each original data sample will eventually fall on the leaf node of the tree, and after the One-Hot encoding is connected, the transformed high-dimensional sparse feature vector is obtained. This paper then uses CNN with Batch Normalization as a further feature extraction tool to find higher-dimensional features to improve classification accuracy. The specific implementation methods are as follows. First, this paper uses BN to standardize the input data of each layer of the network to ensure that the mean and variance of the input distribution are stable within a certain range. While alleviating the Internal Covariate Shift problem in the network, it also alleviates the disappearance of the gradient to a certain extent and accelerates the convergence of the model. Second, BN makes the network more robust to parameters and activation functions and reduces the complexity of training and tuning of the neural network model. Third, the BN training process uses the Mini Batch mean and variance as the overall sample statistics estimation and introduces random noise. To a certain extent, they have a regularization effect on the model and enhance the robustness of the model.

After extracting the characteristics of the original data through the GBDT-CNN method, the classification model is then used to identify and discriminate the untrustworthy enterprises. LR (logistic regression) is a kind of generalized linear model. The output is the weighted sum of the input features, and the final result is output by the Sigmoid function so that it lies between 0 and 1, which conforms to the meaning of probability. The credit evaluation of an enterprise is to conclude whether to lend or not after comprehensively inspecting various financial and operating indicators of the enterprise. Therefore, the logistic regression

Table 1: Comparison of evaluation indexes of different models.

Model	Accuracy	f1_score	Recall_score
GDBT-LR	0.9349	0.9565	0.9445
GDBT-CNN-LR	0.997	0.9782	0.9558
Random Forest Classifier	0.9491	0.9495	0.9448
Decision Tree Classifier	0.9357	0.9364	0.9354
Logistic regression	0.8177	0.8027	0.7327
SVM	0.5107	0.4439	0.386
MLP	0.5107	0.7404	0.6416
GaussianNB	0.5107	0.7647	0.9636
KNN	0.5107	0.7805	0.8456

model can be better applied to the problem of enterprise credit evaluation, and it is easy to explain the importance of each evaluation index to the final evaluation result.

Based on the analysis and discussion above, this paper aims to establish a GBDT-CNN-LR-based credit risk assessment model for SMEs. The frame diagram is shown in Figure 1.

For the use of integrated learning methods for enterprise credit evaluation, we need to consider two factors: (1) whether the model can effectively identify untrustworthy companies from the sample, that is, the accuracy requirements; (2) whether the weak learning model of the model can produce a difference, to avoid the degradation of the model effect, that is, the requirement of diversity. Regarding the first question, using the GBDT-LR model to solve the prediction of Facebook ad clicks in previous studies, the GBDT-LR model can better solve the prediction problem and achieve higher accuracy, which is sufficient to explain that the GBDT-CNN-LR model has a certain application basis, and it is possible to achieve certain recognition accuracy. For the second aspect, GBDT draws on the idea of Boosting in the training process. Every training reduces the residual of the previous training model so that the residual is reduced in the gradient direction, and each classification tree constructed reduced the error in the previous step. Thus, GBDT pays more attention to those samples with larger gradients. It can be considered that each classification decision tree constructed afterward only pays attention to some of its subsamples. Compared with the forecast of ad clicks, enterprise credit evaluation requires a higher accuracy rate. If the evaluation result is wrong, it may cause huge economic losses to the bank. In actual experiments, the traditional GBDT-LR model is still difficult to achieve the expected high accuracy rate. The accuracy rate of LR is limited by the previous feature engineering. Therefore, this paper proposes to use CNN on the basis of the feature vector generated by GBDT. The intention is to find higher-dimensional features as input data to improve the prediction accuracy of LR regression.

4. Experiments and Discussion

4.1. Datasets. The experimental dataset contains the credit records of 14,366 small- and medium-sized enterprises and 14 characteristics, including company stock price, foreign investment, registered capital, corporate assets, income,

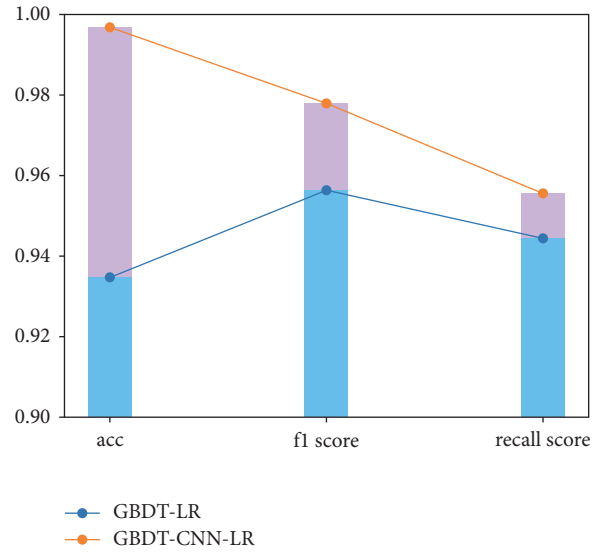

Figure 2: Before and after adding CNN convolution.

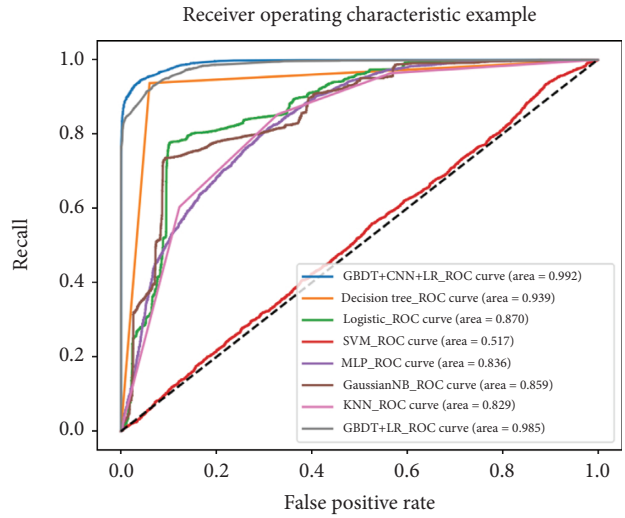

Figure 3: ROC_AUC curve.

expenses, liabilities, and taxation, which are selected as the credit evaluation indicators of small- and medium-sized enterprises.

4.2. Evaluation Index. The accuracy is used as the most important evaluation index, that is, the number of samples that are predicted correctly divided by the total number of samples, and the f1_score coefficient and recall_score are used as auxiliary evaluation indicators.

4.3. The Result of the Experiment. First of all, this paper conducts statistical analysis on the missing values of each feature in the sample set. Most of the features in the sample set used in this paper have 60% or more missing data, which verifies the universality of the problem that this paper aims to solve. Therefore, this paper uses the proposed GBDT-CNN model to search for the distribution and information

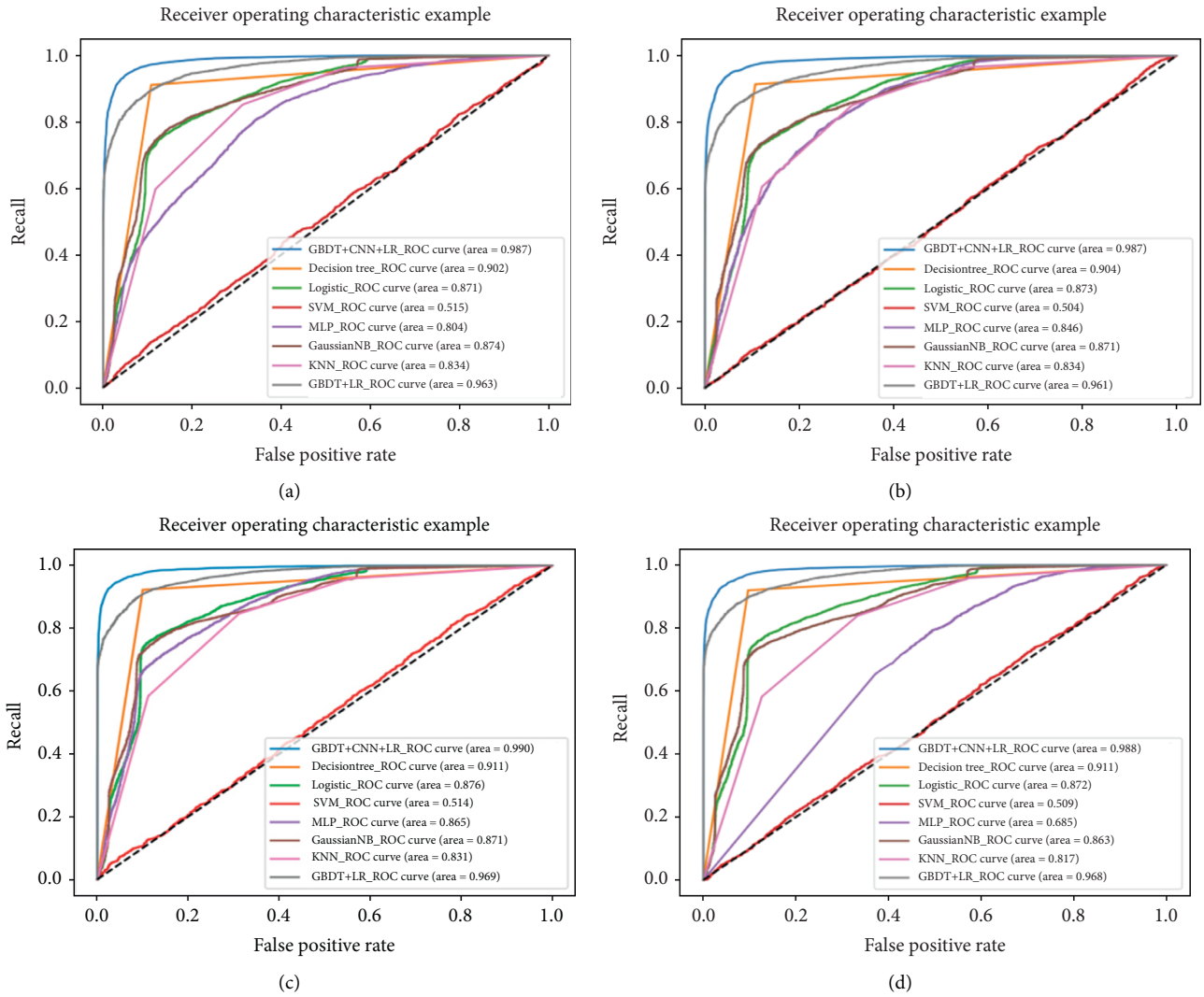

Figure 4: ROC_AUC graph of the subdataset. (a) Sample (a). (b) Sample (b). (c) Sample (c). (d) Sample (d).

of the data itself and automatically fill in the missing data. The new feature vector generated is substituted into the Logistic model as an input index to output the discrimination result.

First, compare the evaluation effects of the single model and the integrated model, and the results are shown in Table 1.

It can be seen from Table 1 that both the tree model and the logistic regression model can achieve better prediction accuracy, but the prediction accuracy rate of the SVM, MLP, NB, and KNN models is only 51.07%. The three evaluation indicators (accuracy, f1_score, and recall_score) of the model after adding CNN to extract features are higher than those of other models.

When CNN has not been added to models to extract features, the effects of random forest, decision tree, and GBDT are significantly better than those of the Logistic model. Since logistic regression is a linear model, random forest, decision tree, and GBDT are all nonlinear models. And they perform better than logistic regression on many nonlinear datasets and linear datasets. Therefore, the linearity of the Logistic model itself limits the predictive ability of the model to explain this phenomenon reasonably.

This paper uses the GBDT model to extract features and then adds the Logistic model for classification, and the prediction accuracy is 93.49%, which is worse than that of a single model such as random forest and decision tree. Therefore, this paper considers further optimization of the model. Since the features automatically filtered out by the GBDT model have high dimensionality and large sparseness, this paper first uses CNN to convolve and sum the features obtained by GBDT and move them from a highly sparse space to a reasonably sparse space, which not only satisfies the certain sparsity required by logistic regression but also maintains the difference between each feature.

The experiment shown in the following figure compares the evaluation effect of the GBDT-CNN-LR model with CNN and that without CNN.

It can be seen from Figure 2 that, after adding CNN to extract features, compared with the GBDT-LR model without adding CNN to extract features, the accuracy is

increased by 4.6%. In addition to the evaluation indicators above, the ROC_AUC curve can more accurately judge the performance of the GBDT-CNN-LR model by the AUC area. Therefore, this paper draws the ROC_AUC curve of different models. As shown in Figure 3, GBDT-CNN-LR's AUC area is 0.992, which is larger than the AUC area of other models. Therefore, it can be considered that the GBDT-CNN-LR model that joins CNN to extract features is reasonable and has higher prediction accuracy for evaluating the credit risk of small- and medium-sized enterprises.

The missing values of the sample data account for a relatively large amount, reaching 42.6% of the total dataset. Using GBDT-CNN to automatically fill missing values has achieved high prediction accuracy, but if the new data does not fit the sample model, the model is very likely to be unstable. Therefore, this paper tests the stability of the model.

The dataset is divided into 4 parts, and each dataset retains the same missing rate as the original dataset. Then, we train each small dataset and draw the corresponding ROC_AUC curve graph, compare the AUC area of the model, and judge the stability of the model. The results are shown in Figure 4.

The results show that the prediction accuracy of the support vector machine model is still poor, and the multilayer perceptron (MLP) fluctuates sharply. The reason may be that the neural network is more sensitive to data, there is too little data, or there are too many missing values. Thus, the training of a neural network has a large error. The AUC area of the GBDT-LR model without the CNN channel showed a downward trend of about 2%–3%, but the AUC area of the GBDT-CNN-LR model using the CNN channel almost did not decrease. Therefore, the GBDT-CNN-LR model can show good generalization ability and stability on both large datasets and small datasets. The GBDT-LR model without the CNN channel also has good generalization ability and stability, but they are lower than those of the GBDT-CNN-LR model numerically.

5. Conclusions

The application of SME credit evaluation based on artificial intelligence algorithms in the bank credit evaluation business is becoming more and more extensive; thus, the accuracy of the evaluation model and algorithm also puts forward higher requirements. This paper proposes the GBDT-CNN-LR evaluation model. The model first uses GBDT to automatically combine and filter the original data features, which can better deal with problems such as the concentration of missing indicator values, and obtain transformed high-dimensional sparse feature vectors. Then, on the basis of the feature vector generated by GBDT, CNN is used for further feature extraction, and finally, these higher-dimensional features are predicted by logistic regression. In the simulation experiment, compared with the Random Forest Classifier, Decision Tree Classifier, Logistic Regression, SVM, and other basic classification algorithms, it can be clearly seen that the accuracy of the GDBT-CNN-LR model is higher than other models. In addition, the model shows good generalization ability and stability in the reliability test, which can effectively reduce the risk of investment and provide reliable technical support for financial institutions, accordingly possessing far-reaching practical significance.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the Group Building Scientific Innovation Project for Universities in Chongqing (CXQT21021) and the Science and Technology Research Project of Chongqing Education Commission (KJQN202100712).

References

[1] Y. Zhu, C. Xie, G. J. Wang, and X. G. Yan, "Comparison of individual ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance," *Neural Computing and Applications*, vol. 28, no. 1, pp. 41–50, 2017.

[2] Z. P. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[3] X. Cai, Y. Qian, Q. Bai, and W. Liu, "Exploration on the financing risks of enterprise supply chain using Back Propagation neural network," *Journal of Computational and Applied Mathematics*, vol. 367, Article ID 112457, 2020.

[4] Y. Wang, Y. Gao, Y. Li, and X. Tong, "A worker-selection incentive mechanism for optimizing platform-centric mobile crowdsourcing systems," *Computer Networks*, vol. 171, Article ID 107144, 2020.

[5] J. P. Bigus, *Ata Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support*, McGraw-Hill, New York, NY, USA, 1996.

[6] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial IoTs," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.

[7] Y. Wang, Z. Cai, Z.-H. Zhan, B. Zhao, X. Tong, and L. Qi, "Walrasian equilibrium-based multiobjective optimization for task allocation in mobile crowdsourcing," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 1033–1046, 2020.

[8] Z. Lu, Y. Wang, Y. Li, X. Tong, C. Mu, and C. Yu, "Data-driven many-objective crowd user selection for mobile crowdsourcing in industrial IoT," *IEEE Transactions on Industrial Informatics*, 2021.

[9] S. Andaryani, V. Nourani, A. T. Haghighi, and S. Keesstra, "Integration of hard and soft supervised machine learning for flood susceptibility mapping," *Journal of Environmental Management*, vol. 291, Article ID 112731, 2021.

[10] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Proceedings of the 2013 fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–7, IEEE, Tiruchengode, India, 2013.

[11] L. Xu, L. Cui, T. Weise, X. Li, Z. Wu, and F. Nie, "Semi-supervised multi-layer convolution kernel learning in credit evaluation," *Pattern Recognition*, vol. 120, 2021.

[12] Z. Liu and Y. Zhang, "Credit evaluation with a data mining approach based on gradient boosting decision tree," *Journal of Physics: Conference Series*, vol. 1848, no. 1, p. 8, Article ID 012034, 2021.

[13] L.-A. Dong, X. Ye, and G. Yang, "Two-stage rule extraction method based on tree ensemble model for interpretable loan evaluation," *Information Sciences*, vol. 573, pp. 46–64, 2021.

[14] X. Huang, X. Liu, and Y. Ren, "Enterprise credit risk evaluation based on neural network algorithm," *Cognitive Systems Research*, vol. 52, pp. 317–324, 2018.

[15] M. S. Uddin, G. Chi, M. A. Al Janabi, and T. Habib, "Leveraging random forest in micro-enterprises credit risk modelling for accuracy and interpretability," *International Journal of Finance & Economics*, 2020.

[16] F. Wang, L. Ding, H. Yu, and Y. Zhao, "Big data analytics on enterprise credit risk evaluation of E-business platform," *Information Systems and E-Business Management*, vol. 18, pp. 1–40, 2019.

[17] C. Luo, D. Wu, and D. Wu, "A deep learning approach for credit scoring using credit default swaps," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 465–470, 2017.

[18] H. Zhong, C. Miao, Z. Shen, and Y. Feng, "Comparing the learning effectiveness of BP,ELM,I-ELM,and SVM for corporate credit ratings," *Neurocomputing*, vol. 128, no. 27, pp. 285–295, 2014.

[19] X. Hu, J. Hu, L. Chen, and Y. Li, "Credit risk assessment model for small, medium and micro enterprises based on RS-PSO-SVM integration," in *Proceedings of the 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)*, pp. 342–345, Chengdu, China, 2021.

[20] J. Sun, J. Lang, H. Fujita, and H. Li, "Imbalanced enterprise credit evaluation with DTE-SBD: decision tree ensemble based on SMOTE and bagging with differentiated sampling rates," *Information Sciences*, vol. 425, pp. 76–91, 2018.

[21] G. Wang and J. Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13871–13878, 2011.

[22] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: an application to credit risk assessment," *Applied Soft Computing Journal*, vol. 86, pp. 1–29, 2019.

[23] Z. Tian, J. Xiao, H. Feng, and Y. Wei, "Credit risk assessment based on gradient boosting decision tree," *Procedia Computer Science*, vol. 174, pp. 150–160, 2020.

[24] Y. LeCun, B. Boser, J. S. Denker et al., "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.