Activity Prediction of Construction Enterprise based on Semi-supervised Learning

Jinhao Zhong[1], Zhen Yu[2], Xinyu Yao[3]

2362765482@qq.com[1], 28431320@qq.com[2],1157422893@qq.com[3]

School of Computer and Network Security, Chengdu University of Technology, Chengdu, China[1],
School of Computer and Network Security, Chengdu University of Technology, Chengdu, China[2],
School of Mathematics and Science, Chengdu University of Technology, Chengdu, China[3]

Abstract. In this paper, the idea of semi-supervised learning is used to propose a method for predicting the activity of construction enterprises based on the SOM-GMM-RF model. This prediction method can help enterprises understand their activity and operation status and make reasonable decisions. The goal is to address the issue that the cost of data annotation is too high in this study, and the prediction effect of the classification model is unsuitable under the condition of only a small amount of annotated data. According to the experimental data, the random forest model trained using the methodology in this study achieves an accuracy of 93.33%, which is 4.29% greater than the random forest model trained with a labeled training set. Additionally, the model has the best prediction effect in the experiment compared to other classification algorithms. The experiment reveals that the method in this paper can effectively use unlabeled data to improve the prediction effect of the model, even when only a small amount of data is labeled. This finding is important for research on predicting construction enterprise activity.

Keywords: construction enterprise activity; self-organizing mapping network; gaussian mixture model; random forest.

1 Introduction

Currently, there is an increasing number of surveys and analyses of enterprise activity. Market regulators in several provinces and cities, such as Nanjing, Guangdong Province, and Shandong Province, have advocated the development of enterprise activity analysis, which has already carried out activity analysis of local enterprises. The analysis of enterprise activity will be applied in more cities and industry sectors.

A number of scholars have conducted relevant studies on enterprise activity analysis. Xiaopeng Luo et al. proposed putting forward the definition of enterprise activity and the corresponding evaluation method based on the quality of life theory, constructed the enterprise big dataset, and validated the results by using two measurement methods[1].Zhang Zhao et al. introduced the "double critical value" method into the enterprise activity measurement method, constructed an index model, and measured the enterprise business activity from three dimensions: business activity, innovation activity, and information sharing[2].With the help of fixed, moderated, and mediated effects models, Lv Lin et al. found that Chinese-style financial decentralization helps to enhance the innovation activity of high-tech firms[3].Tao Lin engineers existing enterprise communication and operation data construct enterprise

BDEDM 2024, January 12-14, Ningbo, People's Republic of China

DOI 10.4108/eai.12-1-2024.2347236

communication activity indicators and trains a CART decision tree model to discriminate enterprise communication activity[4].

There is little machine learning-based enterprise activity prediction research specifically for construction enterprises; instead, most existing research focuses on the design of evaluation techniques, the construction of measurement models for analysis, and the selection of indicators.

Currently, the construction sector is experiencing significant growth, with a dynamic market environment and escalating competition among construction businesses. In order to facilitate enterprises in comprehending their operational performance and making informed decisions, assist government departments in promoting enterprise development and assessing the market landscape, and enable the general public to gauge the growth potential of enterprises and make decisions aligned with their individual requirements, it is imperative to conduct research on activity prediction for construction enterprises utilizing machine learning algorithms.

2 Related theories

2.1 Semi-supervised learning

Semi-supervised learning has gradually become an important research area in machine learning with many applications.Semi-supervised learning techniques that rely on classification, regression, clustering, and dimensionality reduction can be categorized depending on various learning scenarios[5]. This paper primarily employs the semi-supervised classification learning approach to effectively utilize a limited number of labeled data in conjunction with a substantial amount of unlabeled data for model training. The objective is to enhance the model's performance and address the issue of suboptimal results caused by the scarcity of labeled data.

2.2 Self-organizing mapping

A self-organizing mapping network is an unsupervised artificial neural network that operates through competitive learning[6]. The self-organizing mapping network resists noisy data and adopts a competitive learning strategy to generate a low-dimensional mapping by learning the data in the input network and mapping the samples to different competing neurons. Figure 1 displays the topology of the Self-Organizing Mapping network, comprising an input layer and an output layer. The input layer encompasses n neurons, while the output layer, the competition layer, comprises neurons organized in a two-dimensional matrix configuration.

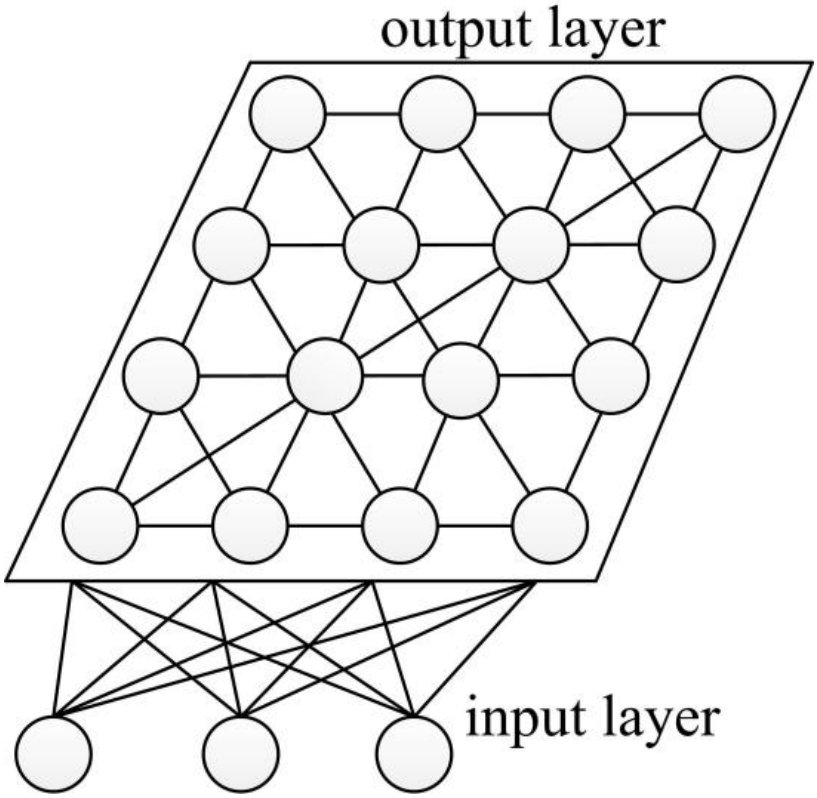

Fig. 1. Self-organizing mapping network structure.

2.3 Gaussian mixture model

The Gaussian mixture model is a clustering approach extensively employed in various domains [7]. The Gaussian Mixture Model is classified as a generative model that uses several Gaussian probability distributions to characterize the data distribution. It can model intricate continuous distributions effectively by utilizing numerous Gaussian model components.

2.4 Random forest

Random Forest Algorithm is a supervised machine learning algorithm based on decision trees and integrated learning[8]. It adopts the idea of Bagging integrated learning, taking the classification regression decision tree as a weak learner, firstly, adopting the Bootstrap self-service sampling method to randomly extract multiple sample data from the data set with put-back, constituting multiple training sets and test sets to train multiple mutually independent decision trees to form a Random Forest; secondly, in the process of the construction of the decision tree, randomly selecting some of the features of the training set to constitute a random feature subset, as the split feature set of the node; finally, the final prediction result of the random forest is obtained by voting. The process of random forest is shown in Figure 2.

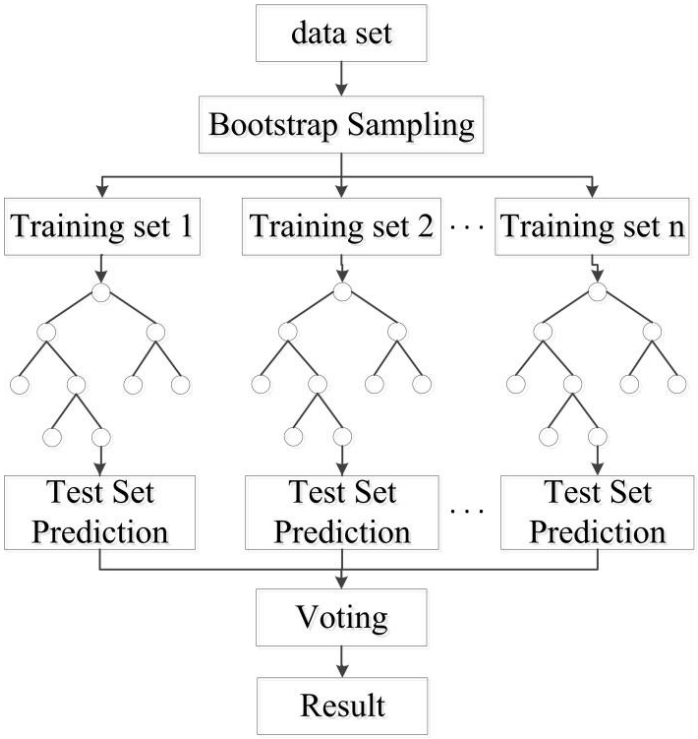

Fig. 2. Random forest processes.

3 Activity prediction method design

The design flow of the prediction approach is comprised of two primary components, as seen in Figure 3.The first part uses the SOM-GMM clustering algorithm to label unlabeled data. Firstly, a small number of labeled datasets are input into the SOM network after data normalization to obtain preliminary clustering results; then, the weights of each competing neuron in the clustering results are used as the clustering center of the GMM and normalized for both labeled and unlabeled data; finally, the Gaussian mixture model implements clustering, and the unlabeled data with the a posteriori probability of greater than 0.9 are selected according to the clustering results. The unlabeled data are labeled with pseudo-labels to form a pseudo-labeled dataset and complete the class classification of unlabeled data.The second part is to train the random forest classification model and parameter optimization and use the test set to evaluate the model prediction effect. Firstly, a small number of labeled datasets are divided into a labeled training set and test set, and pseudo-labeled datasets are added to the tagged training set to form a semi-supervised training set. The expanded semi-supervised training set is used to train the Random Forest classification model, and the Bayesian optimization method combined with five-fold cross-validation is used to obtain the optimal parameters and optimal model; finally, the prediction effect of the model on the test set is evaluated by using a variety of indicators.

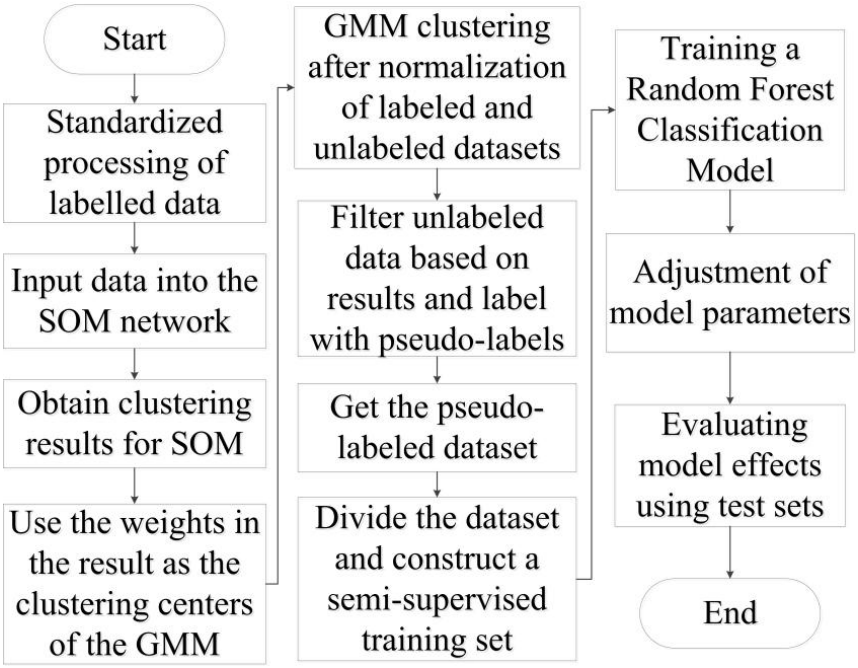

Fig. 3. Activity prediction process.

4 Experiment

4.1 Experimental data

Construction enterprises in Sichuan Province were selected as the research object, and a variety of information about the enterprises on the Enterprise Search and Love Enterprise Search websites was automatically crawled by writing programs. According to the collected data, six influencing factors, namely, the number of change records, the number of administrative licenses, the number of tenders, the number of qualification certificates, the number of construction industry qualifications, and the total number of personnel registration certificates, are selected as the indicators to analyze the activity of enterprises. For the information on six dimensions of enterprises, six kinds of data are counted after removing duplicated and invalid data, and the original data set is formed after data preprocessing, with a total of 3239 data samples.

Referring to the index weights in the enterprise activity analysis index system issued by the former State Administration for Industry and Commerce and combining the importance of construction enterprises to the data volume of the above six indexes as well as the experience of the experts, the enterprise activity is classified into three categories of high, average, and low, which are labeled as 2, 1, and 0, respectively, and 1,049 pieces of data from the sample dataset are marked to form the labeled dataset. In contrast, the remaining samples create the unlabeled dataset. The labeling of some of the data is shown in Table 1, where F1 to F6 represents the six features of the data in turn.

Table 1. Example of labeled data.

F1	F2	F3	F4	F5	F6	Label
11	6	2	3	5	2	0
18	6	4	2	2	11	0
63	27	62	3	12	70	1
76	16	34	4	20	69	1
118	46	125	3	20	232	2
104	253	28	6	16	156	2

4.2 Model training and parameter optimization

We used a division scheme during our experiments where 80% of the labeled dataset was divided into a labeled training set. In contrast, the remaining 20% was designated as the test set. Based on the process of predicting the activity of construction enterprises using SOM-GMM-RF, 2129 unlabeled data were labeled with pseudo-labels. Subsequently, the pseudo-labeled data were incorporated into the tagged training set to form a semi-supervised training set. After constructing the random forest classification model, the model was trained using the semi-supervised training set.

Bayesian optimization is an efficient global optimization algorithm for finding the global optimal[9]. Compared to grid search tuning, Bayesian optimization can tune multiple parameters simultaneously, with fewer iterations of the algorithm and faster operation[10].K-fold cross-validation is a technique that partitions the dataset into numerous subsets, where one subset is designated as the test set, and the remaining subsets are used as training sets to train the model for prediction purposes[11]. To improve the performance of the random forest classification model, the parameters of the classification model are adjusted using the Bayesian optimization technique combined with the five-fold cross-validation method, and the range of parameter values of the model and the tuning results are shown in Table 2.

Table 2. Parameter adjustment results.

Parameter name	Parameter range	Parameter result
n_estimators	(100,200)	151
criterion	["gini","entropy"]	"gini"
max_depth	(3,10)	9
max_features	['auto', 'sqrt','log2']	'sqrt'
min_samples_split	(2,10)	3
min_samples_leaf	(1,10)	1

4.3 Experimental results and analysis

The experiments involve selecting the ideal combination of parameters, as mentioned earlier. The model undergoes training using a semi-supervised training set, followed by the utilization of a test set to evaluate the predictive performance of the model. Based on the conducted experiments, it was determined that the model achieved an accuracy of 93.33%, precision of

85.20%, recall of 93.47%, and F_1 score of 0.8815.In order to verify the effectiveness of this paper's algorithm to improve the model prediction effect using unlabeled data, the random forest classification model under supervised learning is constructed, and the labeled dataset is used for model training and parameter tuning. The comparison of the prediction effect of the models trained using the two methods on the test set is shown in Table 3. The table shows that the algorithm in this paper is more effective, compared with the accuracy, precision, recall, and F_1 score improved by 4.29%, 7.41%, 13.43%, and 0.093, respectively. It proves the algorithm can effectively utilize unlabeled data to improve the model prediction effect.

Table 3. Experimental results.

Model	Accuracy	Precision	Recall	F_1 score
SOM-GMM-RF	93.33%	85.20%	93.47%	0.8815
RF	89.04%	77.79%	80.04%	0.7885

In order to further evaluate the effectiveness of the prediction of construction enterprise activity achieved by this paper's algorithm, the GBDT, XGBoost, and Light GBM classification models are trained using a semi-supervised training set. Then, Bayesian optimization is used to regulate the parameters and the final prediction results of the four models on the test set, as shown in Table 4. Table 4 shows that the random forest model used in this paper's method has the highest accuracy, precision, recall, and F_1 score. The accuracy is improved by 1.91% compared to GBDT, 1.43% to Light GBM, and 0.95% to XGBoost.The F_1 score is enhanced by 0.0238 compared to GBDT, 0.0181 to Light GBM, and 0.0123 to XGBoost.The results of the comparative experiments confirm that this paper's algorithm is effective in the prediction problem of construction enterprise activity, and the prediction effect is better than that of other algorithms.

Table 4. Effects of the four models.

Model	Accuracy	Precision	Recall	F_1 score
SOM-GMM-RF	93.33%	85.20%	93.47%	0.8815
SOM-GMM-GBDT	91.42%	82.81%	92.54%	0.8577
SOM-GMM-Light GBM	91.90%	83.35%	92.77%	0.8634
SOM-GMM-XGBoost	92.38%	83.93%	93.00%	0.8692

5 Conclusions

In this paper, we introduce a semi-supervised learning-based method for predicting the activity of construction enterprises and construct a SOM-GMM-RF model for activity prediction. The main work includes collecting enterprise data, selecting indicators to construct the enterprise activity dataset, and then manually labeling a small amount of data and splitting it into a labeled training set and test set; labeling part of the unlabeled data based on the SOM-GMM clustering algorithm, and adding it to the training set to constitute a semi-supervised training set; A random forest classification algorithm combined with Bayesian optimization method and five-fold cross-validation method is used to train a model using the optimal parameters,

and the prediction effect of the model is evaluated and compared through the test set to realize the multi-classification prediction of the activity of construction enterprises.

Through the experiment, it can be concluded that compared with the random forest model trained using the labeled training set, the prediction effect of the model trained using the method of this paper has a certain degree of improvement and can be used to expand the training set with unlabeled data under the condition of only a small amount of labeled data, to improve the prediction effect of the model. Compared with the models constructed by other classification algorithms, the algorithm model in this paper has a better prediction effect, which can provide an accurate and practical reference for the research on the prediction of the activity of construction enterprises.

References

[1] Luo, X.P., Qi, J.Y., Fu, X.L.(2018)Research on enterprise activity evaluation method for big data regulation. Enterprise Economy (07), 120-128. doi:10.13529/j.cnki.enterprise.economy.2018.07.017.

[2] Zhao, Z., Li, A.Y., Zhu, J.X.(2017)Research on the measurement of business activity in the environment of "Internet+". Statistics and Information Forum (10), 76-83.

[3] Lv, L., Liu, R.Z.(2023)Chinese-style financial decentralization and innovation activity of high-tech firms-theoretical mechanism and empirical test. China Circulation Economy (07), 67-77. doi:10.14089/j.cnki.cn11-3664/f.2023.07.007.

[4] Lin, T.(2020)Analysis of Enterprise Communication Activity Based on CART Decision Tree. Network Security Technology and Application (04), 53-54.

[5] Li, Y.G., Xu, C.Y., Tang, X., Li,X.Y.(2023)A review of research on semi-supervised learning methods. World Scientific and Technological Research and Development (01), 26-40. doi:10.16507/j.issn.1006-6055.2022.07.001.

[6] Nan,F., Li,Y., Jia, X.Y., Dong, L.Y., Chen Y.J.Application of improved SOM network in gene data cluster analysis[J]. Measurement,2019,145.

[7] Zhu, X.Z., Gao, D.L., Yang, C., Yang, C.J.(2023)A blast furnace fault monitoring algorithm with low false alarm rate:Ensemble of greedy dynamic principal component analysis-Gaussian mixture model.Chinese Journal of Chemical Engineering(05),151-161.

[8] Jiang, X.L., Xiong, Y.L., Guo, H.M., Zhao, Z., Zhang, Y., Meng, Y.T. (2023)Application of improved random forest model in population spatialization. Mapping Bulletin (06), 155-160. doi:10.13474/j.cnki.11-2246.2023.0186.

[9] Cui, J.X., Yang, B.(2018)A review of Bayesian optimization methods and applications. Journal of Software (10), 3068-3090. doi:10.13328/j.cnki.jos.005607.

[10] Ding, J.L., Sun, Y.(2021)LightGBM-based multiclassification prediction of flight delays. Journal of Nanjing University of Aeronautics and Astronautics (06), 847-854. doi:10.16356/j.1005-2615.2021.06.003.

[11] Wen, B.M., Zhao, L.W., Huang, L.(2022)Proof of asymptotic equivalence for cross-validation of AIC criterion and leave-one-out method. Statistics and Decision Making (06), 40-43. doi:10.13546/j.cnki.tjyjc.2022.06.008.