Design and Study of Zombie Enterprise Classification and Recognition Systems Based on Ensemble Learning

Shutong Pang
School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, China
1806334268@qq.com

Ziwei Yang
School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, China
2940289012@qq.com

Chengyou Cai
School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, China
chengyou-cai@foxmail.com

Zhimin Li
School of Medical Technology and Information Engineering, Zhejiang Chinese Medical University, China
lzm@zcmu.edu.cn
corresponding author

Abstract: *The existence of a large number of zombie enterprises will affect the economic development and hinder the transformation and upgrading of economic industries. To improve the accuracy of zombie enterprise identification, this paper takes multidimensional enterprise data as the original data set, divides it into training set and validation set, and gives the corresponding data pre-processing methods. Combined with 14 standardized features, an integrated learning model for zombie enterprise classification and recognition is constructed and studied based on three pattern recognition algorithms. By using the idea of integration and the cross-validation method to determine the optimal parameters, the Gradient Boosting Decision Tree (GBDT), linear kernel Support Vector Machine (SVM) and Deep Neural Network (DNN) algorithms with classification accuracies of 95%, 96% and 96%, respectively, are used as sub-models, and a more comprehensive strong supervision model with a classification accuracy of 98% is obtained by the stacking method in combination with the advantages of multiple sub-models to analyze the fundamental information of 30885 enterprises. The study improves the accuracy of zombie enterprise identification to 98%, builds enterprise portraits based on this, and finally visualizes the classification results through the platform, which provides an auxiliary means for zombie enterprise classification and identification.*

Keywords: *Integrated learning, corporate portrait, classification and identification, strong supervision model.*

Received July 12, 2021; accepted December 12, 2022
https://doi.org/10.34028/iajit/20/5/3

1. Introduction

A zombie enterprise is an economic concept put forward by Kane [5]. This type of enterprise loses its self-development ability and must rely on nonmarket factors, namely, government subsidies or bank loan renewals, to survive. Zombie enterprises may easily emerge from industries with overcapacity and low-end industries. Although zombie enterprises do not arise as a result of other zombie enterprises, they still occupy land, utilize labor and have capital and other essential resources, which seriously hinders the growth of new technology and new industries. The accurate and automatic classification and identification of zombie enterprises is conducive to economic and social development [6].

At present, domestic scholars researching the classification and identification of zombie enterprises mainly focus on the analysis of evaluation rules and enterprise fundamentals. Xiaoyang and Qiang [10] considered the relationships between zombie enterprises and the natures of controlling shareholders and industries with overcapacity and built an empirical model by introducing liquidity, leverage, scale, cash flow adequacy and profitability. Shaoqing and Yan [7] modified the Fukuda Nakamura (FN) identification method and provide a comprehensive method to identify zombie enterprises. Xiaoyan [9] introduced the identification standards of the State Council of China and the National Development Research Institute of Renmin University of China on the basis of the Caballero Hoshi Kashyap (CHK), FN and HK methods and added new quantitative indicators such as enterprise Research And Development (R & D) investment to build a zombie enterprise identification model in line with China's national conditions. The above studies verify the usability of traditional evaluation models such as the price sensitivity model, and the CHK, FN-CHK and continuous loss methods. At the same time, this review also points out that these methods or standards are based on only some indicators and have the limitation of weak generalization ability in practical applications [2].

Zombie enterprise classification and recognition problems can be abstracted as dual classification

problems. There are many pattern recognition algorithms for dual classification problems, including Decision Trees (DTs), Support Vector Machines (SVMs), and Multilayer Perceptron (MLPs). Foreign scholars such as Bargagli-Dtoffi *et al*. [1] have defined zombie enterprises empirically through the financial and production resource allocation of enterprises and realized a Bayesian additive regression tree with missing merger attributes under the background of enterprises' undisclosed financial accounts. However, the first developed mock exam is difficult to describe accurately in a modern enterprise because of its complicated operating environment and the basic information of enterprises. The relationships between variables are hard to describe accurately. Because of the strong adaptability of the pattern recognition model, in the actual construction of the botnet classification model, the single model easily causes overfitting.

Ensemble learning is the first mock exam strategy to combine different sub-models by rule [3]. It can improve the generalization ability of the resulting model by using the differences between sub-models, improve the error through communication between the sub-models, and prevent the single model's feature preference from causing overfitting. This paper proposes a classification and recognition method for zombie enterprises based on ensemble learning. This method adopts the idea of integration and combines three supervision models, a Gradient Boosting Decision Tree (GBDT), a linear kernel SVM, and a Deep Neural Network (DNN) through the stacking method to obtain a more comprehensive strong supervision model [8]. According to the basic information about 30885 enterprises from 2015 to 2017 provided by a competition platform, 14 features are extracted to train the integrated verification model. The results show that the ensemble learning model combines the advantages of different pattern recognition models and has higher accuracy and generalization ability than a single model. Based on this, this paper focuses on the accumulation of enterprise big data, data acquisition and processing, the extraction of feature tags, the construction of the model, enterprise portraits and the recognition of the classification results through a visual interface.

2. Dataset Construction and Processing

2.1. Dataset Construction

The data needed for the construction of enterprise portraits are composed of static data and dynamic data [4]. Static data mainly include registered capital, financing amounts, financing costs, total profits, total operating incomes, etc., There are three main sources of data: first, the data of enterprise operations; second, the data of security disclosures; and third, the data obtained by using open-source web crawler tools. Dynamic data refers to the behavior of users in a visual interface, including data imports. Finally, the acquired data are summarized into a dataset. More than 20000 pieces of basic enterprise information data, tax credit data, listing information data, etc., are obtained from the platform.

2.2. Data Preprocessing

The index features include basic information, intellectual property information, financial information and financing information for three consecutive years. The diversity of data sources results in differences in data quality. To ensure high data quality, it is necessary to pre-process the original collected data through data dimensionality reduction, data cleaning, data normalization, mean value substitution and time serialization.

2.2.1. Data Dimensionality Reduction

Data dimensionality reduction is divided into two steps. First, because the collected data include financial information and financing information for three consecutive years, it is necessary to flatten the data for three consecutive years according to the enterprise IDs to reduce the overlapping financial and financing characteristic data in a given year. Second, because the data come from multiple channels and the data from different sources are stored in different data frames, it is necessary to fuse the data frames through the associated enterprise IDs to reduce the number of redundant enterprise IDs.

2.2.2. Data Cleaning

For non-fixed distance data such as “patent”, “trademark”, “copyright”, “registration time”, “industry”, “region”, “enterprise type” and “controller type”, the missing values are interpolated by the mode according to the data in the column of the corresponding feature.

For “financing amount” and “financing cost”, if one item is zero and the other item is missing, the missing item is set to zero. If both items are missing, both items are set to zero. If one item is a non-zero real number and the other is missing, the missing items are calculated and filled in by using the following equation:

$$\frac{\text{Amount of financing}_j}{\text{financing cost}_j} = k_j,$$

$$j \text{ is the financing amount and cost of item } j \ (j = 1, 2, \ldots, \text{n}) \quad (1)$$

$$k_1 \approx k_2 \approx \cdots \approx k_n \approx K, \ K \text{ is a constant}$$

For “main business income” and “total business income”, the proportions of the two data values are calculated by line, the average proportion is calculated, and the average proportion is used to interpolate the missing items of “main business income” and “total business income”:

$$\bar{r} = \left(\sum_{i=1}^{n} \frac{main\ business\ income_i}{total\ operating\ revenue_i}\right)/n,$$

$$\begin{gathered} i\ denotes\ the\ data\ in\ line\ i\ (i = 1, 2, \ldots, n) \\ Nan(main\ business\ income_i) \\ = \bar{r} \times total\ operating\ revenue_i \end{gathered} \quad (2)$$

After the above data cleaning steps are completed, the missing items in the characteristic data, such as the "number of employees", "total assets", "total liabilities", "owner's equity", "total profit", "registered capital" and "shareholding ratio of controller", are interpolated by mean values.

2.2.3. Data Normalization

The Z-score method is used to normalize the data, where x is the vector of a column of feature data [14]:

$$x^* = \frac{x - mean(X)}{\sigma} \quad (3)$$

2.2.4. Mean Substitution and Time Serialization

The average growth rates of the total financing amount (including the debt financing amount, equity financing amount, internal financing and trade financing amounts, and project financing and policy financing amounts) and total financing cost (the total financing cost includes the debt financing cost, equity financing cost, internal financing and trade financing costs, and project financing and policy financing costs) are calculated. Taking the average growth rates of the total financing amount and total financing cost as the characteristic data of the enterprise, the formula is as follows:

$$\Delta D_i = \frac{\frac{D_i^{s+1} - D_i^s}{D_i^s} + \frac{D_i^{s+2} - D_i^{s+1}}{D_i^{s+1}} + \cdots + \frac{D_i^{s+c} - D_i^{s+c-1}}{D_i^{s+c-1}}}{c-1} \quad (4)$$

In the above formula, I is the characteristic of column I, s is the Initial Public Offering (IPO) year of the enterprise, and C is the listing year. To determine whether an enterprise is a zombie enterprise, it should have zombie characteristics for many years. Thus, this model only analyses enterprises that have been listed for more than three years, that is, $C \geq 3$.

The mean values of other characteristic data since the listing dates of enterprises are calculated, and the mean values of these characteristics are taken as the characteristic data of enterprises. The formula is as follows:

$$\overline{D_i} = \frac{D_i^s + D_i^{s+1} + \ldots\ldots + D_i^{s+c}}{c} \quad (5)$$

Furthermore, the extracted enterprise characteristics include the registered capital, shareholding ratio of the controller, innovation index, growth rate of the financing amount, growth rate of the financing cost, number of employees, total assets, total liabilities, total operating income, main business income, total profit, net profit, total tax payment and total owner equity.

3. Model Construction

Ensemble learning uses different training sets with different data distributions to train different sub-models and combines them to obtain meta models with better performance. Because an ensemble learning model improves the generalization ability according to the differences among the sub-models, the greater the differences between sub-models, the better the performance of the meta models. Therefore, not only the classification performance of each sub-model but also the model differences should be considered when modelling. Based on the above considerations, to determine the optimal sub-model combination scheme, this paper tests three kinds of pattern recognition algorithms (a DT, an SVM, and a perceptron).

3.1. Sub-Model Selection

3.1.1. Sub-Model based on a DT

In this paper, three algorithms are selected for testing DT branches, including an ordinary DT, a GBDT and a Random Forest (RF). The GBDT is a boosting algorithm based on a DT, and the RF is a bagging algorithm based on a DT [11]. The specific test results are shown in Table 1.

Table 1. Sub-models based on a DT.

	Average accuracy	Average recall rate
Common DT	0.94	0.94
GBDT	0.95	0.94
RF	0.90	0.89

The above results show that the ordinary DT and the GBDT perform well on the verification set. Because the GBDT can process all kinds of data flexibly, it has high prediction accuracy under a relatively shorter parameter adjustment time. The GBDT algorithm can be implemented from the XGBoost open-source project, which is an engineering optimization that provides better speed and efficiency in practical applications. In this paper, the XGBoost package is used to build a gradient boosting decision tree as the sub-model of a branch of the decision tree.

3.1.2. Sub-Model based on an SVM

In this paper, a linear kernel SVM and a Gaussian kernel SVM are selected for testing. The specific test results are shown in Table 2.

Table 2. Sub-models based on an SVM.

	Average accuracy	Average recall rate
Linear kernel SVM	0.96	0.95
Gaussian kernel SVM	0.71	0.71

According to the test results, this paper selects the linear kernel SVM as the sub-model of the SVM branch. The linear kernel SVM uses the linear separability of the linear kernel function to solve high-dimensional problems directly, and the loss of sample data is small. Its linear kernel further improves the differences between sub-models and enhances the linear operation ability of the ensemble learning model.

3.1.3. Sub-Model Based on A Perceptron

In this paper, two algorithms, a MLP and a DNN (the number of hidden layers is greater than 2), are selected for testing. The specific test results are shown in Table 3.

Table 3. Sub-models based on a perceptron.

	Average accuracy	Average recall rate
MLP	0.97	0.96
DNN	0.96	0.96

The test results show that the two algorithms based on a perceptron have good adaptability on the verification set, and the accuracy and recall are not less than 0.96. The deep learning property of the DNN causes it to consume more resources during the training process, but at the same time, it also has the advantages of strong adaptability to nonlinear features and can construct complex abstract models. To combine the advantages of deep learning with an ensemble learning model, a DNN is selected as the sub-model of the perceptron branch.

3.2. Model Integration

Stacking, as an integration method, can be used to train a model to combine multiple sub-models and to optimize the performance of the combination. In this paper, we combine a GBDT, a linear kernel SVM and a DNN with the stacking method to obtain a more comprehensive integrated learning and classification model. The pseudo code is as follows:

Training set:

$$D_1\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$$

Test set:

$$D_2 = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$$

Submodel: $\varphi_1, \varphi_2, \cdots, \varphi_n$
Metamodel: φ_{meta}

$for\ i = 1,2, \cdots, n$
$do\ \varphi_i \cdot fit(D_1)$ //Training sub model
end

$for\ i = 1,2, \cdots, n$
$do\ P_i = \varphi_i \cdot predict(D_2)$ //Output forecast
end

$P = \cup_1^n f(P_i)$
//Training set of generator model
$\varphi_{meta} \cdot fit(P)$ //Training meta model

The above process can be divided into three steps. First, each sub-training set is obtained by bootstrap sampling from the whole training dataset. Then, the sub-models are trained, and the prediction results are obtained. Finally, the meta model is trained with the prediction results of the sub-models [12]. To train the meta model, this paper uses the prediction categories output from the sub-models as the training set of the meta model.

This paper utilizes three sub-models (a DT, an SVM, and a perceptron). The steps below should be followed:

1. Fit the first group of data with the three sub-models.
2. Make each of the three sub-models perform predictions for the observed data in the second group of data.
3. Fit the meta model on the second set of data using the predictions made by the sub-models as the inputs. The model construction steps are shown in Figure 1.

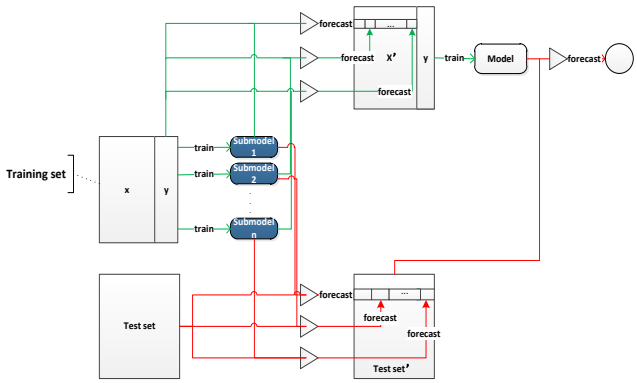

Figure 1. Model construction diagram.

At the end of the first training iteration, an RF is used to calculate the gain of each tree to obtain the corresponding feature scores. According to the feature scores, redundant features are deleted and retrained to make the meta model achieve better classification and recognition results.

4. Platform Display

The visual interface includes data importing, enterprise judgement, result exporting, portrait display and other functions. After importing the enterprise data in the web interface, the data are analyzed according to the above steps to form an enterprise portrait and display the enterprise information visually.

First, the front-end web page based on HTML, CSS, JavaScript and other technologies obtains the user's requests and data files and transmits the user's requests and other related information to the back end through the web server. After Django [13] obtains the data, it uses Numpy, pandas and other Python data processing tools to pre-process and normalize the data. Then, the data table of the fully connected neural network model is formed. Second, the above model is used to build enterprise portraits. Finally, through the calculation of the output enterprise portrait data by the model, Django renders these data to the front-end page to form the final visual interface for users' reference.

5. Conclusions

In this study, a botnet-based enterprise classification and recognition method is established based on ensemble

learning. The classification model is obtained by integrating a GBDT, an SVM and a DNN. By combining with the advantages of the small sample data loss of the linear classifier and the strong learning ability and good adaptability of the deep learning model, the Submodel is stacked to prevent overfitting, the subjectivity of the model is reduced, and the accuracy and recall of the model on the verification set are both greater than 0.98. In summary, the research methods and results proposed in this paper can transform the knowledge and experience of zombie enterprise identification into objective quantitative indicators and realize the automatic requirements of zombie enterprise classification and identification.

References

[1] Bargagli-Dtoffi F., Riccaboni M., and Rungi A., "Machine Learning for Zombie Hunting. Firms Failures and Financial Constraints," Working Papers, 2020.

[2] Jiliang C., "Research on the Influence of Zombie Enterprises on Resource Allocation in China," Shenzhen University, 2019.

[3] Juan H., Weifeng Z., Wei F., and Heng Z., "Study on the Prediction Model of Albacore Tuna Fishing Ground in the South Pacific Based on Ensemble Learning," *Southern Fisheries Science*, vol. 16, no. 5, pp. 194-200, 2020. doi: 10.12131/20200022

[4] Juan T., Dingju Z., and Wenhan Y., "Research on Enterprise Portraits Based on Big Data Platforms," *Computer Science*, vol. 45, no. S2, pp. 58-62, 2018.

[5] Kane E., "Dangers of Capital Tolerance: The Case of the FSLIC and Zombie S&Ls," *Contemporary Economic Policy*, vol. 05, no. 1, pp. 77-83, 1987. DOI: 10.1111/j.1465-7287.1987.tb00247.x

[6] Qin W. and Tan S., "Zombie Enterprise Characteristics, Causes and Identification Disposal Suggestions," *Finance and Economy*, vol. 04, pp. 93-119, 2019.

[7] Shaoqing H. and Yan C., "The Distribution Features and Classified Disposition of China's Zombie Firms," *China's Industrial Economy*, vol. 03, pp. 24-43, 2017.

[8] Wang H. and Changgang L., "Application of Stacking Integrated Learning Method in Sales Forecasting," *Computer Applications and Software*, vol. 37, no. 8, pp. 85-90, 2020. DOI: 10.3969/j.issn.1000-386x.2020.08.016

[9] Xiaoyan L., "Research on Recognition of Chinese Zombie Enterprise under Supply-Side Reform," *Economic Restructuring*, vol. 03, pp: 194-200, 2019.

[10] Xiaoyang L. and Qiang Q., "China's Zombie Firms: Identification and Classification," *International Financial Research*, vol. 08, pp. 3-13, 2017.

[11] Yuping J., Cheng Y., Yanrong L., Haiyan S., Di L., Jiang Z., Hao W., and Hao C., "An Early Screening Method of Chronic Kidney Disease Based on Ensemble Learning Algorithm," *Journal of Southwest University* (*Natural Science Edition*), vol. 42, no. 10, pp, 17-24, 2020.

[12] Yu S., Zhali L., and Xuewen L., "Ensemble Learning and Feature Integration in Prediction for College Students Grant," *Journal of Xi'an University of Science and Technology*, vol. 40, no. 4, pp. 744-750, 2020. 10.13800/j.cnki.xakjdxxb.2020.0424

[13] Zheng J., Qin L., Liu K., Tian B., Tian C., Li B., Chen G., "Django: Bilateral Coflow Scheduling with Predictive Concurrent Connections," *Journal of Parallel and Distributed Computing*, vol. 152, pp. 45-56, 2021. https://doi.org/10.1016/j.jpdc.2021.01.006

[14] Zhihua Z., "Machine Learning," Tsinghua University Press, Beijing, 2016.

Shutong Pang, an undergraduate student from Zhejiang Chinese Medical University, Zhejiang, China. She is currently studying for a master's degree at the Medical College of Guangxi University, Guangxi, China. Her research interests are data processing and analysis, and image segmentation.

Ziwei Yang, an undergraduate student from Zhejiang Chinese Medical University, Zhejiang, China. Her research interests are data processing and analysis.

Chengyou Cai, an undergraduate student at Zhejiang Chinese Medical University, China. He is currently studying for a master's degree at Shanghai University, Shanghai, China. His research interests are data mining and time series analysis.

Zhimin Li, currently an associate professor of Zhejiang Chinese Medical University, China. focuses on data processing and analysis.