An Enterprise Competitiveness Assessment Method Based on Ensemble Learning

Yaomin Chang[1,2], Yuzheng Li[1,2], Chuan Chen[1,2](✉), Bin Cao[3], and Zhenxing Li[3]

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
{changym3,liyzh23}@mail2.sysu.edu.cn, chenchuan@mail.sysu.edu.cn

[2] Guangdong Key Laboratory for Big Data Analysis and Simulation of Public Opinion, Sun Yat-sen University, Guangzhou, China

[3] HuaXia iFinance (Beijing) Information Technology Co., Ltd., Beijing, China
bincao@hxifin.com

Abstract. It is of great significance to assess the competitiveness of enterprises based on big data. The current methods cannot help corporate strategists to judge the status quo and prospects of enterprises' development at a relatively low cost. In order to make full use of big data to evaluate enterprise competitiveness, this paper proposes an enterprise competitiveness assessment method based on ensemble learning. The experimental results show that our method has a significant improvement in the task of the enterprise competitiveness assessment.

Keywords: Enterprise competitiveness assessment · Data mining · Ensemble learning · Machine learning

1 Introduction

Effective and quick assessment of enterprise competitiveness can create huge economic value. It can help people identify more competitive enterprises in bank credit risk management and capital investment, thus to achieve a better allocation of funds. Traditional methods of enterprise assessment are mainly based on the tedious and lengthy investigation, analysis and report, which suffer from strong subjectivity, massive cost and poor generalization.

With the help of the rapid development of data mining techniques, artificial intelligence is widely used in lots of fields [3,9]. A large number of models armed with machine learning methodology have proven to be efficient, reliable

The work described in this paper was supported by the National Key Research and Development Program (2016YFB1000101), the National Natural Science Foundation of China (61722214,11801595), the Natural Science Foundation of Guangdong (2018A030310076), and CCF Opening Project of Information System.

W. Ni et al. (Eds.): WISA 2019, LNCS 11817, pp. 79–84, 2019.
https://doi.org/10.1007/978-3-030-30952-7_9

and powerful. However, there remain several major challenges in assessing enterprise competitiveness with artificial intelligence technology. Firstly, it is difficult to extract important features to evaluate the enterprise, while a huge amount of data is generated in business operations. In addition, it requires a good combination with the time series analysis techniques to consider the enterprises' data of the past years. Moreover, unlike traditional machine learning researches, which have relatively well-formulated problems and objective functions, there cannot be a standard to assess enterprise competitiveness comprehensively. Therefore, it's hard to select an appropriate target to train the model.

In this paper, we propose a method called Enterprise Assessment with Ensemble Learning (EAEL) to assess the enterprise competitiveness based on ensemble learning. Firstly, this method extracts enhanced features from the corporate annual financial data, basic registration information and the national macroeconomic data. Subsequently, we apply the idea of ensemble learning to train the model in four different dimensions, including profitability, operation competency, liquidity and growth opportunity, to achieve a more comprehensive enterprise competitiveness assessment. Finally, we evaluate and validate our model on a real-world dataset collected from Chinese A-share stock market. The experiments demonstrate that our method based on ensemble learning has a significant improvement compared with traditional methods.

2 Related Work

In the subject of management science, methods of assessing enterprise competitiveness are mainly listed into two categories: qualitative analysis and quantitative assessment. Qualitative methods such as Michael Porter's Five Forces Model [7] studied the factors that have a great influence on enterprise competitiveness, which is based on domain-relevant knowledge. However, these approaches rely heavily on the subjective judgment of experts and are so costly. Therefore, a larger number of researches turn to use financial data as indicators of enterprise competitiveness, which is more objective and quantifiable. Edward Altman [2] proposed the Z-Score model after a detailed investigation of the bankrupt and non-bankrupt enterprises. This work selected 5 indicators from 22 financial ratios by mathematical statistics methods and thus to make the model simpler, more effective and more intuitive. Although the Z-Score model has a long history, it still has a good performance on enterprise competitiveness assessment in the past 25 years, according to Vineet Agarwal's research [1].

3 Methodology

In this section, we introduce the feature engineering techniques used in our method, as well as the details of our ensemble learning sub-models.

3.1 Feature Engineering

Incremental Features. Since the enterprises' data from the past few years can reflect their development trend, we extract *Incremental Features* from original data under the management advice. *Incremental Features* are defined as the difference and the growth ratio between the current period and the previous period, both on a yearly and quarterly basis. *Quarterly Incremental Features* can reflect the effect of the enterprise's short-term business strategy and its development trend. On the other hand, *Yearly Incremental Features* can exclude the influence of seasonal factors, thus reflecting the long-term competitiveness of enterprises.

Unit Features. In addition to finding large companies with outstanding market performance, we also need to identify smaller companies that have high growth potential. Competitive enterprises of smaller size can perform well in the market with fewer employees. Less competitive enterprises of larger size can still have a large number of market share in the same industry, despite that their per capital benefit and asset utilization level have already located in a lagging position. To achieve a better evaluation on small and competitive enterprises, we construct *Unit Features* that are defined as the ratio of each feature to the number of employees, total assets, gross liability and book value of equity, which intuitively reflect the corresponding output on these certain features.

3.2 Ensemble Learning

Based on the previous researches [4–6], we propose Enterprise Assessment with Ensemble Learning (EAEL) to assess the competitiveness of enterprises. Our method contains two specific sub-models, i.e., the n-year prediction model and the annual series prediction model, as well as a collective model that combines the predictions of the former two sub-models.

XGBoost. XGBoost [4] is a scalable end-to-end tree boosting system that is used widely by data scientists to solve many machine learning problems in a variety of scenarios [10]. XGBoost is adopted in our EAEL method, where we use the data of the enterprise and the macro-economy x_i to predict a target assessment of this enterprise's competitiveness y_i. The original objective function is Eq. (1).

$$Obj^t(X,Y) = \Sigma_{i=1}^n L(y_i, \hat{y}_i^{t-1}) + \Omega(f_t) \tag{1}$$

where $\hat{y}_i^{t-1}$ represents the model's prediction of round $t-1$ in the training phase, and $L(y_i, \hat{y}_i^{t-1})$ is the self-defined cost function between y_i and $\hat{y}_i^{t-1}$. $\Omega(f_t)$ is the regularization term to control the complexity of the model, i.e., the L2 norm of the leaf scores.

To approximate the objective function, XGBoost performs the second-order Taylor expansion on Eq. (1), using both g_i (the first derivative) and f_i (the second derivative) in Eq. (2).

$$Obj^t(X,Y) \simeq \Sigma_{i=1}^n [L(y_i, \hat{y}_i^{t-1} + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{2}$$

The *n*-year Prediction Model. As we have collected several years data of each enterprise, we use Ent_i^k and C_i^k to denote the i^{th} enterprise's data and its corresponding competitiveness in year k, respectively. The n-year prediction model refers to the idea that we use the data Ent_i^k to predict the enterprise's market performance in the n^{th} years, i.e., C_i^{k+n}. The n-year prediction model tries to dig out the implicit relationship between current year data and the enterprise's future competitiveness.

The Annual Series Prediction Model. In order to mine the annual sequence information of each company, we use time-based linear regression that performs on each feature to figure out the enterprise's development trend over the last few years. We take the results of linear regression as the features of the enterprises' general performance over the past years and train a new XGBoost model.

Collective Model. According to the research on model stacking [8], taking the output of multiple sub-models as the input features of the collective model can improve the generalization and fitting ability of the model. Therefore, we train T of the n-year prediction model and the annual series prediction model each and use their predictions as new features to train the collective XGBoost model.

4 Experiment

4.1 Experiment Setup

The following experiments use a real dataset crawled from different sources, including Chinese A-share stock markets, the Chinese National Bureau of Statistics and the Chinese National Bureau of Commerce and Industry. Totally the financial data and registration information of 3573 enterprises and the macroeconomic data are collected, covering from 2011 to 2018. Firstly we annotate each record in four dimensions (profitability, operation competency, liquidity and growth opportunity) according to experts' advice, turning this assessment problem into a classic binary classification problem. Then we split our dataset into training set, test set and validation set by 6:2:2. To evaluate the proposed method, we compare the predicted result with the traditional Z-Score model in all dimensions. Additionally, several verification experiments are presented to illustrate the effectiveness and necessity of feature engineering and different ensemble methods.

4.2 Experiment Results

Comparative Experiments with Traditional Methods. In this experiment, we compare our proposed method EAEL with the traditional method in different dimensions (profitability, operation competency, liquidity and growth opportunity). As the Z-Score model can only get numeric scores of enterprise competitiveness, we rank the scores predicted by the Z-Score model in descending order, to which we assign corresponding binary labels. The experimental results are listed in Table 1. We can see from Table 1 that the predictive ability of our EAEL method certainly outperforms the Z-Score model in all dimensions, especially on the comprehensive metrics F_1-*score*.

Table 1. Comparison results with traditional methods.

Dimension	Model	Accuracy	Precision	Recall	F_1-*score*
Profitability	Z-Score	0.5791	0.6421	0.6703	0.5765
	EAEL	**0.6939**	**0.7069**	**0.8569**	**0.6775**
Operation Competency	Z-Score	0.4654	0.4234	**0.6027**	0.4590
	EAEL	**0.6592**	**0.6350**	0.5117	**0.6528**
Liquidity	Z-Score	0.5219	0.5258	0.6436	0.5144
	EAEL	**0.6394**	**0.6338**	**0.6832**	**0.6386**
Growth Opportunity	Z-Score	0.4631	0.3772	**0.6145**	0.4620
	EAEL	**0.6684**	**0.6156**	0.3408	**0.6407**

Analysis of Feature Engineering. In this experiment, we try to verify how our feature engineering techniques affect the model's performance. In short, we only display the experimental results in the dimension of profitability. We use different kinds of features to train the model and the results are listed in Table 2. We can figure out that different techniques improve the performance of enterprise competitiveness assessment with a different degree. When we apply all feature engineering techniques to the training data, we can train the best-performing model.

Table 2. The effects of different feature engineering techniques

Data	Accuracy	Precision	Recall	F_1-*score*
Data (raw)	0.5641	0.5110	0.7914	0.4568
Data (with Unit Features)	0.6724	0.6854	**0.8631**	0.6485
Data (with Incremental Features)	0.6432	0.6584	0.8541	0.6141
Data (with Unit & Incremental Features)	**0.6939**	**0.7069**	0.8569	**0.6775**

Analysis of Ensemble Learning. In this experiment, we train the three submodels mentioned in Sect. 3.2 and compare their predictions in the dimension of profitability. From Table 3 we can see that the performance of the n year prediction model decreases as the hyperparameter n increases. That suggests that the correlation between the current year's data and the enterprise's competitiveness in future years is declining. And the annual series prediction model is slightly worse than the 3-year prediction model. It also indicates that the competitiveness of enterprises is more related to the enterprises' data in recent years, while the long series of enterprises' data may disturb the model and reduce its evaluation performance.

Table 3. The performance of different sub-models

Model	Accuracy	Precision	Recall	F_1-*score*
The 1-year prediction model	0.6939	0.7069	0.8569	0.6775
The 2-year prediction model	0.6664	0.6726	0.8328	0.6503
The 3-year prediction model	0.6459	0.6604	0.8258	0.6203
The annual series prediction model	0.6456	0.6508	**0.9444**	0.5676
The collective model	**0.6964**	**0.7319**	0.8047	**0.6905**

5 Conclusion

In the era of big data, it is significantly important to fully explore the implicit information of the enterprises' data. In this paper, we propose an ensemble learning-based method EAEL to assess enterprise competitiveness. Firstly, this method performs feature engineering to get informative *Incremental Features* and *Unit Features*. Then, EAEL trains three sub-models to predict enterprise competitiveness. Finally, experimental results demonstrate that our method has significant improvement in enterprise competitiveness assessment.

References

1. Agarwal, V., Taffler, R.J.: Twenty-five years of the Taffler z-score model: does it really have predictive ability? Account. Bus. Res. **37**(4), 285–300 (2007)
2. Altman, E.I.: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finan. **23**(4), 589–609 (1968)
3. Carneiro, N., Figueira, G., Costa, M.: A data mining based system for credit-card fraud detection in e-tail. Decis. Support Syst. **95**, 91–101 (2017)
4. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: KDD-16, pp. 785–794 (2016)
5. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**, 1189–1232 (2001)
6. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. J. Artif. Intell. Res. **11**, 169–198 (1999)
7. Porter, M.E.: The five competitive forces that shape strategy. Harv. Bus. Rev. **86**(1), 78–93 (2008)
8. Wolpert, D.H.: Stacked generalization. Neural Networks **5**(2), 241–259 (1992)
9. Wu, S., Ren, W., Yu, C., Chen, G., Zhang, D., Zhu, J.: Personal recommendation using deep recurrent neural networks in netease. In: ICDE-16, pp. 1218–1229 (2016)
10. Xi, Y., Zhuang, X., Wang, X., Nie, R., Zhao, G.: A research and application based on gradient boosting decision tree. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 15–26. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_2