The Application of Machine Learning Models in Company Bankruptcy Prediction

Hanxu Chang
Wuhan University of Technology,
NO. 122, Luoshi Road, Hongshan District, Wuhan,
18507193427, China
changhanxu@163.com

ABSTRACT

With the economy increasing, the number of enterprises increases gradually, and it is accompanied by the growth in quantity of bankruptcies.Therefore, predict bankruptcy is becoming more and more important. It could not only help make the correct decision, but also reduce losses. There are several traditional methods are commonly used to predict the corporate bankruptcy conditions. But the traditional methods for bankruptcy prediction are mainly based on human subjective judgment and lack quantitative analysis. So it is for traditional methods to compete with the new and advanced Machine learning algorithms. Machine learning algorithms explore patterns based on objective data analysis, which develop rapidly and have strong learning ability. So we're going to apply machine learning to the prediction of corporate bankruptcy. We use data on the bankruptcy situation of Polish companies in 2007-2013 and construct a model by SVM and random forest algorithm separately. And then, we further use weighted methods to solve the problem of sample imbalance. According to the research, Random forest performs better than SVM in company bankruptcy prediction with accuracy higher than 70% in different years.

CCS Concepts

• **Applied computing → Enterprise computing →Business process management →Business intelligence**

Keywords

Machine learning; Support-vector machines; Random forest; weighted SVM; Weighted random forest

1. INTRUCTION

With the development of the times, more and more companies are constantly being created and bankrupt. For investors and companies, the prediction of corporate bankruptcy is very important for themselves. For example, Ma Yun, a Chinese e-commerce giant, suffered heavy losses due to the bankruptcy of the company he invested. This shows that the wrong prediction of company bankruptcy will lead to a catastrophic consequence. Therefore, an effective method to predict the company bankruptcy is necessary.

The traditional approach for bankruptcy prediction is based on the subjective judgments about a company of the expertise and experienced analysts, bosses or industry experts, which lacks impartiality to some extent. Besides, there are also some quantitative measurements for bankruptcy probability [1][2]. For example, Altman Z-score uses a linear combination of four or five common business ratios to evaluate the bankruptcy risk, where the coefficients are estimated by identifying a set of firms which had declared bankruptcy. Its shortcoming is that all factors are linear, which is too simple for the prediction of corporate bankruptcy. In this case, we need some special methods to achieve ourgoals, such as machine learning.

Machine learning is a method of researching algorithms and statistical models. It can make computer systems effectively perform a specific task by patterns and inference, instead of relying on the explicit instructions. We think of it as a subset of artificial intelligence [3]. In simple terms, it is a method to establish models by classifying and sorting existing data, and then we can make predictions about the rest of the data. For instance, we can execute trades in financial transactions based on parameters set by the trading company. It is a good way which can predict the rise and fall of stocks more accurately. That is why we used machine learning methods, including random forest and SVM models, in this article to solve the problem.

In machine learning, support-vector machines, for short SVMs, are supervised learning models which can analyze data used for classification and regression analysis [4]. A SVM model is a method that transforms the examples to points in space, mapped so that the different categories of samples are divided by a clear and as wide as possible gap. And then, new examples are mapped into this space and predicted to belong to a category which is according to where they fall. Furthermore, it can transform the nonlinear problem to the feature space which have higher dimension. The linear discriminant functions are based on these high dimensional space, and avoid the problem of the number of dimensions are large and the algorithm is complicated.

Besides, we also use random forests in our research. Random forests or random decision forests are an ensemble learning method which can complete the tasks of classification and regression.First, we take a part of the raw data as training data to build decision trees. The rest of the data are predicted as a test set through the decision tree. It can through construct a multitude of decision trees at training time and outputting the class to get the mode of the classes or mean prediction of the individual trees [5]. Furthermore, it can also remain a high degree of accuracy, because of its strong anti-interference ability, when we have a set

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
ICSEB 2019, December 9–11, 2019, TOKYO, Japan
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-7649-5/19/12…$15.00
https://doi.org/10.1145/3374549.3374550

of data which miss numerous elements. Therefore, random forests are more accurate than other models sometimes.

Hence, we are ready to solve the issue of corporate bankruptcy prediction with the machine learning models. We use the data of Polish companies which have different bankruptcy situation in 2007-2013 with the SVM and random forest algorithm to construct models, and solve the problem of sample imbalance by weighting on this basis. The correct rates are 77.6%, 77.9%, 79.9%, 76.1% and 79.4% in 5 different cases respectively.

Here is the structure of this article. In Section 2, we will describe the structure and content of the data in detail. The SVM and its weighted model, the Random Forest and its weighted model are presented in sections 3 and 4 respectively, which also includes their detailed introduction and analysis. Finally, we will summarize these results and look forward to the future in section 5.

2. DATA DESCRIPTION

The company bankruptcy data set is obtained through the analysis of the information of the bankrupted companies in the period 2000-2012, and the still operating companies were evaluated from 2007 to 2013. This data is public available at http://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data, which is theUCI (University of California Irvine) Machine Learning Reposito-ry. The information of name, starting time, ending time, bankrupt-cy status and number of samples is presented in Table 1. It can beseen that the ratio of bankrupt companies is extremely lower than non-bankrupt companies. But in fact, the prediction of bankrupt companies are far more important than non-bankrupt companies, so we would like to reduce the accuracy of non-bankrupt compa- nies to some extend to increase the forecast accuracy of bankrupt ones. Furthermore, Table 2 shows 64 variables which indicate the financial status of these companies. The two variables sales (n) / sales (n-1) and (current assets - inventories) / long-term liabilities are removed because of too many missing values, and the left mis-sing values are filled with average. And then we divide samples into training set and testing set by 70% and 30%.

Table 1. The bankruptcy of companies

Data	Starting time	Ending time	bankruptcies	samples
Year 1	2007	2008	271	7027
Year 2	2007	2009	400	10173
Year 3	2007	2010	495	10503
Year 4	2007	2011	515	9792
Year 5	2007	2012	410	5910

Table 2. The financial condition of companies

X1	net profit / total assets
X2	total liabilities / total assets
X3	working capital / total assets
X4	current assets / short-term liabilities
X5	[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
X6	retained earnings / total assets
X7	EBIT / total assets
X8	book value of equity / total liabilities
X9	sales / total assets
X10	equity / total assets
X11	(gross profit + extraordinary items + financial expenses) / total assets
X12	gross profit / short-term liabilities
X13	(gross profit + depreciation) / sales
X14	(gross profit + interest) / total assets
X15	(total liabilities * 365) / (gross profit + depreciation)
X16	(gross profit + depreciation) / total liabilities
X17	total assets / total liabilities
X18	gross profit / total assets
X19	gross profit / sales
X20	(inventory * 365) / sales
X21	sales (n) / sales (n-1)
X22	profit on operating activities / total assets
X23	net profit / sales
X24	gross profit (in 3 years) / total assets
X25	(equity - share capital) / total assets
X26	(net profit + depreciation) / total liabilities
X27	profit on operating activities / financial expenses
X28	working capital / fixed assets
X29	logarithm of total assets
X30	(total liabilities - cash) / sales
X31	(gross profit + interest) / sales
X32	(current liabilities * 365) / cost of products sold
X33	operating expenses / short-term liabilities
X34	operating expenses / total liabilities
X35	profit on sales / total assets
X36	total sales / total assets
X37	(current assets - inventories) / long-term liabilities
X38	constant capital / total assets
X39	profit on sales / sales
X40	(current assets - inventory - receivables) / short-term liabilities
X41	total liabilities / ((profit on operating activities + depreciation) * (12/365))
X42	profit on operating activities / sales
X43	rotation receivables + inventory turnover in days
X44	(receivables * 365) / sales
X45	net profit / inventory
X46	(current assets - inventory) / short-term liabilities

X47	(inventory * 365) / cost of products sold
X48	EBITDA (profit on operating activities - depreciation) / total assets
X49	EBITDA (profit on operating activities - depreciation) / sales
X50	current assets / total liabilities
X51	short-term liabilities / total assets
X52	(short-term liabilities * 365) / cost of products sold)
X53	equity / fixed assets
X54	constant capital / fixed assets
X55	working capital
X56	(sales - cost of products sold) / sales
X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X58	total costs /total sales
X59	long-term liabilities / equity
X60	sales / inventory
X61	sales / receivables
X62	(short-term liabilities *365) / sales
X63	sales / short-term liabilities
X64	sales / fixed assets

3. SUPPORT VECTOR MACHINE

3.1 SVM

SVM (Support Vector Machine) is a kind of supervised learning which can be used for classification and regression analysis [6]. It could turn samples into spots in space to get intuitive results. In order to make the non-linearly indivisible data set separable, we map the training data set into a high dimensional feature space. And then, we build an optimal separation hyper plane with the largest distance in the feature space [7]. According to the characteristics that it can transform the nonlinear separability problem in low dimensional space to higher dimension, we apply it to build model to solve problems.

We build SVM model with radial basis functions. Five groups are shown in Table 3. In the five sets of data, the error rate of all the first types is 0% and the second types is 100%. Due to the imbalance between bankruptcy and no bankruptcy in these samples, the SVM model predicts that none of them to be bankrupt. This is a meaningless model although our predictions for all no bankrupt companies are correct. In reality, predicting bankrupt companies is much more important than no bankrupt ones. In this case, we give up the old model and replace it with new weighted one.

Table 3. Error rate of SVM model prediction

	Year1	**Year2**	**Year3**	**Year4**	**Year5**
Training error rate	3.62%	3.88%	4.65%	5.08%	6.70%
Testing error rate	4.22%	4.00%	4.73%	5.55%	6.77%
0 class error rate	0.00%	0.00%	0.00%	0.00%	0.00%
1 class error rate	100.00%	100.00%	100.00%	100.00%	100.00%

3.2 Weighted SVM

Obviously, our SVM model has big flaws. According to the trade-off principle, we can only improve the accuracy of the second category by decreasing the accuracy of the first category. Because we care more about the bankruptcy of companies rather than non-bankruptcy, we give higher priority to the second category. Therefore, we use the weighted method to improve the prediction accuracy of the second type, and sacrifice the prediction accuracy of the first category inevitably. The SVM model is established by radial basis function, and each set of data has different weights, which are shown in Table 4 respectively. The error rates of five sets of data are shown in Table 5. In different sets, the prediction error of bankrupt companies declined significantly, while the increase of the prediction error of non-bankrupt companies was also in the acceptable range. It shows that weighting can improve the problems caused by sample imbalance.

Table 4. Weighted sample

	Year 1	**Year 2**	**Year 3**	**Year4**	**Year5**
weight	1:25	10:241	1:20	1:21	1:19

Table 5. Error rate of weighted SVM model prediction

	Year 1	**Year 2**	**Year 3**	**Year4**	**Year5**
Testing error rate	33.02%	42.66%	33.48%	29.88%	27.66%
0 class error rate	32.99%	43.45%	33.78%	29.66%	30.58%
1 class error rate	33.71%	23.77%	27.52%	33.74%	27.86%

4. RANDOM FOREST

4.1 Random Forest

Random forests is an ensemble learning method for classification, regression and other tasks. The method of model implementation is to construct a large number of decision trees during training and output classes, that is, classes or mean prediction of a single tree [8][9]. In addition, because the strong anti-interference ability of random forest model, it can maintain high precision when we have a set of data which miss many elements. Hence in order to improve the accuracy of our prediction, we can use it to construct the model.

Let's take the year 1 data as an example to show the steps of modeling. We used training data to construct random forests with different numbers, and the variation curve of the error of oob and test with the number of trees is shown in Figure 1.

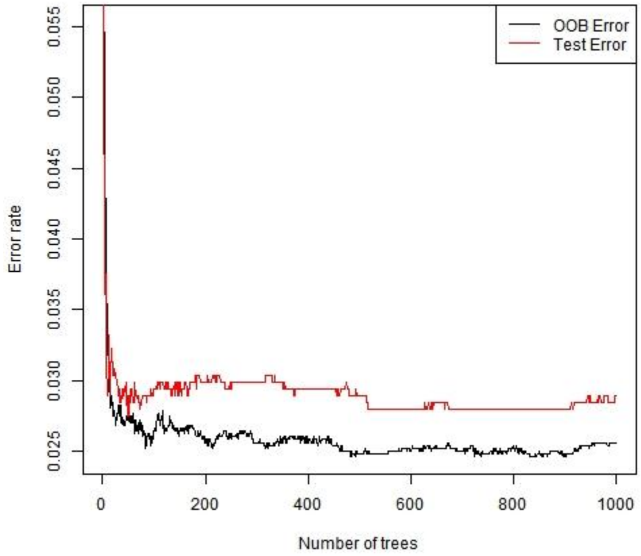

Figure 1. Error rate of random forest with different number of trees

At the beginning, the error drops rapidly with the increase of the number of tree, and then the change is small. According to Figure 1, we let n equal to 200. Table 6 presents the error rate of five sets of data. We can see that the error rate is low overall and is high in the second category. Similarly, other groups of data also exhibit such a phenomenon.

Table 6. Error rate of random forest model prediction

	Year 1	Year 2	Year 3	Year4	Year5
Testing error rate	2.85%	3.41%	4.06%	4.73%	5.81%
0 class error rate	0.15%	0.14%	0.27%	0.18%	1.15%
1 class error rate	64.04%	81.97%	80.54%	82.21%	69.42%

4.2 Weighted Random Forest

Since we are more concerned about the second type of data, which has a higher prediction error rate, we need to find ways to improve our model. Therefore, we use weighted methods to solve problems caused by sample imbalance [10]. The different sets of data are weighted separately to build the random forest, and the weights are shown in Table 7. Then, we used the weighted training data to construct the random forest. Take year 1 data as an example as well, Figure 2 presents the variation curve of oob error with the number of trees and the test results.

Table 7. Weighted sample

	Year 1	Year 2	Year 3	Year4	Year5
weight	120:150	120:150	140:150	125:150	108:150

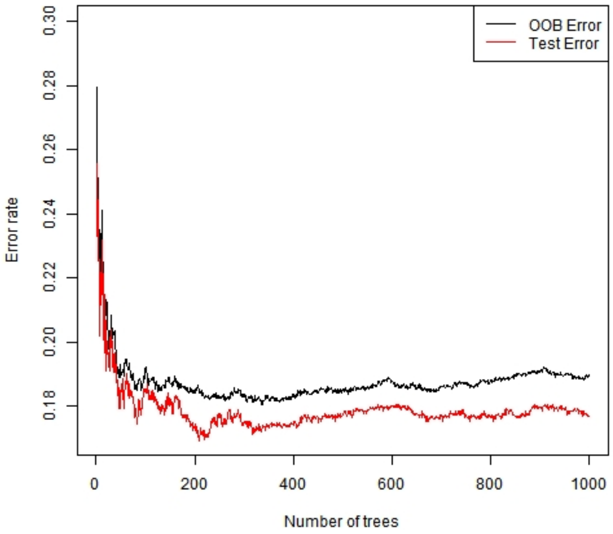

Figure 2. Error rate of weighted random forest with different number of trees

According to the figure above, the final selection of n is 200, and the corresponding confusion matrix is shown in Table 8. Although the overall error rate has increased, the error rate of the second category has dropped. The prediction error rate of all group data is less than 23%.

Table 8. Error rate of weighted random forest model prediction

	Year 1	Year 2	Year 3	Year4	Year5
Testing error rate	17.98%	25.39%	26.44%	24.13%	17.77%
0 class error rate	17.78%	25.53%	26.75%	24.14%	17.55%
1 class error rate	22.47%	22.13%	20.13%	23.93%	20.66%

5. SUMMARY

Predicting bankruptcy is a key issue in this era of rapid economic growth and corporate creation. Because of the drawbacks of traditional methods and the advancement of machine learning methods, we take the SVM and the Random Forest methods, which are commonly used in machine learning to build prediction models for Polish company bankruptcy data in this article. The initial model has low bankruptcy prediction accuracy due to sample imbalance. We further use the method of weighting the sample to effectively solve this problem and improve the accuracy of bankruptcy prediction. And we can see that the accuracy of the weighted Random Forest model is the highest, and the accuracy of all data is more than 70%. As the research progresses, we may consider using other machine learning models or combinatorial models to build prediction models in the future.

6. REFERENCE

[1] Michalski T, Gołebiowska E. Taxonomy methods in credit risk evaluation[J]. *International Advances in Economic Research*, 1996, 2(4):409-412.

[2] Dardac N. Credit Institutions Management Evaluation using Quantitative Methods[J]. *Theoretical & Applied Economics*, 2006, 2(497)(2(497)):35-40.

[3] Koza J.R., Bennett F.H., Andre D., Keane M.A. (1996) Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming. In: Gero J.S., Sudweeks F. (eds) *Artificial Intelligence in Design '96*. Springer, Dordrecht

[4] *Machine Learning*, 1995, Volume 20, Number 3, Page 273, Corinna Cortes, Vladimir Vapnik

[5] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning (2nd ed.).* Springer. ISBN 0-387-95284-5.

[6] James G, Witten D, Hastie T, et al. *An Introduction to Statistical Learning: with Applications in R* [M], *An Introduction to Statistical Learning*. 2013.

[7] Haoran Zhang, Zhengzhi Han, Changgang Li. Support Vector Machine[J]. *Computer Science*, 2002(12):135-137.

[8] Ho, Tin Kam (1995). Random Decision Forests (PDF). *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 14–16 August 1995. pp. 278–282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[9] Ho TK (1998). "The Random Subspace Method for Construc ting Decision Forests" (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 20 (8): 832–844. doi:10.1109/34.709601.

[10] Li Hang. *Statistical learning method* [M]. Beijing: Tsinghua University Press, 2012.