A Financial Early Warning Algorithm Based on Ensemble Learning

Chenggang Zhang and Jingqing Jiang
College of Computer Science and Technology
Inner Mongolia University for Nationalities
Tongliao, China
e-mails: zhangcg86@gmail.com, jiangjingqing@aliyun.com

***Abstract*—The design of effective financial early warning algorithm is of great significance to the financial management of the company. The weak classification algorithm can be improved to a high classification algorithm with high recognition rate through the ensemble learning. The algorithm can overcome the drawback of low classification accuracy of single classifier. Therefore, this paper combines decision stumps with Adaboost algorithm, and proposes DS_Adaboost algorithm. The experimental results show that compared with the traditional C4.5 algorithm, the proposed algorithm can significantly improve the accuracy of financial earning warning.**

Keywords—financial earning warning; ensemble learning; Adaboost; C4.5

I. INTRODUCTION

Company's financial early warning is an alarm mechanism built to prevent the company's financial system from deviating from the expected objectives, it has a targeted and predictive characteristics. It provides intellectual support for decision makers and helps them reach a scientific decision through evaluating comprehensively the company's indicators and predicting the company's financial situation, development trends and changes. The researchers have established a number of financial crisis early warning forecasting models, which mainly focus on statistical analysis and data mining, including univariate and multivariate early warning method[1], multiple linear discriminant method[2], logistic regression model[3] and neural network discriminant method[4]. A single variable model is first proposed by William Beaver[5], using the statistical method. He uses five financial ratios to predict the company's financial crisis and finds that the financial ratio called the debt security rate, i.e., the cash *flow/total* liabilities, can best predict the failure of the enterprises. Edward I. Altman[6] establishes a financial-ratio-based multi-variable financial early warning model to predict corporate bankruptcy, using the difference analysis method in multivariate statistical analysis. But the above algorithms do not consider the relationship between the indicators and lack comprehensive evaluation.

Ensemble learning, which learns through a series of learners, is a machine learning method that achieves a better learning effect than a single learner by integrating different learning results with a certain rule. The present paper, through the ensemble learning methods, fully studies the correlation between the indicators, thereby improving the accuracy of data classification.

Financial crisis early warning indicators can be divided into profitability indicators, solvency indicators, and growth indicators. Each indicator can be further divided into several small indicators. For instance, Profitability indicators can be divided into return on net assets, return on total assets, earnings per share, main business profit margins and cost margins and so on. In the case of financial early warning forecast, if all the indicators are evaluated comprehensively, the model can be too complex. And as the correlation between the indicators is strong, indicators need to be filtered before the establishment of the model.

The present study researches on the financial early warning algorithm based on ensemble learning. The structure is organized as follows: Section 2 pre-processes all the indicators, and selects 10 indicators that can best represent the company's financial status. Section 3 introduces the related algorithms in the study. Section 4 proposes an improved algorithm DS_Adaboost. Section 5 compares DS_Adaboost and traditional algorithm and analyzes the result through experiment. Finally, Section 6 concludes the paper.

II. INDICATOR SCREENING

The primary condition of constructing an effective early warning model of financial crisis is to select the early warning indicators which can reflect the characteristics of the company. To accurately predict the company's financial crisis, the selected indicators must cover all aspects of the company. By researching on the predictor variables that have a significant contribution to the final forecast model and the financial indicators that are widely used in empirical researches at present, the current study preliminarily decides and selects a total of 26 indicators including such 4 financial indicators as profitability, solvency, operating capacity and growth capacity and 22 non-financial indicators such as total assets and types of audit opinions as the initial variables to build the indicator system of company's financial crisis early warning. As shown in Table 1

The present study uses the independent sample T test provided by SPSS statistical analysis software to apply the factor analysis and carry out the initial screening of the 26 financial and non-financial indicators listed above. And those indicators whose total variance of the T test is greater than or equal to 0.06 and whose significant differences are not significant will be filtered out. Then, the principal component eigenvalues of the indicators selected through significant analysis are calculated, and the indicators with largest eigenvalues are taken as the final evaluation indicators of the early warning method of company crisis.

978-1-5386-2030-4/17/$31.00 ©2017 IEEE

TABLE I. ADABOOST ALGORITHM

Indicator Classification	Indicator Code	Indicator Name	Indicator Classification	Indicator Code	Indicator Name
Profitability	X1	Total assets profit margin	Operating capacity	X13	Inventory turnover
	X2	Return on equity		X14	Accounts receivable turnover
	X3	Gross profit margin		X15	Turnover ratio of current assets
	X4	Net selling rate		X16	Turnover of fixed assets
	X5	Cost profit margin		X17	Turnover of total assets
	X6	Earnings per share		X18	Days sales outstanding
Solvency	X7	Current ratio	Growth capacity	X19	Business cycle
	X8	Quick ratio		X20	Sales growth rate
	X9	Equity ratio		X21	Growth rate of net profit
	X10	Interest coverage ratio		X22	Growth rate of net assets
	X11	Operating cash flow		X23	Growth rate of total assets
	X12	Asset liability ratio		X24	Revenue growth rate
	X25	Total assets		X26	Types of audit opinions

Finally, ten indicators including return on net assets are selected as evaluation indicators, as they can comprehensively reflect the company's financial situation. As shown in Table II.

TABLE II. EVALUATION INDICATOR TABLE

Indicator Code	Indicator Name
X2	Return on equity
X6	Earnings per share
X7	Current ratio
X11	Operating cash flow
X12	Asset liability ratio
X13	Inventory turnover
X17	Turnover of total assets
X23	Growth rate of total assets
X25	Total assets
X26	Types of audit opinions

III. RELATED ALGORITHM INTRODUCTION

A. Adaboost Algorithm

Let's consider a two classification problem, in which the training data set consists of input vectors $x_1,...,x_m$. Each vector corresponds to a target value $y_1,...,y_m$, where y_i is {-1,1}. Each data point is given a corresponding weight *D(i)*, and all sample weights are initialized to 1/m. Suppose that a given function h(x) can be learned by weighted data, where *h(x)* is {-1,1}, and it is the base classifier used to train. At each step of the algorithm, Adaboost learns a new classifier, which is learned from the last trained classifier weighted data. Weighting is a process that assigns a higher weight to the wrong sample. Finally, when the appropriate number of base classifiers are trained, they make predictions through a method of adding different weights to different base classifiers[7][8]. Adaboost algorithm is shown in Table III.

TABLE III. ADABOOST ALGORITHM

Adaboost Algorithm

1. **Given:** $(x_1, y_1)\ldots(x_m, y_m)$ **where** $x_i \in X, y_i \in Y = \{-1,+1\}$
2. **Initialize** $D_1(i) = 1/m$.
3. **For** $t = 1\ldots T$:
   - 3.1 Train weak learner using distribution D_t
   - 3.2 Get weak hypothesis $h_t : X \to \{-1,+1\}$ with error
   $$\varepsilon_t = \Pr_{i\sim D_t}\left[h_t(x_i) \neq y_i\right].$$
   - 3.3 **Choose** $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$.
   - 3.4 **Update:**
   $$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}.$$
   Where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).
4. **Out put** the final hypothesis:
$$H(x) = \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

We can see that the first base classifier *h(x)* is trained by all equal weights *D(i)*, which uses standard training process of a single classifier. In the subsequent iteration, the weight coefficient α_t is added to the misclassified data points, and the weight of the correctly-classified data points is reduced. So the subsequent classifiers will be forced to focus more on samples that were previously misclassified, thereby the misclassified samples would be given higher weights. represents a measurement method of the error rate of each base classifier. The weight coefficient α is used to assign a higher weight to classifiers with higher accuracy, and finally *H(x)* outputs the results together.

B. Decision Tree Algorithm

Decision tree is a common algorithm in machine learning, whose core idea is to make predictions of new samples by constructing a tree model. A decision tree contains a root node, a number of internal nodes and a number of leaf nodes. Leaf nodes correspond to the decision results, and each internal node corresponds to an attribute