Computational Statistics (2020) 35:491–514
https://doi.org/10.1007/s00180-019-00916-9

ORIGINAL PAPER

Improving accuracy of financial distress prediction by considering volatility: an interval-data-based discriminant model

Rong Guan[1] · **Huiwen Wang**[2,3] · **Haitao Zheng**[2,4]

Received: 12 October 2017 / Accepted: 9 August 2019 / Published online: 20 August 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Financial distress prediction models are much challenged in identifying a distressed company two or more years prior to the occurrence of its actual distress, on the grounds that the distress signal is too weak to be captured at an early stage. The paper innovatively proposes to predict the distressed companies by a factorial discriminant model based on interval data. The main idea is that we use a new data representation, i.e., interval data, to summarize four-quarter financial data, and then build a interval-data-based discriminant model, namely *i*-score model. Interval data makes both average and volatility information comprehensively included in the proposed prediction model, which is expected to improve prediction performance on the distressed companies. A comparison based on a real data case from China's stock market is conducted. The *i*-score model is compared with five commonly used models that are based on numerical data. The empirical study shows that *i*-score model is more accurate and more reliable in identification of companies in high risk of financial distress in advance of 2 years.

Keywords Financial distress · Prediction · Interval data · Quarterly financial ratio · China's stock market

✉ Haitao Zheng
zhenghaitao@buaa.edu.cn

1 School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China

2 School of Economics and Management, Beihang University, Beijing 100191, China

3 Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing 100191, China

4 MoE Key Laboratory of Complex System Analysis and Management Decision, Beijing 100191, China

1 Introduction

An accurate identification of impending-to-distress companies is particularly important for investors, since such companies will be very likely to suffer from great losses or even declare bankruptcy in the coming years. Due to its importance, the last four decades have witnessed growing research efforts in this topic (Altman 1968; Merton 1974; Zavgren 1985; Tsai and Wu 2008; Xu et al. 2011). A common defect of these models exists in that the predictive power for distressed companies is relatively lower than that for healthy companies. For instance, Altman's z-score model well identified 94% healthy companies but only recognized 72% distressed companies in a lead time of 2 year. In other words, such models focus on picking up companies to construct a potential list for investment portfolio. Unlike this, however, this paper pays more attention to establishing a prediction model that recognize distressed companies 2 years prior to their distress as accurately as possible. Our prediction model is expected to provide a negative list for investment, which includes companies in high risk of distress in the coming 2 years, and thus plays an important role in risk avoidance for investors.

The lack of competence on distinguishing the distressed companies is probably arisen from the fact that the indications for impending distress is less than obvious, especially on condition of longer prediction lead time. To better capture the distress signals of companies, some recent literatures recommend to add financial volatility information into the modeling process. As well accepted to all, a higher degree of volatility in stock price or equity is likely to raise the cost of capital and thus to increase the likelihood of financial distress (Campbell and Taksler 2003). A number of research works, such as Dichev and Tang (2009) and Minton et al. (2002), provided evidence that a prediction model with volatility included as an explanatory variable has greater accuracy and lower bias than that without volatility information. When concerned with financial distress prediction, variability and unexpected deviation (Meyer and Pifer 1970), as well as the volatility of asset price (Merton 1974) have been demonstrated to play a significant role in measuring company default risk. More recently, Campbell et al. (2008) pointed out that distressed stocks experience higher standard deviation of stock returns than stocks with a low risk of failure. Chen et al. (2010) investigated the link between the idiosyncratic volatility and distress by sequential sorting. The results offered support to the conjecture that the idiosyncratic volatility exists conditional on distress risk. Although good efforts have been spent on demonstrating the merits of applying volatility in financial distress prediction, these research have not circumvented the following two limitations. First, only volatility of stock price, return or equity was taken into account. We conjecture that fluctuation in other financial ratios also contributes to forecasting distress. Second, volatility information is included as an explanatory variable, which will eventually increase the number of parameters to be estimated.

To overcome these limitations, we suggest a novel data representation of volatility information, i.e., interval data. To be more specific, we firstly provide a list of financial ratio indicators, and then use an interval data to summarize four-quarter records for each financial ratio. As a result, both the average level and the fluctuation range of the concerning quarterly records are bundled together into an interval (Bock and Diday

2000; Billard and Diday 2003; Diday and Noirhomme-Fraiture 2008). In such an innovative way, the supplement of volatility information does not take the price of increasing the difficulty of parameters estimation. Notably, we are not the first to measure volatility or variability by interval data [see for instance Wang et al. (2012), He and Hu (2007)]. But to the best of our knowledge, this paper is the first to propose a prediction model with volatility information of financial ratios included in interval data representation. It will be proved in later sections that the proposed model has a strengthened predictive power of distressed companies.

In this paper, we provide a detailed introduction to the key related methodology of an interval-data-based prediction model using quarterly financial ratios, namely *i*-score model. Importantly, its merits in recognizing distressed companies are demonstrated by a comparative study with five commonly seen models that are based on numerical data, including Fisher's Discriminant Analysis, Logit Regression, Support Vector Machine, Classification Tree, and Random Forests. In an empirical case from China's stock market, the *i*-score model shows a remarkable superiority in pre-identification of financially distressed companies. Specifically, the inspiring results have corroborated the contributions of our work in the following two aspects. (1) In a single experiment based on pair-wise matching distressed and healthy companies, 83.67% of distressed companies in a testing dataset are correctly predicted by *i*-score model. Within the numerical-data-based models, only Classification Tree achieves the same high level of accuracy, but mistakenly identifies more healthy companies as distressed companies than *i*-score model does. Similar results are also seen in repeated experiments. (2) When dealing with an imbalanced dataset, with distressed and healthy companies mixed in 1:3 ratio, *i*-score models still performs better in terms of accurately identifying distressed companies than the five numerical-data-based models. This again reveals the strong competence of *i*-score model on capturing less-than-obvious distress signals.

1.1 Overview

The remaining of this paper is organized as follows. A brief introduction to interval data as well as the method to transform quarterly financial records into an interval data is presented in Sect. 2. In Sect. 3, we propose the *i*-score model and describe its main steps for financial distress prediction. Afterwards, an empirical case concerning China's listing companies is introduced in Sect. 4. We provide a detailed description on how to apply *i*-score model as well as a mechanism interpretation of the model, carry out a cost-benefit analysis, and make a comparative study between *i*-score model and five commonly used models that are based on numerical data. The paper ends up with some conclusions in Sect. 5.

2 Interval data, quarterly financial ratio and volatility

In this section, we firstly introduce some basic concepts and notations of interval data, and then present a novel representation that summarizes quarterly financial ratios in the form of interval data.

In mathematics, an interval data, say $x = [\underline{x}, \overline{x}]$, refers to a set of real numbers that lie between the lower boundary $\underline{x}$ and the upper boundary $\overline{x}$, where a constraint that $\underline{x} \leq \overline{x}$ must be satisfied. Generally speaking, interval data are often used to record data with measuring errors (Moore 1966; Sunaga 2009), or to represent data with uncertainty information (Lee and Huang 2009; Li 2013). According to Symbolic Data Analysis (Billard and Diday 2003; Diday and Noirhomme-Fraiture 2008), interval data can also be used to summarize large-scaled samples of a certain category. In this way, the information content of both central tendency and variability of the concerning sample data can be preserved, which will then benefit the subsequent statistical modeling. Enlightened by this, we adopt interval data to represent multi-phase financial data in this paper. For the sake of discussion, we hereinafter called a number in the real space as a numerical data.

Given a company that we concern, suppose that one of its financial ratios, say *Return on Total Assets* (*RoTA*), has been observed quarterly. Denote the four-quarter numerical data as $x(1)$, $x(2)$, $x(3)$ and $x(4)$, respectively. We propose to transform these four numbers into an interval data $x = [\underline{x}, \overline{x}]$, where the lower boundary $\underline{x} = \min\{x(1), x(2), x(3), x(4)\}$ and the upper boundary $\overline{x} = \max\{x(1), x(2), x(3), x(4)\}$. To facilitate understanding of the transformation, some examples are listed in Table 1.

By using interval data, statistical modeling will benefit in the following two aspects. On one hand, the information content of volatility of financial ratios can be taken into account. According to our definition, the length of an interval reflects the fluctuation range of four quarterly records. For better interpretation, both quarterly data and interval data in Table 1 are visualized in Fig. 1. As can be seen in Fig. 1a, each company is drawn as a curve, which fluctuates over four quarters. When described by interval data, each company is presented as a vertical line segment, whose bottom/top represents the lower/upper bound of the interval (see Fig. 1b). Obviously, a larger fluctuation in quarterly data corresponds to a longer line segment. In other words, interval data is indeed a good representation for the information of fluctuation. It is not difficult to understand that the fluctuation range conveys information of volatility. Therefore, the information content of volatility of financial ratios can be taken into account, when interval data is adopted as data representation. More importantly, there is no need to add an extra variable for volatility information in the model. It is due to the fact that an interval data can comprehensively summarize the overall information, including both average and volatility, of the concerning financial ratios. The average level is represented as

Table 1 Quarterly data of *RoTA* of five companies and their transformed interval data

Company	Quarterly data				Interval data
	Q1	Q2	Q3	Q4	
A	− 0.12	1.67	− 1.96	− 5.44	[− 5.44, 1.67]
B	1.08	1.97	0.85	− 0.72	[− 0.72, 1.97]
C	1.4	6.2	5.38	0.66	[0.66, 6.2]
D	1.76	− 4.71	− 4.03	− 8.75	[− 8.75, 1.76]
E	− 2.24	− 2.47	− 2.41	− 3.09	[− 3.09, − 2.24]

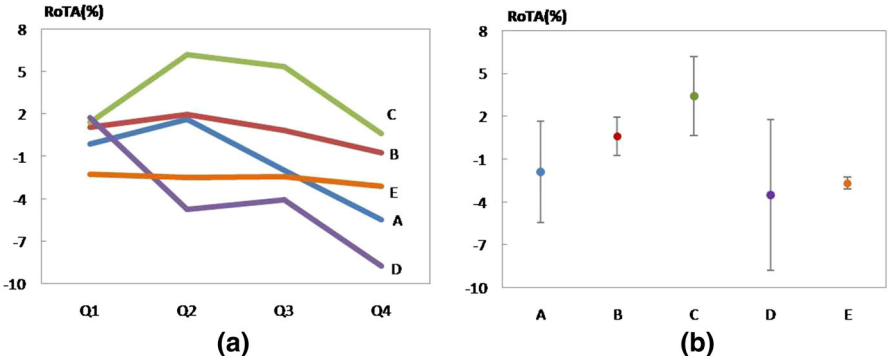

Fig. 1 Quarterly data and interval data of five company examples

the location of the midpoint in each line segment (see Fig. 1b). In consideration of the above-mentioned points, prediction performance based on interval data is expected to be better than that based on numerical data, which will be examined in later sections. **REMARK.** In fact, transforming the quarterly financial data into an interval would ignore the fact that the records are ordered. Therefore, another idea is to incorporate time factor into intervals, which might be useful for the classification of the healthy/distressed companies. Since this is out of the scope of this article, we leave this as a future topic here.

3 Method

Given a sample of companies along with their quarterly financial ratios of the current period, our model aims to distinguish as accurately as possible whether or not each company will be trapped into financial distress 2 years later, according to the volatility information of financial ratios. A common way to realize this goal is to accomplish a discriminant analysis, which provides a combination of a few financial ratios that best separates two groups of observations (Altman 1968; Altman et al. 1977). In this paper, we will set up an interval-data-based discriminant analysis model to classify companies into a healthy group or a distressed group. The last two decades have witnessed a number of research work on the methodology of discriminant analysis model of interval data, a majority of which are about Factorial Discriminant Analysis of interval data (iFDA). In an early study from Lauro et al. (2000), a generalization of factorial discriminant analysis on interval data is proposed. The method heavily relies on a recoding process that transforms interval data into numerical data. Afterwards, Silva and Brito (2006) established a distributional approach for iFDA, hereinafter called as D-iFDA, which extends the framework of factorial discriminant analysis by assuming a distribution in each observed interval. It is demonstrated that D-iFDA shows a superiority over the method proposed by Lauro et al. (2000) in terms of classification accuracy. More recently, some parametric methods, mainly evolved from D-iFDA, have been discussed and assessed (Silva and Brito 2015). In consideration of its advantages, we choose D-iFDA to accomplish financial distress prediction,

details of which will be introduced in what follows. In order to improve prediction performance, a minor modification to its classification rule will be provided.

For sake of convenience, some notations should be firstly presented. For any given company, we assume that it comes from a population of distressed companies (denoted as π_1) or a population of healthy companies (denoted as π_2). Suppose that we have a sample of n company observations, with n_h companies coming from the h-th ($h = 1, 2$) population. Concerned with the issue of financial distress prediction, each company is known as either distressed or healthy in year T. To predict the health status in a lead time of $L(L > 0)$ years, we collect four-quarter data of p financial ratios in year $T - L$. After all quarterly data are transformed into interval data by the transformation method described in Sect. 2, we obtain an $n \times p$ matrix $\mathbf{X}^I$, whose superscript I stands for *Intervals*. Each unit of $\mathbf{X}^I$ is an interval data, i.e., $x_{ij} = [\underline{x}_{ij}, \overline{x}_{ij}]$, $1 \le i \le n$, $1 \le j \le p$. The i-th company observation is denoted as $\mathbf{x}_i'$, and the j-th financial ratio is denoted as X_j. The index set of observations from the h-th population is denoted as C_h.

According to the well-known theory of FDA, a multivariate observation $\mathbf{x}_i'$ is transformed into a univariate observation, denoted as z_i, such that z_i is derived from populations π_1 and π_2 that were separated as much as possible. Noticing that $\mathbf{x}_i'$ is described by interval data in our case, z_i is also an interval-valued observation defined by

$$z_i = \mathbf{a}' \otimes \mathbf{x}_i, \tag{1}$$

where $\mathbf{a} = \left(a_1, a_2, \ldots, a_p\right)' \in \mathbb{R}^p$ is called as combination coefficient vector, and $z_i = [\underline{z}_i, \overline{z}_i]$, namely a projection of the i-th observation on $\mathbf{a}$, such that

$$\underline{z}_i = \sum_{a_j > 0} a_j \underline{x}_{ij} + \sum_{a_j \le 0} a_j \overline{x}_{ij}, \tag{2}$$

$$\overline{z}_i = \sum_{a_j \le 0} a_j \underline{x}_{ij} + \sum_{a_j > 0} a_j \overline{x}_{ij}. \tag{3}$$

According to Silva and Brito (2006), $\mathbf{a}$ can be solved by the first eigenvector of matrix $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W} = [w_{jj'}]_{p\times p}$ and $\mathbf{B} = [b_{jj'}]_{p\times p}$ respectively denote the within-class inertia matrix and the between-class inertia matrix, such that

$$w_{jj} = \frac{1}{3n}\sum_{i=1}^{n}(\underline{x}_{ij}^2 + \underline{x}_{ij}\overline{x} + \overline{x}_{ij}^2) - \sum_{h=1}^{2} n_h \left(m_j^{(h)}\right)^2, \tag{4}$$

$$w_{jj'} = \frac{1}{4n}\sum_{i=1}^{n}(\underline{x}_{ij} + \overline{x}_{ij})(\underline{x}_{ij'} + \overline{x}_{ij'}) - \sum_{h=1}^{2} n_h m_j^{(h)} m_{j'}^{(h)}, \quad j \ne j', \tag{5}$$

$$b_{jj'} = \sum_{h=1}^{2} \frac{n_h}{n} m_j^{(h)} m_{j'}^{(h)} - m_j m_{j'}, \tag{6}$$

where $m_j = \frac{1}{2n}\sum(\underline{x}_{ij} + \overline{x}_{ij})$ and $m_j^{(h)} = \frac{1}{2n_h}\sum_{i\in C_h}(\underline{x}_{ij} + \overline{x}_{ij})$, for $j = 1, 2, \ldots, p$.

After the combination coefficient vector is obtained, we should determine a classification rule next. As proposed in Silva and Brito (2006), a p-dimensional interval-valued observation, say $\mathbf{x}_0$, is classified to π_{h^*} such that

$$h^* = \arg\min_{h=1,2} \left[\frac{1}{n_h} \sum_{i \in C_h} \sqrt{\lambda} D\left(z_0, z_i\right) \right], \tag{7}$$

where $D(\cdot, \cdot)$ is a distance function for any two interval data. Here, we adopt Wasserstein distance of interval data (Irpino and Verde 2008), i.e.,

$$D(z_0, z_i) = \sqrt{(c_0 - c_i)^2 + \frac{1}{3}(r_0 - r_i)^2}, \tag{8}$$

where $c_0 = \frac{1}{2}(\underline{z}_0 + \overline{z}_0)$, $c_i = \frac{1}{2}(\underline{z}_i + \overline{z}_i)$, $r_0 = \frac{1}{2}(\overline{z}_0 - \underline{z}_0)$, and $r_i = \frac{1}{2}(\overline{z}_i - \underline{z}_i)$. Apparently, this classification rule of D-iFDA requires massive calculations by traversing the computation of distance between the to-be-classified observation z_0 and each of the n_h observations in C_h. Besides, the classification result may be badly biased if there exist some outliers in C_1 or C_2.

To avoid these pitfalls, we modify Eq. (7) as

$$h^* = \arg\min_{h=1,2} D_W(z_0, z_h^{pro}), \tag{9}$$

where $z_h^{pro} = [\underline{z}_h^{pro}, \overline{z}_h^{pro}]$ represents the prototype of observations from the h-th population, such that

$$\underline{z}_h^{pro} = \frac{1}{n_h} \sum_{i \in C_h} \underline{z}_i, \quad \overline{z}_h^{pro} = \frac{1}{n_h} \sum_{i \in C_h} \overline{z}_i. \tag{10}$$

Notably, the revised rule in Eq. (9) has sharply cut down the amount of computations, compared with the previous rule in Eq. (7). On account of this minor modification, we call it as i-score model. The main steps of i-score model, including a modeling part and a predicting part, are listed as below.

(1) Modeling part:

(a) For a sample of companies, collect their four-quarter financial data of p financial ratios in year $T - L$, as well as their health status in year T, i.e., healthy or distressed. Transform their quarterly data into interval data and thus obtain an interval-valued matrix $\mathbf{X}^I$.

(b) Estimate the combination coefficient vector $\mathbf{a}$ by an eigen-decomposition of $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$, as defined in Eqs. (4)–(6), are respectively the within-variance matrix and the between-variance matrix for $\mathbf{X}^I$.

(2) Predicting part:

(a) For any company to be predicted, summarize its quarterly financial ratios into an interval data and construct an observation vector $\mathbf{x}_0'$ of interval data. Transform $\mathbf{x}_0'$ into $z_0 = [\underline{z}_0, \overline{z}_0]$ by $z_0 = \mathbf{a}' \otimes \mathbf{x}_0$, where $\underline{z}_0$ and $\overline{z}_0$ can be obtained according to Eqs. (2) and (3).

(b) Predict the health status of this company by allocating it to the h^*-th group according to Eq. (9)

So far, we have described the methodology of *i*-score model in details. The next step is to assess its prediction performance. Two evaluation indexes will be used. The first index is *accuracy of total* (*AOT*), defined as the ratio of the correctly-predicted observation number to the total observation number. The second index, called as *accuracy of distress* (*AOD*), is defined as the ratio of the correctly-predicted distressed observation number to the total distressed observation number. Clearly, the index of *AOT* provides an assessment of overall predictive accuracy, and *AOD* shows the predictive power for distressed companies. Since our primary goal is to create a negative list for investment, the index of *AOD* is worth of our special attention.

4 Data and results

Does *i*-score model provide an accurate detection of distress status of companies in advance, according to their current performance on certain financial ratios? Furthermore, is *i*-score model more capable of identifying companies in high risk of distress, due to its novel representation of volatility information that is contained in four-quarter records of financial ratios? To address these issues, an empirical study on China's stock market is presented in this section.

To start with, we introduce a set of financial ratios used for distress prediction. Then, we describe the method of sample selection as well as our data source. In order to guarantee overall performance, exploratory data analysis are first conducted. Afterwards, we show how to apply our *i*-score model, and display a summary of the predicting outcomes. Meanwhile, a cost-benefit analysis is provided to justify the focus on identifying the distressed companies. To examine the superiority of *i*-model, five numeircal-data-based models commonly used in previous research are included for comparison. Ultimately, the superiority of the *i*-score model is supported by the comparative analysis.

4.1 Variables

Generally speaking, at least five aspects of financial performance, including liquidity, profitability, solvency, leverage, and activity, should be taken into account when concerned with distress prediction. We carefully choose several financial ratios in each aspect, as listed below.

- Liquidity: *Working Capital/Total Asset* (*WC/TA*),*Current Ratio* (*CR*), *Quick Ratio* (*QR*).

- Profitability: *Return on Total Assets* (*RoTA*), *Retained Earnings/Total Asset* (*RE/TA*), *Return on Equity* (*RoE*),*Operating Margin* (*OM*), *Total Profits/ Operating Revenue* (*TP/OR*), *Net Profit Margin* (*NPM*).
- Solvency: *Debt-to-Equity Ratio* (*DtER*).
- Leverage: *Debt Ratio* (*DR*).
- Activity: *Operating Cash Flow/Operating Revenue* (*OCF/OR*), *Operating Revenue Growth* (*ORG*).

4.2 Sample selection

The definition of finance-distressed companies shall be given before predictive modeling. According to the special treatment scheme of China's securities market, a listed company is labeled with *ST* (short for "Special Treatment") before its stock name, if it has reported negative net profit in its Annual Financial Report during the past two consecutive years. Generally speaking, such companies, hereinafter called as *ST* companies, are highly likely to experience a sharp decrease in their stock price, or even face high risk of delisting. Therefore, *ST* companies are deemed as financially distressed companies in this paper, like previous studies (Sun and Li 2008; Xu et al. 2011). In the period from 1 January 2004 to 31 December 2011, a total number of 165 listed companies have been subject to *ST* in China's stock market. To construct a paired sample, we carefully select one healthy company for each chosen *ST* company. We recognize that the distressed company group are not homogenous with respect to asset size and industry. Therefore, the healthy company group are chosen on a stratified random basis. The selection criterion is intuitive and simple: pick up a healthy company with the most similar asset size to its matched *ST* company, on the premise that both companies come from the same industry sector.

Next, financial records of the sample companies are extracted from their financial disclosure statements. Our data source is Wind Financial Terminal (www.wind.com.cn/en/wft.html), a financial database disclosing information of listed companies in China's stock market. For each company pair, we use financial data L annual periods prior to the year when the distressed company is subject to *ST*. For instance, if the distressed company is labeled with *ST* in the year 2004, financial data in the year of $2004 - L$ for both companies in this pair will be gathered. In this paper, we set the prediction lead time $L = 2$. The reasons for this are two-fold. On one hand, as mentioned above, a listed company will become an *ST* company after it suffers negative profits for two consecutive years. Therefore, we attempt to capture the forthcoming investment risk from the beginning of becoming *ST*. On the other hand, setting a 2-year prediction lead time is commonly seen in previous studies concerning China's stock market.

Through preliminary checks, we find out that there are 12 *ST* companies with incomplete financial records and 6 *ST* companies with extremely high or low records. To guarantee overall performance, we simply delete the corresponding 18 company pairs. So far, we have constructed a sample with two groups, i.e., both healthy and distressed, in a total amount of 147 company pairs. Table 2 presents a summary of the asset size for company pairs in different industries. For reader's convenience, we

Table 2 A summary of asset size of company pairs in different industries (in 10,000 Yuan)

Industry	# of company pairs	Distressed company			Healthy company		
		Mean	Min.	Max.	Mean	Min.	Max.
Agriculture, forestry, livestock farming, fishery	5	209684.6	33199.0	708785.0	76449.2	51000.0	113979.0
Comprehensive	9	303860.0	68950.0	1488023.0	206732.6	77774.0	417178.0
Electronics	6	158215.0	66223.0	311515.0	295131.7	51482.0	468170.0
Food and beverage	10	137706.2	15577.0	390625.0	155307.2	64992.0	450184.0
Information and technology	12	229720.4	21439.0	1579680.0	131178.2	33902.0	314285.0
Machinery	19	102864.2	11119.0	423144.0	148989.6	52560.0	457340.0
Metals and non-metals	10	168638.0	75257.0	253661.0	159858.5	108058.0	314426.0
Mining	2	656970.5	35299.0	1278642.0	124112.0	92247.0	155977.0
Other manufacturing	1	41950.0	41950.0	41950.0	228784.0	228784.0	228784.0
Paper and printing	5	67562.2	19464.0	182870.0	124571.4	50638.0	235514.0
Petrochemicals	25	184982.7	27562.0	1834413.0	194588.4	37087.0	999348.0
Pharmaceuticals	9	88894.8	17982.0	166139.0	110092.9	44272.0	257245.0
Real estate	7	105214.3	24518.0	298451.0	232304.4	59765.0	649175.0
Textiles and apparel	10	82395.1	33012.0	199145.0	140680.7	35794.0	482635.0
Timber and furnishings	1	85704.0	85704.0	85704.0	92388.0	92388.0	92388.0
Transportation	4	2001121.8	88376.0	7297294.0	2226446.8	105522.0	7928000.0
Utilities	5	312351.6	88336.0	560721.0	391479.4	107165.0	829613.0
Wholesale and retail trade	7	170548.9	75711.0	635257.0	102102.3	77282.0	168466.0

provide the stock ID information of the selected company pairs, which are left in the Appendix for the clarity of the context.

4.3 Exploratory data analysis

To ensure the prediction accuracy, financial ratios with too-weak predictive power shall be excluded. Besides, multicollinearity, mainly shown as high correlations between financial ratios, may have some disturbance on the overall performance of *i*-score model. In consideration of these two points, we carry out exploratory data analysis prior to predictive modeling.

First, we estimate the correlation coefficients for 13 financial ratios (see Table 3). According to Billard and Diday (2003), i.e.,the correlation coefficient of any two interval-valued variables, say $\mathbf{X}_j^I$ and $\mathbf{X}_{j'}^I$ $(j, j' = 1, 2, \ldots, p)$, is defined by

$$corr(\mathbf{X}_j^I, \mathbf{X}_{j'}^I) = \begin{cases} \frac{COV(\mathbf{X}_j^I, \mathbf{X}_{j'}^I)}{\sqrt{D(\mathbf{X}_j^I)} \cdot \sqrt{D(\mathbf{X}_{j'}^I)}}, & j \neq j' \\ 1, & j = j' \end{cases} \tag{11}$$

where

$$\begin{aligned} COV(\mathbf{X}_j^I, \mathbf{X}_{j'}^I) =& \frac{1}{4n} \sum_{i=1}^{n} (\underline{x}_{ij} + \overline{x}_{ij})(\underline{x}_{ij'} + \overline{x}_{ij'}) \\ & - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (\underline{x}_{ij} + \overline{x}_{ij}) \right] \left[\sum_{i=1}^{n} (\underline{x}_{ij'} + \overline{x}_{ij'}) \right], \quad j \neq j' \\ D(\mathbf{X}_j^I) =& \frac{1}{3n} \sum_{i=1}^{n} (\underline{x}_{ij}^2 + \underline{x}_{ij} \overline{x}_{ij} + \overline{x}_{ij}^2) - \frac{1}{4n^2} \left[\sum_{i=1}^{n} (\underline{x}_{ij} + \overline{x}_{ij}) \right]^2. \end{aligned}$$

Generally speaking, a correlation with its absolute value greater than 0.8 is interpreted as very strong, and 0.6 up to 0.8 as strong. Remarkably, four variable groups in Table 3 appear strong or very strong association, including (1) *WC/TA*, *RE/TA* and *DR*, (2) *CR*, *QR* and *DtER*, (3) *OM* and *TP/OR*, (4) *NPM* and *OCF/OR*. In order to avoid severe multicollinearity problem, only one financial variable shall be chosen from each of these groups.

Then, for each financial ratio, we apply *i*-score method, and calculate the index of *AOT* (see the lower part of Table 3). We select from each group only one financial ratio with the highest value of *AOT*. Ultimately, 7 ratios are preserved for the subsequent modeling, including *Working Capital/Total Asset* (*WC/TA*), *Current Ratio* (*CR*), *Return on Total Assets* (*RoTA*), *Return on Equity* (*RoE*), *Total Profits/ Operating Revenue* (*TP/OR*), *Net Profit Margin* (NPM), and *Operating Revenue Growth* (ORG).

Next, descriptive analysis is performed. Figure 2 shows both distressed companies (light brown crosses in left subfigure and light brown rectangles in right subfigure) and healthy companies (black circles in left subfigure and black rectangles in right

Table 3 Correlation coefficients and *AOT* of 13 financial ratios

Financial ratios	Correlation coefficient						
	WC/TA	*CR*	*QR*	*RoTA*	*RE/TA*	*RoE*	*OM*
WC/TA	1.0000						
CR	0.5011	1.0000					
QR	0.4354	0.9220	1.0000				
RoTA	0.1867	0.1061	0.0885	1.0000			
RE/TA	0.6820	0.3559	0.3129	0.3251	1.0000		
RoE	0.0885	0.0290	0.0234	0.1391	−0.0090	1.0000	
OM	−0.0841	−0.0244	−0.0185	−0.0722	−0.0763	−0.0616	1.0000
TP/OR	0.0156	0.0094	0.0066	0.2625	0.0475	0.0906	−0.6786
NPM	−0.1494	−0.0314	−0.0252	0.4849	−0.0534	0.0857	−0.0043
DtER	0.3252	0.7603	0.7391	0.0758	0.2768	0.0154	−0.0163
DR	−0.7511	−0.4359	−0.3858	−0.2853	−0.8294	0.0107	0.0623
OCF/OR	0.1584	0.0302	0.0243	−0.4459	0.0660	−0.0689	−0.1516
ORG	0.0187	−0.0067	−0.0015	0.0171	−0.0845	0.0031	−0.0085
AOT (%)	72.04	67.43	58.22	73.36	69.08	65.79	50.33

Financial ratios	Correlation coefficient					
	TP/OR	*NPM*	*DtER*	*DR*	*OCF/OR*	*ORG*
TP/OR	1.0000					
NPM	0.3095	1.0000				
DtER	0.0078	−0.0174	1.0000			
DR	−0.0278	0.0704	−0.3871	1.0000		
OCF/OR	−0.1545	−0.7188	0.0082	−0.0768	1.0000	
ORG	0.0035	−0.0103	−0.0073	−0.0277	−0.0077	1.0000
AOT(%)	50.66	50.99	58.55	69.41	49.34	51.64

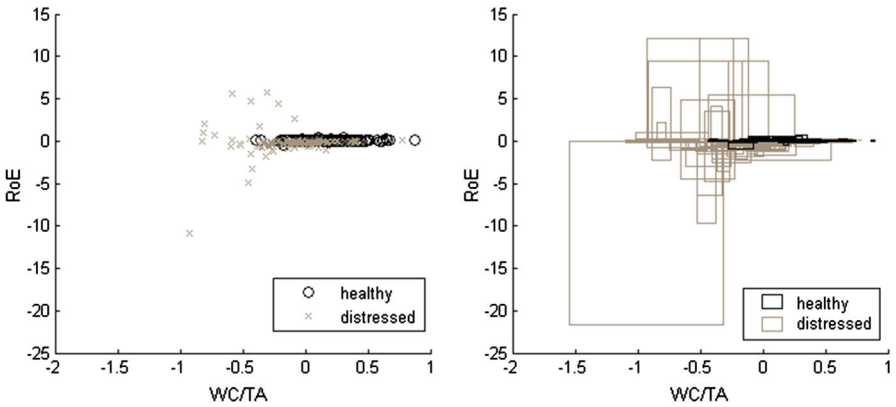

Fig. 2 Visualizations of distressed and healthy companies by two financial ratios of *WC/TA* and *RoT A*. The left-hand-side subfigure is based on numerical data, and the right-hand-side subfigure is based on interval data (color figure online)

subfigure) in terms of *Working Capital/Total Asset* (*WC/TA*) and *Return on Equity* (*RoE*). Some distinguishable patterns exist in these two subfigures. The left subfigure is a scatterplot based on numerical data, i.e., the mean value of the four-quarter. Apparently, a large part of distressed companies (light brown crosses) appear mixed with healthy companies (black circles). In other words, the two groups do not clearly distinguish from each other. Notably, this observation is a good evidence to support the conclusion raised by Altman (1968), which suggests that the indications for impending distress may become less clear when the prediction lead time exceeds 1 year. In the right subfigure, each rectangle represents a company, which is a common way to visualize samples described by two interval-valued variables [see Wang et al. (2012)]. The centers of the rectangles correspond to the midpoints of the intervals, and the width and height respectively represent the lengths of the intervals in the horizontal axis (i.e., *WC/TA*) and the vertical axis (i.e., *RoE*). By this easily understood visualization, we know that a bigger rectangles correspond to longer interval(s) in one/two dimension(s), and thus embody greater volatility in financial ratio(s). Remarkably, the two rectangle groups mainly differ from each other in terms of size. More specifically, black rectangles (healthy companies) tend to appear in smaller size, whereas most of light brown rectangles (distressed companies) cover larger areas. That is, the company groups are more distinct to each other in size, rather than in location. Similar findings can also be seen within other variables (see Figs. 4, 5 in the "Appendix"). Accordingly, prediction accuracy is expected to improve when interval data is used to describe volatility information of financial ratios of companies.

4.4 Predicting outcomes

In what follows, we present how to apply *i*-score model in this example. We randomly divide the whole sample into two subsets, i.e., two thirds as a training dataset (with 98 company pairs) and the remaining one third as a testing dataset (with 49 company

pairs). We first use the training dataset to train the model, as stated in the modeling part of the *i*-score model. In this example, the combination coefficient vector is estimated as (0.9989, 0.0092, 0.0371, −0.0003, 0.0284, 0.0018, 0.0010). And the prototype of the distressed group and that of the healthy group, defined by Eq. (10), are $z_1^{pro} = [-0.8495, 0.1088]$ and $z_2^{pro} = [0.3373, 0.6923]$, respectively. In the testing step, the health status of each company observation in the testing dataset is predicted according to the predicting part of the *i*-score model.

To demonstrate the merits of *i*-score model in financial distress prediction, we carry out a comparison between *i*-score model and five commonly used models, including Factorial Discriminant Analysis (FDA), Logit Regression (LR), Support Vector Machine (SVM), Classification Tree (CT), and Random Forests (RF). While the *i*-score model is applied on the interval-valued dataset, the five models participating comparison use numerical data, i.e., means of intervals. All experiments are based on Matlab. We use the ready-made functions for the five numerical-data-based models, with the parameters being set by default.

Table 4 reports a summary of predicting outcomes in the testing dataset. There exist some remarkable observations in this table. (1) Concerning the index of *AOT*, there is no evident difference between *i*-score model and the five models. Our *i*-score model correctly predicts 83 out of 98 companies in total ($AOT = 81.63\%$). Within the five models using numerical data, SVM reaches the highest values of *AOT* (84.69%), whereas the model of Classification Tree only achieves 77.55%. (2) When the evaluation index of *AOD* is concerned, however, *i*-score model shows an advantage over the other five models. Remarkably, *i*-score model has attained a fairly high level of *AOD*, i.e., 83.67%, as high as the Classification Tree model, followed by Random Forests (81.63%) and SVM (79.59%). It is consistent with our expectation that *i*-

Table 4 A comparison within factorial discriminant analysis (FDA), logit regression (LR), support vector machine (SVM), classification tree (CT), random forests (RF), and *i*-score model in terms of *accuracy of all (AOT)* and *accuracy of distress (AOD)* in the testing dataset

Models	Original groups	Predicted groups		AOT(%)	AOD(%)
		Distressed	Healthy		
FDA	Distressed	38	11	82.65	77.55
	Healthy	6	43		
LR	Distressed	36	13	81.63	73.47
	Healthy	5	44		
SVM	Distressed	39	10	84.69	79.59
	Healthy	5	44		
CT	Distressed	41	8	77.55	83.67
	Healthy	14	35		
RF	Distressed	40	9	80.61	81.63
	Healthy	10	39		
i-score	Distressed	41	8	81.63	83.67
	Healthy	10	39		

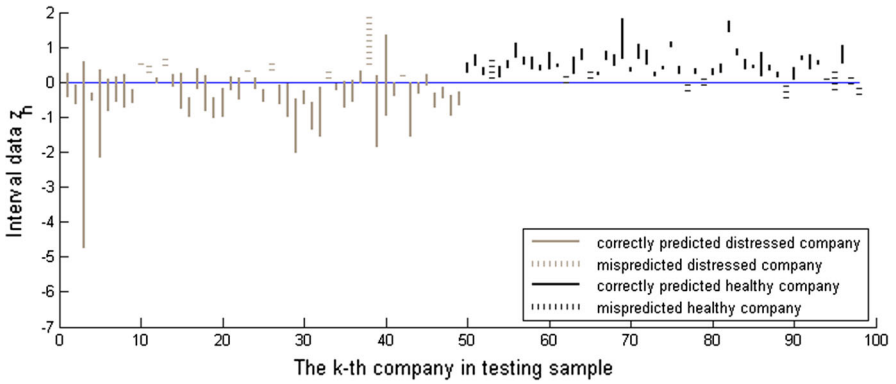

Fig. 3 Univariate projection z_k in the testing dataset

score model pays more attention to recognizing distressed companies. (3) Taking a further comparison between *i*-score model and Classification Tree, we find out that Classification Tree mistakenly identifies 14 healthy companies as distressed companies in the concerning sample, whereas this number is only 10 for *i*-score model. Therefore, both models achieve the goal of investment risk avoidance by correctly identifying distressed companies, but Classification Tree may miss out of some investment returns.

Why does the *i*-score model have a strong predictive power for distressed companies? Does the *i*-score model identify some patterns that could distinguish between the distressed companies and the healthy companies based on the selected seven financial ratios? The visualization provided in Fig. 3 is helpful to answer these questions, which further enables us to better understand the *i*-score model. In the figure, each vertical line represents an interval-valued projection z_k in the testing dataset. Its bottom and top are respectively the lower and the upper bound of the interval data. According to the original labels, distressed and healthy companies are respectively drawn as light brown and black lines. Those mistakenly predicted observations are represented by dashed lines. In addition, we draw an auxiliary line to facilitate understanding (see the horizontal line whose value is taken as 0).

Apparently, most distressed companies appear as longer lines, whereas most lines of healthy companies generally look shorter. Besides, it is obvious that the midpoints of most light brown lines lie below the auxiliary line, while a majority of black lines are located above the line. In other words, after projected on the combination coefficient vector by the *i*-score model, the two company groups distinguish each other in terms of their midpoint values and interval ranges. Taking a closer look at those mistakenly predicted observations, represented by dashed lines in this figure, we find out some interesting patterns. That is, light brown lines standing above the auxiliary line and black lines lying below the auxiliary line are more likely to be mistakenly predicted. Accordingly, we know that a company in label of distress will be probably mis-predicted as a healthy company, if it is converted into a short interval, or with a high value of midpoint. On the contrast, a healthy company will also be mistaken for a company in high risk of distress, if it experience bad performance and/or large fluctuations in financial ratios. According to the modeling outputs, the prototype of

the distressed group is $z_1^{pro} = [-0.8495, 0.1088]$, whereas that of the healthy group is $z_2^{pro} = [0.3373, 0.6923]$. As defined by the *i*-score model, each univariate observation is predicted according to its similarity with the prototypes of the two groups. Therefore, it is not difficult to understand that univariate observations with longer intervals as well as lower values of midpoints are more likely to be predicted as distressed, considering that it has a closer distance to the prototype of the distressed group than to that of the healthy group. From this, we know that the mechanism of the *i*-score model, i.e., to recognize distressed company is to construct an interval-valued prototype by summarizing the distress signals of some known observations according to their financial performance, especially their financial volatility. Furthermore, according to Eqs. (1)–(3), all of which describe how to project a p-dimensional interval-valued observation $\mathbf{x}_i'$ into a univariate observation z_i, we know that the combination coefficient vector $\mathbf{a}'$=(0.9989, 0.0092, 0.0371, −0.0003, 0.0284, 0.0018, 0.0010) is useful to identify patterns that could distinguish between the distressed companies and the healthy companies based on the selected 7 financial variables. To be specific, three variables, including *WC/TA*, *RoTA*, and *TP/OR*, have relatively more important influence on the prediction, since the absolute values of their coefficients are relatively higher than those of other four variables. Companies have lower values or larger volatility on quarterly data of these three variables are more likely to be predicted as distressed.

4.5 Cost-benefit analysis

To justify the focus on correctly identifying distressed companies, we provide a cost-benefit analysis below. Let us consider two scenarios of mistaken prediction. When a distressed company is wrongly predicted as healthy, investors will buy the stock of the company and will then be very likely to suffer great losses, which is hereinafter called as a True-Negative (TN) cost. On the contrary, when a healthy company is misjudged as a distressed company, investors will decide not to buy the stock of the company and will then be prone to miss out on some benefits. Such benefits are actually an opportunity cost, which is called as a False-Positive (FP) cost in what follows. If the TN cost is greater than the FP cost, it demonstrates that we should focus on correctly identifying distressed companies.

For each model, we can calculate its TN cost and FP cost based on the predicting outcomes. To be specific, we use its predicting outcomes on the testing dataset to construct two investment portfolios, i.e., an TN portfolio and an FP portfolio. For simplicity, all stocks in each portfolio are equally weighted. To obtain the yield for each portfolio, we need to determine when to buy and when to sell each stock in the portfolio. For an actually distressed company, we calculate its 3-month return from March 31st to June 30th in the year when it was labeled with ST. As for its matched healthy company, the calculation period is the same. Started from March 31st is due to the fact that listed companies are required to announce their annual statements before that date and their actual labels are known by then. After obtaining returns for all stocks, we then calculate the average return for each portfolio. Results based on different models are shown in Table 5.

Table 5 Returns of portfolio based on predicting outcomes in the testing dataset by different models

Original label	Predicted label											
	Distressed *(positive)*						Healthy *(negative)*					
Distressed *(true)*	− 14.69						i− score	FDA	LR	SVM	CT	RF
							− 22.72	− 13.83	− 16.36	− 20.25	− 22.76	− 25.60
Healthy *(false)*	i-score	FDA	LR	SVM	CT	RF	− 3.34					
	8.76	− 1.70	− 19.67	− 15.58	− 3.86	− 9.93						

Numbers in the table represent the 3-month returns (%) of investment decisions by God's view or according to predicting outcomes based on different models. By the phrase of "God's view", we mean that we are assumed to know the actual label of each company

In the whole sample, buying distressed companies and buying healthy companies will respectively obtain returns of −16.00% and −3.56% according to God's view

Apparently, according to God's view, buying distressed companies does suffer from greater losses than buying healthy companies, whether in the testing sample (−14.69% vs. −3.34%) or the whole sample (−16.00% vs. −3.56%). However, people can only make investment decisions according to the prediction of a certain model. In Table 5, the upper right corner shows the TN costs by different models, whereas the lower left corner presents the FP costs. We do not compare the TN cost (or the FP cost) within these six models, since the different investment times of portfolios makes the comparison less meaningful. Instead, attention should be focused on the comparison between the TN cost and the FP cost of the same model. Remarkably, regardless of models, the TN cost is much higher than the FP cost, except for the LR model. So far, we have demonstrated that it is more important to correctly identify distressed companies than to correctly recognize healthy companies.

4.6 Robust analysis

Now, we move on to some robust analysis. The results will enable us to further assess the effectiveness of the *i*-score model in predicting financial distress. To be specific, we aim to seek for answers to the following questions: (1) We use a balanced sample in Table 4, but in the actual market there are much fewer distressed companies and much more healthy companies. Will the prediction outcomes be affected by the proportions of distressed companies and healthy companies? What if we use an imbalanced sample? (2) The current results presented in Table 4 are predicated on a specific selection of variables. Will the prediction outcomes be affected by variable selection? What if we use all variables or other selections of variables?

For the answers to these questions, we carry out some experiments. First, we compare the predicting outcomes between using a balanced dataset and an imbalanced dataset. As mentioned in previous sections, the balanced dataset is a pair-wise sample. In the imbalanced dataset, distressed companies and healthy companies are mixed in 1:3 ratio. We randomly choose one third of companies from the distressed group in the balanced dataset, and keep all companies from the healthy group. Second, to explore the effect of variable selection for the *i*-score model, we add two more experiments, i.e., (a) *i*-score model with all of the 13 variables, and (b) *i*-score model with 7 randomly-selected variables. To obtain a random selection of 7 variables, we simply choose variables at random from each of the variable clusters as mentioned in Sect. 4.3. Besides, since Random Forests would not as much suffer from multi-collinearity, an extra experiment that utilizes all of the 13 variables to make a prediction is provided. In order to avoid any occasional conclusions resulted from any single-sampling-based analysis, all experiments are based on random sampling. To be specific, each experiment will be run for 100 times. In each run, we randomly select two-thirds of the companies from the distressed group and the healthy group, respectively, to construct a training dataset, and use the remaining companies as a testing dataset.

Table 6 displays the mean values and the corresponding standard deviations (in the parentheses) of the evaluation index of *AOD* for different models under different settings. There are some remarkable findings. (1) The sample proportion does have an effect on the prediction performance. Regardless of models, the mean value of

Table 6 Comparative results based on repeated experiments

Models	Selection of variables	Balanced dataset	Imbalanced dataset
FDA	7 specifically selected variables	0.739 (0.052)	0.456 (0.111)
LR	7 specifically selected variables	0.778 (0.052)	0.566 (0.107)
SVM	7 specifically selected variables	0.795 (0.047)	0.670 (0.114)
CT	7 specifically selected variables	0.771 (0.054)	0.644 (0.130)
RF	7 specifically selected variables	0.773 (0.052)	0.609 (0.105)
	13 variables without selection	0.810 (0.049)	0.599 (0.113)
i-score	7 specifically selected variables	0.816 (0.045)	0.758 (0.098)
	7 randomly selected variables	0.834 (0.043)	0.781 (0.112)
	13 variables without selection	0.827 (0.041)	0.774 (0.107)

Numbers in parentheses are the standard deviations

AOD in the imbalanced dataset is lower than that in the balanced dataset. We also notice that the standard deviation in the imbalanced dataset is about double of that in the balanced dataset. Therefore, both the accuracy and the stability of predicting distressed companies are affected by the sample proportion. Fortunately, *i*-score model, regardless of variable selection, is superior to other models whether in balanced or imbalanced datasets. (2) By comparing results in the last three rows, we find out that the selection of variables has an effect on the predictive power of *i*-score model. First, when using a random selection of variables, we obtain an improvement in the accuracy. In other words, there is probably a selection of variables more powerful for prediction, which remains to be proved. Second, by adding five more variables, the mean value of *AOD* increases from 81.6 to 82.7% in the balanced dataset and from 75.8% to 77.4% in the imbalanced dataset. Therefore, it is hopeful to improve the predicting accuracy of *i*-score model by adding more variables. As for Random Forests, a slight difference can be observed. In the balanced dataset, the increase of variables brings a positive change on the mean value of *AOD*. But the change is not evident in the imbalanced dataset.

5 Conclusions

This paper has launched a new endeavor in the research topic of financial distress prediction. Unlike previous research, we attempt to achieve an improvement in predicting companies in high risk of financial distress. In light of this, we propose an *i*-score model, which packages four-quarter financial records into an interval and thereby allows volatility information involved in prediction modeling. In such an innovative way, the *i*-score model is expected to better capture the not-too-clear distress signals in a prediction lead time of 2 years.

We provide a detailed introduction to the key methodology of the *i*-score model. To demonstrate the merits of *i*-score model, some comparative studies have been carried out between *i*-score model and five commonly used models based on numerical

data. Encouragingly, the results show that *i*-score model is superior to other models in predicting financially distressed companies. We consider it extremely valuable, since investors will be beneficial to prevent great losses from investing the stocks of distressed companies. Robust analysis also verify the reliability of *i*-score model. As a consequence, it is highly recommended to adopt *i*-score model to build up a negative list for investment. More importantly, it makes a rolling prediction possible when quarterly financial ratios are adopted, which can be considered as another benefit from *i*-score model.

Acknowledgements The research of Rong Guan is supported by National Natural Science Foundation of China (Grant No. 71401192), the Fundamental Research Funds for the Central Universities (QL18009), and the Program for Innovation Research in Central University of Finance and Economics. The research of Huiwen Wang is supported by National Natural Science Foundation of China (Grant No. 71420107025). The research of Haitao Zheng is partially supported by National Natural Science Foundation of China (Grant Nos. 71371021, 71873012), and Humanities and Social Sciences Planning Fund of Ministry of Education (Grant No. 17YJA790097).

Appendix

A Data and Figures

For reader's convenience, we provide the stock ID information of the 147 selected company pairs in Table 7. Besides, descriptive figures of the samples of the interval-valued data and numerical data are respectively shown in Figs. 4 and 5. Each of the 21 subfigures corresponds to two of the seven financial ratios. In Fig. 5, distressed companies are shown as light brown crosses, whereas healthy companies are drawn as black circles. In Fig. 4, rectangles in light brown and black correspond to distressed and healthy companies, respectively.

Table 7 Stock ID of the selected companies and the year when the distressed company in the pair was labeled with *ST*

Distressed	Healthy	Year	Distressed	Healthy	Year	Distressed	Healthy	Year
000691	600366	2004	000035	000938	2005	600429	600238	2007
000040	000573	2004	600335	600580	2005	000780	000752	2007
600159	000848	2004	600681	600069	2005	600076	600677	2007
600695	600127	2004	600053	600966	2005	600657	600658	2007
600737	600186	2004	600615	600141	2005	600609	000550	2007
600139	600687	2004	000950	600731	2005	000880	000617	2007
600503	600756	2004	000719	600636	2005	600213	000868	2007
000805	600845	2004	000587	600337	2005	600516	600255	2007
400054	600718	2004	400052	600896	2005	000928	000926	2007
000736	600480	2004	600891	600830	2005	600173	600558	2007

Table 7 continued

Distressed	Healthy	Year	Distressed	Healthy	Year	Distressed	Healthy	Year
000980	600686	2004	600515	600327	2005	000408	600992	2007
400037	600208	2004	000863	600785	2005	600419	002103	2007
000766	600867	2004	600313	600540	2006	000650	600803	2007
000505	600393	2004	600844	000998	2006	000979	600746	2007
000005	000511	2004	600199	600809	2006	600645	002007	2007
600781	600233	2004	000596	600702	2006	600671	002019	2007
000529	000158	2004	000892	600797	2006	600614	600422	2007
600864	600292	2004	000862	000070	2006	600890	600696	2007
600766	600128	2004	000887	000404	2006	600136	600463	2007
600738	600824	2004	000791	600470	2006	000779	600241	2007
600234	600822	2004	600767	600325	2006	000681	002029	2007
600203	000881	2005	600369	600317	2006	000018	000045	2007
600735	600305	2005	600209	600621	2007	600003	600020	2007
600248	600975	2008	600421	600222	2008	000576	002143	2010
600084	600051	2008	000605	002038	2008	600130	600804	2010
600080	600724	2008	600568	000661	2008	600728	002148	2010
000716	000810	2008	600757	000726	2008	600372	600501	2010
600207	600602	2008	600381	600987	2008	600340	600379	2010
600198	000636	2008	000692	002039	2008	002113	002108	2010
000058	600747	2008	600817	600723	2008	000955	002015	2010
600608	000909	2008	600225	600965	2009	000720	000875	2010
600706	600588	2008	600800	600643	2009	000415	600770	2011
600988	600405	2008	600149	002050	2009	600355	002214	2011
600984	600592	2008	600854	600336	2009	000981	002095	2011
000922	600243	2008	600604	600806	2009	600860	600843	2011
600716	600586	2008	600678	600459	2009	000908	002284	2011
600217	600720	2008	000935	000055	2009	000676	600499	2011
002075	600390	2008	600187	002103	2009	600539	600318	2011
000578	000762	2008	600259	600458	2009	600769	002061	2011
600714	600971	2008	002145	002064	2009	600179	000637	2011
600462	600103	2008	000818	000755	2009	000953	000615	2011
600722	000830	2008	000633	601126	2009	600727	002109	2011
600579	600260	2008	600275	000043	2009	600538	600796	2011
600223	000792	2008	000971	600107	2009	600299	600309	2011
600656	600589	2008	600115	600029	2009	600885	600146	2011
600699	600227	2008	600868	000767	2009	600301	000510	2011
002002	000565	2008	600506	600265	2010	600077	000667	2011
600771	600789	2008	000068	600584	2010	002072	600493	2011
600466	600666	2008	000995	002124	2010	000958	600310	2011

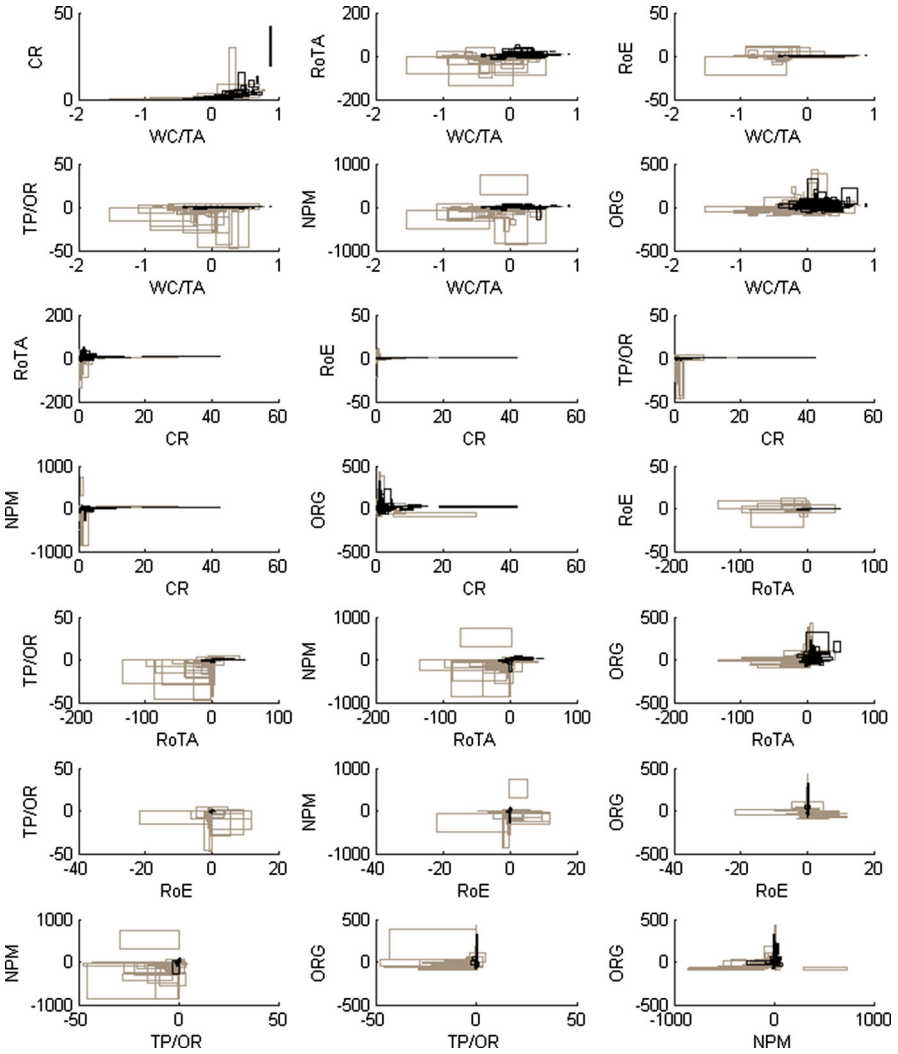

Fig. 4 Pair-wise visualizations of interval data of the selected 7 financial ratios

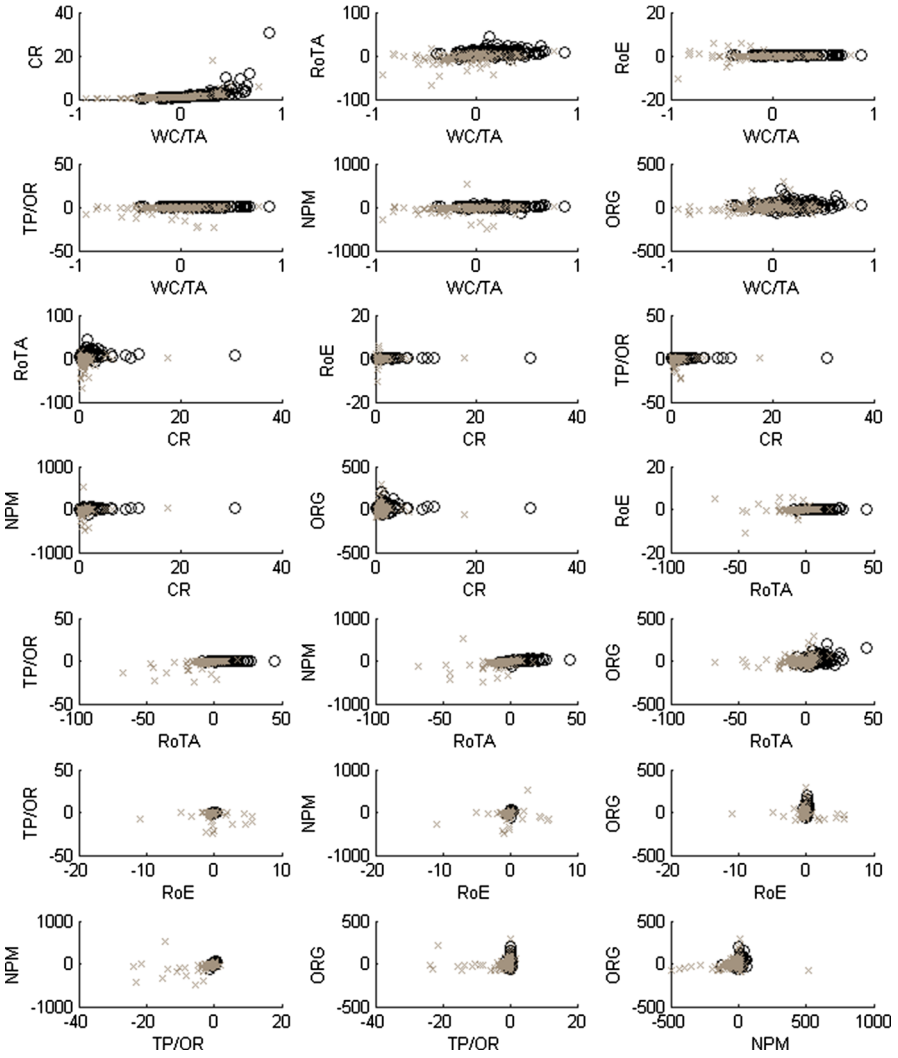

Fig. 5 Pair-wise visualizations of numerical data of the selected 7 financial ratios

References

Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23(4):589–609
Altman EI, Haldeman RG, Narayanan P (1977) Zeta analysis: a new model to identify bankruptcy risk of corporations. J Bank Finance 1(1):29–54
Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487
Bock H-H, Diday E (2000) Analysis of symbolic data: exploratory methods for extracting statistical information from complex data. Springer, Berlin
Campbell JY, Taksler GB (2003) Equity volatility and corporate bond yields. J Finance 58(6):2321–2350
Campbell JY, Hilscher J, Szilagyi J (2008) In search of distress risk. J Finance 63(6):2899–2939
Chen J, Chollete L, Ray R (2010) Financial distress and idiosyncratic volatility: an empirical investigation. J Financ Mark 13(2):249–267
Dichev ID, Tang VW (2009) Earnings volatility and earnings predictability. J Account Econ 47(1–2):160–181
Diday E, Noirhomme-Fraiture M (2008) Symbolic data analysis and the SODAS software. Wiley, Chichester
He LT, Hu C (2007) Impacts of interval measurement on studies of economic variability: evidence from stock market variability forecasting. J Risk Finance 8(5):489–507
Irpino A, Verde R (2008) Dynamic clustering of interval data using a wasserstein-based distance. Pattern Recognit Lett 29(11):1648–1658
Lauro N, Verde R, Palumbo F (2000) Factorial discriminant analysis on symbolic objects. In: Bock HH, Diday E (eds) Analysis of symbolic data, exploratory methods for extracting statistical information from complex data. Springer, Heidelberg, pp 212–333
Lee YC, Huang SY (2009) A new fuzzy concept approach for kano's model. Exp Syst Appl 36(3):4479–4484
Li Q (2013) A novel likert scale based on fuzzy sets theory. Exp Syst Appl 40(40):1609–1618
Merton RC (1974) On the pricing of corporate debt: the risk structure of interest rates. J Finance 29(2):449–470
Meyer PA, Pifer HW (1970) Prediction of bank failures. J Finance 25(4):853–868
Minton B, Schrand C, Walther B (2002) The role of volatility in forecasting. Rev Account Stud 7(2):195–215
Moore RE (1966) Interval analysis. Prentice-Hall, Upper Saddle River
Silva APD, Brito P (2006) Linear discriminant analysis for interval data. Comput Stat 21:289–308
Silva APD, Brito P (2015) Discriminant analysis of interval data: an assessment of parametric and distance-based approaches. J Classif 32(3):516–541
Sun J, Li H (2008) Data mining method for listed companies' financial distress prediction. Knowl Based Syst 21(1):1–5
Sunaga T (2009) Theory of interval algebra and its application to numerical analysis. Jpn J Ind Appl Math 26(2):125–143
Tsai C-F, Wu J-W (2008) Using neural network ensembles for bankruptcy prediction and credit scoring. Exp Syst Appl 34(4):2639–2649
Wang H, Guan R, Wu J (2012) Cipca: complete-information-based principal component analysis for interval-valued data. Neurocomputing 86:158–169
Xu X, Chen Y, Zheng H (2011) The comparison of enterprise bankruptcy forecasting method. J Appl Stat 38(2):301–308
Zavgren CV (1985) Assessing the vulnerability to failure of american industrial firms: a logistic analysis. J Bus Finance Account 12(1):19–45

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.