Knowledge-Based Systems 63 (2014) 59–67

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Financial ratio selection for business failure prediction using soft set theory

Wei Xu [a], Zhi Xiao [a,*,1], Xin Dang [b], Daoli Yang [a], Xianglei Yang [c]

[a] School of Economics and Business Administration, Chongqing University, Chongqing 400044, PR China
[b] Department of Mathematics, University of Mississippi, University 38677, USA
[c] Survey Office of the National Bureau of Statistics in Yongchuan, Chongqing 402160, PR China

ARTICLE INFO

Article history:
Received 20 September 2013
Received in revised form 13 March 2014
Accepted 15 March 2014
Available online 25 March 2014

Keywords:
Business failure prediction
Financial ratios
Logistic regression
Parameter reduction
Soft set theory

ABSTRACT

This paper presents a novel parameter reduction method guided by soft set theory (NSS) to select financial ratios for business failure prediction (BFP). The proposed method integrates statistical logistic regression into soft set decision theory, hence takes advantages of two approaches. The procedure is applied to real data sets from Chinese listed firms. From the financial analysis statement category set and the financial ratio set considered by the previous literatures, our proposed method selects nine significant financial ratios. Among them, four ratios are newly recognized as important variables for BFP. For comparison, principal component analysis, traditional soft set theory, and rough set theory are reduction methods included in the study. The predictive ability of the selected ratios by each reduction method along with the ratios commonly used in the prior literature is evaluated by three forecasting tools support vector machine, neural network, and logistic regression. The results demonstrate superior forecasting performance of the proposed method in terms of accuracy and stability.

1. Introduction

Business failure prediction (BFP) is one of the most essential problems in the field of economics and finance. It has been a subject of great interest to practitioners and researchers over decades. Being able to forecast potential failure provides an early warning system so that timely decisions are allowed to be made and appropriate adjustment in resource allocation can be taken place. There are three major tasks involved in the process of BFP, as shown in Fig. 1.

First, researchers need to determine their research objects [13]. Business failure prediction is a broad subject. Specific fields such as bank failure prediction, tourism failure prediction, small business failure prediction may take different approaches. Here we are interested in the failure prediction of Chinese listed firms from the Shenzhen Stock Exchange and Shanghai Stock Exchange.

Second, researchers need to select variables for BFP. Since the operational business environment changes quickly, BFP must be done in a timely fashion to provide early warnings. It has been shown that financial ratios have more forecasting power than other types of variables in such dynamic settings [20]. Hence we focus on financial ratio variables. Available ratios could not be used indiscriminately because some ratios could prove to be more powerful in their predictive ability than others. Predictive ability presents in two aspects. One is the forecasting accuracy (ACC). The other is the forecasting stability. If a forecasting system includes too many nonsignificant financial ratios, it will produce results low in forecasting accuracy and stability [10]. The goal of this paper is to select important financial ratios for BFP.

Third, researchers need to select forecasting models for BFP. The forecasting method, in particular the forecasting classifier used for a qualitative response, has a significant impact on the forecasting performance. Since the early empirical work on methods adopted by large USA banks, there has been a large number of literatures on the forecasting methods. Those methods include discriminant analysis [2], logistic regression (LR) [28], neural networks (NN) [3], probit method [35], rough set theory (RS) [11], support vector machine (SVM) [25], case-based reasoning [17], combination methods [7,20,34] and others. For a detailed review, one can refer to Dimitras et al. [10] and Zopounidis [36]. In this paper, we will use LR, SVM and NN to evaluate performance of variable selection methods rather than to study the selection of those forecasting models.

* Corresponding author. Tel.: +86 13808345199.
E-mail address: xiaozhicqu@163.com (Z. Xiao).

[1] Support by the National Science Foundation of China (Grant No. 71171209) and support by Major Consulting Research Project of Chinese Academy of Engineering (Grant No. 2012-ZD-12).

http://dx.doi.org/10.1016/j.knosys.2014.03.007
0950-7051/

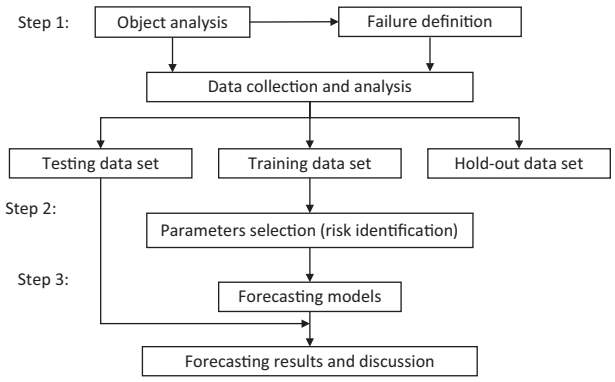

Fig. 1. The process of business failure prediction.

Differing from the development on prediction models, not much progresses have been made in variable selection for BFP. Beaver [4] selected six financial ratios including debt ratios. Altman [2] employed five financial ratios including sales to total assets. Deakin [8] made an attempt to identify variables useful in BFP. Ohlson [28] adopted nine different features. Recently, scholars [12,21,26,31,34] proposed additional financial ratios for BFP. Most popular financial ratios adopted in the prior literatures are summarized in Table 1. However, most of those financial ratios are selected either by the expert system method or by statistical approaches. Expert system relies heavily on users' knowledge and ability, which imposes difficulty to make it widely used. Statistical approaches have disadvantages for variable selection on their stringent model assumptions, which are often not met in practice. Small departures from the assumed model may make the statistical methods yielding unreliable even unacceptable results.

On the other hand, as we mentioned before, BFP must to be done in a dynamic setting. We shall include more factors or variables in the model such that information loss on the nature of firms in a dynamic operational environment is minimal. Inevitably we deal with BFP problem based on high-dimensional data. Soft set theory (SS), initiated by Molodtsov [27], has advantages to deal with high-dimension data sets. It also has been proved theoretically to be an effective tool for dimension reduction [9]. We expect SS a good performance on financial ratio selection for BFP. However, the prior literatures on SS are either purely theoretical or applied only on simple situations [24,6,16,37,14,30,23,1]. The available algorithms are rarely useful to be applied directly to the BFP problem. This motivates us to develop a novel method based on SS (NSS) to select financial ratios for BFP.

Table 1
Financial ratios adopted in prior literatures for BFP.

No.	Financial ratio	No.	Financial ratio
x_1	Tax rates	x_2	Equity value per share
x_3	No-credit interval	x_4	Operating earnings per share
x_5	Equity growth ratio	x_6	EPS
x_7	Current ratio	x_8	Cash flow/total debt
x_9	Cash flow/total asset	x_{10}	Cash flow/sales
x_{11}	Debt ratio	x_{12}	Working capital/total asset
x_{13}	Market value equity/total debt	x_{14}	Current assets/total asset
x_{15}	Quick asset/total asset	x_{16}	Sales/total asset
x_{17}	Current debt/sales	x_{18}	Quick asset/sales
x_{19}	Working capital/sales	x_{20}	Net income/total asset
x_{21}	Retained earnings/total asset		
x_{22}	Earnings before interest and taxes/total asset		
x_{23}	Continuous 4 quarterly EPS (earning per share)		
x_{24}	Log (total asset/GNP price-level index)		
x_{25}	One if total liabilities exceeds total assets, zero otherwise		
x_{26}	One if net income was negative for the past 2 years, zero otherwise		
x_{27}	$(NI_t - NI_{t-1})/(\lvert NI_t\rvert + \lvert NI_{t-1}\rvert)$, NI_t: Latest net income		

We first propose a general way to transfer the complex real-life data to 0–1 data frame so that SS or RS methods can be applied. Using LR, the importance of each variable is measured by its influence on predicting whether the firm will fail or not. A critical parameter involved in this step is determined optimally by a cross-validation procedure. Then the *uni-int* decision making on the SS is employed to obtain an optimal set of significant financial ratios. In such a way, our method utilizes the flexibility and efficiency of soft set theory and in the same time takes advantages of the statistical method without worrying about justifications of the underlying assumptions.

For comparison, principle component analysis (PCA) [15], traditional soft set (TSS) [16], rough set (RS) [29] are reduction methods included in the study of real data sets from Chinese listed firms along with evaluations of the financial ratio set proposed in previous literatures. TSS and RS use the same tabular representation data as NSS. Comparing with TSS, the *uni-int* decision making method is developed based on the redefined operations that exploits available tabular information more fully.

The remainder of this paper is organized as follows. Section 2 reviews the classical SS theory and introduces the proposed parameter reduction method. Section 3 describes the application to a real data set. In Section 4, we present the empirical results and compare performance of the proposed method with other methods. We conclude and discuss possible future work in Section 5.

2. Soft set oriented parameter reduction methods

Originated by Molodtsov [27], soft set theory deals with uncertainty in a non-parametric manner. It has been extended to effectively select parameters [16]. In this section, we first review soft set theory, the *uni-int* decision making method and the traditional reduction method proposed, then propose our novel method.

2.1. Soft set theory

Let U be a non-empty initial universe of objects, E be a set of parameters to objects in U, $\mathcal{P}(U)$ be the power set of U and $A \subseteq E$.

Definition 2.1. A soft set F_A on U is defined by the set of ordered pairs

$$F_A = \{(x, f_A(x)) : x \in E, f_A(x) \in \mathcal{P}(U)\},$$

where $f_A : E \to \mathcal{P}(U)$ such that $f_A(x) = \emptyset$ if $x \notin A$.

Here f_A is called approximate function of the soft set F_A. The soft set F_A, in other words, is a parameterized family of subsets of the set U. Every set $f_A(x)(x \in E)$ from this family may be considered as the set of x-elements of the soft set F_A. Denote the collection of all soft sets on U as $\mathcal{S}(U)$.

Çağman and Engioğlu [5] redefined operations on soft sets. They defined product operations as binary operations of soft sets depending on an approximation function of two variables to exploit information of soft sets more fully. Then, they proposed the *uni-int* decision making method as follows.

Definition 2.2. If $F_A, F_B \in \mathcal{S}(U)$, then the $\wedge$-product (and-product) of two soft sets F_A and F_B is a soft set $F_A \wedge F_B$ defined by the approximation function

$$f_{A\wedge B} : E \times E \to \mathcal{P}(U), f_{A\wedge B}(x, y) = f_A(x) \cap f_B(y).$$

Denote $\wedge(U)$ as the set of all $\wedge$-products of the soft sets over U.

Definition 2.3. If $F_A \wedge F_B \in \wedge(U)$, then *uni-int* operations denoted by $uni_x int_y$ and $uni_y int_x$ are defined, respectively, as

$$uni_x int_y : \wedge(U) \to \mathcal{P}(U),\ uni_x int_y(F_A \wedge F_B) = \cup_{x\in A}(\cap_{y\in B}(f_{A\wedge B}(x,y))),$$

$$uni_y int_x : \wedge(U) \to \mathcal{P}(U),\ uni_y int_x(F_A \wedge F_B) = \cup_{y\in B}(\cap_{x\in A}(f_{A\wedge B}(x,y))).$$

Finally, the *uni-int* decision set is defined as the union of two *uni-int* operation sets. That is,

$$uni - int(F_A \wedge F_B) = uni_x int_y(F_A \wedge F_B) \cup uni_y int_x(F_A \wedge F_B).$$

2.2. Traditional parameter reduction method of soft set (TSS)

There are many literatures in soft set that focus on the parameter reduction problem. One of the most popular algorithms, denoted as TSS, is proposed by Kong et al. [16]. The authors use the jackknife (leave one out) idea to define an importance degree of a variable by a measure of change in decision partition sets. The following is a brief review of TSS.

Suppose $U = \{h_1, h_2, \ldots, h_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. (F, E) is a soft set represented by a table $\{h_{ij}\}$. Define $f_A(h_i) = \sum_{j:e_j\in A} h_{ij}$ for $A \subseteq E$. Clearly, $f_E(h_i) = \sum_{j=1}^{m} h_{ij}$. Then according to all possible values of f_E, U is partitioned to $C_E = \{H_{f_1}, H_{f_2}, \ldots, H_{f_s}\}$, where $f_1 \geqslant f_2 \geqslant \ldots \geqslant f_s$, and any $h_i \in H_{f_j}$ if and only if $f_E(h_i) = f_j$. In other words, objects with the value of $f_E(\cdot)$ are partitioned into a same subclass.

Definition 2.4. $A \subset E$ is dispensable if $f_A(h_1) = f_A(h_2) \ldots = f_A(h_m)$, otherwise, A is indispensable.

In other words, A is dispensable if $C_A = \{U\}$, which means A having no use to partition U.

Definition 2.5. $B \subset E$ is a normal parameter reduction of E if B is indispensable and $E - B$ is dispensable. $G \subset E$ is called a pseudo reduction of E if $C_G = C_E$.

One can define an importance degree of e_i by a measure of change between C_E and C_{E-e_i}. C_{E-e_i} is the decision partition after deleting e_i. For presentation brevity, we write $f_{E-e_i}(h_j)$ as $g_i(h_j), C_E = \{H_{f_1}, \ldots, H_{f_s}\}$ as $C = \{H_1, H_2, \ldots, H_s\}$ and C_{E-e_i} as $C_{-i} = \{H_{g_{i1}}, H_{g_{i2}}, \ldots, H_{g_{it}}\}$. Then the importance degree of e_i is

$$r_i = \frac{1}{|U|}\sum_{k=1}^{s}\alpha_{k,i},$$

where $|\cdot|$ denotes the cardinality of a set and

$$\alpha_{k,i} = \begin{cases} |H_k - H_{g_{il}}| & \text{if } g_{il} = f_k \quad \text{for } 1 \leqslant l \leqslant t,\ 1 \leqslant k \leqslant s; \\ |H_k| & \text{otherwise.} \end{cases}$$

The algorithm of TSS is described as follows.

1. Input the soft set (F, E) and the parameter set E.
2. Compute parameter importance degree r_{e_i}, for $1 \leqslant i \leqslant m$.
3. Search for the maximal subset $A = \{e_{i_1}, e_{i_2}, \ldots e_{i_p}\}$ in E whose sum of r_{i_j} is a nonnegative integer.
4. Check A whether it is dispensable. If it is, $E - A$ is the normal parameter reduction and A is saved to the feasible parameter reduction set.
5. Find the set A with the maximum cardinality in the feasible reduction set.
6. Compute $E - A$ as the optimal normal parameter reduction.

The above algorithm has a clear motivation and works well for small data sets with categorical parameters. However, for BFP with a large number of numerical financial ratio variables, a direct application of TSS is infeasible. Besides, there is no construction way in TSS on how to get a tabular representation of SS. In the next, we propose a novel parameter reduction method based on SS for BFP.

2.3. A novel parameter reduction method of soft set (NSS)

For BFP, it seems natural that U is the collection of firms in the study and E is the set of financial ratio variables. However, such a setup imposes tremendous difficulties for our task. First, most of variables are numerical. Although a discretization technique can be used to convert those continuous variables to be nominal, information might be lost largely. Second, how to use the label information (failure or not) in the real-world data set is unclear. Third, with a large number of variables, traditional methods are very computationally intense.

To overcome those difficulties, we let $U = \{x_1, \ldots, x_m\}$ be the set of financial ratio variables and $E = \{h_1, \ldots, h_n\}$ be the set of companies under consideration. In such a way, the parameter reduction problem is transferred to be the problem of selecting an optimal decision, which can be scaled up for a large m. Also the status of each firm is categorical, either failure or normal. That can be easily incorporated into a decision procedure. Now we focus on using numerical information of financial ratio variables to construct the tabular representation of SS.

To assess the relationship between financial ratios and status of firms, we apply the logistic regression model (LR) [28] because of its simplicity and interpretability. Use binary variable Y as the status of firms, that is,

$$Y_j = \begin{cases} 1 & \text{if } j^{th} \text{ firm is in normal status;} \\ 0 & \text{otherwise} \end{cases} \qquad j = 1, \ldots, n.$$

LR models the logit of odds as a linear function of financial ratios as follows.

$$\log\frac{P(Y=1)}{P(Y=0)} = \boldsymbol{\beta}^T\boldsymbol{x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$$

or equivalently,

$$P(Y=1) = \frac{1}{1+\exp(-\boldsymbol{\beta}^T\boldsymbol{x})}.$$

Coefficient β_i is the amount of log odd change when x_i increases 1 unit, hence it measures relative effect of x_i on Y. We estimate β_i by the maximum likelihood estimator (MLE) $\hat{\beta}_i$ based on the training data set $(Y_j,\ \boldsymbol{x}_j)$ for $j = 1, \ldots, n$ where $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$. $\hat{\beta}_i$ takes information of the training data set and measures relative effect of the ith variable on status Y. Hence for $Y_j, exp(\hat{\beta}_i x_{ij})$ represents the contribution of the ith variable of the jth firm to its odd $P(Y_j = 1)/P(Y_j = 0)$. If $exp(\hat{\beta}_i x_{ij})$ is large (greater than c), the ith variable is helpful for predicting the j^th firm to be in normal status, otherwise, it is useful for the failure prediction. Here, c is the critical value that is used to differentiate the role of x_{ij} in predicting the status of j^th firm. In other words, if $exp(\hat{\beta}_i x_{ij}) \geqslant c$, x_{ij} is useful to predict the jth firm in normal status. If $exp(\hat{\beta}_i x_{ij}) < c, x_{ij}$ is useful to predict the jth firm in failure status. Then, we use the ground truth of Y_j to verify the prediction. If the prediction from x_{ij} is the jth firm in normal status and Y_j is 1, we are sure that x_{ij} is really useful for BFP. If the prediction from x_{ij} is the jth firm in failure status and also Y_j is 0, we confirm that x_{ij} is really useful for BFP. Otherwise, if the prediction does not match the true status, variable x_{ij} is not helpful for BFP. This idea provides us a way to construct 0–1 tabular representation for SS. The algorithm of NSS is illustrated as follows.

1. Fit LR on the training data set to obtain MLE $\hat{\beta}_i$.
2. Input $U = \{x_1, \ldots, x_m\}$ to be the set of financial ratios and $E = \{h_1, \ldots, h_n\}$ to be the set of firms.
3. Construct the tabular representation of soft sets. Its entry v_{ij} for $i = 1, \ldots, m; j = 1, \ldots, n$ is

$$v_{ij} = \begin{cases} 1 & \text{if } \exp(\hat{\beta}_i x_{ij}) \geqslant c \text{ and } Y_j = 1; \\ 1 & \text{if } \exp(\hat{\beta}_i x_{ij}) < c \text{ and } Y_j = 0; \\ 0 & \text{otherwise}, \end{cases}$$

where c is a critical value that is determined optimally in a procedure illustrated later. Then our approximation function of the soft set is $f_A(h_j) = \{x_i : h_j \in A, v_{ij} = 1\}$.

4. Let $NST = \{h_j : Y_j = 1\}$ to be the set of "Not Specially Treated" firms and $ST = \{h_j : Y_j = 0\}$ to be the set of "Specially Treated" firms. Without loss of generality, $NST = \{h_1, h_2, \ldots, h_{n_1}\}$ and $ST = \{h_{n_1+1}, \ldots, h_n\}$.
5. Obtain soft sets $F_{NST} = \{(h_j, f_{NST}(h_j)), j = 1, \ldots, n_1\}$ and $F_{ST} = \{(h_j, f_{ST}(h_j)), j = n_1 + 1, \ldots, n\}$
6. Find $F_{NST} \wedge F_{ST}$, the and-product of F_{NST} and F_{ST}.
7. Apply the *uni-int* decision making rule on $F_{NST} \wedge F_{ST}$.
8. Obtain the optimal decision set of U, which is the optimal financial ratios set for BFP.

The performance of NSS is largely determined by the critical value c. Here we use grid search through cross-validation on the training data set to obtain c such that the mean value of validation accuracy is the highest. The procedure can be described as follows and shown in Fig. 2.

Step 1. We do a grid search for c in the range where $\ln c$ changes from -3 to 3 with a step 0.01. For each c, we do the following steps 2–4.
Step 2. The training data set is randomly divided into two groups evenly. One is used to train $\hat{\beta}_i$. The other one is used to evaluate the forecasting accuracy (ACC, the definition is described in Section 4.2.). $\hat{\beta}_i$ is obtained through LR on the first group data set. Along with the value of c, we obtain SS table and select financial ratio sets using NSS.
Step 3. SVM is employed as the verification tool. With the selected financial ratios from Step 2, we implement SVM on the second group data set of the year $(t-1)$ for ACC. The Gaussian radical basis function (RBF) is set as the kernel [22]. Grid-search and cross-validation are used to search for optimal parameters values of RBF on the first group data set. ACC is computed.
Step 4. Repeat 5 times of the steps 2 and 3. The mean value of ACC is obtained.
Step 5. The optimal c is the value that attains the highest mean value of forecasting accuracy.

With the optimal c, NSS selects the optimal financial ratio variables that are helpful to BFP. It integrates LR into SS and hence takes advantages of two methods. We expect a good performance of the proposed NSS.

3. Empirical experiment

3.1. Sample and data

According to the benchmark of China Securities Supervision and Management Committee (CSSMC), listed firms are categorized into two classes: Specially Treated firms (ST) and Not Specially Treated firms (NST). The criteria are either negative net profits in recent consecutive two years or announcement on purpose about serious financial misstatements. Here, we consider ST companies as firms that have had negative net profits in recent two years.

We use real financial data sets from Chinese listed firms. The data were collected from the Shenzhen Stock Exchange and Shanghai Stock Exchange in China. We randomly selected a sample of 120 NST and 120 ST firms in the period between the year 2000–2012. Financial ratios in the financial analysis statement category set (FASCS) (see Table 12 at Appendix A) and in the prior literatures financial ratio sets (PLFRS) (see Table 1] were calculated for each firm in the sample.

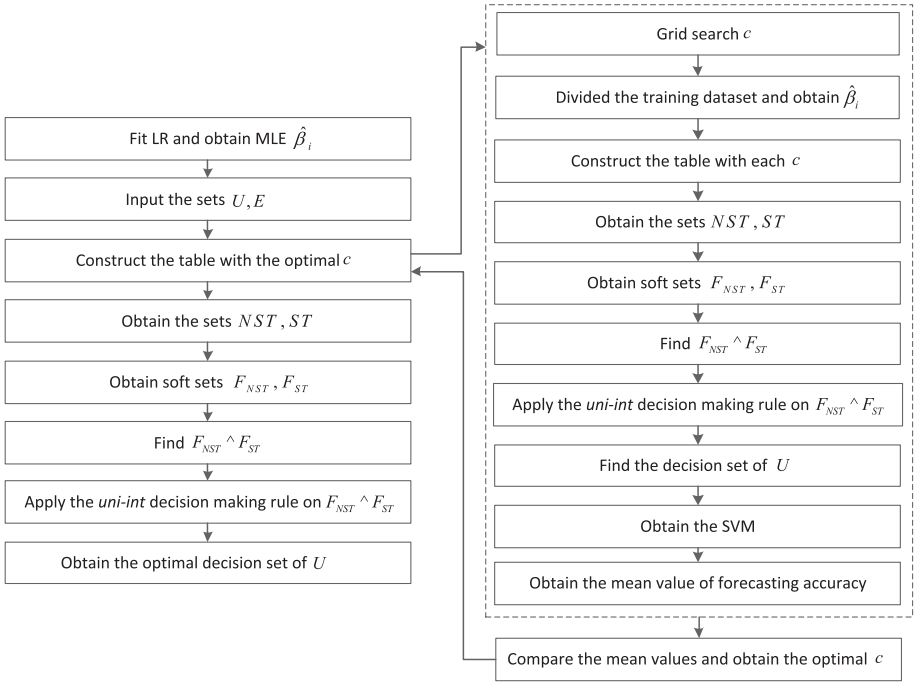

Fig. 2. NSS Algorithm and the procedure for optimal c.

Table 2
Financial ratios of FRPBA.

No.	Financial ratio	No.	Financial ratio
F_1	Working capital/total asset	F_2	Market value equity/total debt
F_3	Sales/total asset	F_4	Retained earnings/total asset
F_5	Current ratio	F_6	Cash flow/total debt
F_7	Debt ratio	F_8	No-credit interval
F_9	Earnings before interest and taxes/total asset		

3.2. Experiment design

As pointed out in the literature [19], forecasting of the year t using the data set of the year $(t-2)$ or $(t-3)$ is more difficult than that using the data set of the year $(t-1)$. In this paper, we tackle this challenge.

To investigate whether the new financial ratio set selected from NSS has a good performance for BFP, three parameter selection methods (PCA, RS, TSS) are employed and three forecasting methods (LR, SVM, NN) are used to evaluate the prediction performance. Also, we include the nine most popular financial ratios used by Beaver [4] and Altman [2] (FRPBA). They are listed in Table 2.

The framework of this experiment is shown in Fig. 3. The details can be illustrated as follows.

Step 1. Collect data and classify NST and ST firms. We randomly split those data into three groups for this experiment. One is the hold-out data set that will be rejected. One is the training data set and the last one is the testing data set.
Step 2. Obtain financial ratios in the list of PLFRS and FASCS. We employ PCA, RS, TSS, and NSS to select parameters on the training data set.
Step 3. Using the benchmark set FRPBA and the financial ratios sets from PCA, RS, TSS, we separately obtain forecasting models using LR, SVM and NN, respectively.
Step 4. The testing data set is used for evaluating each forecasting model. Obtain ACC.
Step 5. Compare the prediction performance and conclude.

4. Experiment results and discussion

Matlab software package (2012) and Statistical Product and Service Solutions (IBM SPSS 20) are employed to obtain the result of financial ratios selection and forecasting performance of different forecasting models.

4.1. Financial ratios selection results

With a grid search procedure described in Section 2.3, we obtain the optimal value $c = 1.27$ on the training data set. It yields the highest mean value 0.967 of the validation accuracy. With this value of c, 0–1 data frame is constructed and used for NSS, RS and TSS. In the next, we present the selected ratios by each parameter reduction method. For comparison purpose, we keep nine financial ratios in order to make the selected model with the same complexity as the others so that their prediction ability is comparable.

4.1.1. Results with PCA

PCA selects five financial ratios as the parameters for BFP because their sum variation contribution to the total variation is 87.62%, which is greater than 85%, a common critical value. Listed in Table 3 are those five variables along with the following four variables whose accumulative variation accounts for 87.93% of the total variation.

4.1.2. Results with RS

The RS algorithm of computing all reduction [33] is employed for RS reduction method. We take the net profit parameter as the decision attribute to do the parameters reduction with RS. With the transformed data from NSS, eight parameters are selected by RS. Again, to keep model complexity the same, nine parameters are listed in Table 4.

4.1.3. Results with TSS

With the procedure described in Subsection 2.2 on the transformed data from NSS, TSS extracts nine financial ratios listed in Table 5.

4.1.4. Results with NSS

Nine financial ratios listed in Table 6 are selected by NSS. Among them, net profit over sales, management fee over total cost, the total asset turnover and comprehensive leverage factor are the first time to be considered useful for BFP.

4.2. Forecasting results

We choose the Gaussian radical basis function (RBF) as the SVMs kernel [22]. Grid-search and cross-validation are used to search for optimal parameters values of RBF based on the training

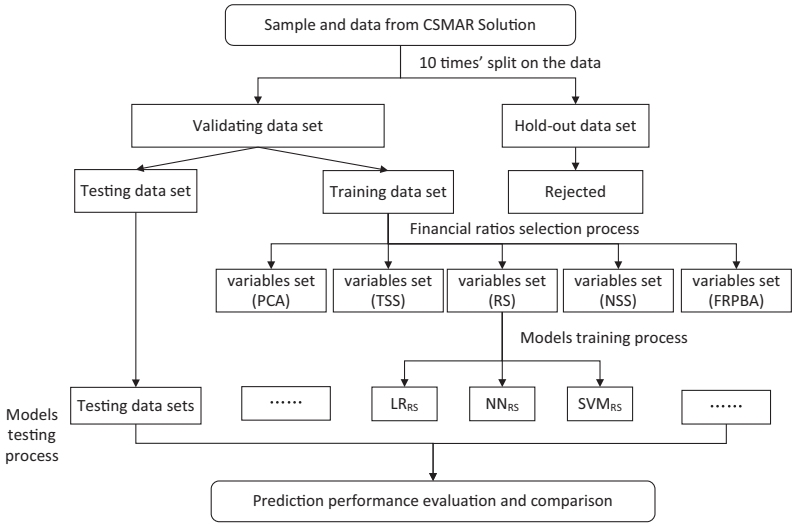

Fig. 3. The framework of the experiment.

Table 3
Financial ratios selected with PCA.

No.	Financial ratio	No.	Financial ratio
P_1	Working capital/total asset	P_2	Net income/total asset
P_3	Income tax	P_4	Cash ratio
P_5	Debt ratio	P_6	Working capital turnover
P_7	Re-investment cash/total cash	P_8	Quick assents/sales
P_9	Continuous 4 quarterly EPS (earning per share)		

Table 4
Financial ratios selected with RS.

No.	Financial ratio	No.	Financial ratio
R_1	Net income/total asset	R_2	Cash ratio
R_3	Working capital/total asset	R_4	Retained earnings/total asset
R_5	Equity value per share	R_6	Quick assents/sales
R_7	Debt ratio	R_8	Continuous 4 quarterly EPS
R_9	Comprehensive leverage factor		

Table 5
Financial ratios selected with TSS.

No.	Financial ratio	No.	Financial ratio
T_1	Working capital/total asset	T_2	Working capital turnover
T_3	Net income/total asset	T_4	Equity value per share
T_5	Debt ratio	T_6	Capital-intensity
T_7	Cash ratio	T_8	Tangible net debt ratio
T_9	The growth rate of net cash flow generated by operating activities		

Table 6
Financial ratios selected with NSS.

No.	Financial ratio	No.	Financial ratio
N_1	Working capital/total asset	N_2	Debt ratio
N_3	Net profit/sales	N_4	Net income/total asset
N_5	Management fee/total cost	N_6	Total asset turnover
N_7	Reinvestment cash/total cash	N_8	Comprehensive leverage factor
N_9	Continuous 4quarterly EPS		

data set. Back propagation neural network (BPNN) is employed as the NN algorithm [34]. We run twenty experiments on the training data set for NN prediction and select the optimal set of experiment results as the output of NN prediction. The testing data set is used to compare the forecasting accuracy of different models. For LR, we forecast a firm as NST if its estimated probability of $Y = 1$ is greater than 0.5.

For BFP, it is a typical binary classification problem, in which the prediction are labeled as either "positive" (P) or "negative" (N). There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual state is also positive, it is called a true positive (TP); if the actual status is negative then it is said to have a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and false negative (FN) occurs when the prediction outcome is negative while the actual value is positive. We use the percentage of correct identification to measure the forecasting accuracy (ACC) of a forecasting model [32], which is

$$\mathrm{ACC} = \frac{(\mathrm{TP} + \mathrm{TN})}{\mathrm{P} + \mathrm{N}}.$$

The 10-fold cross-validation approach is used to perform the experiment. The forecasting comparison of parameter reduction methods evaluated by LR, SVM and NN for data sets of the year $(t-2)$ are listed in Table 7 and for data sets of the year $(t-3)$ are listed in Table 8.

4.3. Comparison and discussion

For each forecasting method, six statistical summary indices of ACC from the 10-fold cross-validation procedure for data sets of the year $(t-2)$ and for the data sets of the year $(t-3)$ are listed in Tables 9–11, respectively. The mean and median are crucial on assessing forecasting accuracy [18]. Those two indices are also demonstrated in Fig. 4. Variance value and coefficient of variance are critical on assessing forecasting stability [34]. Both are illustrated in Fig. 5.

4.3.1. Discussion on forecasting accuracy

From Tables 9–11 and Fig. 4, one can easily find out that NSS is an effective tool to select important variables for improving the forecasting performance for BFP. The model with the financial ratios selected by NSS uniformly performs the best no matter what evaluation method is used or which year of data set is used for forecasting. It has the highest mean, median and minimum forecasting accuracy comparing to those with financial ratios selected by PCA, RS, TSS and FRPBA. LR, SVM and NN have a similar forecasting accuracy with financial ratios selected by RS, TSS and FRPBA. With financial ratios selected by PCA, LR, SVM and NN have the lowest forecasting accuracy.

Without a surprise, LR, SVM and NN consistently have a higher forecasting accuracy with data sets of the year $(t-2)$ than those with data sets of the year $(t-3)$ for all selected variable models. Prediction on a long term time horizon is more difficult than a short term prediction.

Table 7
Results of 10-fold cross-validation and summaries of ACC on average (ave), variance (var) and variance coefficient (cv) using data sets of the year $(t-2)$ of Chinese listed firms.

	PCA			RS			TSS			NSS			FRPBA		
	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN
1	0.750	0.875	0.813	0.813	0.938	0.875	0.875	1.000	0.938	1.000	1.000	0.938	0.875	0.875	1.000
2	0.938	0.875	0.750	0.938	0.813	0.813	0.813	0.813	0.813	0.875	0.813	0.938	0.563	0.938	0.875
3	0.625	0.688	0.750	0.688	0.625	0.750	0.625	0.688	0.750	0.688	0.750	0.813	0.875	0.750	0.875
4	0.813	0.813	0.938	0.875	0.875	0.938	0.813	0.875	0.938	0.813	1.000	0.938	0.625	1.000	0.813
5	0.688	0.563	0.563	0.625	0.625	0.625	0.625	0.688	0.688	0.688	0.688	0.688	0.813	0.625	0.625
6	0.563	0.813	0.875	0.625	0.875	0.938	0.688	0.813	0.938	0.750	0.875	1.000	0.625	0.688	0.750
7	0.688	1.000	0.688	0.750	1.000	0.750	0.813	0.938	0.688	0.875	0.938	0.750	0.813	0.625	0.688
8	0.813	0.875	0.625	0.875	0.875	0.625	0.875	0.813	0.688	0.875	0.875	0.750	0.938	0.875	0.813
9	1.000	0.750	0.938	0.938	0.813	0.875	1.000	0.750	0.813	1.000	0.813	0.875	1.000	0.938	0.938
10	0.625	0.938	0.625	0.625	0.938	0.813	0.688	0.938	0.688	0.750	0.938	0.750	0.750	0.938	0.688
Ave	0.750	0.819	0.756	0.775	0.838	0.800	0.781	0.831	0.794	0.831	0.869	0.844	0.788	0.825	0.806
Var	0.020	0.016	0.018	0.017	0.016	0.013	0.015	0.011	0.012	0.013	0.011	0.012	0.021	0.020	0.014
Cv	0.027	0.020	0.023	0.022	0.019	0.016	0.019	0.014	0.015	0.016	0.012	0.014	0.027	0.024	0.018

Table 8
Results of 10-fold cross-validation and summaries of ACC on average (ave), variance (var) and variance coefficient (cv) using data sets of the year $(t-3)$ of Chinese listed firms.

	PCA			RS			TSS			NSS			FRPBA		
	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN	LR	SVM	NN
1	0.938	0.875	0.875	0.688	0.938	0.875	0.750	1.000	0.813	0.938	0.813	0.875	0.750	0.938	0.875
2	0.813	0.750	0.625	1.000	0.688	0.625	0.938	0.813	0.875	0.813	0.875	1.000	0.875	0.625	0.938
3	0.500	0.563	0.813	0.938	0.500	0.625	0.688	0.625	0.625	0.563	0.625	0.813	0.688	0.625	0.625
4	0.875	0.688	0.875	0.750	0.750	1.000	0.875	0.938	0.813	0.813	1.000	0.750	0.938	0.938	0.625
5	0.563	0.500	0.438	0.500	0.625	0.500	0.625	0.563	0.563	0.688	0.563	0.563	0.625	0.813	0.563
6	0.438	0.688	0.875	0.563	0.875	0.813	0.563	0.563	0.813	0.625	0.750	0.938	0.563	0.563	0.875
7	0.875	0.938	0.563	0.750	1.000	0.938	0.688	0.813	0.625	1.000	0.938	0.875	0.625	0.875	0.625
8	0.750	0.875	0.500	0.813	0.938	0.563	0.938	0.875	1.000	0.875	0.875	0.563	0.938	0.875	0.938
9	0.875	0.500	0.938	0.938	0.688	0.875	1.000	0.625	0.875	0.938	0.750	0.688	1.000	0.625	0.875
10	0.500	1.000	0.563	0.625	0.813	0.688	0.563	0.938	0.563	0.563	0.938	0.875	0.625	0.938	0.563
Ave	0.713	0.738	0.706	0.756	0.781	0.750	0.763	0.775	0.756	0.781	0.813	0.794	0.763	0.781	0.750
Var	0.037	0.033	0.035	0.028	0.025	0.030	0.027	0.028	0.023	0.026	0.020	0.023	0.026	0.024	0.026
Cv	0.051	0.044	0.049	0.037	0.033	0.039	0.035	0.036	0.030	0.034	0.025	0.028	0.034	0.030	0.035

Table 9
Statistics of forecasting accuracy using LR.

Year	Method	Mean	Median	Minimum	Maximum	Variance	Coef. var.
$(t-2)$	PCA	0.7500	0.7188	0.5625	1.0000	0.0200	0.0266
	RS	0.7750	0.7813	0.6250	0.9375	0.0167	0.0215
	TSS	0.7813	0.8125	0.6250	1.0000	0.0150	0.0192
	NSS	0.8313	0.8438	0.6875	1.0000	0.0131	0.0157
	FRPBA	0.7875	0.8125	0.5625	1.0000	0.0210	0.0267
$(t-3)$	PCA	0.7125	0.7813	0.4375	0.9375	0.0366	0.0514
	RS	0.7563	0.7500	0.5000	1.0000	0.0282	0.0372
	TSS	0.7625	0.7188	0.5625	1.0000	0.0267	0.0351
	NSS	0.7813	0.8125	0.5625	1.0000	0.0263	0.0336
	FRPBA	0.7625	0.7188	0.5625	1.0000	0.0259	0.0339

Table 10
Statistics of forecasting accuracy using SVM.

Year	Method	Mean	Median	Minimum	Maximum	Variance	Coef. var.
$(t-2)$	PCA	0.8188	0.8438	0.5625	1.0000	0.0160	0.0196
	RS	0.8375	0.8750	0.6250	1.0000	0.0158	0.0189
	TSS	0.8313	0.8125	0.6875	1.0000	0.0113	0.0136
	NSS	0.8688	0.8750	0.6875	1.0000	0.0108	0.0124
	FRPBA	0.8250	0.8750	0.6250	1.0000	0.0198	0.0240
$(t-3)$	PCA	0.7375	0.7188	0.5000	1.0000	0.0328	0.0445
	RS	0.7813	0.7813	0.5000	1.0000	0.0254	0.0325
	TSS	0.7750	0.8125	0.5625	1.0000	0.0280	0.0361
	NSS	0.8125	0.8438	0.5625	1.0000	0.0200	0.0246
	FRPBA	0.7813	0.8438	0.5625	0.9375	0.0237	0.0303

SVM consistently has a higher forecasting accuracy than LR and NN. The result agrees with the conclusion from prior literatures [19,25]. SVM has many advantages on two-class classification problems [25] and the optimal choice on kernel does improve the performance. LR and NN yield a similar forecasting accuracy.

Table 11
Statistics of forecasting accuracy using NN.

Year	Method	Mean	Median	Minimum	Maximum	Variance	Coef. var.
$(t-2)$	PCA	0.7563	0.7500	0.5625	0.9375	0.0178	0.0235
	RS	0.8000	0.8125	0.6250	0.9375	0.0128	0.0161
	TSS	0.7938	0.7813	0.6875	0.9375	0.0122	0.0154
	NSS	0.8438	0.8438	0.6875	1.0000	0.0115	0.0136
	FRPBA	0.8063	0.8125	0.6250	1.0000	0.0143	0.0177
$(t-3)$	PCA	0.7063	0.7188	0.4375	0.9375	0.0348	0.0492
	RS	0.7500	0.7500	0.5000	1.0000	0.0295	0.0394
	TSS	0.7563	0.8125	0.5625	1.0000	0.0230	0.0304
	NSS	0.7938	0.8438	0.5625	1.0000	0.0226	0.0285
	FRPBA	0.7500	0.7500	0.5625	0.9375	0.0260	0.0347

4.3.2. Discussion on forecasting stability

Results on forecasting stability show a very similar pattern as that of forecasting accuracy. From Tables 9–11 and Fig. 5, we can conclude that the model with the ratios selected by NSS has the highest forecasting stability. No matter which data sets of the year $(t-2)$ or $(t-3)$ are used and which method is used, the financial ratios selected by NSS have the lowest variance and coefficient of variance of forecasting accuracy comparing to those with the ratios selected by PCA, RS, TSS and FRPBA. LR, SVM and NN have a similar forecasting stability with financial ratios selected by RS, TSS. With financial ratios selected by PCA, LR, SVM and NN have the worst forecasting stability, especially for $(t-3)$ case.

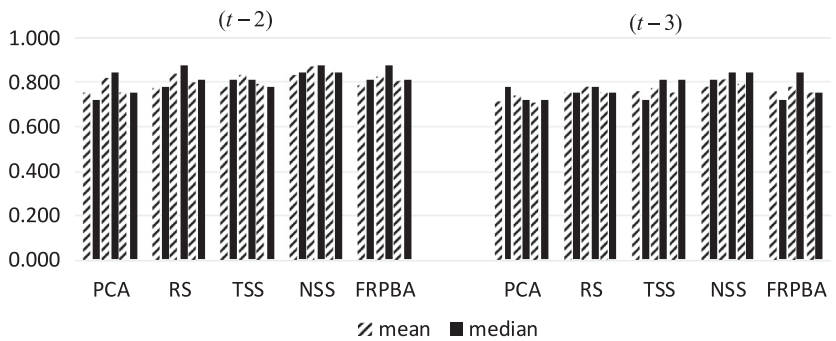

Fig. 4. Mean and median of forecasting accuracy.

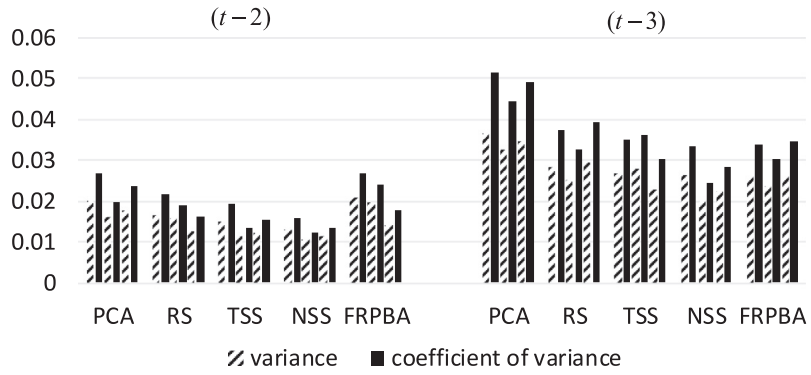

Fig. 5. Variance and coefficient of variance of forecasting accuracy.

Again, model forecasting stability using data sets of the year $(t-2)$ is uniformly better than that with data sets of the year $(t-3)$.

Among three evaluation methods, SVM has the best forecasting stability for variable sets selected by PCA, TSS and NSS. NN shows a good stability for the models with variables from FRPBA or RS in $(t-2)$ forecasting. LR produces the lowest stability.

4.4. Summary

For BFP, experimental results above show that the model with the financial ratios selected by NSS can effectively improve the forecasting performance of BFP. It has a higher average and minimum prediction accuracy, more often to obtain the perfect prediction accuracy, a lower variance and coefficient of variation than those from other financial ratios sets. In other words, NSS proposed in this paper improves the forecasting performance of BFP in terms of accuracy and stability.

5. Conclusion

In this paper, we extended research of financial ratios for BFP by proposing a novel parameters reduction method based on soft set theory. It constructs a tabular representation of SS from LR, then uses optimal decision on SS to select significant financial ratio variables. It inherits advantages and in the same time avoids disadvantages of both methods. From PLFRS and FASCS, NSS chooses nine financial ratios for BFP. Among them, four financial ratios are the first time to be selected as parameters for BFP. Compared with PCA, RS, TSS and FRPBA, our method has demonstrated its superior forecasting performance for BFP of Chinese listed firms.

Though results are satisfactory, there are some limitations in this paper. As a key component for success of NSS, the method of determining v_{ij} is intuitive; it should be studied in a more systematical way. In other words, systematical construction method for tabular representation of SS is one of our future research directions. Besides, although we have used LR, SVM and NN, more research can be done about forecasting methods to validate efficiency of the financial ratios set selected by NSS.

Also, the NSS obtains a good forecasting performance for BFP of Chinese listed firms. We do not know its forecasting performance on financial data sets from other countries. It definitely deserves a further exploration.

Acknowledgements

The authors are highly grateful to referees and editors for their valuable comments and suggestions that helped in improving this paper.

Appendix A

Table 12

Table 12
Financial ratios from the CSMAR Solution financial ratio set

No.	Financial ratio	No.	Financial ratio
CS_1	Operating profit ratio	CS_2	Net profit/sales
CS_3	Operating cost ratio	CS_4	Dividend cover
CS_5	P/ CF	CS_6	Book to market ratio
CS_7	Income tax	CS_8	OCFPS
CS_9	EEOI	CS_{10}	Cash ratio
CS_{11}	Financial cost/total cost	CS_{12}	Operating cost/total cost
CS_{13}	Management fee/total cost	CS_{14}	Price/sale
CS_{15}	Income before tax	CS_{16}	Operating tax ratio
CS_{17}	Profit per shares before tax	CS_{18}	Surplus reserves per share
CS_{19}	The profitability of ordinary share	CS_{20}	Cash/debts
CS_{21}	Cash dividend cover	CS_{22}	Surplus cash/cover
CS_{23}	Capital expenditure/ deprecation amortization	CS_{24}	Long term liabilities/operating cash
CS_{25}	Accounts receivable/total income	CS_{26}	Reinvestment cash/total cash
CS_{27}	Capital intensity	CS_{28}	Operating cycle
CS_{29}	Long term borrowing/total asset	CS_{30}	Working capital/total net asset
CS_{31}	Total asset turnover	CS_{32}	A/R turnover per days
CS_{33}	Inventory turnover	CS_{34}	A/R & N/R turnover
CS_{35}	Fixed asset turnover	CS_{36}	Equity turnover
CS_{37}	Inventory turn per days	CS_{38}	A/P turnover per days
CS_{39}	Account payable turnover	CS_{40}	Working capital turnover
CS_{41}	Working capital/loan capital	CS_{42}	Interest coverage
CS_{43}	Fixed changes reimbursement ratio	CS_{44}	Sustainable growth ratio
CS_{45}	Profit before interest, tax, depreciation/ debts	CS_{46}	Long term liabilities/total liabilities
CS_{47}	Financial leverage factor	CS_{48}	Operating leverage factor
CS_{49}	Comprehensive leverage factor	CS_{50}	Accrued items
CS_{51}	Equity value per share	CS_{52}	EPS
CS_{53}	Continuous 4 quarterly EPS	CS_{54}	Yield of cash
CS_{55}	Yield of dividend	CS_{56}	Tangible net debt ratio
CS_{57}	Account receivable turnover	CS_{58}	Growth ratio of total asset
CS_{59}	Growth ratio of major sales operation	CS_{60}	Capital maintenance and appreciation
CS_{61}	Growth ratio of net profit	CS_{62}	Cash flow/current liability
CS_{63}	Cash/main business income	CS_{64}	Net operating cash flow per share
CS_{65}	Net cash flow of investing activities per share	CS_{66}	ROA (C) before tax interest and depreciation

Table 12 (*continued*)

No.	Financial ratio	No.	Financial ratio
CS_{67}	Net asset per share	CS_{68}	ROC (A) after tax, before interest
CS_{69}	ROC (B) after tax, before interest and depreciation	CS_{70}	Net growth ratio of cash folw per share generated by business
CS_{71}	Equity multiplier	CS_{72}	Equity/debt
CS_{73}	Owner's equity ratio	CS_{74}	Right ratio of long term asset
CS_{75}	Bad debt/ income	CS_{76}	Retention ratio
CS_{77}	TCRI	CS_{78}	PBR
CS_{79}	The growth ratio of net cash flow generated by operating activities		
CS_{80}	The growth ratio of net cash flow generated by investing activities		
CS_{81}	The growth ratio of net cash flow generated by financing activities		

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.knosys.2014.03.007.

References

[1] M.I. Ali, Another view on reduction of parameters in soft sets, Appl. Soft Comput. 12 (2012) 1814–1821.
[2] E.L. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Finance 23 (1968) 589–609.
[3] E.L. Altman, G. Marco, F. Varetto, Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience), J. Bank. Finance 28 (1994) 505–529.
[4] W.H. Beaver, Financial ratios and predictions of failure, J. Account. Res. 4 (1966) 71–111.
[5] N. Çağman, S. Engioğlu, Soft set theory and uni-int decision making, Euro. J. Oper. Res. 207 (2010) 848–855.
[6] D. Chen, E.C.C. Tsang, D.S. Yeung, X. Wang, The parameterization reduction of soft sets and its applications, Comput. Math. Appl. 49 (2005) 757–763.
[7] M.Y. Chen, A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering, Inform. Sci. 220 (2013) 180–195.
[8] E.B. Deakin, A discriminant analysis of predictors of business failure, J. Account. Res. 10 (1972) 167–179.
[9] T.Q. Deng, X.F. Wang, Parameter significance and reductions of soft sets, Int. J. Comput. Math. 89 (2012) 1979–1995.
[10] A.I. Dimitras, S.H. Zanakis, C. Zopounidis, A survey of business failure with an emphasis on prediction method and industrial application, Euro. J. Oper. Res. 90 (1996) 487–513.
[11] A.I. Dimitras, R. Slowinski, R. Susmaga, Business failure prediction using rough sets, Euro. J. Oper. Res. 114 (1999) 263–280.
[12] Y.S. Ding, X.P. Song, Y.M. Zen, Forecasting financial condition of Chinese listed companies based on support vector machine, Expert Syst. Appl. 34 (2008) 3081–3089.
[13] M.C. Gupta, R.J. Huefner, A cluster analysis study of financial ratios and industry characteristics, J. Account. Res. 10 (1972) 77–84.
[14] Q.H. Hu, S.A. An, D.R. Yu, Soft fuzzy rough sets for robust feature evaluation and selection, Inform. Sci. 180 (2010) 4384–4400.
[15] I.T. Jolliffe, Principle Component Analysis, Springer, New York, 2002.
[16] Z. Kong, L. Gao, L. Wang, S. Li, The normal parameter reduction of soft sets and its algorithm, Comput. Math. Appl. 56 (2008) 3029–3037.
[17] H. Li, J. Sun, Ranking order case-based reasoning for financial distress prediction, Knowel.-Based Syst. 21 (2008) 868–878.
[18] H. Li, J. Sun, Forecasting business failure in china using case-based reasoning with hybrid case representation, J. Forecast. 29 (2010) 486–501.
[19] H. Li, J. Sun, forecasting business failure: the use of nearest-neighbor support vectors and correcting imbalanced samples-Evidence from the Chinese hotel industry, Tourism Manage. 33 (2012) 622–634.
[20] H. Li, J. Sun, Predicting business failure using an RSF-based case-based reasoning ensemble forecasting method, J. Forecast. 32 (2013) 180–192.
[21] F.Y. Lin, D.R. Liang, E.C. Chen, Financial ratio selection for business crisis prediction, Expert Syst. Appl. 38 (2011) 15094–15102.
[22] F.Y. Lin, C.C. Yeh, M.Y. Lee, The use of hybrid manifold learning and support vector machines in the prediction of business failure, Knowl.-Based Syst. 24 (2011) 95–101.
[23] X.Q. Ma, N. Sulaiman, H.W. Qin, T. Herawan, J.M. Zain, A new efficient normal parameter reduction algorithm of soft set, Comput. Math. Appl. 62 (2011) 588–598.
[24] P.K. Maji, R. Biswas, A.R. Roy, Soft set theory, Comput. Math. Appl. 45 (2003) 555–562.
[25] J.H. Min, Y.C. Lee, Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters, Expert Syst. Appl. 28 (2005) 603–614.
[26] S.H. Min, J. Lee, I. Han, Hybrid genetic algorithms and support vector machines for bankruptcy prediction, Expert Syst. Appl. 31 (2006) 653–660.
[27] D. Molodtsov, Soft set theory – first results, Comput. Math. Appl. 37 (1999) 19–31.
[28] J.A. Ohlson, Financial ratios and the probabilistic prediction of bankruptcy, J. Account. Res. 18 (1980) 109–131.
[29] Z. Pawlak, Rough sets, Int. J. Comput. Inform. Sci. 11 (1982) 341–356.
[30] N. Senan, R. Ibrahim, N.W. Nawi, I.T.R. Yanto, T. Herawan, Soft set theory for feature selection of traditional Malay musical instrument sounds, in: Zhu et al. (Eds.), Information Computing and Applications, Publishing, Berlin, 2010, pp. 253–260.
[31] T.S. Shin, T.S. Lee, H.J. Kim, An application of support vector machine in bankruptcy prediction model, Expert Syst. Appl. 28 (2005) 127–135.
[32] J.A. Swets, Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected papers, Lawrence Erlbaum Association, Mahwah, NJ, 1996.
[33] K. Thangavel, A. Pethalakshmi, Dimensionality reduction based on rough set theory: a review, Appl. Soft Comput. 9 (2009) 1–12.
[34] Z. Xiao, X.L. Yang, Y. Pang, X. Dang, The prediction for listed companies' financial distress by using multiple prediction methods with rough set and Dempster–Shafer evidence theory, Knowl.-Based Syst. 26 (2012) 196–206.
[35] H.J. Zmijewski, Methodological issues related to the estimation of financial distress prediction models, J. Account. Res. 22 (1984) 59–82.
[36] C. Zopounidis, Evaluation du risque de défailance de l'ebterprise: Méthods et cas d'application, Economica, Parias, 1995.
[37] Y. Zou, Y. Chen, Research on soft set theory and parameters reduction based on relational algebra, in: Proceeding of Second International Symposium on Intelligent Information Technology Application, 2008, pp. 152–156.