Financial distress prediction using support vector machines: Ensemble vs. individual

Jie Sun*, Hui Li

School of Economics and Management, Zhejiang Normal University, Jinhua City, Zhejiang Province 321004, China

A R T I C L E   I N F O

Article history:
Received 27 July 2009
Received in revised form 26 July 2011
Accepted 18 March 2012
Available online 30 March 2012

Keywords:
Financial distress prediction
Support vector machine ensemble
Multiple classifiers

A B S T R A C T

Financial distress prediction (FDP) is of great importance to both inner and outside parts of companies. Though lots of literatures have given comprehensive analysis on single classifier FDP method, ensemble method for FDP just emerged in recent years and needs to be further studied. Support vector machine (SVM) shows promising performance in FDP when compared with other single classifier methods. The contribution of this paper is to propose a new FDP method based on SVM ensemble, whose candidate single classifiers are trained by SVM algorithms with different kernel functions on different feature subsets of one initial dataset. SVM kernels such as linear, polynomial, RBF and sigmoid, and the filter feature selection/extraction methods of stepwise multi discriminant analysis (MDA), stepwise logistic regression (logit), and principal component analysis (PCA) are applied. The algorithm for selecting SVM ensemble’s base classifiers from candidate ones is designed by considering both individual performance and diversity analysis. Weighted majority voting based on base classifiers’ cross validation accuracy on training dataset is used as the combination mechanism. Experimental results indicate that SVM ensemble is significantly superior to individual SVM classifier when the number of base classifiers in SVM ensemble is properly set. Besides, it also shows that RBF SVM based on features selected by stepwise MDA is a good choice for FDP when individual SVM classifier is applied.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Financial distress is the situation that an enterprise’s financial condition becomes worse according to its profitability, solvency, cash flow, and so on. Specifically, financial distress may include continuous or great decrease of profit, deferred payment of liability or interest, deferred payment of preferred dividends, default on other financial bills, and even bankruptcy. For this study, we focus on financial distress from the aspect of profitability. Namely, financial distress is defined as the state that a company has had negative net profit in consecutive two years, or its stockholder’s equity per share becomes lower than the stock’s face value due to one year’s great loss.

For decades, financial distress prediction (FDP) has been a very important topic in both practical and academic fields of corporate finance. From practical view, stockholders, creditors, senior management, auditors, et al. are all interested in FDP because it has great impact on their decision making. Besides, financial distress also brings serious social problems such as unemployment, economic depression and financial crisis if many companies run into financial distress in the same period. Hence, there is urgent demand for accurate and stable FDP tool in practice, to which many academic researchers has been devoted.

The rest of the paper is divided into five sections. Section 2 gives the literature review and contribution of this study. Section 3 states the theoretical framework of SVM ensemble for FDP. Section 4 is the experimental design and Section 5 illustrates the experimental results and analysis. Section 6 makes conclusion.

2. Literature review and contribution of this study

2.1. FDP based on single classifier

Fitzpatrick [1] pioneered the research on FDP based on analysis of financial ratios, which was followed by Beaver [2] with univariate model. After that, FDP research began to enter multi-variate stage. Altman [3] firstly attempted multiple discriminant analysis (MDA) method to construct the Z-score model, which was considered as a milestone for FDP research. Afterwards, various statistical models and artificial intelligence methods were applied to FDP. To improve prediction accuracy, Altman [4] further put forward Zeta-model with newer data and more variables still based on MDA. Ohlson [5] broke through the linear separation and utilized Logit model to FDP since the relationship between financial distress probability and
financial ratios is more likely to be nonlinear, because the financial market tends to be chaos and nonlinear [6]. Zavgren [7] continued the study of Logit FDP model, but he obtained the input variables by factor analysis. In early 1990s, the artificial intelligence method of neural networks (NNs) started to be tried in FDP and its performance was often compared with that of MDA and Logit. Odom and Sharda [8] employed back propagation NNs with the same financial ratios of Altman and compared it with MDA. Many studies of FDP based on NNs were carried out in that stage. Though minority of them made the conclusion that NNs FDP model did not outperform statistical ones [9,10], most literatures provided the supporting evidence that NNs FDP model was superior to statistical ones in terms of prediction accuracy [11–16]. Besides, Back et al. [17] and Anandarajan et al. [18] tried to integrate genetic algorithm (GA) and NNs, aiming at further enhancing prediction accuracy of NNs by the wrapper feature selection based on GA. Atiya [19] improved the performance of NNs for FDP through indicators extracted from the equity markets.

Later on, artificial intelligence classification techniques based on GA, rough sets (RS), and case-based reasoning (CBR) were also applied to FDP. Shin and Lee [20] and Kim and Han [21] respectively utilized GA to draw quantitative rules and qualitative rules for bankruptcy prediction, but their rules’ coverage were both relatively low. Dimitras et al. [22] and McKee [23] both developed a RS-based bankruptcy prediction model and respectively applied them to Greece data set and USA data set. CBR based FDP methods usually use k-nearest neighbor principle. Jo and Han [24,25] concluded that there was no significant difference between traditional CBR model and MDA model. Park and Han [26] enhanced the prediction accuracy of CBR based FDP method through weighting k-nearest neighbors by analysis of hierarchical process (AHP). Sun and Hui [27] put forward a FDP method based on similarity weighted voting CBR with application to Chinese listed companies’ FDP.

As a relatively new artificial intelligence method, support vector machine (SVM) has shown very inspiring results in its application to FDP. Shin et al. [28] and Min and Lee [29] both applied SVM to Korean bankruptcy prediction and drew the conclusion that SVM outperforms MDA, Logit and NNs. Hui and Sun [30] and Ding et al. [31] attempted SVM model for Chinese listed companies’ FDP and obtained similar conclusions. SVM can often get better result than other classification and prediction techniques when the number of samples is relatively small. This makes SVM particularly suitable for FDP problem in that the number of financially distressed companies is limited in the real world. Therefore, some other techniques began to be integrated with SVM to form SVM-based hybrid methods for FDP. Min et al. [32] and Wu et al. [33] both integrated GA with SVM to further improve the prediction ability of SVM model for FDP. The former used GA to optimize both input features and model parameters, and the latter only use GA to optimize model parameters. Hua et al. [34] interpreted and modified the outputs of the SVM classifiers according to the result of logistic regression analysis, which was named as integrated binary discriminant rule by them.

2.2. FDP based on classifier ensemble

Study on classifier ensemble has been carried out for decades. Different ensemble methods for constructing diversified base classifiers, including bagging, boosting, stacking, random subspace, etc., and different combination decision mechanisms including majority voting, weighted majority voting, product rule, minimum or maximum, Borda count, simple average, etc. have been widely studied [35]. Friedman [36] explored the property of bagging by Taylor expansion and found that bagging can reduce the nonlinear variation. However, researches on classifier ensemble for FDP just began to rise in recent years. Jo and Han [24] combined the bankruptcy prediction results of CBR, NNs and MDA, to find that integration of three classifiers produced higher prediction accuracy than individual models. Sun and Li [37] put forward a FDP method based on weighted majority voting combination of diversified multiple classifiers, which was constructed by different classification algorithms on the same dataset. As a result, it got higher average accuracy and lower variance and coefficient of variation than base classifiers. Alfaro et al. [38] carried out an empirical comparison of AdaBoost and NNs for bankruptcy and found that the former outperformed the latter. Nanni and Luminini [39] made a comparison among several ensemble methods for bankruptcy prediction and credit scoring, and showed that method of Random Subspace outperformed the other ensemble methods. Tsai and Wu [40] compared the performance of the single NNs classifier with the (diversified) multiple NNs classifiers over three datasets for the bankruptcy prediction and credit scoring problems. They found that NNs ensemble did not outperform a single best NNs classifier in many cases, which was attributed to too little training dataset for diversified single classifiers and the binary classification problem of FDP. Hung and Chen [41] proposed a selective ensemble for bankruptcy prediction, which selects some suitable classifiers based on an expected probability of each individual classifier to replace the voting policy.

2.3. Contribution of this study

The previous researches acted an important role in promoting the performance of FDP, and FDP study based on classifier ensemble has become a new trend of this field. Though single SVM classifier or hybrid method based on it has been proved to have promising performance and shown much superiority to other classification algorithms in FDP, whether combination of multiple SVM classifiers or SVM ensemble can bring even better FDP performance still remains to be found out. It is of great significance if SVM ensemble can further enhance FDP performance. The main contribution of this paper is to propose a SVM ensemble for FDP, whose diversified single SVM classifiers are to be produced by different kernel functions and different feature subspaces. It is applied to Chinese listed companies’ FDP and empirically tested with real world dataset.

3. Theoretical framework of SVM ensemble for FDP

3.1. Framework of SVM ensemble

The motivation of ensemble is to take advantage of various information output by multiple different base classifiers instead of using result of a conventional single classifier, so that classification performance can be improved and biased decision can be better avoided [35]. As consensus of the domain researchers, the effectiveness of the ensemble approach depends on the accuracy and diversity of the base classifiers [42–44]. If base classifiers in the ensemble system are of low accuracy, the combination based on them is meaningless due to incorrect input information. Similarly, without diversity, the combination of multiple classifiers can only increase the system’s complexity instead of improving performance. Therefore, SVM ensemble system should be composed of diverse SVM classifiers with acceptable individual performance. Thus, when some base classifiers output wrong predicted labels, other base classifiers may still have possibility to output the right predicted label [45]. On the above theoretical analysis, framework of SVM ensemble for FDP is designed as Fig. 1, which mainly focuses
on how to construct diverse base SVM classifiers with relatively high individual accuracy.

Generally, there are mainly three approaches to generate diverse base classifiers. One is to apply different learning algorithms (with heterogeneous model representations) to a single dataset. Another is to apply a single learning algorithm with different parameters settings to a single dataset. The last is to apply a single learning algorithm to different versions of a given dataset [46]. This paper integrates two of the above three methods, which means that several SVM algorithms with different kernel functions will be applied on different feature subsets of a given data set with the optimal parameters, to obtain a number of diversified candidate SVM classifiers. Which candidate SVM classifiers will become the members of SVM ensemble is up to their individual performances and the diversity degree among them.

For a specific FDP task, the SVM ensemble system obtains the final prediction result through four steps. Step 1 is the process of selecting features from initial feature set by some feature selection or feature extraction methods, to construct several feature sets that will be used to train SVM classifier. Step 2 is to construct candidate SVM classifiers with different kernel functions based on the different feature sets selected in step 1. In step 3, selection of base classifiers for SVM ensemble is carried out through accuracy and diversity analysis. In the last step, weighted majority voting is utilized to combine the results of base classifiers to obtain a final prediction result, which can be used for financial decision aiding. The above four steps are discussed in detail in the following subsections.

3.2. Feature selection methods

Feature selection may influence the quality of an ensemble in several ways, namely by reducing model complexity, promoting diversity of ensemble members, and affecting the trade-off between the accuracy and diversity of ensemble members [47]. Usually, feature selection methods can be classified into two categories based on its relation with base learner: filter and wrapper.

Filter approaches make use of human’s experiential expertise or statistical information of the dataset to carry out feature selection and are independent of the base classifier. Filter approaches can be further divided into transformation-based feature selection methods such as principal component analysis (PCA) and relevant measure-based ones which rely on the definition of a relevant measure, for example, correlation coefficient, mutual information, similarity measure and separability measure [48]. Filter is generally computationally more efficient than a wrapper approach, but its major drawback is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm that is used to construct the classifier [49]. Another limitation of filter approaches based on relevant measure is that they require users to determine the number of features selected, rather than providing stopping criteria [48].

Wrapper approaches, on the other hand, use classification accuracies or criteria derived from the classifier to rank the discriminative power of all (or part) of the possible feature subsets and select the subset that produces the best performance. Since wrapper usually involves search, choosing a proper way of searching is
also an important issue. Because the number of feature subsets is usually very large, a simple exhaustive search is computationally infeasible in practice. Therefore, suboptimal yet practical search method is desired. The well known stepwise regression is a wrapper approach based on a greedy search, which is computationally efficient yet achieves suboptimal solution. Though randomized or probabilistic search such as genetic algorithm (GA) promises a solution close to the optimal solution, it is also computationally more involved [49].

Above all, both filter and wrapper methods have their own merits and limitations. As a relatively computationally efficient wrapper method, stepwise regression such as stepwise MDA or stepwise logistic regression belongs to wrapper when the base learner is just trained with MDA or logistic regression, but they become filter feature selection when other classification algorithms are used to train base classifiers on their selected feature subset. Because of the complexity of the ensemble system itself, the feature selection methods used for the step 1 of the above SVM ensemble system for FDP should be informative as well as simple and fast. Therefore, it is more reasonable to use filter feature selection idea for the SVM ensemble. Thus, PCA, stepwise MDA and stepwise logistic regression are chosen to select different feature subsets for training SVM learner. The filter methods based on relevant measures are excluded due to the difficulty of determining the optimal number of features selected.

3.3. SVMs with different kernel functions

SVM was put forward by Vapnik in 1990s. As a relatively new machine learning technique, it is developed on the basis of statistical learning theory. Basic idea of SVM is to construct an optimal separating hyperplane with high classification accuracy. A simple description of the SVM algorithm is provided as follows [28–29,50–52].

Consider a training dataset \( D = \{x_i, y_i\}_{i=1}^{N} \), in which \( N \) is the total number of training samples. \( x_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(n)})^T \in \mathbb{R}^n \) is a \( n \) dimension input vectors and \( y_i \in \{-1, +1\} \) represents target labels. In condition of linear separation with two classes, SVM is to find an optimal separating plane which completely separates the two classes and makes the two parallel bounding planes obtain the biggest margin width at the same time (as shown in Fig. 2).

To solve non-linear classification tasks, a nonlinear function \( \Phi(x) \) is usually employed to map input space to a higher-dimensional feature space. This allows the SVM to fit the maximum-margin hyperplane in the transformed feature space by the same technique as linear model. The instances nearest to the hyperplane are called support vectors, and the other instances are irrelevant to the bounding hyperplanes.

Because the margin width between both bounding hyperplanes equals to \( 2/(\|w\|^2) \), the constraint optimization model of soft-margin based SVM is as follows:

\[
\begin{align*}
\min & \frac{1}{2}w^T w + C \sum_{i=1}^{N} \xi_i \\
\text{s.t.} & \quad y_i [w^T \Phi(x_i) + b] \geq 1 - \xi_i \quad (i = 1, 2, \ldots, N) \\
& \quad \xi_i \geq 0 \quad (i = 1, 2, \ldots, N)
\end{align*}
\]

in which \( \xi_i \) are slack variables, which allows a small part of misclassification. In many cases, feature space cannot be linearly separated. If SVM only allows constructing perfect separating hyperplane without any error, over-fitting is easy to appear. \( C \in \mathbb{R}^+ \) is a tuning parameter, weighting the importance of classification errors with the margin width. So, the above optimization problem becomes a trade-off between the margin maximization and training errors minimization. It is transformed into its dual problem as follows:

\[
\begin{align*}
\max & \frac{1}{2}a^T Qa - e^T a \\
\text{s.t.} & \quad 0 \leq a_i \leq C \quad (i = 1, 2, \ldots, N) \\
& \quad y^T a = 0
\end{align*}
\]

\( e^T \) is the \( N \)-dimension vector of all ones. \( Q \) is a \( N \times N \) positive semi-definite matrix, and \( Q_{ij} = y_i y_j K(x_i, x_j) \), in which \( K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \) is called kernel function. \( a_i \) is Lagrange multiplier. A multiplier exits for each training data instance and data instances corresponding to non-zero \( a_i \) are support vectors.

Do this optimization problem and the ultimate SVM classifier is constructed as follows:

\[
\text{sgn} \left( \sum_{i=1}^{N} a_i y_i K(x_i, x) + b \right)
\]

Some typical SVM kernels include linear kernel function, polynomial kernel function, radial basis kernel function and sigmoid kernel function.

(1) Linear function with no extra parameter:

\[
K(x, y) = x \cdot y
\]

(2) Polynomial function with degree \( d \), \( \gamma > 0 \) and \( c \in \mathbb{N} \):

\[
K(x, y) = [\gamma (x \cdot y) + c]^d
\]

(3) Gaussian radial basis function (RBF) with \( \gamma > 0 \):

\[
K(x, y) = \exp[-\gamma (x-y)^2]
\]

(4) Sigmoid function with \( \gamma > 0 \) and \( c \in \mathbb{N} \):

\[
K(x, y) = \tanh(\gamma (x \cdot y) + c)
\]

As shown by the SVM theory and different kernel functions, the tuning parameter \( C \) and the associated kernel parameters such as \( \gamma \) and \( d \) should be properly set before training candidate SVM classifiers. This process is often influential on how an SVM classifier behaves, because improper setting of these parameters will lead to training problems such as over-fitting or under-fitting. Usually some kinds of search methods such as grid search and genetic algorithm can be applied to find proper values for these parameters.

3.4. Criteria for selecting base SVM classifiers

How to select base classifiers from candidate classifiers set is one of the critical problems in an ensemble system. Since individual
performance and diversity are the basic properties to guarantee the ensemble's effectiveness, the method for selecting base classifiers is designed by two criteria: cross validation accuracy on training dataset and diversity measure. The specific algorithm is described as follows:

**Input.** (1) Candidate classifiers set, denoted as \( C_C = \{f_1, f_2, \ldots, f_l\} \); (2) training data set, denoted as \( \text{Train data} \); (3) the ideal number of base classifiers, denoted as \( K (K < r) \).

**Output.** Base classifiers set denoted as \( B_C \).

**Algorithm.**

1. Obtain the cross validation accuracy of each candidate classifier and the corresponding prediction result.
   For each candidate classifier \( f_i \) in \( C_C \)
   \[ [A_{C_C}, \text{Predicted class}] = \text{Cross Validation}(f_i, \text{Train data}) \]

2. Rearrange the candidate classifiers set by the descending order of cross validation accuracy.
   \[ C_C = \text{Rearrange}(C_C) = \{f_1, f_2, \ldots, f_l\} \]

3. Select base classifiers from candidate classifiers initially.
   \[ B_C = \{f_i, f_j, \ldots, f_k\} \]

4. Calculate the diversity measure between each pair of base classifiers.
   For each pair of base classifiers \( f_i \) and \( f_j \) in \( B_C \):
   \[ i < j \]
   \[ D_{MC} = \text{Diversity Measure}(f_i, f_j) \]
   If \( D_{MC} \) indicates no diversity at all
   Delete \( B_C = \{f_i, f_j\} \) from base classifiers set \( B_C \)

5. If \( C_C \) is not empty:
   Add \( \{f_i, f_j\} \) from \( C_C \) to base classifiers set \( B_C \)
   \[ C_C = C_C \cup \{f_i, f_j\} \]

In the above algorithm, diversity degree between two classifiers can be measured by one of the diversity measures, including Q statistic, correlation coefficient, disagreement measure, and double-fault measure [44,53,54]. For example, Q statistic is defined as follows.

\[
Q_0 = \frac{M_{ij}^1M_{ji}^{01} - M_{ij}^{01}M_{ji}^{10}}{M_{ij}^{11}M_{ji}^{00} + M_{ij}^{01}M_{ji}^{10}}
\] (8)

\( M_{ij} \) is the number of samples which are correctly recognized by both classifier \( f_i \) and classifier \( f_j \), \( M_{ij}^{01} \) is the number of samples which are recognized by both classifier \( f_i \) and classifier \( f_j \), \( M_{ij}^{10} \) is the number of samples which are correctly recognized by classifier \( f_i \) and misclassified by classifier \( f_j \), \( M_{ij}^{01} \) is the number of samples which are misclassified by classifier \( f_i \) and correctly recognized by classifier \( f_j \). The value range of \( Q_0 \) is \([-1, 1]\). \( Q_0 = 1 \) means completely positive correlation between classifier \( f_i \) and classifier \( f_j \), and \( Q_0 = -1 \) means completely negative correlation between classifier \( f_i \) and classifier \( f_j \).

### 3.5. Combination mechanism of the ensemble

Various combination methods have been proposed for the fusion of output information of ensemble members and their usefulness has been experimentally demonstrated in different domains, for example, majority vote, Borda count, Bayesian method, behavior-knowledge space (BKS), and Dempster–Shafer theories of evidence [55,56]. Among them, majority voting idea is the most simple but also very effective one and needs no extra memory, so it has been applied to FDP in some former researches [37,40,57,58]. Majority voting method selects the class that is supported by the majority of the classifiers as the final decision. Each classifier has one vote that can be cast for any one class in the majority voting. However, it treats classifiers equally without considering their differences in performance. Thus, what emerged as extension of majority voting method is weighted majority voting method, which assigns each classifier different weight according to its performance [59]. The proposed SVM ensemble for FDP weighs each base classifier in terms of their training performance, which is described as follows:

\[
F(x_i) = c_j \quad \text{if} \quad \exists \exists_{c_j} \in C \quad \mathcal{A} \quad T_F(x_i \in c_j) = \max_{j=1}^{K} T_F(x_i \in c_j)
\] (9)

\[
T_F(x_i \in c_j) = \sum_{k=1}^{K} w_k(x_i \in c_j) \quad (j = 1, 2, \ldots, N)
\] (10)

\[
w_k(x_i \in c_j) = \begin{cases} A_{C_k} & \text{if } f_k(x_i) = c_j \\ 0 & \text{else} \end{cases}
\] (11)

\( k = 1, 2, \ldots, K; \quad j = 1, 2, \ldots, N \)

In above formula (9)–(11), \( C = \{c_1, c_2, \ldots, c_K\} \) is class set which is composed of \( N \) different classes; \( w_k(x_i \in c_j) \) represents the weight of classifier \( f_k \) when voting to class \( c_j \) for the given sample \( x_i \); \( K \) is the number of base classifiers in ensemble system; \( A_{C_k} \) is the cross validation accuracy of classifier \( f_k \) on the training dataset.

### 3.6. Theoretical analysis on the performance of SVM ensemble

Suppose there are \( K \) independent base classifiers in SVM ensemble and each has the same error rate \( p \), the probability of getting \( t \) incorrect votes by ensemble is \( P(r) \), which is a binomial distribution as formula (12) shows.

\[
P(r) = \frac{K!}{r!(K-r)!} p^r (1-p)^{K-r}
\] (12)

For example, when \( K = 11 \) and \( p = 0.3 \), the error probability of SVM ensemble by voting is represented as follows.

\[
P(r \geq 6) = \frac{\sum_{r=6}^{11} \frac{11!}{r!(11-r)!} 0.3^r \times 0.7^{11-r}}{0.0782}
\] (13)

Error probability of ensemble is much lower than error probability of base classifiers on the assumption of independency among them. However, certain degree of dependency among base classifiers is inevitable for an ensemble system. Littlestone and Warmuth [59] proved that mistake bounds for weighted majority voting is closely related to the mistake bounds of the best algorithms of the pool. Breiman [60] also concluded that the error bound of a voting ensemble is denoted as formula (14).

\[
P_B(F(x) \neq y) \leq \hat{p}(1 - s \hat{p}) \quad \frac{1}{s^2}
\] (14)

In which, \( \hat{p} \) is the average pairwise correlation between the errors of individual models, and \( s \) measures the extent to which the average number of votes for the correct class exceeds the average vote for any other class. Small \( \hat{p} \) and high \( s \) results in small ensemble error.

Therefore, the proposed SVM ensemble is expected to produce good FDP performance for the following two reasons. On one hand, the most accurate individual is definitely included in the SVM ensemble and more accurate individual is more likely to become base classifier, according to the criteria for selecting base SVM classifiers in Section 3.4. On the other hand, correlation among base classifiers is reduced by the iteration process of deleting and adding base classifiers in terms of diversity measure. Such expectation of SVM ensemble's performance for FDP is to be validated in the following parts of empirical experiment.

### 4. Experimental design

#### 4.1. Experimental assumption and objective

Though the pattern recognition problem like cancer diagnosis definitely has different cost for two types of misclassifying, whether
such difference of misclassification cost exists in FDP is still disputed. Misclassifying distressed company as non-distressed one will delay implement of measures for distress prevention. At the same time, misclassifying non-distressed company as distressed one will also bring the company disadvantageous public evaluation, which may become a block for the company's development. Hence, it is assumed that two types of misclassification have the same cost in this study.

On the above assumption, the primary objective of the experiment is to validate the effectiveness of the proposed SVM ensemble method for FDP, by comparing its experimental results with those of single SVM classifiers. In detail, the experiment should be designed to test whether SVM ensemble improves the prediction accuracy. For this purpose, some statistic analysis should be done on the experimental results, and results coming from only one time of hold-out experiment are certainly not enough for statistic analysis. Therefore, the primary experimental objective is to be realized by 30-time hold-out training and testing.

Additionally, we set a subsidiary experimental objective to explore the appropriate range of the number of base classifiers in SVM ensemble for FDP. In spite of many literatures which attempted to validate the ensemble's advantage over single classifiers, the number of base classifiers in an ensemble system is mostly up to the constructor's subjective experience. Although it is an important parameter for ensemble, almost no research has involved empirical analysis on such a topic to our knowledge, especially in the domain of FDP. Our experiment attempts to find the number range of base classifiers in ensemble that helps achieve good performance for FDP, thereby at least providing some guidance for the structure of ensemble in further study or practical application.
Table 1

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable and feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Profitability ratios</td>
<td>V01: Gross income to sales</td>
</tr>
<tr>
<td></td>
<td>V03: Earning before interest and tax to total asset</td>
</tr>
<tr>
<td></td>
<td>V05: Net profit to current assets</td>
</tr>
<tr>
<td></td>
<td>V07: Profit margin</td>
</tr>
<tr>
<td>Activity ratios</td>
<td>V09: Account receivables turnover</td>
</tr>
<tr>
<td></td>
<td>V11: Account payable turnover</td>
</tr>
<tr>
<td>Liability ratios</td>
<td>V13: Current assets turnover</td>
</tr>
<tr>
<td></td>
<td>V15: Current ratio</td>
</tr>
<tr>
<td>Growth ratios</td>
<td>V21: Interest coverage ratio</td>
</tr>
<tr>
<td>Structure ratios</td>
<td>V24: Proportion of current assets to total assets</td>
</tr>
<tr>
<td>Per share ratios</td>
<td>V26: Proportion of equity to fixed assets</td>
</tr>
<tr>
<td></td>
<td>V28: Earning per share</td>
</tr>
<tr>
<td></td>
<td>V30: Operating cash flow per share</td>
</tr>
<tr>
<td></td>
<td>V02: Net profit to sales</td>
</tr>
<tr>
<td></td>
<td>V04: Net profit to total assets</td>
</tr>
<tr>
<td></td>
<td>V06: Net profit to fixed assets</td>
</tr>
<tr>
<td></td>
<td>V08: Net profit to equity</td>
</tr>
<tr>
<td></td>
<td>V10: Inventory turnover</td>
</tr>
<tr>
<td></td>
<td>V12: Total assets turnover</td>
</tr>
<tr>
<td></td>
<td>V14: Fixed assets turnover</td>
</tr>
<tr>
<td></td>
<td>V16: Operating cash flow to current liability</td>
</tr>
<tr>
<td></td>
<td>V18: Equity to debt ratio</td>
</tr>
<tr>
<td></td>
<td>V20: Ratio of liabilities to market value of equity</td>
</tr>
<tr>
<td></td>
<td>V22: Growth rate of primary business</td>
</tr>
<tr>
<td></td>
<td>V25: Proportion of fixed assets to total assets</td>
</tr>
<tr>
<td></td>
<td>V27: Proportion of current liability to total liabilities</td>
</tr>
<tr>
<td></td>
<td>V29: Net assets per share</td>
</tr>
</tbody>
</table>

4.2. Design of experimental process

The main experimental process is designed as Fig. 3. Firstly, we select sample companies including financially distressed ones and financially healthy ones and construct initial features set which is composed of different financial ratios. All feature values of each sample company are collected and then preprocessed to generate an initial dataset.

Secondly, three feature selection methods, i.e., PCA, stepwise MDA and stepwise Logit, are respectively applied to obtain factors set or features subsets with more discriminative ability and fewer dimensions. As a result, three corresponding experimental datasets based on them can be built by extracting information from initial dataset, and they are foundation of the following experimental processes.

To realize the experimental objectives, the core part of the whole experimental process is carried out by thirty times of hold-out training and testing, so as to produce thirty groups of prediction accuracies for statistic analysis. In each round of hold-out training and testing, each of the three experimental datasets based on PCA, stepwise MDA and stepwise Logit is divided into training dataset and testing dataset by the proportion of 2:1 randomly, and then SVM algorithms with different kernel functions are used to model candidate classifiers. That is to say, in the division of training and testing datasets, two thirds of financially healthy companies and two thirds of financially healthy companies in each experimental dataset are respectively selected to form training samples by simple random sampling without replacement, and the rest one third are used to form testing samples. Because we use five SVM algorithms based on kernel functions including linear kernel, polynomial kernel with degree of two (i.e. $d = 2$), polynomial kernel with degree of three (i.e. $d = 3$), sigmoid kernel and RBF kernel, fifteen candidate SVM classifiers can be modeled based on the three experimental datasets.

In the experiment, the method integrating grid search and five-fold cross validation on training dataset is used to find the proper parameter values for each SVM algorithm on each experimental dataset, so as to ensure the acceptable individual performance. According to formula (4)–(7) in Section 3.3, there are four parameters that should be determined when training SVM classifiers, namely the tuning parameter $C$ and the kernel parameter of $\gamma$, $d$, and $c$. To reduce the number of searching parameters in the experiment, the polynomial degree parameter $d$ is respectively valued as two and three, and constant term $c$ uses the default value of zero in LIBSVM. The parameter $C$ is searched in the range of $\{2^{-5}, 2^{-4}, \ldots, 2^{10}\}$, and the parameter $\gamma$ is searched in the range of $\{2^{-10}, 2^{-9}, \ldots, 2^2\}$, and the pair of parameter values for $C$ and $\gamma$ that leads to the highest cross validation accuracy on training dataset is selected. Such kind of parameter searching is repeated for each SVM algorithm (linear, RBF, polynomial with the degree of three, polynomial with the degree of two, and sigmoid) and each experimental dataset (PCA, stepwise MDA and stepwise Logit) in the thirty times of experimental training, namely totally four hundred and fifty times. For limitation of paper length, the specific parameter values are not listed here.

Then, according to the criteria for selecting base SVM classifiers which is described in Section 3.4, candidate classifiers with relatively high cross validation accuracy on training dataset and certain degree of diversity are chosen to form the base classifiers set. To explore the appropriate range of the number of base classifiers, or to supply empirical support for the determination of the number of base classifiers, the number of base classifiers ranging from two to fifteen are all tried out in our experiments.

At last, experimental analysis should be carried out based on thirty times of holdout prediction results.

4.3. Selection of sample companies

According to the definition of financial distress in Section 1, listed companies that are specially treated (ST) by Chinese Stock Exchange due to abnormal financial status are selected as financial distress samples. Matching method is utilized to choose normal non-financial-distress companies according to the following rules.

(1) The matched two companies have the same or similar industry.
(2) The chosen normal company should never be specially treated.
(3) The matched two companies should have similar asset size in the range of $[−30\%, +30\%]$. But if no normal company exists in such range of asset size, the condition is relaxed to choose a normal company that has as small asset difference as possible. According to these sampling criteria and Chinese ST information from 2000 to 2005, the initial samples consist of 135 pairs of financial distress and non-financial-distress companies listed in Shenzhen Stock Exchange and Shanghai Stock Exchange. Namely, the class distribution is not biased.

4.4. Selection of initial features

The initial features selected for the empirical experiment involve six aspects, i.e. profitability, activity, solvency, growth,
structural ratios and per share ratios, which can provide comprehensive indication of firms’ financial state. Each aspect is composed of several financial ratios, as listed in Table 1.

4.5. Data collection and preprocessing

Let’s take company’s ST year as the benchmark year (t – 0). Then, our empirical experiment is done based on year (t – 2), which represents two years before ST. It means that financial distress is tried to be predicted according to financial information two years in advance. The original data used in this research are obtained from China Stock Market and Accounting Research Database.

The preprocessing to eliminate missing and outlier data is carried out by the following steps. (1) Sample companies in case of missing financial ratio data were eliminated. (2) Sample companies with financial ratios deviating from the mean value as much as three times of standard deviation are excluded. Thus the initial datasets (so called in Fig. 3) can be built.

4.6. Experimental data sets constructed by three feature selection methods

Three experimental datasets are constructed by PCA, stepwise MDA and stepwise Logit. For PCA, nine principal components are extracted and their cumulative variance percentage is over 75%. The component matrix is listed in Table 2. It is shown that the first important component mainly covers financial ratios of profitability, which are followed by solvency and activity ability. The features selected by stepwise MDA and stepwise Logit are listed in Table 3. Although the financial ratios selected by the two stepwise methods are almost same, stepwise MDA is biased in favor of profitability and stepwise Logit is biased towards activity ratio and solvency. PCA, stepwise MDA and stepwise Logit all show that profitability, activity and solvency are the most important features for the prediction of financial distress defined in this paper.

5. Results and analysis

5.1. Performance of different candidate SVM classifiers

Each candidate SVM classifier’s testing accuracies based on 30 times of hold-out validation are listed in Table 4, in which the bold values mean the highest testing accuracies in each time of hold-out validation among all candidate SVM classifiers and the bold italic value in the last line means the lowest mean testing accuracy in 30 times of hold-out validation among all candidate SVM classifiers. T statistics and significance level of one-tailed mean comparison between each pair of candidate SVM classifiers are listed in Table 5.

From Table 4, R_M (RBF SVM based on features selected by stepwise MDA) has the highest average testing accuracy of 81.16%, but S_M (sigmoid SVM based on features selected by stepwise MDA) gets the most (9) times of highest testing accuracy in the total 30 times. Among the fifteen SVM classifiers, R_M, S_P, R_P, P_3, P_4, and P_5 are the top five ones with relatively high average testing accuracy. According to Table 5, though R_M achieves the highest average testing accuracy, there exist no significant difference between R_M
and the following eight candidate classifiers of S_P, R_P, P_3,P, P_3,M, S_M, P_3,M, P_3,P and L_P.

The L_M (linear SVM based on features selected by stepwise MDA) has the lowest average testing accuracy, and its difference with the other candidate classifiers are all significant at the level of 1% except for L_L. The other two linear SVM classifiers respectively based on features selected by stepwise Logit or PCA (namely L_L and L_P) both have average testing accuracy lower than 0.8%. It indicates that linear SVM is not very suitable for FDP for the possible reason that there exists non-linear relationship between financial distress probability and financial ratios. This result is consistent with the conclusion of some previous researches, in which nonlinear NNS outperformed linear SVM in [11,13].

For a given SVM algorithm with specific kernel function, the feature selection method of stepwise Logit has the relatively worse performance compared with stepwise MDA and PCA. Tables 4 and 5 show that stepwise Logit gets the lowest testing accuracies among the three feature selection methods for SVM with RBF, polynomial, and sigmoid kernels, and the differences are all statistically significant at the level of 1%. For linear SVM, stepwise Logit is also significantly worse than PCA at the level of 1%, but it is not significantly different from stepwise MDA at the level 5% though it has higher testing accuracy than stepwise MDA. Hence, stepwise Logit is not an appropriate feature selection method for SVM, while stepwise MDA and PCA are much better choices. To some extent, these results agree with former researches on the assertion that RBF-SVM often outperforms SVM with other kernels. Here, it gets the highest average testing accuracy. Furthermore, our results indicate that features selected by stepwise MDA and PCA are more suitable for RBF-SVM than stepwise Logit.

Table 5

<table>
<thead>
<tr>
<th>Feature Selection Method</th>
<th>Accuracy</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_M</td>
<td>83.33</td>
<td>1%</td>
</tr>
<tr>
<td>L_L</td>
<td>84.72</td>
<td>1%</td>
</tr>
<tr>
<td>L_P</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>R_M</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>R_L</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>R_P</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>P_3,M</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>P_3,P</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>P_3,M</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>S_M</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>S_L</td>
<td>87.50</td>
<td>1%</td>
</tr>
<tr>
<td>S_P</td>
<td>87.50</td>
<td>1%</td>
</tr>
</tbody>
</table>

Significance levels: 1% = ***, 5% = **, 10% = *.
Therefore, if individual SVM classifier is to be applied to FDP, RBF-SVM based on features selected by stepwise MDA is a good choice, followed by the other nonlinear SVM with feature selection methods of stepwise MDA or PCA.

5.2. Performance of SVM ensembles with different number of base classifiers

Testing accuracy of SVM ensembles with different number of base classifiers based on 30 times of hold-out validation are listed in Table 6. The numbers in the column titles represent the different numbers of base classifiers. For example, 2BCs means SVM ensemble with 2 base classifiers and 15BCs means SVM ensemble with 15 base classifiers. In Table 6, the bold values mean the highest testing accuracies in each time of hold-out validation among all SVM ensembles with different number of base classifiers and the bold italic value in the last line means the lowest mean testing accuracy in 30 times of hold-out validation among all SVM ensembles with different number of base classifiers.

Table 6 shows that SVM ensemble with 11 base classifiers obtains the highest average testing accuracy of 82.55%, and it also gets the most (14) times of highest testing accuracy in the total 30 times. While, SVM ensemble with 2 base classifiers leads to the lowest average testing accuracy of 79.81%, and it gets the least (1) time of highest testing accuracy in the total 30 times. Besides, when the number of base classifiers is in the range of [10,14], the ensemble’s average testing accuracy is always higher than 82%. When the number of base classifiers is less than 4, the SVM ensemble’s average testing accuracy is relatively lower, i.e. <81.2%.

Table 7 shows the significance levels of the tests. The significance levels are represented in bold and italic. The p-values are calculated using the t-test. The significance level is set at 0.05 for all tests. The table also includes the pair-wise comparisons of SVM ensembles and individual SVM classifiers.

<table>
<thead>
<tr>
<th>id</th>
<th>2BCs</th>
<th>3BCs</th>
<th>4BCs</th>
<th>5BCs</th>
<th>6BCs</th>
<th>7BCs</th>
<th>8BCs</th>
<th>9BCs</th>
<th>10BCs</th>
<th>11BCs</th>
<th>12BCs</th>
<th>13BCs</th>
<th>14BCs</th>
<th>15BCs</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_{M}</td>
<td>-7.50**</td>
<td>-10.41***</td>
<td>-10.57***</td>
<td>-10.31***</td>
<td>-10.33***</td>
<td>-10.31***</td>
<td>-11.07***</td>
<td>-12.28***</td>
<td>-12.15***</td>
<td>-12.68***</td>
<td>-12.33***</td>
<td>-10.82***</td>
<td>-11.20***</td>
<td>-10.75***</td>
</tr>
<tr>
<td>R_{M}</td>
<td>1.73**</td>
<td>0.05</td>
<td>0.00</td>
<td>-0.61</td>
<td>-0.81</td>
<td>-0.41</td>
<td>-0.79</td>
<td>-0.99</td>
<td>-1.31</td>
<td>-2.48**</td>
<td>-2.23**</td>
<td>-1.83**</td>
<td>-1.75**</td>
<td>-1.23**</td>
</tr>
<tr>
<td>P_{M}</td>
<td>0.97</td>
<td>0.82</td>
<td>0.75</td>
<td>-1.44</td>
<td>-1.67</td>
<td>-1.23</td>
<td>-1.65</td>
<td>-1.96</td>
<td>-2.21**</td>
<td>-3.14**</td>
<td>-3.04**</td>
<td>-2.76**</td>
<td>-2.59**</td>
<td>-2.15**</td>
</tr>
<tr>
<td>D_{L}</td>
<td>1.40**</td>
<td>-0.95</td>
<td>-0.67</td>
<td>-2.00**</td>
<td>-2.45**</td>
<td>-1.48**</td>
<td>-2.23**</td>
<td>-1.74**</td>
<td>-2.13**</td>
<td>-2.61**</td>
<td>-2.59**</td>
<td>-2.04**</td>
<td>-2.02**</td>
<td>-1.42**</td>
</tr>
<tr>
<td>D_{P}</td>
<td>0.78**</td>
<td>-1.05</td>
<td>-0.94</td>
<td>-1.61</td>
<td>-1.79**</td>
<td>-1.36**</td>
<td>-1.83**</td>
<td>-2.02**</td>
<td>-2.60**</td>
<td>-3.86**</td>
<td>-3.72**</td>
<td>-3.05**</td>
<td>-2.92**</td>
<td>-2.32**</td>
</tr>
<tr>
<td>S_{L}</td>
<td>0.83</td>
<td>1.62**</td>
<td>1.73**</td>
<td>-2.78</td>
<td>-3.43</td>
<td>-2.09**</td>
<td>-2.84**</td>
<td>-2.11**</td>
<td>-2.51**</td>
<td>-2.54**</td>
<td>-2.54**</td>
<td>-2.06**</td>
<td>-2.09**</td>
<td>-1.59**</td>
</tr>
<tr>
<td>S_{P}</td>
<td>0.75**</td>
<td>-0.76</td>
<td>-0.69</td>
<td>-1.25</td>
<td>-1.41</td>
<td>-1.09**</td>
<td>-1.42**</td>
<td>-1.75**</td>
<td>-2.02**</td>
<td>-3.61**</td>
<td>-3.56**</td>
<td>-3.39**</td>
<td>-3.12**</td>
<td>-2.78**</td>
</tr>
<tr>
<td>S_{Q}</td>
<td>2.45**</td>
<td>-1.27</td>
<td>-2.70**</td>
<td>-4.30**</td>
<td>-4.30**</td>
<td>-4.07</td>
<td>-4.46**</td>
<td>-2.97**</td>
<td>-4.61**</td>
<td>-6.94**</td>
<td>-6.37**</td>
<td>-6.88**</td>
<td>-6.17**</td>
<td>-5.13**</td>
</tr>
<tr>
<td>S_{R}</td>
<td>1.51**</td>
<td>-0.43</td>
<td>-0.31</td>
<td>-1.27</td>
<td>-1.64**</td>
<td>-0.97</td>
<td>-1.51**</td>
<td>-1.51**</td>
<td>-1.88**</td>
<td>-2.11**</td>
<td>-2.02**</td>
<td>-1.61**</td>
<td>-1.68**</td>
<td>-1.11**</td>
</tr>
</tbody>
</table>

* Significance level of 10%.
** Significance level of 5%.
*** Significance level of 1%.
Therefore, SVM ensemble with three or less base classifiers is more likely to result in relatively low average testing accuracy and SVM ensemble with more than ten base classifiers is more likely to result in relatively high average testing accuracy. When SVM ensemble system for FDP is to be constructed, it should include more than nine base classifiers ideally and more than three base classifiers at least.

5.3. Performance comparison between SVM ensemble and individual SVM classifiers

According to Tables 4 and 6, the average testing accuracy of R_M, the best individual SVM classifier, is not beyond almost all SVM ensembles with different number of base classifiers except the one with two base classifiers. This shows that SVM ensemble helps improve the performance of FDP based on SVM except when the number of base classifiers is too small and the information to be integrated by ensemble is too limited.

To test whether statistically significant difference exists between the performance of SVM ensemble and that of individual SVM classifier, mean comparison is carried out by 7 test, with the results listed in Table 7. It indicates that SVM ensemble with more than three base classifiers can definitely outperform the relatively worse individual SVM classifiers such as L,M, L,L, R,L, P,S, L,P, S,L and L,P, at the significance level of at least 5%. However, as for the two candidate SVM classifiers that obtain the highest or second highest average testing accuracies, R_M and S,P, only SVM ensembles with certain number of base classifiers significantly improve FDP performance. In detail, only the SVM ensembles with eleven to fourteen base classifiers have significant better performances than R_M and the SVM ensembles with ten to twelve base classifiers have significant better performances than S,P at the level of 5% or 1%. This further proves that the proposed SVM ensemble tends to achieve good FDP performance when the number of base classifiers varies from 10 to 14.

Therefore, we consider that SVM ensemble can improve the FDP performance to some extent. That is, SVM ensemble has better predictive ability than individual SVM classifier in FDP if the parameters are properly set. It demonstrates that the FDP information drawn by base SVM classifiers can be effectively integrated through the SVM ensemble proposed in this paper. As a result, SVM ensemble can fully take the role of information complementation in FDP.

6. Conclusions

Due to the importance of FDP in corporate management, many academic researches compare performance of different single classifiers and find that SVM is a relatively better one in FDP. However, ensemble method for FDP just began to rise in recent years and needs to be further studied. This paper contributes to put forward a new SVM ensemble method for FDP, in which the criteria for selecting base SVM classifiers and the combination mechanism of the SVM ensemble are designed. Individual SVM classifiers’ predictive ability and their diversity degree are considered as the criteria for selecting base SVM classifiers from candidate ones. This helps build an effective SVM ensemble for FDP, and avoids just adding complexity to the SVM-based FDP system without performance improvement. Weighted majority voting is applied as combination mechanism according to base classifiers’ cross validation accuracy on training dataset. The candidate SVM classifiers for the ensemble are constructed through different SVM kernels and different features subset. The applied SVM kernels include linear kernel, polynomial kernel, RBF kernel and sigmoid kernel, and the feature selection/extraction methods include stepwise MDA, stepwise logistic regression and PCA. In the empirical experiment based on Chinese listed companies, 30-time hold-out method is designed. Experimental results indicate that the performance of the proposed SVM ensemble for FDP is significantly better than that of individual SVM classifiers when the number of base SVM classifiers is properly set. Empirically, SVM ensemble with more than nine base classifiers tends to result in acceptable prediction performance in our experiment, and at least more than three base classifiers are needed to avoid an invalid SVM ensemble. If single SVM classifier is used for FDP, RBF-SVM with features selected by stepwise MDA is a good choice.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (Grant No. 70801054), the Zhejiang Provincial Natural Science Foundation of China (Grant No. Y6090392) and the Zhejiang Provincial Philosophy and Social Science Foundation (11ZQN082YB).

References
