Expert Systems with Applications 41 (2014) 2472–2483

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Novel feature selection methods to financial distress prediction

Fengyi Lin [a], Deron Liang [b,*], Ching-Chiang Yeh [c], Jui-Chieh Huang [c]

[a] Department of Business Management, National Taipei University of Technology, Taipei, Taiwan
[b] Department of Computer Science and Information Engineering, and Software Research Center, National Central University, Taoyuan, Taiwan
[c] Department of Business Administration, National Taipei College of Business, Taipei, Taiwan

ARTICLE INFO

Keywords:
Financial distress prediction
Genetic algorithm
Wrappers
Integrated prediction model
Feature selection

ABSTRACT

Financially distressed prediction (FDP) has been a widely and continually studied topic in the field of corporate finance. One of the core problems to FDP is to design effective feature selection algorithms. In contrast to existing approaches, we propose an integrated approach to feature selection for the FDP problem that embeds expert knowledge with the wrapper method. The financial features are categorized into seven classes according to their financial semantics based on experts' domain knowledge surveyed from literature. We then apply the wrapper method to search for "good" feature subsets consisting of top candidates from each feature class. For concept verification, we compare several scholars' models as well as leading feature selection methods with the proposed method. Our empirical experiment indicates that the prediction model based on the feature set selected by the proposed method outperforms those models based on traditional feature selection methods in terms of prediction accuracy.

1. Introduction

Financially distressed prediction (FDP) is a challenging problem that generates extensive studies over the past decades (Altman, 1968; Hua, Wang, Xu, Zhang, & Liang, 2007; Ohlson, 1980; Tam & Kiang, 1992). Recent outbreaks of corporate financial crises worldwide have intensified the need to reform the existing financial architecture. It is generally believed that symptoms and alarms can be observed prior to a business encountering financial difficulty or crisis. The overall objective of business crisis prediction is to build models (or predictors) that can extract knowledge of risk evaluation from past observations and to evaluate business crisis risk of companies with a much broader scope (Altman, 1968; Beaver, 1966; Zmijewski, 1984).

The FDP problem is a typical binary classification problem in the context of the pattern recognition theory, where a predictor attempts to assign one of the two labels, distressed company (D) vs. non-distressed company (ND), to an input sample (Tsai & Wu, 2008). There are two basic issues that have profound impacts on the performance of a predictor (Ben-Bassat, 1982; Lin, Liang, & Chen, 2011): choosing the right feature selection algorithm to find the optimal feature set and selecting the right classification algorithm to build the predictor using that feature set (shown in Fig. 1).

Numerous scholars have conducted research into business crisis prediction. In the early 1960s, scholars applied statistical methods such as Multiple Discriminate Analysis (MDA) (Altman, 1968; Beaver, 1966) and Logit (Hua et al., 2007; Ohlson, 1980; Zmijewski, 1984). In recent years various machine learning algorithms have been used. Examples are the Decision Tree (DT) (Tam & Kiang, 1992), Neural Network (Ozkan-Gunay & Ozkan, 2007; Tam & Kiang, 1992), Support Vector Machine (SVM) (Chandra, Ravi, & Bose, 2009; Chen & Hsiao, 2008; Ding, Song, & Zen, 2008; Hua et al., 2007; Shin, Lee, & Kim, 2005; Wu, Tzeng, Goo, & Fang, 2007), and Case-Based Reasoning (CBR) (Jo & Han, 1996; Li & Sun, 2008; Li & Sun, 2009; Li, Sun, & Sun, 2009). Table 1 summarizes the classifiers used in financial prediction. These previous studies all focused on the second issue of the pattern recognition problem, i.e. to explore better ways of predictor construction based on a chosen classification algorithm. Over the decades, it seems very difficult to significantly improve the forecasting accuracy following the existing approaches (Cho, Kim, & Bae, 2009). In this research, we turn our focus to the feature selection method, which is the other design issue toward the construction of a good predictor.

The feature selection methods adopted in previous FDP studies fall in one of two approaches: expert recommendation and statistical methods. The advantage of relying on expert recommendation for knowledge and experience is its ability to cope with the complex and unstructured nature of the business problems. As the rapid changes in business environment and government regulations, more aspects need to be considered and the size of the feature set quickly grows beyond the limit of human comprehension. Recently scholars have applied statistical methods, such as T-test (Gudmund & Helmut, 1987), discriminant analysis (Chen & Hsiao, 2008; Wiliam, 1980), Stepwise Selection (Jo & Han, 1996; Kleinbaum, Klein,

* Corresponding author. Address: No. 300, Jhongda Road, Jhongli City, Taoyuan County 320, Taiwan. Tel.: +886 3 422 7151.
E-mail address: drliang@csie.ncu.edu.tw (D. Liang).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.09.047

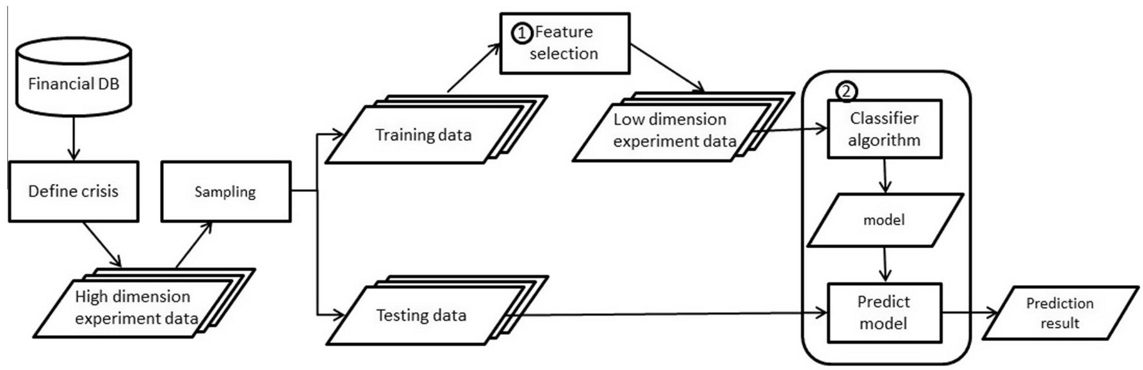

Fig. 1. Two major factors influencing financial distressed prediction.

Table 1
Classifiers used in financial prediction studies.

Classifiers	Paper studied
MDA	Altman (1968), Beaver (1966) and Chuvakhin and Gertmenian (2003)
Logit Regression	Ohlson (1980), Hua et al. (2007), Tam and Kiang (1992) ad Zmijewski (1984)
Neural Network	Tam and Kiang (1992), Lee, Han, and Kwon (1996) , Huang et al. (2004) and Ozkan-Gunay and Ozkan, 2007
Decision Tree	Tam and Kiang (1992)
Support Vector Machine (SVM)	Chandra et al. (2009), Ding et al. (2008), Hua et al. (2007), Shin et al. (2005) and Wu et al. (2007)
Case-Based Reasoning	Jo and Han (1996), Li and Sun (2008), Li et al. (2009) and Li et al. (2009)

& Pryor, 2002; Sun & Li, 2009), or combination of these methods (Cho et al., 2009; Ding et al., 2008), in order to select the most "relevant" features from a larger feature set. These statistic methods are categorized as the *filter* approach by the machine learning theory (Blum & Langley, 1997; Kohavi & John, 1997). The advantage of using the filter approach is its computational and statistical scalability (Guyon & Elisseeff, 2003). A common disadvantage is that this approach ignores the interaction with the classification algorithm used to build the predictor, and thereby, may lead to less accurate prediction performance (Saeys, Inza, & Larrañaga, 2007).

Recently, various approaches, such as the wrapper approach and the embedded approaches, have been proposed in an attempt to find more effective subsets of features (Blum & Langley, 1997; Kohavi & John, 1997). In contrast to the filter approach, these methods assess the subsets of features according to their usefulness to a given predictor. Consequently, these methods usually involve a searching process for a good feature subset, and usually require massive amount of computation (Guyon & Elisseeff, 2003; Liu & Yu, 2005). Therefore, they are applicable to problems with a smaller feature size because the search space becomes intractable as the size increases.

Kim, et al. suggested that the business problems are unstructured in nature; therefore combining machine-learning driven predictors with human-driven predictors may be a better approach (Kim, Min, & Han, 2006). In light of their observation, we propose an integrated approach to feature selection for the FDP problem that embeds expert knowledge with the wrapper method. We categorize the financial features into seven classes according to their financial semantics based on experts' domain knowledge from literature. We then apply the wrapper method to search for "good" feature subsets consisting of top candidates from each feature class. The search space of the wrapper method therefore effectively shrinks because features have been pre-classified before the wrapper method is applied. We conducted the case studies of listed companies in Taiwan's Stock Exchanges. This study shows that the proposed integrated method significantly outperforms the existing feature selection approaches commonly found in literature.

2. Related works

Business crisis prediction is a challenging problem that has stimulated numerous studies over the past few decades. Early studies tend to treat financial ratios measuring profitability, liquidity and solvency as significant indicators for the detection of financial difficulties. However, reliance on these financial ratios can be problematic. The order of their importance, for example, remains unclear as different studies suggest different ratios as the major indicators of potential financial problems.

2.1. Financial crises and financial features

Despite the numerous definitions of business crises, the general meaning should include some narrower definitions like bankruptcy and shut-down and some broader definitions like failure, decline and distress. According to Beaver (1966), a business crisis occurs when a firm announces its bankruptcy, bond default, over-drawn bank account or nonpayment of preferred stock dividends. As financial factors are mostly backward-looking, point-in-time measures, prediction models examining only financial features are inherently constrained. This paper accordingly would like to further explore the role of non-financial features in corporate business crisis prediction.

The pioneering study of Beaver (1966) introduces a univariate approach of discriminant analysis to predict financial distress. The method was later expanded into a multivariate framework by Altman (1968). Discriminant analysis had been the primary method of business failure prediction until 1980s during which the use of logistic regression method was emphasized. Standard discriminant analysis procedures assume that the variables used to characterize the members of the groups under investigation are in multivariate normal distribution. However, in real life, deviations from the normality assumptions are more likely to take place, and this violation may result in biased results. A non-linear logistic function is preferred over multivariate discriminant analysis (MDA), and there are researchers (Altman, 1968; Gunther & GrUning, 2000; Huang, Chen, Hsu, Chen, & Wu, 2004) claiming that even when all the assumptions of MDA hold, a Logit model is virtually as efficient as a linear classifier. Considerable discrepancy is observed in the prediction accuracy reached by the three methods since using different methods leads to different prediction models that adopt different financial ratios.

Major financial features selected for financial distress prediction include financial leverage, long-term and short-term capital intensiveness, return on investment, EPS and debt coverage stability, etc. Selection of these features, however, is seldom based on a theory capable of explaining why and how certain financial factors are linked to corporate bankruptcy (Gunther & GrUning, 2000; Huang et al., 2004). However, the selected features could have huge impact on financial prediction. The financial features considered in this paper are summarized in Table 2. The features are selected because they are frequently used in previous studies dealing with bankruptcy prediction and/or business failure prediction, as well as because of their availability in the dataset.

2.2. Categories of financial ratios

Financial ratio analysis groups the ratios into categories which tell us about different facets of a company's finances and operations. An overview of some of the categories of ratios is given below.

- Leverage ratios which show the extent that debt is used in a company's capital structure.
- Liquidity ratios which give a picture of a company's short term financial situation or solvency.
- Operational ratios which use turnover measures to show how efficient a company is in its operations and use of assets.
- Profitability ratios which use margin analysis and show the return on sales and capital employed.
- Solvency ratios which give a picture of a company's ability to generate cash flow and pay it financial obligations.

A closer examination of the above 21 constructed features reveals some interesting patterns of how domain knowledge is represented through different combinations of raw accounting variables. For example, among the 21 constructed features, 8 of them are constructed by dividing raw accounting variables by total assists, 2 of them are constructed by dividing the variables of loan-specific assets by gross loans and 4 of them are constructed by dividing total sales. These constructed features in some sense reflect preliminary domain knowledge of normalization. The goal of normalization is to eliminate the effects of some irrelevant factors in describing a company's financial condition (Zhao, Sinha, & Ge, 2009). While the 21 financial features are all in relatively simple forms, they constitute important domain knowledge, which is not explicitly captured in, and cannot be automatically learned from, the raw accounting data. Without some knowledge of the financial domain, even a data mining specialist would not know how to combine different raw accounting variables in meaningful ways to construct such intermediate concepts.

3. Business crisis prediction model: the background

Substantial literature can be found on business crisis prediction. We briefly review methods used in this research, i.e., the Iterative Relief and Support Vector Machine (SVM).

3.1. Feature selection

Feature selection is one of the two important factors contributing to the performance of a prediction model for any classification problem. The objectives of feature selection are three-fold: (a) better performance, (b) faster and more cost-effective models, and (c). deeper insight into the underlying processes. (Guyon & Elisseeff, 2003)

The feature selection methods adopted in previous studies of the FDP includes T-test (Gudmund & Helmut, 1987), discriminant

Table 2
Selected financial ratios listed by categories.

Variable	Meaning	Variable	Meaning
Liquidity ratios		*Operational ratios*	
X_1	Current ratio	X_{23}	Payable turnover ratio
X_2	Acid test	X_{24}	Total assets turnover
X_3	Quick assets/total assets	X_{25}	Receivable turnover ratio
X_4	Current assets/total assets	X_{26}	Fixed assets turnover
X_5	Working capital/total assets		
X_6	Working capital/sales	*Profitability ratios*	
X_7	No-credit interval	X_{27}	Operating income after tax/equity
		X_{28}	Operating income after tax per share
Solvency ratios		X_{29}	1 if net income was negative for the last two years, otherwise, 0
X_8	Interest expenses/equity	X_{30}	Pre-tax income per share
X_9	Market value equity/book value of total debt	X_{31}	Retained earnings/total assets
X_{10}	Cost of interest-bearing debt	X_{32}	Operating income before tax/total assets
X_{11}	Interest expense/revenue	X_{33}	Operating income after tax/total assets
		X_{34}	Operation income per employee
Growth ratios		X_{35}	Gross profit/net sales
X_{12}	Total equity growth	X_{36}	Realized gross profit/net sales
X_{13}	Total assets growth	X_{37}	Sales per employee
X_{14}	Ordinary income growth	X_{38}	Net income/total assets
X_{15}	Return on total asset growth	X_{39}	Net income/equity
X_{16}	Net income growth		
X_{17}	Sales growth	*Capital structure ratios*	
		X_{40}	Equity/total assets
Cash-flow ratios		X_{41}	Fixed assets per employee
X_{18}	Cash flow/total assets	X_{42}	Liabilities/total assets
X_{19}	Cash flow/total liabilities	X_{43}	One if total liabilities exceeds total assets, zero otherwise
X_{20}	Cash flow/equity		
X_{21}	Cash re-investment ratio	*Other ratios*	
X_{22}	Funds provided by operations/total liabilities	X_{44}	Size

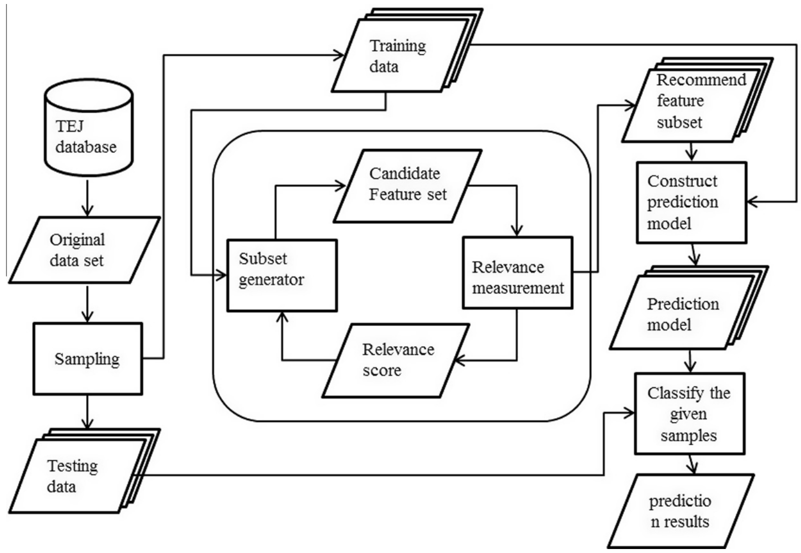

Fig. 2. The filter approach to feature selection.

analysis (Chen & Hsiao, 2008; William, 1980), Stepwise Selection (Jo & Han, 1996; Kleinbaum et al., 2002; Sun & Li, 2009), or combination of these methods (Cho et al., 2009; Ding et al., 2008). These methods are categorized as the *filter* approach in the context of machine learning taxonomy (Blum & Langley, 1997; Kohavi & John, 1997). These statistic methods are categorized as the *filter* approach by the machine learning theory (Blum & Langley, 1997; Kohavi & John, 1997). As shown in Fig. 2, the filter approach selects the most "relevant" features from a larger feature set based on some relevance indices; and this selection process is independent from the classifier used to build the prediction model.

The methods based on the filter approach are usually computationally efficient and statistically scalable when the feature set under consideration is large in size (Blum & Langley, 1997; Guyon & Elisseeff, 2003). A common disadvantage of this approach is that the performance of the prediction model may be inferior (Saeys et al., 2007). One of the reasons is that the filter approach ignores the interaction with the classification algorithm used to build the predictor. Recent studies have demonstrated that relevance does not imply optimality, and vice versa (Kohavi & John, 1997). Another reason is that the filter methods mentioned above do not model the feature dependency. Guyon and Elisseeff pointed out that a feature that is completely irrelevant by itself can provide a significant performance improvement when taken with others (Guyon & Elisseeff, 2003). Furthermore, perfectly correlated variables are truly redundant in the sense that no additional information is gained by adding them; however, two features with high correlation do not mean they are not complimentary (Guyon & Elisseeff, 2003).

In contrast to the filter approach, wrappers are proposed as an alternative approach to the feature selection problem (Blum & Langley, 1997; Kohavi & John, 1997). A wrapper assesses the subsets of features according to their usefulness to a given classifier (shown in Fig. 3).This approach also takes the feature dependency issue into consideration. Consequently, the methods of this approach usually involve a searching process for a good feature subset, and therefore require massive amount of computation (Guyon & Elisseeff, 2003; Liu & Yu, 2005). The search space quickly becomes intractable when the size of the feature set increases if exhausted search is applied. Many wrapper methods adopts heuristic search strategy in order to make it feasible; examples are

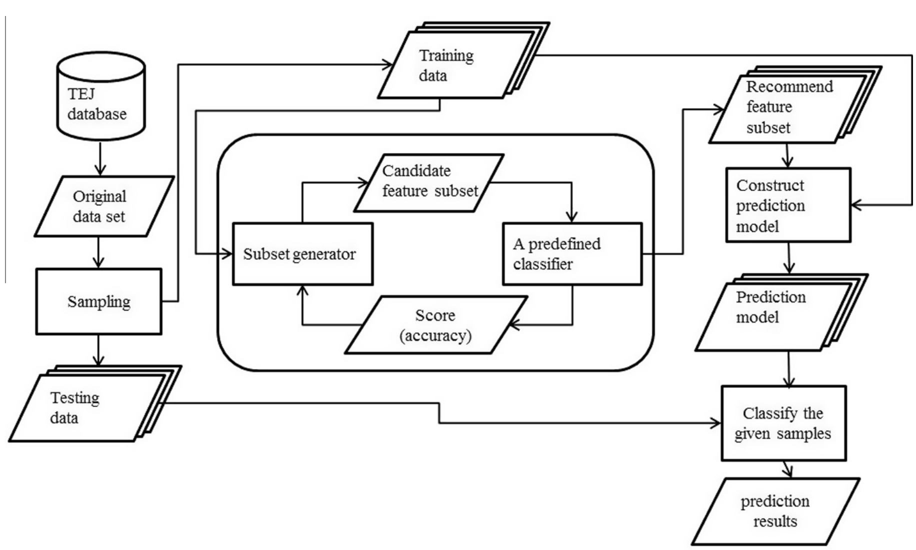

Fig. 3. The wrapper approach to feature selection.

Table 3
Characteristics summary of the feature selection approaches.

	Filter	Wrapper
Advantages	1. Fast 2. Scalable 3. Independentof classifier	1. Interacts with classifier 2. Models feature dependencies
Disadvantages	1. Ignores feature dependencies 2. Ignores interaction with the classifier	1. Computationally intensive 2. Classifier dependent selection 3. Risk of overfitting
Examples	ANOVA (Gudmund & Helmut, 1987) Stepwise logit regression (Kleinbaum et al., 2002) Discriminant analysis (William, 1980) Iterative RELIEF (Sun, 2007)	Sequential forward selection (Kittler, 1978) Sequential backward selection (Kittler, 1978) Randomized hill climbing (Kohavi & John, 1997) Genetic algorithms (Goldberg, 1989) Recursive feature elimination (Guyon, Weston, Barnhill, & Vapnik, 2002)

sequential forward selection (Kittler, 1978), sequential backward selection (Kittler, 1978), randomized hill climbing (Kohavi & John, 1997), and genetic algorithms (or GA) (Goldberg, 1989).

Sexton et al. found that the performance of the prediction model using GA-based feature selection algorithm is better than the model using CATLRN after they studied 137 bankrupted US banks (Sexton, Sriram & Etheridge, 2003). Jeong et al., on the other hand, reported that there were no significant performance difference between models using GA and those using other feature selection methods in a study involves 1271 bankrupted Korean firms (Jeong, Min & Kim, 2012). Both studies were based on privately collected datasets. Huang and Liu studied the TEJ dataset, a public dataset that contains companies listed in the stock exchanges of Taiwan and China (Huang & Liu, 2006). Their study indicates that GA is a better algorithm for feature selection than gain-ratio. However, they remind reader that this result is inconclusive due to an ill experiment design that "may result in a favorable influence on the results observed".

Another well-known problem of the wrapper approach is that they are prone to overfitting problem. A general framework of a wrapper method consists of start point selection, search strategy, evaluation function, and halting criteria (Blum & Langley, 1997). Table 3 summaries the characteristics of both filter and wrapper approaches.

3.2. Genetic algorithm

Genetic algorithms (GAs) have been widely used to solve various optimization problems (Goldberg, 1989; Grefenstette, 1986; Holland, 1992). Mimicking the evolutionary processes in nature, a GA algorithm typically starts from a set of solutions, either randomly created or manually selected. This set of solutions is referred to as population, and every individual in the population is referred to as a chromosome. Within every generation, a fitness function should be used to evaluate the quality of every chromosome to determine the probability of it surviving to the next generation; usually, the chromosomes with larger fitness values have a higher survival probability. In order to form a new group of population, the population reproduction is done by using operations like selection, crossover and mutation on the current population. This reproduction goes through one generation to another, until it converges on the individual generation with the most fitness values for goal functions or the required number of generations was reached. The optimal solution is then determined.

Several issues shall be considered when it comes to the design of an effective GA; this includes the population size, genetic operators (selection, crossover and mutation), and the stopping criteria. The size of the population has impacts on both the performance as well as the efficiency of the GA, which is usually set from 30 to 200 (Srinivas & Patnaik, 1994). The reproduction of the current population to the next generation starts with the chromosome selection. Roulette wheel method and tournament method are two standard methods to select those chromosomes that can survive to the next generation from the current population (Bäck, 1996). All chromosomes that survive to the next generation are placed in a matting pool for crossover and mutation.

The chromosomes are randomly selected in pairs from the matting pool for crossover. This probability is referred to as the crossover rate, which typically ranges from 0.5 to 1.0. Commonly used crossover methods are single-point, two-point and uniform crossover (Srinivas & Patnaik, 1994). The newly crossed chromosomes are then combined with the rest of the chromosomes to generate a new population. Following the crossover, the mutation operator produces small changes to the bit string by choosing a single bit at random, then changing its value. The probability that a chromosome is mutated, or the mutation rate, ranges typically from 0.001 to 0.05. Commonly used mutation methods are point mutation, polynomial mutation and uniform mutation. The condition with which the evolution process stops is called the stopping criteria. Commonly applied criteria can be either the convergence to a good solution or a preset number of the evolution rounds.

3.3. SVM model

As a relatively new algorithm in machine learning, Support Vector Machine (SVM) was first developed by Boster, Guyon, and Vapnik, (1992) to provide better solutions than other traditional classifiers such as neural networks. SVM belongs to the type of maximal margin classifier, in which the classification problem can be represented as an optimization process.

The basic procedure for applying SVM to a classification model can be summarized as follows (Chen & Hsiao, 2008). First, the input vector is mapped into a feature space, which is possible with a higher dimension. Then, within the feature space, the approach proceeds to seek an optimized division, i.e., to construct a hyper-plane that separates two (or more) classes. The hyper-plane determined by a SVM is composed of a set of support vectors that are a subset of training data used to define the boundary between two classes. To find the optimal hyper-plane, the boundary margin between the two classes should be maximized.

As suggested by Vapnik (1999), SVM can be generalized well even in high-dimensional spaces under small training sample conditions, indicating a learning ability independent of the feature space dimensionality. SVM seeks to minimize an upper bound of the generalization error rather than minimizing the training error. Using the structural risk minimization rule, the training of SVMs always seeks a globally optimized solution and avoids over-fitting. These characteristics make SVM a strong candidate to tackle the challenging FDP problem. The SVM approach has been put into several financial applications recently (Ding et al., 2008; Shin et al., 2005; Wu et al., 2007).

4. HARC: the proposed wrapper algorithm based on the genetic algorithm

Fig. 4 depicts the basic flows of the proposed HARC algorithm, which consists of three major steps. In Step 1, ratios are clustered by ratio categories as discussed in Section 2. This step reflects expert knowledge in financial domain over the years. Ratios classified

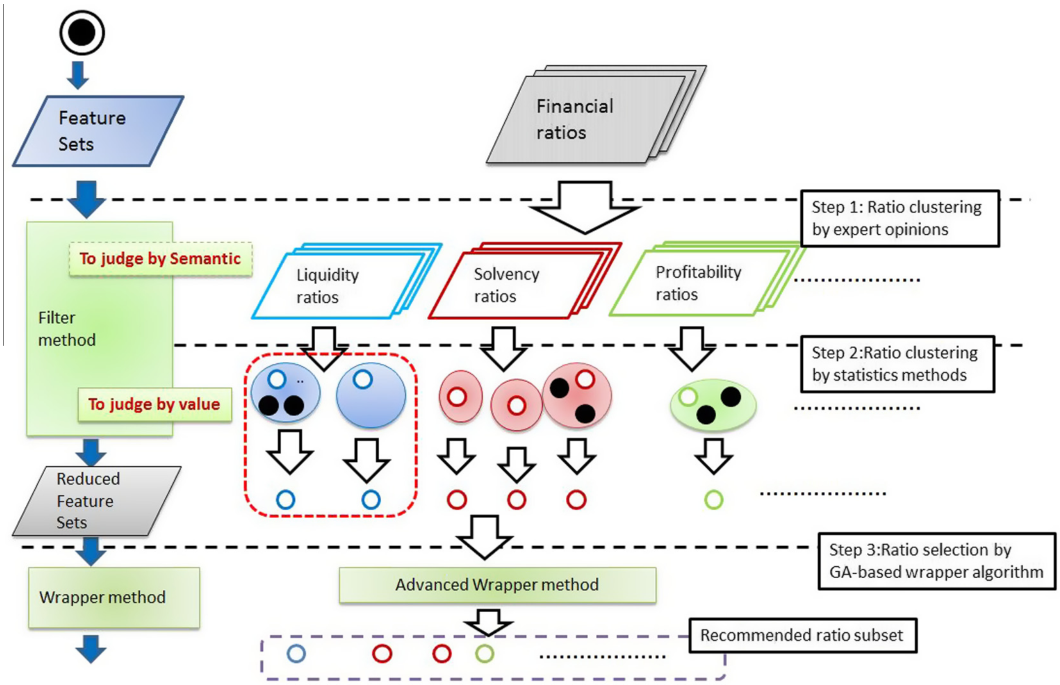

Fig. 4. The concepts of the proposed feature selection algorithm: HARC.

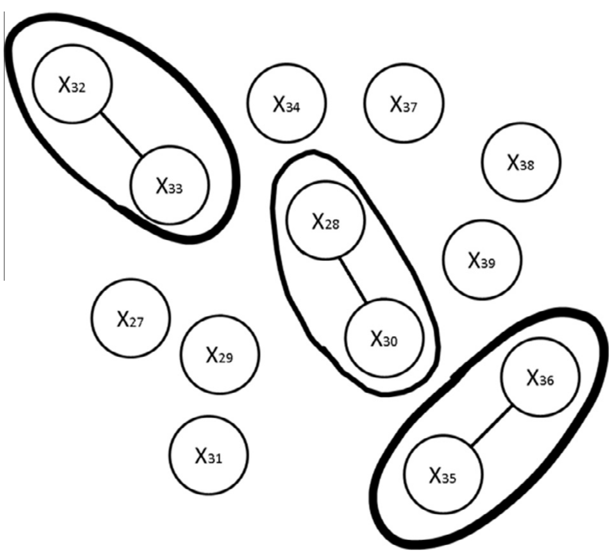

Fig. 5. The selection of representative ratios from a ratio category: an example.

into the same category usually share common semantics in certain aspect; for example, the *current ratio* and the *acid test* both reflect the *liquidity* of a company.

In Step 2, ratios in the same category are pair-wisely examined according to their correlation. Any two ratios with high correlation are grouped into the same sub-category. Previous studies showed that two features with high correlation in general do not mean they are not complimentary (Guyon & Elisseeff, 2003). For two ratios in the same sub-category, however, it is likely that no information is gained by adding both of them into the feature set for two reasons; first, they belong to the same ratio category thus they are semantically correlated, and secondly, the high level of correlation indicates that they are statistically correlated. Therefore, one ratio from each sub-category is designated as the *representative ratio* for the rest, and is added to the representative ratio set.

Fig. 5 demonstrates how the representative ratios are selected using the *Profitability* ratio category as an example. Given a ratio category, we first examine the correlation for each pair of ratios in this category using t-test (Box, 1987). As shown in Fig. 5, a vertex in the graph represents a ratio, and a solid arc connecting two ratios if their t-test score exceeds the preset threshold.[1] In this example, an arc connecting the ratios [X_{28}] (*Operating Income after Tax per Share*) and [X_{30}] (*Pre-Tax Income per Share*) implies that these two ratios are highly correlated in statistical sense. Next, we find connected components of the graph (or the sub-categories of the ratio category) using an efficient algorithm from graph theory (Hopcroft & Tarjan, 1973). In this example, ratios [X_{28}] and [X_{30}] form a connected component; likewise, ratios [X_{32}] and [X_{33}] forms another component. In the third step, we select a ratio from each connected component to be the *representative ratio* of this sub-category. In this paper, we pick the ratio with the highest connectivity and/or with the highest prediction score to be the representative ratio of this sub-category. For example, ratios [X_{28}] and [X_{32}] are each selected as the representative of their own group. Other measures for the selection of the representative ratio are possible. We discuss this further in Section 6. Table 4 presents the pseudo codes of this algorithm, named *Identify_Representative_Ratios*.

In the final step, a GA-based wrapper algorithm, GA-Wrapper, is proposed to select the optimal ratio subset from the representative ratio set. We present the design details of this algorithm since these issues have profound impacts on the performance of this GA-based algorithm as discussed in the previous section. The pseudo codes of GA_Wrapper and HARC is shown in Table 5 and Table 6, respectively.

4.1. Chromosome coding and initial population

The traditional binary coding scheme is used in this algorithm. The chromosome X is represented as $X = \{x_1, x_2, \ldots, x_n\}$ where $x_i = 1$ if the ratio r_i is selected and $x_i = 0$ otherwise. The initial population

[1] In this paper, the Z score threshold is set to 1.04, which marks the correlation level (P-value) of 30%.

Table 4
The pseudo codes of the algorithm: Identify_Representative_Ratios.

Identify_Representative_Ratios (S,τ)
Input: S the input ratio set
τ the t-test threshold
Output: $S_{rep} = \emptyset$ the set of the representative ratios from S
Initialize the graph $G = (S,E)$ where the vertex set of G is the ratio set S and let the edge set $E = \emptyset$;
//Construct the graph G
for each pair (r_i,r_j) where $r_i, r_j \in S$ do
Compute the t-test score t_{ij} of (r_i,r_j);
if (t_{ij} is greater than τ) then $E = E \bigcup \{(r_i,r_j)\}$;
endif
endfor
Find the connected components $\{C_1, C_2, \ldots, C_N\}$ in G;
//Find the representative ratio for each connected component in this category.
for $i = 1$ to N do
Find the ratio r_{ij} in C_i with the highest connectivity;
$S_{rep} = S_{rep} \bigcup \{r_{ij}\}$;
endfor
return S_{rep};

Table 5
The pseudo code of genetic algorithm based wrapper for feature selection.

GA_Wrapper
Input: S Input ratio set δ number δ of iterations
γ rate γ of mutation β rate β of crossover
α size α of population
Output: S^* best solution
//Initialization
Generate α feasible solutions randomly;
Save them in the population *Pop*;
//Loop until the terminated condition
for $i = 1$ to δ do
//Chromosome selection
for $j = 1$ to α do
Fitness Value = average Accuracy X − Standard Deviation X;
endfor
Delete the bad solutions in *Pop*;
Copy the good solutions in *Pop* to replace the bad solutions;
Save them in the population *Pop*;
//Crossover
number of crossover $n = \alpha/2$;
for $j = 1$ to n do
randomly select two solutions X and X from *Pop*;
if random < β
Generate X and X by one-point crossover to X and X;
Save X and X to *Pop*;
endif
endfor
//Mutation
for $j = 1$ to n do
Select a solution X from *Pop*2;
Mutate each bit of X under the rate γ and generate a new solution X';
Update X with X' in *Pop*;
endfor
//Updating
endfor
//Returning the best solution
return the recommended ratio subset S^* based on the best subset X in *Pop*;

is generated randomly, and the size is set to 60 so that the convergence time and the population divergence are kept balanced.

4.2. The fitness function

The average accuracy is used here in as the fitness value of the chromosomes under consideration. The hold-out method is a typical cross-validation method to obtain the accuracy of the training. After a few experimental tests, we set the hold-out number to 100, the minimum number with which we can obtain a result with a satisfactory confidence level.

Table 6
The pseudo codes of the HARC algorithm.

HARC Algorithm
Input: α size α of population β rate β of crossover
γ rate γ of mutation δ number δ of iterations
Output: τ rate of *t-test* threshold
$D(R_1, R_2, \ldots, R_N)$ a training data set with N financial ratios
$S^* = \emptyset$ an optimal feature subset selected from D
//Step 1:Partition the input ratio set S into K ratio categories based on domain knowledge
$S = \bigcup_{i=1}^{k} S_i$;
$S_{rep} = \emptyset$;
//Step 2: Find the set of representative ratios, S_{rep}, from S
for $i = 1$ to K do
S_i' = Call Identify_Representative_Ratios(S_i, τ);
$S_{rep} = S_{rep} \bigcup S_i'$;
endfor;
//Step 3: Search the optimal ratio subset, S^*, from S_{rep}
Call S^* = GA_Wrapper(S_{rep}, α, β, γ, δ);
return S^*;

4.3. Genetic operations

A standard tournament method is used to retrieve 30 pairs of chromosomes, i.e., half the size of the current population. A uniform crossover is applied to create new chromosomes with the possibility of 0.7. The mutation operation follows the crossover operation, and uniform mutation is used in this study. Every chromosome in the population should come with the possibility of 0.005 for mutation.

4.4. Stopping criteria

Within the evolution process, every round is arranged with 120 generations. The algorithm halts when it either reaches convergence or all generations have completed.

5. Experiment framework and design

In order to validate the proposed ratio selection algorithm HARC, we designed an experiment, which we will now discuss. As depicted in Fig. 6, the experiment process is presented in three parts: the data sampling, the financial ratios selected by HARC, and, finally, the training and testing of underlying classifier used to build the prediction model. The following is the detailed discussion.

5.1. Sample variables

In this section, we present the experiment framework and design of our proposed model. A publicly listed firm is regarded to encounter business crisis and turns into a distressed company when declared for any one of the following conditions: full-value delivery, stock transaction suspension, re-construction, bankruptcy or withdrawal from the stock market. Based on the above criteria, we selected 240 distressed and 240 non-distressed (as matched samples) companies from TEJ database range year 2000 to 2008. The matched samples (non-distressed) are selected via the stratified random sampling (Altman, 1968).

For the variables, in addition, TEJ financial database for general industry is divided into twelve categories: 1. Balance sheet (60 + financial accounts such as total asset, total debt, etc.); 2. Income Statement (40 + financial accounts such as operating costs,

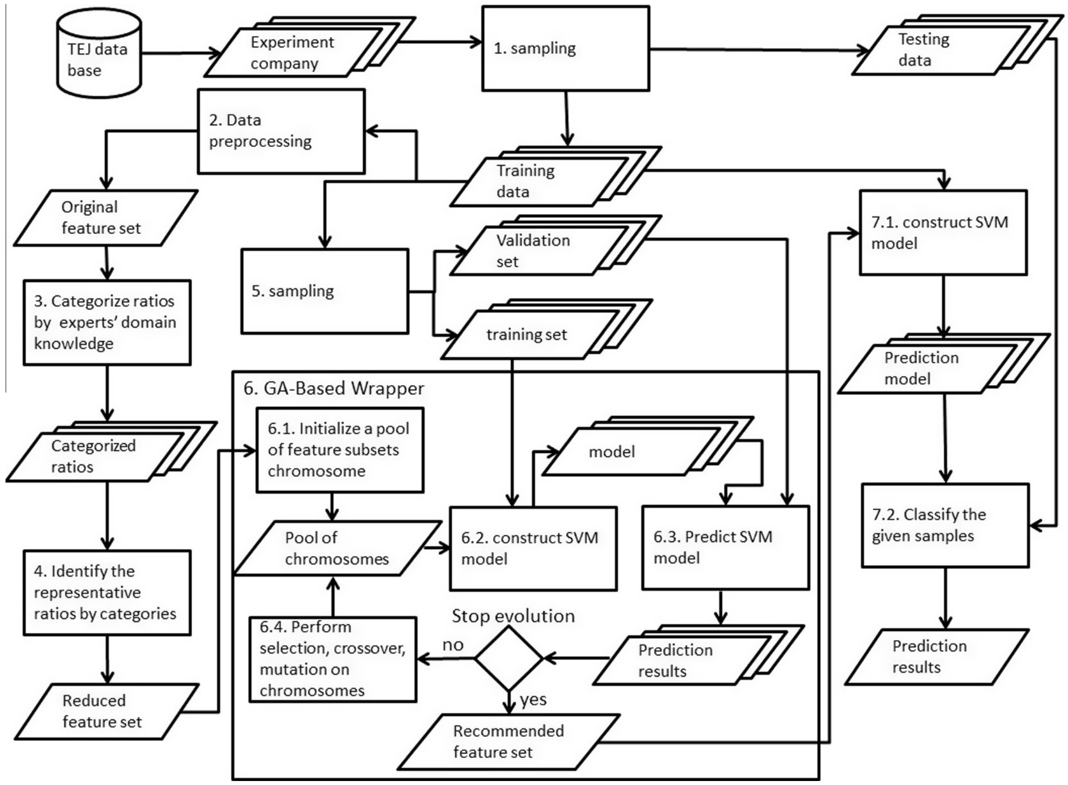

Fig. 6. The framework of the FDP prediction system based on the HARC algorithm.

interest expense, etc.); 3. Earning distributions; 4. Cash flow statement (50 + financial accounts such as depreciation, etc.); 5. Related Party Sales; 6. Notes and supplementary; 7. Operating costs; 8. Manufacturing expenses; 9. Operating expenses; 10. Retirement pay; 11. Warrant and employee cost; and 12. Financial ratios. The total number of features in 12 categories is more than 500 items, some features are from financial statements, some features are calculate from well-known financial ratios, and some features are defined by experts in data collecting, economy analysis, and computer science. We choose Financial Ratios category as our experiment feature set. We used 44 ratios (see Table 2) as experiment variables.

We collect data which was one to three years before the year when the financial distresses took place in order to analyze the prediction accuracy multiple years ahead (Ding et al., 2008). Some of the financial data are missing for a given company in the TEJ database. This makes our matched-pair companies less than we actually planned for. We collected 1260 instances of companies; only valid instances are used for experiments as show in Table 7.

5.2. Selection of the SVM kernel and parameters

As discussed in Section 3.3, SVM-based prediction models perform well for problems with high-dimensional spaces under small training sample condition. In this paper, we use LIBSVM 2.9 (Chang & Lin, 2011), to build our SVM-based verification model. The selection of kernel and the corresponding parameters plays a crucial role in the prediction quality of the SVM-based models. However, there is no general guideline for this selection process. In general, the radial basis function (RBF) is suggested for SVM. Here, we use RBF as the SVM kernel function that can determine an optimal hyper-plane for classifying two classes of data. Using RBF as a SVM kernel function, users have to set parameters C and gamma for optimizing a verification (prediction) model. To find optimal parameters, we use particle swarm optimization (PSO) (Kennedy & Eberhart, 1995; Kennedy, Kennedy, & Eberhart, 2001) since it can simply and rapidly find an optimal C and gamma in a continuous scope.

Table 7
The instance number used in 1 to 3 years-ahead forecasts.

	Number of distressed firms	Number of non-distressed firms
1-Year-ahead	231	231
2-Year-ahead	229	229
3-Year-ahead	228	228

5.3. Ratios selected by HARC

Seventeen ratios are selected by HARC from 44 ratios listed in Table 2. As shown in Table 8, the selected ratios are somewhat uniformly distributed over the seven categories within which Profitability and Growth categories have more ratios been selected than others. We analyze the profile of these ratios with their means, standard deviations, and differences. The analysis is further tested via *P*-value analysis of their statistical significance. The results are shown in Table 8.

From Table 8, it is clear that the selected ratios have significant differences between the distressed firms and non-distressed firms in terms of the mean values and standard deviations. The lower the standard deviation, the higher stability is in ratios (or there will be less fluctuation). Taking the profitability ratio [X_{32}] (operating income before tax/total assets) as an example, the non-distressed firms' mean value is 6.91, which is significantly different from the mean value (−6.92) of the distressed firms. The difference of selected ratios between the distressed firms and non-distressed firms are all significant at the level of 5% except two ratios; i.e.,

Table 8
Profile analysis of the feature selected by the proposed algorithm HARC.

Features		Firm type				Difference		T-test p-value ($p < 0.05$)
		Distressed firms		Non-distressed firms				
Variable	Meaning	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.	
Liquidity ratios								
X_1	Current ratio	1.32×10^2	1.73×10^2	2.33×10^2	2.19×10^2	1.01×10^2	4.63×10^1	$5.9 \times 10^{-8*}$
X_2	Acid test	7.69×10^1	1.65×10^2	1.57×10^2	1.78×10^2	8.04×10^1	1.27×10^1	$7.2 \times 10^{-7*}$
X_5	Working capital/total assets	2.25×10^{-2}	2.12×10^{-1}	2.24×10^{-1}	2.2×10^{-1}	2.02×10^{-1}	7.19×10^{-3}	$1.4 \times 10^{-21*}$
X_6	Working capital/sales	2.66×10^{-1}	1.91	3.95×10^{-1}	7.02×10^{-1}	1.29×10^{-1}	−1.21	$3.4 \times 10^{-1*}$
X_7	No-credit interval	−5.84	5.54×10^1	-1.94×10^{-1}	1.33×10^1	5.65	-4.21×10^1	$1.3 \times 10^{-1*}$
Solvency ratios								
X_8	Interest expenses/equity	2.1×10^2	5.76×10^2	4.69×10^1	6.25×10^1	-1.63×10^2	-5.14×10^2	$2.9 \times 10^{-5*}$
X_9	Market value equity/book value of total debt	1.14	3.81	2.29	2.26	1.15	−1.56	$9.4 \times 10^{-5*}$
X_{11}	Interest expense/revenue	5.24	4.2×10^2	−9.5	3.39×10^2	-1.47×10^1	-8.09×10^1	6.78×10^{-1}
Growth ratios								
X_{13}	Total assets growth	−7.6	2.32×10^1	1.14×10^1	3.12×10^1	1.9×10^1	8	$6 \times 10^{-13*}$
Cash flow ratios								
X_{18}	Cash flow/total assets	-1.8×10^{-2}	8.47×10^{-2}	1.02×10^{-2}	7.77×10^{-2}	2.8×10^{-2}	-7×10^{-3}	$2.4 \times 10^{-4*}$
X_{19}	Cash flow/total liabilities	-8.8×10^{-2}	7.28×10^{-1}	3.81×10^{-2}	3.73×10^{-1}	1.26×10^{-1}	-3.6×10^{-1}	$1.9 \times 10^{-2*}$
X_{20}	Cash flow/equity	-9.4×10^{-2}	4.89×10^{-1}	1.6×10^{-2}	1.29×10^{-1}	1.1×10^{-1}	-3.6×10^{-1}	$1.1 \times 10^{-3*}$
X_{21}	Cash re-investment ratio	−5.04	3.57×10^1	3.65	1.68×10^1	8.69	-1.89×10^1	$9.1 \times 10^{-4*}$
Operational ratios								
X_{24}	Total assets turnover	6.93×10^{-1}	5.75×10^{-1}	8.8×10^{-1}	6.37×10^{-1}	1.87×10^{-1}	6.2×10^{-2}	$1 \times 10^{-3*}$
Profitability ratios								
X_{28}	Operating income after tax per share	−2.3	2.77	1.01	2.97	3.31	2×10^{-1}	$1 \times 10^{-30*}$
X_{31}	Retained earnings/total assets	-2.3×10^{-1}	3.1×10^{-1}	1.35×10^{-2}	2.29×10^{-1}	2.39×10^{-1}	-8.2×10^{-2}	$3 \times 10^{-19*}$
X_{32}	Operating income before tax/total assets	−6.92	1.47×10^1	6.91	1.19×10^1	1.38×10^1	−2.8	$1.5 \times 10^{-25*}$
X_{34}	Operation income per employee	-6.39×10^2	1.82×10^3	9.67×10^2	3.45×10^3	1.61×10^3	1.63×10^3	$1.09 \times 10^{-9*}$
X_{35}	Gross profit/net sales	8.13	1.88×10^1	1.92×10^1	1.82×10^1	1.11×10^1	-5.5×10^{-1}	$2.76 \times 10^{-10*}$
X_{39}	Net Income/equity	−1.1	4.83	1.56×10^{-2}	2.79×10^{-1}	1.11	−4.56	$5.85 \times 10^{-4*}$
Capital structure ratios								
X_{42}	Liabilities/total assets	6.05×10^1	1.66×10^1	3.96×10^1	1.66×10^1	-2.09×10^1	2.64×10^{-4}	$1.8 \times 10^{-35*}$
Other ratios								
X_{44}	Size	3.06×10^1	8.5×10^1	2.75×10^1	6.75×10^1	−3.1	-1.75×10^1	6.64×10^{-1}

* Significant at the level of 5%.

Table 9
Financial Ratios used in the benchmark models.

Model	Reference	Ratios used
M_1	Altman (1968)	$[X_5][X_9][X_{24}][X_{31}][X_{32}]$
M_2	Beaver (1966)	$[X_1][X_5][X_7][X_{19}][X_{38}][X_{42}]$
M_3	Ohlson (1980)	$[X_1][X_5][X_{16}][X_{22}][X_{29}][X_{38}][X_{42}][X_{43}][X_{44}]$
M_4	Stepwise logistic regression (SLR)	$[X_8][X_{32}][X_{42}]$
M_5	Stepwise discriminant analysis (SDA)	$[X_8][X_{15}][X_{16}][X_{19}][X_{23}][X_{32}][X_{33}][X_{36}][X_{38}][X_{39}][X_{42}]$

Table 10
Statistics of prediction accuracies of different models.

	M_1 (Altman) (%)	M_2 (Beaver) (%)	M_3 (Ohlson) (%)	M_4 (SLR) (%)	M_5 (SDA) (%)	M_6 (HARC) (%)
Minimum	65.22	69.57	67.39	69.57	69.57	75.00
Maximum	83.70	88.04	86.96	88.04	93.48	92.39
Mean	75.34	79.68	79.37	79.96	80.04	81.36
Median	75.00	79.35	79.35	80.43	80.43	81.52
S.D	3.89	3.43	3.85	4.12	4.09	3.23

$[X_{11}]$ (interest expense/revenue) and $[X_{44}]$ (Size). It is interesting to notice that these two ratios are selected by HARC but not by other traditional feature selection approaches such as stepwise logistic regression and discriminant analysis.

Table 11
Statistics of the Type I error rate of different models.

	M_1 (%)	M_2 (%)	M_3 (%)	M_4 (%)	M_5 (%)	M_6 (%)
Minimum	10.87	8.70	6.52	6.52	4.35	6.52
Maximum	43.48	32.61	36.96	39.13	36.96	34.78
Mean	25.86	20.14	22.01	19.61	21.52	19.26
Median	26.09	19.57	21.74	19.57	21.74	19.57
S.D	6.57	4.54	6.09	5.52	5.92	5.59

Table 12
Performance summary of various models.

	M_1 (%)	M_2 (%)	M_3 (%)	M_4 (%)	M_5 (%)	M_6 (%)
Type I	25.86	20.14	22.01	19.61	21.52	19.26
Type II	23.41	20.48	19.13	20.48	18.39	16.63
Brier Score (BS)	24.64	20.32	20.63	20.04	19.96	18.64
Accuracy	75.34	79.68	79.37	79.96	80.04	81.36

6. Experiment results and discussion

6.1. Performance of HARC

To verify the effectiveness of the proposed method, models based on prior scholars as well as known feature selection algorithms are used as benchmarks for comparison. Model 1 (M_1) to Model 3 (M_3) is based on the ratio set proposed by Altman (1968), Beaver (1966) and Ohlson (1980), respectively; whereas

Table 13
McNemar value (P-value) for pairwise comparison of the average accuracy, Type I and Type II errors.

M_6	M_1	M_2	M_3	M_4	M_5
Accuracy	$5.47 \times 10^{-24**}$	$1.45 \times 10^{-3**}$	$1.06 \times 10^{-4**}$	$8.04 \times 10^{-3**}$	$1.25 \times 10^{-2*}$
Type I	$2.1 \times 10^{-12**}$	2.53×10^{-1}	$7.65 \times 10^{-4**}$	6.58×10^{-1}	$5.99 \times 10^{-3**}$
Type II	$3.06 \times 10^{-11**}$	$1.78 \times 10^{-3**}$	1.34×10^{-1}	$2.39 \times 10^{-3**}$	6.2×10^{-1}

* Significant at the level of 5%.
** Significant at the level of 1%.

Model 4 (M_4) and 5 (M_5) are based on the feature selection results of stepwise logistic regression (SLR) and stepwise discriminant analysis (SDA). Table 9 summarizes the ratios adopted in these models. Model M_6 is based exclusively on the proposed algorithm, HARC. To assess the predictive performance more precisely, we apply the 100 times hold-out method in this study. For each time of a hold-out, 80% of the whole data samples are used for training, and the remaining 20% are used for testing. Secondly, the comparisons of the prediction results between models M_1 to M_6 are made by conducting the same SVM.

Different types of errors result in different penalty costs. As presented earlier, 240 distressed firms in the years of 2000–2008 are analyzed against 240 non-distressed counterparts. We first compare the prediction accuracy of the six models using the financial data one year prior to the distressed year of the firm in distress. This prediction is also known as the 1-year-ahead forecast (Ding et al., 2008).

Tables 10–12 give full statistics of the performance metrics. The model based on HARC (M_6) outperforms all other models in all accounts. The mean accuracies of the six models range from 75.34% to 81.36%. The standard deviations fall in the range of 3.23% to 4.12%. As summarized in Table 10, the average accuracy of the 1-year-ahead forecast of M_6 is 81.36%, superior to those of M_1 (75.34%), M_2 (79.68%), M_3 (79.37%), M_4 (79.96%), and M_5 (80.04%). M_6 also yields lower average Brier Score[2] (BS) of 18.64%, Type I error rate (19.26%) and Type II error rate (16.63%) than other models as shown in Table 12. The superiority of M_6 is further confirmed after we run a series of McNemar[3] to compare M_6 against other 5 models, as shown in Table 13. M_6 yields better performance than other models at the significance level of 5% or better, except the Type I errors of M_4. Regarding the average accuracy, in particular, the performance improvement of M_6 over other models is significant at an impressive level of 1%.

Taking a closer look at Table 13, we find that either M_4 or M_5 performs better than M_1, but comparable to M_2 and M_3. It seems to suggest that, for a given FDP problem, it is wiser to find proper ratio set by using feature selection methods such as SLR than to adopt directly ratio sets reported in the literature. Compared with other models, on the other hand, model M_1 seems to have lowest mean value and highest standard deviation. However, her Type I error occurs with a less frequency than most models. In actual practice, the cost of misclassifying a failed firm into a healthy one (Type I error) is likely to be much greater than that of misclassifying a healthy firm into a failed one (Type II error). As indicated above, the Type I error rate of M_4 is the second best next to the proposed model M_6, and is much lower than that of other models. Empirical results indicate that M_4 can serve as a promising alternative for existing financial distress prediction models.

6.2. Performance comparison of SVM model against models based on various classifiers

For benchmark purpose, we conducted experiments with Logit, MDA and RBFN models as their SVM counterparts. These models are all built with the same feature set (shown in Table 9) used in M_6. Table 14 indicates the statistical description results. The standard deviations of the models are between 3.23% and 4.27%.

The RBFN, MDA and Logit models consistently fall short of their SVM counterpart at the significance level of 5% or above as summarized in Table 14. For example, SVM yields 81.36% accuracy that is the highest accuracy rate among her peers.

Moreover, we conduct McNemar test to assess the significance of the difference between different models. As shown in Table 15, the SVM model is superior to other models at the significant level of 5% or above, yet there is no significant difference between the model Logit, MDA and KNN.

6.3. The analysis of predictive accuracy for longer-term forecast

We conduct additional experiments to observe the effect of the prediction capability of models M_1–M_6 for longer term forecasts. Table 16 shows the results of applying these models from 1-year-ahead forecast to 3-year-ahead prediction. For example, M_6

Table 14
Statistics of predictive accuracies of various classifiers.

Statistical indices	SVM (%)	Logit (%)	MDA (%)	KNN (%)
Minimum	75.00	71.74	72.83	68.48
Maximum	92.39	86.96	86.96	86.96
Mean	81.36	80.36	80.04	79.47
Median	81.52	80.43	80.43	80.43
S.D.	3.23	3.61	3.45	4.27

Table 15
McNemar value (*P*-value) of accuracy comparison for pairwise models.

	SVM	Logit	MDA	KNN
SVM	*X*	$4.05 \times 10^{-2*}$	$5.91 \times 10^{-3**}$	$5.18 \times 10^{-4**}$
Logit		*X*	5.29×10^{-1}	1.13×10^{-1}
MDA			*X*	2.95×10^{-1}
KNN				*X*

* Significant at the level of 5%.
** Significant at the level of 1%.

Table 16
The 1-year ahead to 3-year ahead forecasts by all models.

	M_1 (%)	M_2 (%)	M_3 (%)	M_4 (%)	M_5 (%)	M_6 (%)
1-Year-ahead forecast	75.34	79.68	79.37	79.96	80.04	81.36
2-Year-ahead forecast	70.26	73.02	73.52	73.89	73.96	75.13
3-Year-ahead forecast	63.43	66.51	64.88	67.07	67.91	66.78

[2] The Brier Score (BS) is a measure of prediction accuracy well-known in meteorology and medical science. It is formulated as $\left[BS = \frac{1}{n}\sum_i^n(\theta_i - p_i)^2 1)\right]$ where θ_i is a binary indicator for the actual realization of the default variable (1 if default, 0 if no default) and p_i, is the estimated probability of default. The difference between the Brier Score and the percentage of correctly classified observations is that the former is more sensitive to the level of the estimated probabilities. The Brier Score takes the estimated probabilities directly into account.

[3] McNemar test is useful for detecting before–after measurement of the same subject as a nonparametric test for two related samples using the chi-square.

Table 17
McNemar value of multiple-year-ahead forecasts.

Accuracy					
M_6	M_1	M_2	M_3	M_4	M_5
1-Year-ahead	5.47×10^{-24}**	1.45×10^{-3}**	1.06×10^{-4}**	8.04×10^{-3}	1.25×10^{-2}*
2-Year-ahead	1.18×10^{-14}**	4.9×10^{-4}**	3.64×10^{-3}**	3.69×10^{-2}*	4.8×10^{-2}*
3-Year-ahead	4.55×10^{-10}**	6.61×10^{-1}	2×10^{-3}**	5.88×10^{-1}*	4.9×10^{-2}*

* Significant at the level of 5%.
** Significant at the level of 1%.

Table 18
The distribution of financial ratios adopted in various models.

	M_1	M_2	M_3	M_4	M_5	M_6
Liquidity ratios	1	3	2	0	0	5
Solvency ratios	1	0	0	1	1	3
Operational ratios	1	0	0	0	2	1
Cash flow ratios	0	1	1	0	1	4
Growth ratios	0	0	1	0	1	1
Profitability ratios	2	1	2	1	5	6
Capital structure ratios	0	1	2	1	1	1
Other ratios	0	0	1	0	0	1

sustains an accuracy of 75.13% for 2-year-ahead forecast and 66.78% for 3-year-ahead forecast. The proposed M_6 outperforms other models for 1-year-ahead forecasts and 2-year-ahead forecasts. Table 17 shows that M_6 outperforms M_5 and M_6 at the significant level of 5%, and M_3, M_2, M_1 at the level of 1%.

Models M_4 and M_5 perform slightly better than M_2 and M_3 for both 1-year-ahead and 2-year-ahead forecasts. For 3-year-ahead forecasts, M_5 outperform other models and M_6 performs roughly the same as M_2 and M_4. M_1 remains the least accurate in any-year-ahead forecasts.

Extending the data period of financial variables from one to three years reduces the accuracy rate of all models, as reported in other studies (Altman, 1968; Ohlson, 1980; Wu et al., 2007). This implies that the most recent year's financial data plays the major role in financial prediction. However, the mixed effect that multi-year data has on financial prediction models requires further study.

7. Discussion

We observe about all the models built with the HARC feature set performs better than M_1 to M_5 expect KNN model. This result indicates that the feature set recommended by HARC is consistently superior to feature sets selected by other approaches. Table 18 summarizes the number of ratios selected by each model (M_1 to M_6) in each ratio category. It appears that the ratios selected by HARC cover all 8 categories. The number of ratios selected from these categories by HARC range from one to five. This suggests that each ratio category provides useful information in certain aspect of the distress prediction. Omitting any category may cause negative impact on the predictive power of that model.

It seems that the ratios from categories solvency, probability and capital structure are effective since ratios from these categories all appear in the selected feature sets of M_4, M_5 and M_6. The ratios [X_8] (interest expenses/equity) are all recommended in M_4, M_5, and M_6, but none of the ratios in this category is included in either M_2 or M_3. M_1 uses [X_9] (market value equity/book value of total debt), which is a solvency ratio. It appears [X_8] is a more effective indicator than [X_9] as the experiment results suggests. All models select some ratios from the probability category. However, [X_{32}] (operating income before tax/total assets) are included in M_1, M_4, M_5 and M_6, whereas [X_{38}] are included in M_2, M_3 and M_5. We notice that [X_{32}] and [X_{38}] are somewhat homogeneous. Semantically, the experiment results seem to indicate that the former one is more effective as far as the Taiwan dataset is concerned.

All models select at least one ratio from the capital structure category except M_1. These models all select [X_{42}] (liabilities/total assets) in their feature subsets. It seems to suggest [X_{42}] is the most effective ratio in this category.

Finally, we find that the cash flow ratios and the growth ratios are omitted from M_1 and M_4, but are included in M_3, M_5, and M_6. It appears that these two categories play mixed roles in predicting distressed Taiwanese firms, since M_1 is the least accurate model, while M_4 and M_6 are the more accurate models.

8. Conclusion

In this study, we propose an integrated approach to feature selection for the FDP problem that embeds expert knowledge with the wrapper method. We categorize the financial features into seven classes according to their financial semantics based on experts' domain knowledge from literature. Ratios in the same category are pair-wisely examined according to their correlation. Any two ratios with high correlation are grouped into the same sub-category. Only one ratio from the same sub-category is chosen to enter the next round of selection because these ratios are highly correlated not only in statistical sense but also in semantic sense. We then apply the wrapper method to search for "good" feature subsets consisting of top candidates from the remaining ratios in each ratio category. The search space of the wrapper method therefore effectively shrinks because features have been "filtered" before the wrapper method is applied. We conducted the case studies of listed companies in Taiwan's Stock Exchanges. This empirical study shows that the proposed integrated method significantly outperforms the existing feature selection approaches commonly found in literature. This study also shows that HARC is able to recommend a ratio subset that is more evenly distributed over all ratio categories than other methods. The two-step selection approach makes HARC an algorithm with both efficiency and performance.

In the current setting of HARC, the threshold of the t-test is set to 1.04, which is rather ad hoc. We plan to look for a more rigorous way to find the optimal threshold value so that the performance can be enhanced. Some of practitioners and scholars have suggested that a ratio may be classified into more than one category. We are interested in studying the effect on the HARC's performance if the current one-category constraint is lifted in the future. This study has conducted a case study using a public dataset that collects listed Taiwanese companies from the year of 2000 to 2008. We plan to perform more experiments with different open datasets to strengthen our conclusion on the performance of HARC.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press on Demand.
Beaver, W. H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 71–111.
Ben-Bassat, M. (1982). Pattern recognition and reduction of dimensionality, In *Handbook of Statistics-II* (pp. 773–910).
Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1–2), 245–271.
Boster, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory* (pp. 144–152). ACM.
Box, J. F. (1987). Guinness, Gosset, Fisher, and small samples. *Statistical Science, 2*(1), 45–52.
Chandra, D. K., Ravi, V., & Bose, I. (2009). Failure prediction of dotcom companies using hybrid intelligent techniques. *Expert Systems with Applications, 36*(3), 4830–4837.
Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST), 2*(3), 27.
Chen, L. H., & Hsiao, H. D. (2008). Feature selection to diagnose a business crisis by using a real GA-based support vector machine: An empirical study. *Expert Systems with Applications, 35*(3), 1145–1155.
Cho, S., Kim, J., & Bae, J. K. (2009). An integrative model with subject weight based on neural network learning for bankruptcy prediction. *Expert Systems with Applications, 36*(1), 403–410.
Chuvakhin, N., & Gertmenian, L. W. (2003). Predicting bankruptcy in the WorldCom age. *Journal of Contemporary Business Practice, 6*(1).
Ding, Y., Song, X., & Zen, Y. (2008). Forecasting financial condition of Chinese listed companies based on support vector machine. *Expert Systems with Applications, 34*(4), 3081–3089.
Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning* (1st ed.). New York: Addison Wesley.
Grefenstette, J. J. (1986). Optimization of control parameters for genetic algorithms. *IEEE Transactions on Systems, Man and Cybernetics, 16*(1), 122–128.
Gudmund, R. I., & Helmut, N. (1987). *Analysis of variance*. Sage Publications Inc..
Gunther, T., & GrUning, M. (2000). Einsatz von Insolvenzprognoseverfahren bei der Kreditwurdigkeitsprufung im Firmenkundenbereich. *Betriebswirtschaft-Stuttgart-, 60*(1), 39–59.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422.
Holland, J. H. (1992). Genetic algorithms. *Scientific American, 267*(1), 66–72.
Hopcroft, J., & Tarjan, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. *Communications of the ACM, 16*(6), 372–378.
Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications, 33*(2), 434–440.
Huang, Z., Chen, H., Hsu, C. J., Chen, W. H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems, 37*(4), 543–558.
Huang, P. W., & Liu, C. L. (2006). Using genetic algorithms for feature selection in predicting financial distress with support vector machines. International Conference on Systems. *Man and Cybernetics*, 4892–4897.
Jeong, P. W., Min, J. H., & Kim, L. S. (2012). A tuning method for the architecture of neural network models incorporating GAM and GA as applied to bankruptcy prediction. *Expert Systems with Applications, 39*, 3650–3658.
Jo, H., & Han, I. (1996). Integration of case-based forecasting, neural network, and discriminant analysis for bankruptcy prediction. *Expert Systems with Applications, 11*(4), 415–422.
Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *IEEE International Conference on Neural Networks, 1995. Proceedings* (pp. 1942–1948). IEEE.
Kennedy, J. F., Kennedy, J., & Eberhart, R. C. (2001). *Swarm intelligence*. Morgan Kaufmann Pub.
Kim, M. J., Min, S. H., & Han, I. (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications, 31*(2), 241–247.
Kittler, J. (1978). Feature set search algorithms. *Pattern Recognition and Signal Processing*, 41–60.
Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2002). *Logistic regression: A self-learning text*. Springer Verlag New York Inc.
Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1), 273–324.
Lee, K. C., Han, I., & Kwon, Y. (1996). Hybrid neural network models for bankruptcy predictions. *Decision Support Systems, 18*(1), 63–72.
Li, H., & Sun, J. (2008). Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems, 21*(8), 868–878.
Li, H., & Sun, J. (2009). Predicting business failure using multiple case-based reasoning combined with support vector machine. *Expert Systems with Applications, 36*(6), 10085–10096.
Li, H., Sun, J., & Sun, B. L. (2009). Financial distress prediction based on OR-CBR in the principle of k-nearest neighbors. *Expert Systems with Applications, 36*(1), 643–659.
Lin, F., Liang, D., & Chen, E. (2011). Financial ratio selection for business crisis prediction. *Expert Systems with Applications, 38*(12), 15094–15102.
Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491–502.
Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109–131.
Ozkan-Gunay, E. N., & Ozkan, M. (2007). Prediction of bank failures in emerging financial markets: An ANN approach. *Journal of Risk Finance, 8*(5), 465–480.
Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507–2517.
Sexton, R. S., Sriram, R. S., & Etheridge, H. (2003). Improving decision effectiveness of artificial neural networks: A modified genetic algorithm approach. *Decision Sciences, 34*(3), 421–442.
Shin, K. S., Lee, T. S., & Kim, H. J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications, 28*(1), 127–135.
Srinivas, M., & Patnaik, L. M. (1994). Genetic algorithms: A survey. *Computer, 27*(6), 17–26.
Sun, J., & Li, H. (2009). Financial distress prediction based on serial combination of multiple classifiers. *Expert Systems with Applications, 36*(4), 8659–8666.
Sun, Y. (2007). Iterative RELIEF for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(6), 1035–1051.
Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science, 38*(7), 926–947.
Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639–2649.
Vapnik, V. (1999). *The nature of statistical learning theory*. Springer Verlag New York Inc.
William, R. K. (1980). *Discriminant analysis*. SAGE Publications, Incorporated.
Wu, C. H., Tzeng, G. H., Goo, Y. J., & Fang, W. C. (2007). A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert Systems with Applications, 32*(2), 397–408.
Zhao, H., Sinha, A. P., & Ge, W. (2009). Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications, 36*(2), 2633–2644.
Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research, 22*, 59–82.