Knowledge-Based Systems 73 (2015) 289–297

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

The effect of feature selection on financial distress prediction

Deron Liang [a], Chih-Fong Tsai [b,*], Hsin-Ting Wu [a]

[a] Department of Computer Science and Information Engineering, National Central University, Taiwan
[b] Department of Information Management, National Central University, Taiwan

ARTICLE INFO

Article history:
Received 17 December 2013
Received in revised form 7 September 2014
Accepted 11 October 2014
Available online 17 October 2014

Keywords:
Financial distress prediction
Bankruptcy prediction
Credit scoring
Feature selection
Data mining

ABSTRACT

Financial distress prediction is always important for financial institutions in order for them to assess the financial health of enterprises and individuals. Bankruptcy prediction and credit scoring are two important issues in financial distress prediction where various statistical and machine learning techniques have been employed to develop financial prediction models. Since there are no generally agreed upon financial ratios as input features for model development, many studies consider feature selection as a pre-processing step in data mining before constructing the models. However, most works only focused on applying specific feature selection methods over either bankruptcy prediction or credit scoring problem domains. In this work, a comprehensive study is conducted to examine the effect of performing filter and wrapper based feature selection methods on financial distress prediction. In addition, the effect of feature selection on the prediction models obtained using various classification techniques is also investigated. In the experiments, two bankruptcy and two credit datasets are used. In addition, three filter and two wrapper based feature selection methods combined with six different prediction models are studied. Our experimental results show that there is no the best combination of the feature selection method and the classification technique over the four datasets. Moreover, depending on the chosen techniques, performing feature selection does not always improve the prediction performance. However, on average performing the genetic algorithm and logistic regression for feature selection can provide prediction improvements over the credit and bankruptcy datasets respectively.

1. Introduction

Financial distress prediction is very critical in enterprise risk management, especially for financial institutions. In particular, financial institutions have to develop various risk management models, such as bankruptcy prediction and credit scoring models [37,43]. For bankruptcy prediction, financial institutions need effective prediction models in order to make appropriate lending decisions. On the other hand, credit scoring models are used for the management of large loan portfolios and/or credit admission evaluation.

Specifically, bankruptcy prediction and credit scoring are two binary classification problems in financial distress prediction, which aim at assigning new observations to two pre-defined decision classes (e.g., ‘good’ and ‘bad’ risk classes) [40]. For example, bankruptcy prediction models are used to predict the likelihood that the loan customers will go bankrupt whereas credit scoring models are used to determine whether the loan applicants should be classified into a high risk or low risk group. In the literature, many supervised machine learning (or classification) techniques have been used for financial distress prediction [2,10,24,29].

Though many novel sophisticated techniques have been proposed for effective prediction, very few have examined the effect of feature selection on financial distress prediction. Feature selection is an important data pre-processing step of knowledge discovery in databases (KDD). The aim is to filter out unrepresentative features from a given dataset [11,17]. As there are no generally agreed financial ratios for bankruptcy prediction and credit scoring, collected variables must be examined for their representativeness, i.e., importance and explanatory power, in the chosen dataset [29]. Therefore, the performance of classifiers after performing feature selection could be enhanced over that of classifiers without feature selection.

Generally speaking, feature selection can be broadly divided into the filter, wrapper, and hybrid approaches [3,31]. The filter based method (usually based on some statistical techniques) evaluates and selects feature subsets by the general characteristics of the given dataset. The wrapper based method is based on a pre-determined mining algorithm and its performance is used as

* Corresponding author. Tel.: +886 3 422 7151; fax: +886 3 4254604.
E-mail address: cftsai@mgt.ncu.edu.tw (C.-F. Tsai).

http://dx.doi.org/10.1016/j.knosys.2014.10.010
0950-7051/© 2014 Elsevier B.V. All rights reserved.

the evaluation criterion to select feature subsets. Specifically, it aims at searching for features that are better suited to the mining algorithm to improve the mining performance. The hybrid method is based on combining these two methods by exploiting different evaluation criteria in different search stages.

In recent studies, the filter and wrapper based feature selection methods have been widely used for bankruptcy prediction [41,13,14,27,25,26,8,4] and credit scoring [18,7,30,16,42,5]. Most studies apply either filter or wrapper based methods for single domain problems, i.e., bankruptcy prediction and credit scoring. One reason for the lack of using hybrid based feature selection methods is because currently there is no standard and representative method. In addition, there are no guidelines for which filter and wrapper based methods should be combined to select the best features for the later prediction performance.

One major limitation of current studies is that each work only considers one specific feature selection method for either bankruptcy prediction or credit scoring problems. In other words, there is no study focusing on comparing both types of feature selection methods for both bankruptcy prediction and credit scoring problems (c.f. Section 2.2). Therefore, the aim of this paper is to examine the effect of the filter and wrapper based feature selection methods on both bankruptcy prediction and credit scoring problems. Moreover, the effect of performing feature selection on different classification techniques will also be investigated.

The contributions of this paper are twofold. First, we provide a comprehensive study of comparing different filter and wrapper based feature selection methods in terms of two financial distress problems, which are bankruptcy prediction and credit scoring. In particular, the most suitable methods for these two specific problems are identified. As a result, the identified methods can also be regarded as the baseline feature selection methods for future related researches. Second, the research findings also allow us to understand which classification technique(s) are more sensitive to feature selection. Therefore, this can provide a guideline for future studies to choose suitable techniques for their prediction models.

The rest of this paper is organized as follows. Section 2 overviews related literature about filter and wrapper based feature selection methods. Moreover, related works are compared in terms of the feature selection methods employed, prediction methods constructed, etc. Sections 3 and 4 present the experimental setup and results, respectively. Finally, Section 5 concludes the paper.

2. Literature review

2.1. Feature selection

As there are no generally agreed factors (i.e., variables) for bankruptcy prediction and credit scoring, some of the collected variables as features may contain noise that could affect the prediction result. On the other hand, if too many features were used for data analysis, it can cause high dimensionality problems [36]. In data mining, feature selection or dimensionality reduction can be approached to reduce irrelevant or redundant features. This is an important data pre-processing technique in data mining, which aims at selecting more representative features having more discriminatory power over a given dataset [11,17].

Feature selection can be defined as the process of choosing a minimum subset of m features from the original dataset of n features ($m < n$), so that the feature space (i.e. the dimensionality) is optimally reduced according to four steps, which are subset generation, subset evaluation, stopping criteria, and result validation [11,31].

In general, subset generation is a search procedure which generates subsets of features for evaluation. Each subset generated is evaluated by a specific evaluation criterion and compared with the previous best one with respect to this criterion. If a new subset is found to be better, then the previous best subset is replaced by the new subset.

2.2. Filter based feature selection

The filter based feature selection methods usually contain the following procedures. Given a dataset, the method based on a particular search strategy initially searches from a given subset, which may be an empty set, a full set, or any randomly selected subset. Then, each generated subset is evaluated by a specific measure and compared with the previous best one. This search process iterates until the pre-defined stopping criterion is met. Consequently, the final output of this method is the last current best subset.

More specifically, the search strategy and evaluation measure can be different depending on the algorithms used. In addition, filter based methods do not involve any mining algorithm during the search and evaluation steps, they are computationally efficient. Some examples of filter based methods that are used in financial distress prediction are based on statistical techniques, such as t-testing, principal component analysis, discriminant analysis, and regression [4,8,37,26,41,45].

2.2.1. Discriminant Analysis

Linear Discriminant analysis (LDA) is used to find a linear combination of features which characterizes or separates two or more classes of objects. The resulting combination can be used for dimensionality reduction. LDA can also be used to express one dependent variable as a linear combination of other features. In other words, LDA looks for the linear combination of features which best explains the given data [34].

LDA involves the determination of a linear equation like regression that will predict which group the case belongs to. The form of the equation or function is

$$D = v_1X_1 + v_2X_2 + v_3X_3 + \cdots v_iX_i + a \tag{1}$$

where D is the discriminant function, v is the discriminant coefficient or weight for that feature, X is the respondent's score for that feature, a is a constant, and i is the number of predictor features.

2.2.2. t-Test

The t-test method is used to determine whether there is a significant difference between two group's means. It helps to answer the underlying question: Do the two groups come from the same population, and only appear differently because of chance errors, or is there some significant difference between these two groups? Three basic factors help determine whether an apparent difference between two groups is a true difference or just an error due to chance [35]:

1. The larger the sample, the less likely that the difference is due to sampling errors or chance.
2. The larger the difference between the two means, the less likely that the difference is due to sampling errors.
3. The smaller the variance among the participants, the less likely that the difference is created by sampling errors.

2.2.3. Logistic regression

Logistic regression (LR) is a type of probabilistic statistical classification model. LR measures the relationship between a categorical dependent variable and one or more independent variables, which are usually continuous, by using probability scores as the predicted values of the dependent variables. LR allows us to look at the fit of the model as well as at the significance of the relationships between dependent and independent variables that are modeled [19].

The LR function can be written as

$$P = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \tag{2}$$

where P is the probability of a 1, e is the base of the natural logarithm and α and β are the parameters of the model.

2.3. Wrapper based feature selection

The wrapper based feature selection methods are similar to the filter based ones except that a pre-defined mining algorithm is utilized for the search strategy and evaluation measure. That is, for each generated subset, the mining algorithm is used to evaluate the goodness of the selected subset in terms of the quality of mined results. Therefore, using different mining algorithms will produce different feature selection results over the same dataset.

Since the mining algorithms are used to select and evaluate the feature subsets, the wrapper based methods are likely to perform better than the filter based methods [20]. However, it is usually more computationally expensive than the filter based methods. Some examples of wrapper based methods used in financial distress prediction are the Bayesian classifier, particle swarm optimization, rough set, and genetic algorithm methods [14,33,30,42].

2.3.1. Genetic algorithms

Genetic algorithms (GA) have become an effective feature selection approach to improve the performance of data mining algorithms. In GA, a population of strings (called chromosomes), which encode candidate solutions (called individuals) to an optimization problem, evolves for better solutions. In general, the genetic information (i.e., chromosome) is represented by a bit string (such as binary strings of 0s and 1s) and sets of bits encode the solution. Then, genetic operators are applied to the individuals of the population for the next generation (i.e., a new population of individuals). There are two main genetic operators, which are crossover and mutation. Crossover creates two offspring strings from two parent strings copying selected bits from each parent. On the other hand, mutation randomly changes the value of a single bit (with small probability) to the bit strings. Furthermore, a fitness function is used to measure the quality of an individual in order to increase the probability that the single bit can survive throughout the evolutionary process [15].

2.3.2. Particle swarm optimization

Particle swarm optimization (PSO) is also a type of evolutionary algorithm. It optimizes a problem by looking at a population of particles (i.e., candidate solutions). The particles are moved around in the search space according to a mathematical formula considering the particle's position and velocity. As a result, it is expected that the swarm will move toward the best solutions [22].

Specifically, both GA and PSO share the following common elements:

- Both initialize a population in a similar manner.
- Both use an evaluation function to determine how fit (good) a potential solution is.
- Both are generational, that is both repeat the same set of processes for a predetermined amount of time.

PSO has two primary operators, which are velocity update and position update. During each generation each particle is accelerated toward its previous best position and the global best position. In each iteration, a new velocity value for each particle is calculated based on its current velocity, the distance from its previous best position, and the distance from the global best position. The new velocity value is then used to calculate the next position of the particle in the search space. This process continues until a minimum error is achieved.

2.4. Comparisons of related works

Table 1 compares related works from the past five years (2009–2013) in terms of the feature selection methods employed, prediction methods constructed, and domain datasets used. Note that the number for the filter and wrapper based methods means the number of methods used in their corresponding works. According to Table 1, we can observe that most studies performing feature selection only consider one specific type of method. In addition, many works only focus on one specific domain problem, i.e., either bankruptcy prediction or credit scoring.

To be specific, wrapper based methods are applied for bankruptcy prediction in three studies and filter based methods in seven studies. On the other hand, wrapper based methods are used for credit scoring in four studies and filter based methods are applied in three studies.

Consequently, this literature review raises two research questions. The first one is: which type of feature selection method is suitable for which problem domain? Second, is there any best combination for combining specific types of feature selection methods and specific classification techniques for bankruptcy prediction and credit scoring? The following experiments are conducted to answer these two questions.

3. Experimental design

3.1. The datasets

Table 2 shows the information for the four chosen datasets utilized in this study. The Australian and German datasets are public sets widely used in the literature. The Taiwanese and Chinese datasets were collected from the Taiwan Economic Journal[1] and the definitions of bankrupt companies are based on the business regulations from the Taiwan Stock Exchange and Shanghai and Shenzhen Stock Exchange, respectively.

Since the Chinese and Taiwanese bankruptcies are real-world datasets, in practice they contain very few bankrupt cases whereas the numbers of non-bankrupt cases are very large. This makes the class imbalance problem of the collected datasets, which is likely to degrade the final prediction performance. Therefore, the method of stratified sampling [1] is used to collect the same numbers of good and bad cases. Moreover, each of the attributes is normalized into the range from 0 to 1. For training and testing each classifier, the 10-fold cross-validation strategy is used to divide each dataset into 10 distinct training and testing subsets.

3.2. Feature selection

3.2.1. Filter based feature selection methods

From the relevant studies reviewed in Section 2, three widely used filter based feature selection methods are chosen for comparison, namely, linear discriminant analysis (LDA), t-test, logistic regression (LR).

Fig. 1 outlines the process of performing filter based feature selection for financial distress prediction. The first step is to divide each dataset into the training and testing sets by 10-fold cross validation. Then, each feature selection method is executed over the training set. Next, the selected features are used as the new

[1] http://www.tej.com.tw/twsite/.

Table 1
Comparisons of related works.

Works	Feature selection methods		Prediction models	Problem domains	
	Filter	Wrapper		BP[a]	CS[b]
Chandra et al. [4]	1		MLP[c]/CART[d]/SVM[e]/RF[f]/LR[g]	v	
Chen [7]	4		Rough sets		v
Chen and Li [5]	4		SVM		v
Cho et al. [8]	2		CBR[h]	v	
Divsalar et al. [13]	1		GEP[i]	v	
Feki et al. [14]		1	SVM	v	
Gonen et al. [16]		2	Probit regression/MKL[j]		v
Hajek and Michalak [18]		2	MLP/RBF neural network/SVM/ NB[k]/RF/LDC[l]/NMC[m]		v
Li and Sun [25]		1	SVM	v	
Li and Sun [26]	1		CBR	v	
Lin et al. [27]	1		SVM	v	
Ling et al. [30]		1	SVM		v
Martin et al. [33]		1	Fuzzy c-means/MARS	v	
Tsai [41]	5		MLP	v	v
Wang et al. [43]		1	RBF neural network/LR/DT[n]		v
Yang et al. [45]	1		SVM	v	

[a] BP: Bankruptcy Prediction.
[b] CS: Credit Scoring.
[c] MLP: multilayer perceptron neural network.
[d] CART: classification and regression tree.
[e] SVM: support vector machines.
[f] RF: random forest.
[g] LR: logistic regression.
[h] CBR: case-based reasoning.
[i] GEP: gene expression programming.
[j] MKL: multiple kernel learning.
[k] NB: naïve Bayes.
[l] LDC: linear discriminant classifier.
[m] NMC: nearest mean classifier.
[n] DT: J48 decision tree.

Table 2
Dataset information.

Dataset	Total cases	Good/bad cases	No. of attributes
Chinese bankruptcies	688	344/344	45
Taiwanese bankruptcies	440	220/220	95
Australian credit	690	307/382	14
German credit	1000	700/300	24

training set (not the original training set) to train a prediction model. Finally, the testing set containing the same selected features as the new training set is used to test the performance of the prediction model.

Note that the threshold to determine representative features by the filter based feature selection methods is based on the feature, which is significant at the 0.05 level. For example, using the t-test method the features having the p values less than 0.05 are kept; otherwise they are filtered out.

3.2.2. Wrapper based feature selection methods

Two representative methods are used for wrapper based feature selection in this paper, which are genetic algorithm (GA) and particle swarm optimization (PSO).

Fig. 2 shows the process of performing wrapper based feature selection for financial distress prediction. First, each dataset is divided into the training and testing sets by 10-fold cross validation. Each training set is further sampled for the training and validation subsets to train the wrapper based feature selection methods. Then, the population pool is initialized where each group of the chromosome or particle represents the selected feature set. Next, each chromosome or particle in the population pool (as the training subset) is used to construct multiple models. After the models are constructed, the validation subset is used to test their accuracy. For GA, the performance of the models constructed by each chromosome is examined, and then the selection, crossover, mutation operations are performed to replace the current population pool. On the other hand, each particle of PSO is examined for its performance and its position and velocity in the feature space are adjusted by the sigmoid function to replace the current population pool. Consequently, the evolutionary process will be terminated until the stopping criterion is met. Then, the chromosome or particle having the highest accuracy over the validation subset is used as the training set to train the prediction model. Finally, the testing set containing the same selected features as the chromosome or particle is used to test the performance of the prediction model.

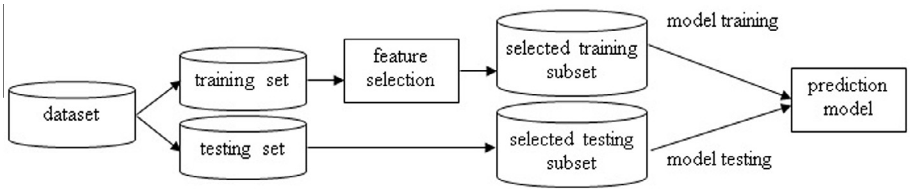

Fig. 1. Filter based feature selection for financial distress prediction.

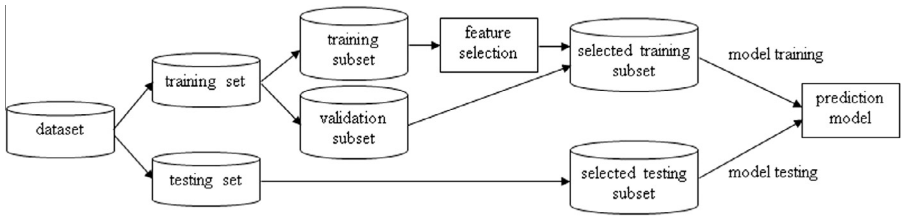

Fig. 2. Wrapper based feature selection for financial distress prediction.

Table 3
The parameters of GA and PSO.

Methods	Parameters	Values
GA	Objective function	Fitness value = average accuracy
	Selection	Roulette wheel selection
	Crossover method	Uniform crossover
	Generation	Generation ≧ 20; average accuracy after 8 repetitions
	Population size	60
	Crossover rate	0.7
	Mutation rate	0.01
	Elite chromosome	2
PSO	Objective function	Fitness value = average accuracy;
	Swarm size	60
	Generation	Generation ≧ 20; average accuracy after 8 repetitions
	C1	2
	C2	2
	Vmax	6

Table 3 lists related parameters of GA and PSO, which are based on some related works, such as Srinivas and Patnaik [38], Ko and Lin [23], Lin et al. [28], and Liu et al. [32]. Note that different values of the population size and swarm size (20–500), crossover rate (0.4–1.0), mutation rate (0.001–0.1), and generations (100–5000) were compared in order to find out the best parameter values.

3.3. The classifiers

For classifier design, six classification techniques are used, namely, linear SVM, RBF SVM, *k*-NN, Naïve Bayes, CART, and MLP. These were identified as the most popular and widely used classification techniques by Wu et al. [44]. Table 4 lists the parameters for constructing these classifiers for comparison.

3.4. Evaluation metrics

To assess the performance of the above mentioned classifiers (i.e., prediction models), two evaluation metrics are used, which are prediction accuracy and the Type I error. They can be measured by a confusion matrix, as shown in Table 5.

Table 5
Confusion matrix.

↓predicted \ actual→	Non-bankruptcy	Bankruptcy
Non-bankruptcy	a (True Positive)	b (False Positive)
Bankruptcy	c (False Negative)	d (True Negative)

Therefore, the average prediction accuracy is obtained by

$$\text{Prediction accuracy} = \frac{a+d}{a+b+c+d} \tag{3}$$

and the Type I error is based on

$$\text{Type I error} = \frac{b}{b+d} \tag{4}$$

In addition to the average accuracy, Type I error is taken into account. This occurs when the classifier incorrectly classifies a bankrupt firm (or member of the high risk group) into the non-bankrupt class (or the low risk group). This is also critical because a higher Type I error rate requires financial institutions to expend larger costs, which can enhance the enterprise risk.

4. Results

4.1. Results on credit scoring datasets

Tables 6 and 7 show the performances of different classifiers over the Australian and German credit datasets, respectively. Note that the baseline means the classifier without feature selection. In addition, the bold numbers indicate performances that significantly outperform the other classifiers (indicated by non-bold numbers). The level of performance significance is measured by the Wilcoxon test [12]. Furthermore, the highest rate of classification accuracy and the lowest rate of the Type I error are underlined.

The results show that filter based feature selection methods generally perform better than wrapper based ones. In particular, the LDA and *t*-test perform the best over the Australian and German datasets, respectively.

On the other hand, on average, for the Australian dataset, performing feature selection makes the classifiers outperform the ones without feature selection in terms of the prediction accuracy and the Type I error. This is different from the German dataset, in that only performing a *t*-test can allow the classifiers to provide

Table 4
Parameters for constructing the classifiers.

Classifier	Parameters
SVM	Kernel functions: linear kernel and RBF kernel; other related parameters are based on the default parameters from the LIBSVM toolbox.[a]
KNN	*K* = 7; distance function: Euclidean distance
CART	The default parameters used are based on the Matlab toolbox
MLP	The number of hidden nodes: 8/16/32/64; learning epochs: 50/100/200/400 [41]
Naïve Bayes	Kernel function: kernel density estimate [21]

[a] http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Table 6
Performances of different classifiers over the Australian credit dataset (# samples: 690; # attributes: 14) ($p < 0.05$).

		Wrapper methods		Filter methods			Baseline (%)
		GA (%)	PSO (%)	*t*-Test (%)	LDA (%)	LR (%)	
Linear SVM	Accuracy	85.52	85.52	85.52	85.52	85.52	85.52
	Type I error	20.08	20.06	20.08	20.08	20.08	20.08
RBF SVM	Accuracy	84.27	84.82	**85.54**	**85.57**	**85.45**	84.47
	Type I error	16.84	16.76	**14.18**	**14.18**	**15.29**	17.52
CART	Accuracy	84.85	84.82	85.25	85.46	85.11	85.20
	Type I error	**15.85**	17.50	19.19	17.68	17.98	18.98
k-NN	Accuracy	84.69	84.64	**86.06**	85.31	84.81	84.58
	Type I error	14.87	14.39	**13.36**	**13.20**	**13.30**	14.97
Naïve Bayes	Accuracy	**86.09**	**85.86**	68.52	67.09	66.74	68.55
	Type I error	**13.61**	**12.41**	21.66	21.24	**20.51**	21.51
MLP	Accuracy	**85.57**	**85.49**	**85.60**	**86.00**	**85.89**	84.15
	Type I error	13.93	14.97	13.82	13.82	13.78	14.86
Avg.	Accuracy	85.17	85.19	82.75	82.49	82.25	82.08
	Type I error	15.86	16.02	17.05	16.70	16.82	17.99

Table 7
Performances of different classifiers over the German credit dataset (# samples: 1000; # attributes: 24).

		Wrapper methods		Filter methods			Baseline (%)
		GA (%)	PSO (%)	*t*-Test (%)	LDA (%)	LR (%)	
Linear SVM	Accuracy	76.54	73.76	76.74	75.72	75.10	77.18
	Type I error	53.07	61.07	54.13	58.93	64.00	50.80
RBF SVM	Accuracy	74.80	72.26	76.40	75.98	75.20	76.30
	Type I error	54.87	59.73	51.07	52.40	59.00	49.80
CART	Accuracy	**75.72**	74.16	74.28	73.52	73.66	74.30
	Type I error	55.80	59.33	59.27	62.33	65.07	57.87
k-NN	Accuracy	72.24	71.60	71.82	71.86	**72.62**	70.86
	Type I error	**59.73**	86.27	63.20	**60.47**	**61.20**	65.33
Naïve Bayes	Accuracy	**71.56**	**74.16**	**72.40**	70.88	**71.44**	70.52
	Type I error	**84.53**	**59.33**	**83.80**	**92.67**	**88.80**	95.93
MLP	Accuracy	**74.03**	72.54	73.28	**73.44**	**73.42**	71.76
	Type I error	57.25	59.07	57.73	57.40	60.93	57.73
Avg.	Accuracy	74.15	73.08	74.15	73.57	73.57	73.49
	Type I error	60.88	64.13	61.53	64.03	66.50	62.91

better performance than the baselines. These results are consistent with Tsai [41], who compared five filter based feature selection methods and found the *t*-test to be the optimal feature selection method.

4.2. Results on bankruptcy prediction datasets

Tables 8 and 9 show the performances of different classifiers over the China and Taiwan bankruptcy datasets, respectively. The results are interesting in that, on average, performing feature selection makes the classifiers outperform the baselines in terms of prediction accuracy. However, the classifiers followed by feature selection do not necessarily mean that they can provide lower Type I errors than the baselines.

On the other hand, the GA based feature selection method can make CART and linear SVM provide the highest rate of prediction accuracy over the China and Taiwan datasets respectively. For the Type I error, LDA and GA perform the best over the China and Taiwan datasets, respectively. These results are somewhat similar to Hua et al.'s [20] conclusion that wrapper based methods could be superior to filter based methods over high dimensional datasets.

4.3. Discussion

4.3.1. The effect of performing feature selection on classifier performances

Based on the above results, the effect of performing feature selection on classification techniques is discussed below.

- Linear SVM: Linear SVM combined with feature selection does not perform significantly better than baseline linear SVM over the four datasets. This may be because some weights are assigned to the input features when constructing the linear SVM classifier. In other words, important features can be identified and have higher weights assigned. Therefore, executing feature selection may not be necessary.
- RBF SVM: In most cases, there is no significant difference between combining feature selection with RBF SVM and the baseline method. However, for the Australian dataset RBF SVM combined with filter based feature selection methods perform significantly better than the baseline method. This implies that the chosen dataset may have a positive impact on the performance of RBF SVM after performing feature selection.

Table 8
Performances of different classifiers over the China dataset (# samples: 688; # attributes: 45).

		Wrapper methods		Filter methods			Baseline (%)
		GA (%)	PSO (%)	*t*-Test (%)	LDA (%)	LR (%)	
Linear SVM	Accuracy	91.33	91.54	91.54	91.48	91.13	91.45
	Type I error	6.58	**5.99**	6.80	**5.41**	**5.98**	6.80
RBF SVM	Accuracy	91.77	91.36	91.48	91.83	**92.14**	91.59
	Type I error	6.70	7.33	6.52	6.58	6.40	6.11
CART	Accuracy	92.98	92.55	93.04	92.43	92.87	93.04
	Type I error	5.99	6.87	5.76	7.09	6.69	5.76
k-NN	Accuracy	**91.28**	91.05	90.69	**91.34**	**91.82**	90.58
	Type I error	7.61	7.21	7.05	6.81	7.05	7.45
Naïve Bayes	Accuracy	**90.72**	**90.26**	**90.70**	89.43	**91.62**	88.61
	Type I error	7.39	7.86	5.99	10.11	8.44	7.32
MLP	Accuracy	91.63	**90.59**	89.67	**91.28**	**91.27**	89.67
	Type I error	**7.71**	9.29	10.41	**7.56**	**6.81**	9.30
Avg.	Accuracy	91.62	91.23	91.19	91.30	91.81	90.82
	Type I error	7.00	7.43	7.09	7.26	6.90	7.12

Table 9
Performances of different classifiers over the Taiwan dataset (# samples: 440; # attributes: 95).

		Wrapper methods		Filter methods			Baseline (%)
		GA (%)	PSO (%)	*t*-Test (%)	LDA (%)	LR (%)	
Linear SVM	Accuracy	82.64	82.09	82.05	79.00	80.86	81.95
	Type I error	16.09	16.91	17.91	22.18	17.09	17.27
RBF SVM	Accuracy	81.91	81.45	80.59	81.41	82.32	82.73
	Type I error	15.00	16.00	15.36	16.09	18.64	13.18
CART	Accuracy	78.27	79.36	79.09	79.09	79.73	79.68
	Type I error	20.18	19.55	19.36	18.64	18.00	17.55
k-NN	Accuracy	**78.95**	78.68	77.09	**79.59**	**82.73**	77.05
	Type I error	22.18	22.45	21.55	22.82	**16.55**	22.00
Naïve Bayes	Accuracy	**78.86**	**78.27**	**81.05**	77.68	**81.00**	76.77
	Type I error	17.91	17.64	17.82	17.73	20.27	13.73
MLP	Accuracy	**78.68**	**77.39**	76.59	**79.59**	**79.86**	74.82
	Type I error	18.66	23.64	20.55	20.82	21.45	21.55
Avg.	Accuracy	79.89	79.54	79.41	79.39	81.08	78.83
	Type I error	44.25	44.84	44.44	44.72	44.91	42.75

- CART: For the four datasets, CART combined with feature selection does not provide significantly better accuracy or Type I error than the baseline method. This may be because the feature selection step is already employed during the construction of the CART classifier. Therefore, similar to linear SVM, feature selection may not help CART for the performance improvement.
- *k*-NN: The *k*-NN classifier combined with feature selection performs significantly better than the baseline *k*-NN for prediction accuracy and the Type I error over the four datasets. Particularly, using *k*-NN with the filter based feature selection methods can provide better performance than with the wrapper based feature selection methods. Since *k*-NN does not include pre-processing of the input features (like linear SVM and CART) and the final output of using *k*-NN is based on the distances in the feature space between the training data samples and the testing ones, performing feature selection first can have a positive impact on the *k*-NN performance.
- Naïve Bayes: The naïve Bayes classifier combined with feature selection can significantly outperform the baseline method in most cases. The only exception is using the filter based feature selection methods over the Australian dataset. Specifically, using the wrapper based methods allows the naïve Bayes classifier to provide better performance than using the filter based methods. Similar to *k*-NN, performing feature selection can positively impact the naïve Bayes performance. This finding is consistent with related works, such as Chen et al. [6], that the naïve Bayes classifier is highly sensitive to feature selection.
- MLP: Combining feature selection with MLP can provide significantly better performance than the baseline MLP over three datasets, with the exception of the German dataset. In particular, using filter based methods make MLP perform better than wrapper based methods. Therefore, performing feature selection improves the performance of MLP. Despite different weights being assigned to the input features during the construction of MLP, overfitting can occur during the classifier training stage. As a result, performing feature selection can reduce the risk of overfitting and thus improve the final accuracy [39].

4.3.2. The best combinations between feature selection methods and prediction models

The best combinations for combining the feature selection methods and classification techniques over the four datasets are discussed below.

- Australian dataset: Looking at both prediction accuracy and Type I error, there are several better combinations that do not have a significant level of difference in performance, which are *t*-test + *k*-NN, GA + naïve Bayes, PSO + naïve Bayes, *t*-test + RBF SVM, GA + MLP, and LDA + *k*-NN. Among them, GA + naïve Bayes and PSO + naïve Bayes provide the highest rate of prediction accuracy and the lowest Type I error rate respectively. In particular, GA + naïve Bayes performs the third best for the Type I error. Therefore, the optimal combination could be GA + naïve Bayes.

- German dataset: In this dataset, the baseline linear SVM and RBF SVM classifiers without feature selection perform the best in terms of prediction accuracy and the Type I error respectively. In addition, the baseline RBF SVM performs the second best in terms of prediction accuracy. Therefore, these indicate that performing feature selection is likely to degrade the models' performances over this dataset, especially for SVM. However, to compare all of the combinations *t*-test + linear SVM and *t*-test + RBF SVM outperform the others for prediction accuracy and the Type I error respectively.
- China dataset: In this dataset, the baseline CART and *t*-test + CART provide the same prediction accuracy and Type I error, which outperform the other models. On the other hand, LDA + linear SVM provide the lowest Type I error rate, whereas the baseline CART and *t*-test + CART perform the second best. These indicate that performing feature selection is not necessary in this dataset. Particularly, the baseline CART is a more suitable model for this dataset.
- Taiwan dataset: LR + *k*-NN and the baseline RBF SVM perform the best in terms of prediction accuracy where they perform the same. On the other hand, the baseline RBF SVM provides the lowest Type I error rate. Specifically, LR + *k*-NN and GA + RBF SVM outperform the other combinations in terms of prediction accuracy and Type I errors respectively.

In summary, there is no exact answer for the best combination of the feature selection method and the classification technique over the four datasets. However, if we compare the average prediction results (including average prediction accuracy and the Type I error) by each feature selection method and the baseline models, we can see that the models' performances can be improved if the feature selection method was carefully chosen. Particularly, the better feature selection methods for credit scoring and bankruptcy prediction are GA and LR respectively (c.f. Tables 6–9).

Although it is difficult to conclude the best feature selection for the financial distress prediction problems, several feature selection methods that can provide relatively better performances can be recommended for future researches. That is, better filter and wrapper feature selection methods are *t*-test, LR, and GA.

Finally, the prediction improvements by performing filter based feature selection over the credit scoring datasets containing small numbers of attributes (i.e. 14 and 24) are small. That is, on average the prediction improvements by using filter based feature selection is only about 0.17–0.67% and 0.08–0.66% over the Australian and German datasets respectively. These may be within the standard deviation of the prediction accuracy obtained from 10-fold cross validation. On the other hand, performing GA can provide about 3.09% and 0.66% prediction improvements over the Australian and German datasets respectively. These results demonstrate the importance of choosing a suitable feature selection for credit scoring and bankruptcy prediction.

5. Conclusion

In this paper, we examine the effect of feature selection in financial distress prediction. Specifically, filter and wrapper based feature selection methods are compared in terms of prediction accuracy and the Type I errors made by six different classifiers.

We found that there is no the best combination of the feature selection method and the classification technique over the four datasets. However, on average GA performs better than the others over the credit scoring datasets whereas LR outperforms the other methods over the bankruptcy prediction datasets. Despite these findings, several feature selection methods have shown some promising results for bankruptcy prediction and credit scoring. In particular, *t*-test and LR as the filter methods and GA as the wrapper method can be used in the future.

It should be noted that performing feature selection does not always improve the models' performances, especially for CART and SVM. This may be because when constructing these models, CART can determine important features like many feature selection methods do during the tree construction process whereas SVM generally assigns some weights to the input features (i.e. attributes). Moreover, since related studies, e.g. Clarke et al. [9], have shown the advantage of SVM for high dimensional data, the dimensionalities of financial distress datasets are relatively small compared with other domains, such as genomic and proteomic problems. Therefore, the need of performing the feature selection step for credit scoring and bankruptcy prediction depends on the chosen classifiers.

For future works, several issues could also be considered. First of all, since the chosen filter and wrapper based feature selection methods are based on the mostly used methods in bankruptcy prediction and credit scoring, other filter (e.g. information gain) and wrapper (e.g. naïve Bayes) methods can also be employed for the feature selection task. Secondly, in addition to using single classification techniques to develop the prediction models, combining multiple classifiers or classifier ensembles by the bagging and boosting combination methods can be developed for further comparison.

References

[1] E.I. Altman, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, J. Fin. 23 (1968) 589–609.

[2] S. Balcaen, H. Ooghe, 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems, Brit. Account. Rev. 38 (2006) 63–93.

[3] A.L. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artif. Intell. 97 (1997) 245–271.

[4] D.K. Chandra, V. Ravi, I. Bose, Failure prediction of dotcom companies using hybrid intelligent techniques, Expert Syst. Appl. 36 (2009) 4830–4837.

[5] F.-L. Chen, F.-C. Li, Combination of feature selection approaches with SVM in credit scoring, Expert Syst. Appl. 37 (2010) 4902–4909.

[6] J. Chen, H. Huang, S. Tian, Y. Qu, Feature selection for text classification with naïve Bayes, Expert Syst. Appl. 36 (2009) 5432–5435.

[7] Y.-S. Chen, Classifying credit ratings for Asian banks using integrating feature selection and the CPDA-based rough sets approach, Knowl.-Based Syst. 26 (2012) 259–270.

[8] S. Cho, H. Hong, B.-C. Ha, A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance for bankruptcy prediction, Expert Syst. Appl. 37 (2010) 3482–3488.

[9] R. Clarke, H.W. Ressom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data sapces: implications for exploring gene and protein expression data, Nat. Rev. Cancer 8 (1) (2008) 37–49.

[10] J.N. Crook, D.B. Edelman, L.C. Thomas, Recent developments in consumer credit risk assessment, Eur. J. Oper. Res. 183 (2007) 1447–1465.

[11] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156.

[12] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.

[13] M. Divsalar, H. Roodsaz, F. Vahdatinia, G. Norouzzadeh, A.H. Behrooz, A robust data-mining approach to bankruptcy prediction, J. Forecast. 31 (2012) 504–523.

[14] A. Feki, A.B. Ishak, S. Feki, Feature selection using Bayesian and multiclass support vector machines approaches: application to bank risk prediction, Expert Syst. Appl. 39 (2012) 3087–3099.

[15] D.E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, Addition Wesley, 1989.

[16] G.B. Gonen, M. Gonen, F. Gurgen, Probabilistic and discriminative group-wise feature selection methods for credit risk analysis, Expert Syst. Appl. 39 (2012) 11709–11717.

[17] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[18] P. Hajek, K. Michalak, Feature selection in corporate credit rating prediction, Knowl.-Based Syst. 51 (2013) 72–84.
[19] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, second ed., Wiley, 2000.
[20] J. Hua, W.D. Tembe, E.R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recogn. 42 (3) (2009) 409–424.
[21] G.H. John, P. Langley, Estimating continuous distributions in Bayesian classifiers, in: International Conference on Uncertainty in Artificial Intelligence, 1995, pp. 338–345.
[22] J. Kennedy, R.C. Eberhart, Swarm Intelligence, Morgan Kaufmann, 2001.
[23] P.-C. Ko, P.-C. Lin, An evolution-based approach with modularized evaluations to forecast financial distress, Knowl.-Based Syst. 19 (1) (2006) 84–91.
[24] P.R. Kumar, V. Ravi, Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review, Eur. J. Oper. Res. 180 (2007) 1–28.
[25] H. Li, J. Sun, Predicting business failure using support vector machines with straightforward wrapper: a re-sampling study, Expert Syst. Appl. 38 (2011) 12747–12756.
[26] H. Li, J. Sun, Principal component case-based reasoning ensemble for business failure prediction, Inform. Manage. 48 (2011) 220–227.
[27] F. Lin, D. Liang, E. Chen, Financial ratio selection for business crisis prediction, Expert Syst. Appl. 38 (2011) 15094–15102.
[28] S.-W. Lin, K.-C. Ying, S.-C. Chen, Z.-J. Lee, Particle swarm optimization for parameter determination and feature selection for support vector machines, Expert Syst. Appl. 35 (2008) 1817–1824.
[29] W.-Y. Lin, Y.-H. Hu, C.-F. Tsai, Machine learning in financial crisis prediction: a survey, IEEE Trans. Syst., Man Cybernet. – Part C: Appl. Rev. 42 (4) (2012) 421–436.
[30] Y. Ling, Q. Cao, H. Zhang, Credit scoring using multi-kernel support vector machine and chaos particle swarm optimization, Int. J. Comput. Intell. Appl. 11 (3) (2012) 1250019 (13 pages).
[31] H. Liu, L. Yu, Toward integrating feature selection algoritms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (4) (2005) 491–502.
[32] Y. Liu, G. Wang, H. Chen, H. Dong, X. Zhu, S. Wang, An improved particle swarm optimization for feature selection, J. Bionic Eng. 8 (2) (2011) 191–200.
[33] A. Martin, V. Gayathri, G. Saranya, P. Gayathri, P. Venkatesan, A hybrid model for bankruptcy prediction using genetic algorithm, fuzzy c-means and MARS, Int. J. Soft Comput. 2 (1) (2011) 12–23.
[34] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley Interscience, 2004.
[35] R.R. Pagano, Understanding Statistics in the Behavioral Sciences, sixth ed., Wadsworth/Thomson Learning, 2001.
[36] W.B. Powell, Approximate Dynamic Programming: Solving the Curses of Dimensionality, Wiley-Interscience, 2007.
[37] C. Serrano-Cinca, B. Gutierrez-Nieto, Partial least square discriminant analysis for bankruptcy prediction, Decis. Support Syst. 54 (2013) 1245–1255.
[38] M. Srinivas, L.M. Patnaik, Genetic algorithms: a survey, IEEE Comput. 27 (6) (1994) 17–26.
[39] T.-C. Tang, L.-C. Chi, Neural networks analysis in business failure prediction of Chinese importers: a between-countries approach, Expert Syst. Appl. 29 (2005) 244–255.
[40] C.-F. Tsai, Financial decision support using neural networks and support vector machines, Expert Syst. 25 (2008) 380–393.
[41] C.-F. Tsai, Feature selection in bankruptcy prediction, Knowl.-Based Syst. 22 (2) (2009) 120–127.
[42] J. Wang, A.-R. Hedar, S. Wang, J. Ma, Rough set and scatter search metaheuristic based feature selection for credit scoring, Expert Syst. Appl. 39 (2012) 6123–6128.
[43] G. Wang, J. Ma, L. Huang, K. Xu, Tow credit scoring models based on dual strategy ensemble trees, Knowl.-Based Syst. 26 (2012) 61–68.
[44] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowl. Inf. Syst. 14 (2008) 1–37.
[45] Z. Yang, W. You, G. Ji, Using partial least squares and support vector machines for bankruptcy prediction, Expert Syst. Appl. 38 (7) (2011) 8336–8342.