Knowl Inf Syst
DOI 10.1007/s10115-017-1061-1

REGULAR PAPER

A novel classifier ensemble approach for financial distress prediction

Deron Liang[1] · Chih-Fong Tsai[2] · An-Jie Dai[1] · William Eberle[3]

Received: 11 April 2016 / Revised: 21 April 2017 / Accepted: 28 April 2017

Abstract Financial distress prediction is very important to financial institutions who must be able to make critical decisions regarding customer loans. Bankruptcy prediction and credit scoring are the two main aspects considered in financial distress prediction. To assist in this determination, thereby lowering the risk borne by the financial institution, it is necessary to develop effective prediction models for prediction of the likelihood of bankruptcy and estimation of credit risk. A number of financial distress prediction models have been constructed, which utilize various machine learning techniques, such as single classifiers and classifier ensembles, but improving the prediction accuracy is the major research issue. In addition, aside from improving the prediction accuracy, there have been very few studies that specifically consider lowering the Type I error. In practice, Type I errors need to receive careful consideration during model construction because they can affect the cost to the financial institution. In this study, we introduce a classifier ensemble approach designed to reduce the misclassification cost. The outputs produced by multiple classifiers are combined by utilizing the unanimous voting (UV) method to find the final prediction result. Experimental results obtained based on four relevant datasets show that our UV ensemble approach outperforms the baseline single classifiers and classifier ensembles. Specifically, the UV ensemble not only provides relatively good prediction accuracy and minimizes Type I/II errors, but also produces the smallest misclassification cost.

Keywords Financial distress prediction · Bankruptcy prediction · Credit scoring · Machine learning · Classifier ensembles · Type I error

✉ Chih-Fong Tsai
cftsai@mgt.ncu.edu.tw

1 Department of Computer Science and Information Engineering, National Central University, Taoyuan City, Taiwan

2 Department of Information Management, National Central University, Taoyuan City, Taiwan

3 Department of Computer Science, Tennessee Technological University, Cookeville, TN, USA

Published online: 11 May 2017

1 Introduction

Financial distress prediction is a necessary capability for financial institutions. In general, the aim is to predict the risk to the institution related to customer loans. Prediction models should be able to accurately identify the potential for failure of the customer applying for the loan, whether an individual or a company. Specifically, bankruptcy prediction and credit scoring are the two most important aspects considered for the prediction of financial distress [24].

Bankruptcy prediction focuses on predicting the likelihood that loan receivers will become bankrupt, while determination of the credit score is important for credit evaluation to determine whether loan applicants should be classified into a high-risk or low-risk group. Both problems can be regarded as binary classification problems and can be solved with a variety of statistical and machine learning techniques [2,8,21,24,33].

One of the earliest works mentioned in the literature was conducted by Fitzpartrick [13] who evaluated company health, examining 13 financial ratios and 40 companies. Since then, many financial distress prediction models have been developed based on different statistical methods, including univariate analysis [3], discriminant analysis [1], and logistic regression [25].

Recently, machine learning techniques have been applied and have shown their superiority over traditional statistical methods. These methods have included support vector machines (SVM) [30], decision trees (DT) [23], and artificial neural networks (ANN) [26,31,37].

It has been proposed that a combination of multiple classifiers (or classifier ensembles) be utilized to improve on the performance of single classifiers in machine learning [19]. This has become one of the main focuses in machine learning [10]. Simply speaking, classifier ensembles are solved based on the "divide-and-conquer" principle, in which a complex problem is divided into subproblems, and each subproblem is solved by a specific classification technique. Then, the outputs generated by multiple classifiers are combined to obtain the final result.

Classifier ensembles, including SVM ensembles [27], DT ensembles [43], and ANN ensembles [38], have also been constructed for financial distress prediction that offer better performance than single classifiers in terms of prediction accuracy.

Despite the improvement in the performance of classifier ensembles, in terms of their average classification (or prediction) accuracy, there have been very few studies which have focused specifically on dealing with Type I and II errors. Type I error, or a false positive, occurs when the classifier incorrectly classifies a bankrupt firm (or member of the high-risk group) as part of the non-bankrupt class (or the low-risk group). Type II error, on the other hand, occurs when the classifier incorrectly classifies a non-bankrupt firm into the bankrupt class. A larger Type I error rate results in higher costs to financial institutions and enhances the enterprise risk. Thus, in practice, decision makers view the Type I error as far more serious than the Type II error [28].

In related studies, the accuracy of the prediction rate is generally based on averaging the Type I and II errors; however, it may be that the classifier that provides the best prediction accuracy has a lower Type II error but higher Type I error, while the classifier that provides the second best prediction accuracy might have the lowest Type I error.

To improve prediction accuracy as well as lower the Type I error rate, we propose a novel classifier ensemble approach, called the unanimous voting (UV) ensemble, which is based on the application of the unanimous voting method to combine the outputs obtained from multiple classifiers. In particular, the UV ensemble produces the final prediction output by choosing the class for which all classifiers agree, whereas the conventional classifier ensembles employ a majority voting method, choosing the class which receives the largest number of votes.

The rest of this paper is organized as follows: Section 2 overviews classifier ensemble techniques and reviews related studies on financial distress prediction. In addition, the limitations of related works are discussed. Section 3 describes the proposed UV ensemble approach, while the experimental setup and results are described in Sect. 4. Finally, some conclusions are offered in Sect. 5.

2 Literature review

2.1 Classifier ensembles

The classifier ensemble is developed by combining a number of classifiers. Ensemble classifiers aim at obtaining highly accurate classifications by combining less accurate ones. They are designed to improve on the classification performance of a single classifier [19]. In other words, the combination is able to make up for the errors made by the individual classifiers on different parts of the input space. As a consequence, the performance of classifier ensembles is likely to be better than that of the best single classifier used in isolation [4]. The four combination methods most widely used for developing classifier ensembles are described below.

2.1.1 Majority voting

The simplest method for combining multiple classifiers is by majority voting. In the case of bankruptcy prediction and credit scoring, the binary outputs of k individual classifiers are pooled together. Then, the class that receives the largest number of votes is selected as the final classification decision. In general, the final classification decision that reaches the majority of $\frac{k+1}{2}$ votes is taken.

2.1.2 Bagging

In bagging, several classifiers are trained independently on different training sets using the bootstrap method [6]. Bootstrapping builds k replicated training datasets to construct k independent classifiers by randomly re-sampling the original given training dataset, but with replacements. Each training example may appear repeatedly or not at all in any particular replicated training dataset of k. Then, the k classifiers are aggregated through an appropriate combination method, such as majority voting.

2.1.3 Boosting

In boosting, like bagging, each classifier is trained using a different training set. However, the k classifiers are trained sequentially rather than in parallel and independently (as in bagging). Adaptive Boosting (AdaBoost) is the most representative boosting approach.

In AdaBoost, each example of a given training set has the same weight; n sets of training samples among S are used to train the kth classifier. The trained classifier is then evaluated by comparison of S in order to identify those examples which cannot be classified correctly. The $k + 1$ classifier is then trained using a modified training set which boosts the importance of those incorrectly classified examples. This sampling procedure is repeated until K training samples are built for constructing the Kth classifier. The final decision is based on the weighted vote of the individual classifiers [14].

2.1.4 Stacking

Stacking or stacked generalization [39] is generally based on constructing multi-level classifiers in a hierarchical way. The first level is composed of several single classifiers with the outputs produced by the first-level classifiers used to train second level of the 'stacked' classifier. The final resultant decision is made based on the output of the stacked classifier. Unlike the previously mentioned combination methods, such as majority voting, which is a 'static' combiner, the stacked classifier is a 'trainable' combiner. In other words, it is a scheme for estimating the errors of a classifier when working on a particular learning dataset and then correcting those errors.

2.2 Discussion of the limitations of related work

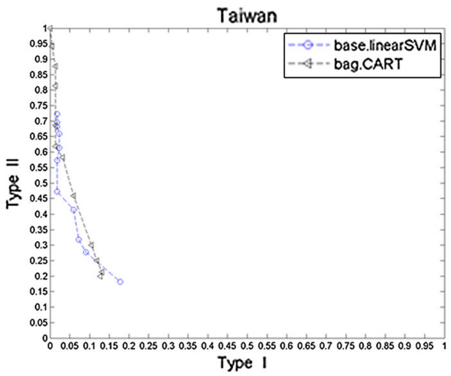
Table 1 shows a comparison between related works based upon classifier ensemble techniques, datasets, and evaluation metrics.

This comparison indicates some of the limitations of past studies. First, although it has been concluded that classifier ensembles can provide better performance than single classifiers, many of them only constructed specific classifier ensembles without making a comprehensive comparison. In Wu et al. [40], the top 10 data mining algorithms are identified. Of these, five are supervised learning-based classification techniques, namely the SVM, KNN, MLP neural network, CART decision tree, and naïve Bayes. However, more reliable conclusions could be reached by comparison of single classifiers and classifier ensembles for all five of these techniques. In this current study, these five single classifiers are constructed as the

Table 1 Comparison of related works

Work	Classifier ensemble	Datasets	Evaluation metrics
Geng [15]	MV[a] of C4.5[b], MLP[c], and SVM	Chinese	Accuracy/Precision/ Recall
Heo and Yang [16]	Boosting C4.5	Korean	Accuracy/Type I & II errors
Wang and Ma [35]	Bagging SVM	Chinese	Accuracy/Type I & II errors
Kim et al. [18]	Bagging CART[d]	German	Accuracy
Lei et al. [22]	Boosting C4.5	German	Accuracy
Shi et al. [29]	Bagging MLP	German	Accuracy
Wang et al. [36]	Boosting SVM	Australian/German	Accuracy
Wang et al. [34]	Bagging SVM/Bagging LR[e]/Boosting SVM	Australian/German	Accuracy/Type I error
West et al. [38]	Bagging MLP	Australian	Accuracy
Yao [42]	Boosting CART	Australian	Accuracy

[a] *MV* majority voting
[b] *C4.5* C4.5 decision tree
[c] *MLP* multilayer perceptron neural network
[d] *CART* classification and regression decision tree
[e] *LR* logistic regression



(a) (b)

Fig. 1 An example of the ROC curves over the Taiwan dataset. **a** Results on specific single points, **b** results on full curves

single baseline classifiers. The majority voting, bagging, boosting, and stacking methods are also used to develop classifier ensembles for comparison based on the five classification techniques.

Second, there have been very few studies where two or more relevant datasets are considered for the validation of performance. Many have used the publicly available Australian[1] and German[2] credit datasets, making data collection and direct comparison across different studies very easy. However, they are often only used for credit scoring problems. There are very few datasets related to the bankruptcy domain problem available. This may be because of the difficulty in collecting information on a sufficient number of bankruptcy cases or that the factors affecting bankruptcy vary between different countries. To make up for this insufficiency, in this paper we not only use the Australian and German datasets, but also collect bankruptcy datasets from China and Taiwan. This should allow us to more fully examine the performance of the constructed classifiers in terms of financial distress prediction.

Third, as discussed previously, classifiers with lower Type I errors allow financial institutions to reduce the risk of spending too much. However, very few past studies have assessed the Type I error of their prediction models. Moreover, the cost ratios and the cost of misclassification [5] have seldom been examined even though they are indicative of the cost of Type I and II errors. Therefore, in addition to examining the prediction accuracy and the Type I/II errors, the cost ratios and misclassification costs of the constructed prediction models are also assessed.

Moreover, the receiver operating characteristic curve (ROC curve) [12], which is a widely used evaluation metric in many pattern recognition problems, is also considered. The ROC curve is a graphical plot that illustrates the performance of a binary classifier as its discrimination threshold is varied. Figure 1 shows examples with a single SVM (base.linearSVM) and a bagging-based DT ensemble (bag.CART).

In our comparison of the average prediction accuracy and Type I/II errors of the two classifiers, we focus on specific points on the ROC curves (as shown in Fig. 1a). In this case, bag.CART performs better than base.linearSVM in terms of prediction accuracy and Type I error. However, examining the full ROC curves of these two classifiers allows us to

1 http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval).

2 http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

determine their performance given different Type I and II conditions. Simply speaking, in a comparison of the two ROC curves, the classifier producing a smaller area under the curve can be said to perform better than another one having a larger area under its corresponding curve. Therefore, the ROC curves show that base.linearSVM outperforms bag.CART.

3 The UV ensemble approach

3.1 Unanimous voting

The UV ensemble approach is based on utilization of the classifier ensemble technique to combine the outputs of multiple classifiers. Figure 2 shows the process of the UV ensemble approach.

The given dataset, composed of data for several different case companies, is divided into the training and testing datasets, based on the tenfold cross-validation strategy [20]. Specifically, the dataset is divided into 10 distinct subsets (or folds) where ninefolds are used as the training sets and the remaining fold for the testing set. Consequently, a classifier is trained and tested 10 times over ten different combinations of training and testing sets. The classifier's performance is based on averaging the 10 different results.

In Fig. 2, it can be seen that k different classifiers (i.e., k base learners) are trained and tested and their outputs are combined based on the UV method to obtain the final prediction result. This result is chosen based on the class on which all classifiers agree. For example, when a testing case and three classifiers are combined (represented by k_1, k_2, and k_3) four possible combination results can be produced (as shown in Table 2).

The UV method is used for lowering the Type I error, meaning that if one of the combined classifiers outputs the class '1', then the final result is '1' for the testing case. Only when

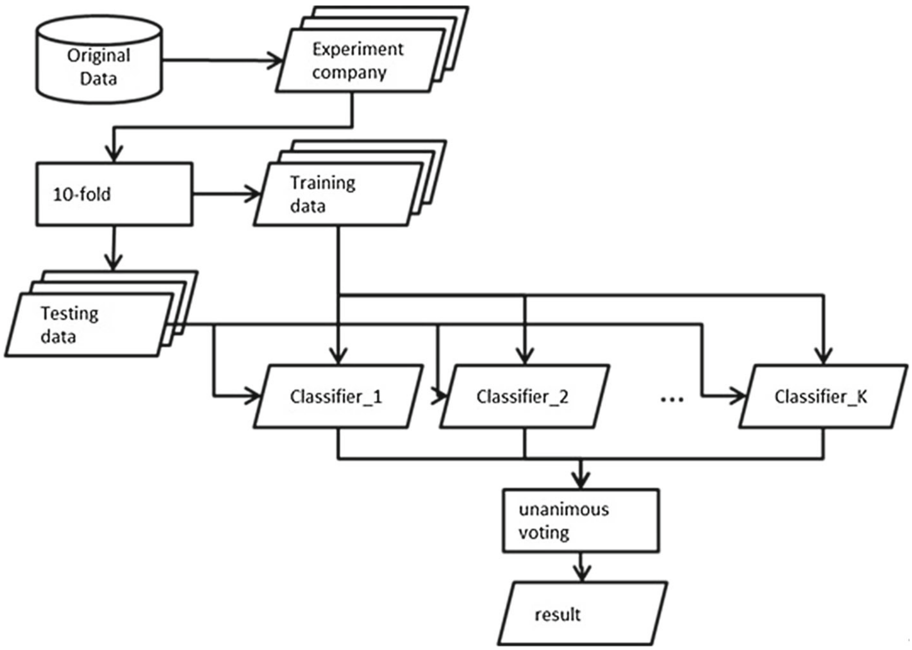


Fig. 2 UV ensemble approach

Table 2 Four possible results produced by the unanimous voting (UV) and majority voting (MV) methods (1 represents bankrupt and 0 non-bankrupt)

	k_1	k_2	k_3	UV	MV
1	0	0	1	1	0
2	1	1	0	1	1
3	1	1	1	1	1
4	0	0	0	0	0

Table 3 Prediction performance of 15 different classifiers over four datasets (accuracy/Type I error; unit: percentage)

	SVM	KNN	MLP	CART	Bayes
Taiwanese dataset					
Single	82.05/17.73	77.27/22.27	76.36/26.36	79.77/15.91	76.36/14.55
Bagging	82.50/16.36	76.59/21.36	83.64/13.64	83.86/14.55	75.91/13.18
Boosting	79.55/18.64	72.05/30.45	77.50/13.64	83.86/14.55	71.59/28.64
Chinese dataset					
Single	91.27/7.00	90.37/7.32	89.38/9.88	92.71/5.83	88.82/6.39
Bagging	91.27/7.00	90.24/7.60	91.42/6.41	92.42/6.42	89.11/6.11
Boosting	89.97/9.04	83.00/15.69	85.60/8.44	92.85/6.42	83.10/17.52
Australian dataset					
Single	85.54/20.01	84.94/14.80	83.05/16.66	84.96/18.18	68.24/22.73
Bagging	85.54/20.01	84.80/15.06	85.67/14.55	86.25/14.55	68.97/23.77
Boosting	84.51/13.78	77.40/21.34	79.72/15.84	86.39/12.99	69.12/22.21
German dataset					
Single	77.60/49.67	69.40/66.33	71.40/61.00	73.20/63.00	70.70/96.00
Bagging	77.20/50.33	70.40/65.33	76.50/59.00	75.80/56.67	70.70/95.67
Boosting	76.60/50.00	66.20/54.00	72.50/42.33	74.50/51.33	73.60/50.67

all of the classifiers output class '0' and agree that the testing case is non-bankrupt, the final result will be '0'. In contrast, the final result from the MV method is based on choosing the class which receives the largest number of votes.

Intuitively, although the UV method can improve on the incidence of Type II error, it may not improve the average prediction accuracy, because the Type I error is certainly low. However, if several best base learners that provide relatively good performance are combined, this limitation can certainly be remedied. The process of choosing the base learners is discussed in the next subsection. Our experimental results (c.f. Sect. 4) demonstrate that the UV ensemble approach is particularly suitable for financial distress prediction, offering optimal performance in terms of prediction accuracy, Type I/II errors, misclassification costs, and ROC curves.

3.2 The base learners

As discussed in Kittler et al. [19], the combination of diverse classifiers should provide better combination results. Thus, the selection of the base learners for the classifiers to be combined is critical for the success of the classifier ensemble. Table 3 shows the initial experimental

Table 4 Dataset information

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

results for three types of classifiers including single-, bagging-, and boosting-based classifiers. Note that the experimental setup is described in more detail in Sect. 4.1.

For our UV ensemble method, we choose the top three diversified classifiers providing the best performance over the four datasets as the base learners. They are SVM, bagging-based MLP, and boosting-based CART.

It should be noted that choosing weak classifiers, which provide less accurate prediction rates for the base learners, could affect the final performance of the UV method. However, since the goal is to make a UV ensemble approach, which outperforms the baseline classifiers, the performance of ensembles comprised of the best and weaker base learners is not considered here. This is because the computational cost of constructing multiple classifiers by the UV ensemble approach is much larger than for the baseline classifiers. For considerations of prediction accuracy and computational cost, combining only the top diversified classifiers as the base learners is the best strategy to see whether the UV ensemble approach is suitable for the financial distress prediction problem.

4 Experiments

4.1 Experimental setup

4.1.1 Databases

The key characteristics of the four chosen datasets utilized in this study are presented in Table 4. The Australian and German datasets are public datasets, which are widely used in the literature. The Taiwanese and Chinese datasets were collected from the Taiwan Economic Journal,[3] and company bankruptcy is defined based on the business regulations of the Taiwan Stock Exchange and Shanghai and Shenzhen Stock Exchange.

Here, the method of stratified sampling [1] is used to collect the same number 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, a tenfold cross-validation strategy is used to divide each dataset into 10 distinct training and testing subsets.

4.1.2 The classifiers

Five single classification techniques are used, SVM, KNN, CART, MLP, and Naïve Bayes. For the classifier ensembles, bagging, boosting, stacking, and majority voting are employed. Table 5 lists the parameters for constructing these classifiers for comparison.

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

Table 5 Parameters for constructing single classifiers and classifier ensembles

Classifier	Parameters
Single classifiers	
SVM	Kernel function: linear kernel; other related parameters are based on the default parameters from the LIBSVM toolbox[a]
KNN	The default parameters used are based on the MATLAB toolbox
CART	The default parameters used are based on the MATLAB toolbox
MLP	The numbers of hidden nodes: 8/16/32/64; learning epochs: 50/100/200/400 [32]
Naïve Bayes	Kernel function: kernel density estimate [17]
Classifier ensembles	
Bagging	Bagging SVM/KNN/CART/MLP/Bayes are constructed; numbers of bootstrap: 25 [7]
Boosting	Boosting SVM/KNN/CART/MLP/Bayes are constructed by Adaboost.M1; number of iterations: 50 [41]
Stacking	Base learners: linearSVM, bag.MLP, and boost.CART (the same as our UV ensemble); meta learner: linearSVM
Majority voting	Base learners: linearSVM, bagMLP, and boost.CART (the same as for our UV ensemble)

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

Table 6 Confusion matrix

↓actual\predicted→	Bankruptcy	Non-bankruptcy
Bankruptcy	(a)	(b)
Non-bankruptcy	(c)	(d)

4.1.3 Evaluation metrics

The four evaluation metrics used to assess the performance of the aforementioned classifiers (i.e., prediction models) are prediction accuracy, Type I/II errors, misclassification cost, and the ROC curve. The first three can be measured using a confusion matrix, as given in Table 6. For a description of the ROC curve, please refer to Sect. 2.2.

Therefore, the average prediction accuracy is obtained by

$$\text{Prediction accuracy} = \frac{a+d}{a+b+c+d} \tag{1}$$

and the Type I/II errors are based on

$$\text{Type I error} = \frac{b}{a+b}, \tag{2}$$

$$\text{Type II error} = \frac{c}{c+d}. \tag{3}$$

On the other hand, the cost for misclassification can be obtained by taking the

$$(\text{Type I error} \times \text{bankrupt firms} \times \text{cost ratio}) + (\text{Type II error} \times \text{non-bankrupt firms}). \tag{4}$$

For example, a cost ratio of 20 indicates that a Type I (II) error is considered to be twenty times more costly than a Type II (I) error. If the cost ratio is one, the errors are equally

Table 7 Prediction accuracy and Type I/II errors for different classifiers over the Taiwan dataset

Taiwan	Cost ratios/penalties									
	1 (%)	2 (%)	3 (%)	5 (%)	7 (%)	10 (%)	15 (%)	20 (%)	25 (%)	30 (%)
UV										
Acc	83.18	**83.18**	**81.36**	**75.00**	**75.00**	**75.00**	**75.00**	**75.00**	**65.23**	**65.23**
Type I	7.27	**7.27**	**5.00**	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.00**	**0.00**
Type II	26.36	**26.36**	**32.27**	**49.09**	**49.09**	**49.09**	**49.09**	**49.09**	**69.55**	**69.55**
Stacking										
Acc	**83.64**	83.64	83.64	83.64	53.64	53.64	53.64	53.64	53.64	53.64
Type I	**13.64**	13.64	13.64	13.64	1.36	1.36	1.36	1.36	1.36	1.36
Type II	**19.09**	19.09	19.09	19.09	91.36	91.36	91.36	91.36	91.36	91.36
MV										
Acc	**84.32**	**84.32**	**84.32**	84.32	84.32	84.32	84.32	84.32	84.32	84.32
Type I	**9.09**	**9.09**	**9.09**	9.09	9.09	9.09	9.09	9.09	9.09	9.09
Type II	**22.27**	**22.27**	**22.27**	22.27	22.27	22.27	22.27	22.27	22.27	22.27
Single										
linearSVM										
Acc	82.05	81.59	75.45	75.45	75.45	75.45	75.45	75.45	75.45	75.45
Type I	17.73	9.09	1.82	1.82	1.82	1.82	1.82	1.82	1.82	1.82
Type II	18.18	27.73	47.27	47.27	47.27	47.27	47.27	47.27	47.27	47.27
KNN										
Acc	77.50	75.68	75.68	75.68	75.68	64.77	64.77	64.77	64.77	64.77
Type I	12.27	4.55	4.55	4.55	4.55	1.82	1.82	1.82	1.82	1.82
Type II	32.73	44.09	44.09	44.09	44.09	68.64	68.64	68.64	68.64	68.64
MLP										
Acc	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36
Type I	26.36	26.36	26.36	26.36	26.36	26.36	26.36	26.36	26.36	26.36
Type II	20.91	20.91	20.91	20.91	20.91	20.91	20.91	20.91	20.91	20.91
CART										
Acc	79.77	79.77	69.55	65.68	65.68	65.68	50.00	50.00	50.00	50.00
Type I	15.91	15.91	5.00	2.73	2.73	2.73	0.00	0.00	0.00	0.00
Type II	24.55	24.55	55.91	65.91	65.91	65.91	100.00	100.00	100.00	100.00
Bayes										
Acc	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36	76.36
Type I	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55
Type II	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73	32.73
Bag										
LinearSVM										
Acc	82.50	80.23	75.68	**75.68**	75.68	75.68	75.68	75.68	**75.68**	**75.68**
Type I	16.36	8.64	2.27	**2.27**	2.27	2.27	2.27	2.27	**2.27**	**2.27**
Type II	18.64	30.91	46.36	**46.36**	46.36	46.36	46.36	46.36	**46.36**	**46.36**

Table 7 continued

Taiwan	Cost ratios/penalties									
	1 (%)	2 (%)	3 (%)	5 (%)	7 (%)	10 (%)	15 (%)	20 (%)	25 (%)	30 (%)
KNN										
Acc	78.18	76.14	76.14	76.14	76.14	67.05	67.05	67.05	67.05	67.05
Type I	11.82	4.55	4.55	4.55	4.55	2.27	2.27	2.27	2.27	2.27
Type II	31.82	43.18	43.18	43.18	43.18	63.64	63.64	63.64	63.64	63.64
MLP										
Acc	**83.64**	83.64	83.64	83.64	83.64	83.64	83.64	83.64	83.64	83.64
Type I	**13.64**	13.64	13.64	13.64	13.64	13.64	13.64	13.64	13.64	13.64
Type II	**19.09**	19.09	19.09	19.09	19.09	19.09	19.09	19.09	19.09	19.09
CART										
Acc	**83.86**	**83.64**	**83.64**	70.23	67.73	67.73	67.73	60.23	50.00	50.00
Type I	**14.55**	**10.91**	**10.91**	2.73	1.82	1.82	1.82	0.91	0.00	0.00
Type II	**17.73**	**21.82**	**21.82**	56.82	62.73	62.73	62.73	78.64	100.00	100.00
Bayes										
Acc	75.91	75.91	75.91	75.91	75.91	75.91	75.91	75.91	75.91	75.91
Type I	13.18	13.18	13.18	13.18	13.18	13.18	13.18	13.18	13.18	13.18
Type II	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00	35.00
Boost										
LinearSVM										
Acc	79.77	79.77	75.23	72.05	72.05	72.05	67.95	67.95	67.95	67.95
Type I	15.45	11.82	5.91	3.18	3.18	3.18	2.27	2.27	2.27	2.27
Type II	25.00	28.64	43.64	52.73	52.73	52.73	61.82	61.82	61.82	61.82
KNN										
Acc	75.23	75.00	75.00	72.95	72.95	72.95	72.95	72.95	72.95	72.95
Type I	21.36	7.73	7.73	5.91	5.91	5.91	5.91	5.91	5.91	5.91
Type II	28.18	42.27	42.27	48.18	48.18	48.18	48.18	48.18	48.18	48.18
MLP										
Acc	77.50	77.50	77.50	77.50	77.50	77.50	77.50	77.50	77.50	77.50
Type I	13.64	13.64	13.64	13.64	13.64	13.64	13.64	13.64	13.64	13.64
Type II	31.36	31.36	31.36	31.36	31.36	31.36	31.36	31.36	31.36	31.36
CART										
Acc	**83.86**	83.86	83.86	83.86	83.86	83.86	83.86	83.86	83.86	83.86
Type I	**14.55**	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55
Type II	**17.73**	17.73	17.73	17.73	17.73	17.73	17.73	17.73	17.73	17.73
Bayes										
Acc	71.59	71.59	71.59	71.59	71.59	71.59	71.59	71.59	71.59	71.59
Type I	28.64	28.64	28.64	28.64	28.64	28.64	28.64	28.64	28.64	28.64
Type II	28.18	28.18	28.18	28.18	28.18	28.18	28.18	28.18	28.18	28.18

The bold-faced numbers mean that they are significantly different from the others ($p < 0.01$)

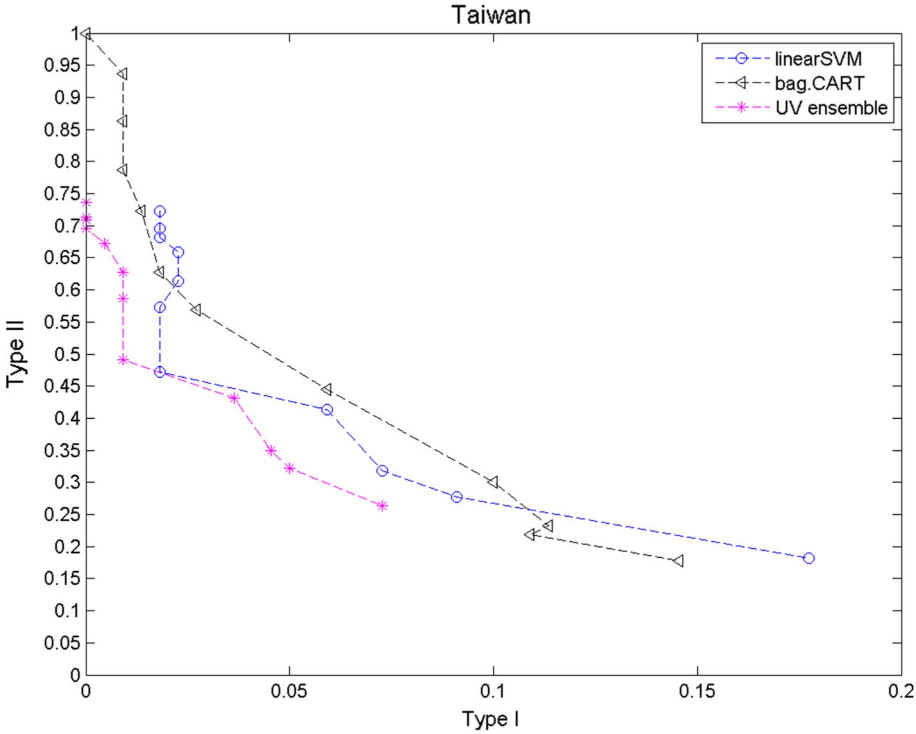


Fig. 3 ROC curves for UV ensemble, linear.SVM and bag.CART over the Taiwan dataset

costly. In this study, different cost ratios are examined, ranging from 1 to 30. Moreover, each classifier is tested using different penalty weights (from 1 to 30), in order to fine-tune the complexity of the prediction model during cross-validation [11].

4.2 Experimental results

4.2.1 Results on the Taiwan dataset

Table 7 shows the prediction accuracy and Type I/II errors for the different classifiers using different cost ratios and the Taiwan dataset. The bold-faced numbers indicate the misclassification costs of those classifiers. For example, when the cost ratio is 1, the stacking, MV, bag.MLP, bag.CART, and boost.CART methods perform significantly better than the other classifiers. Note that the level of significance is measured by the Wilcoxon test [9].

According to these results, our UV ensemble approach outperforms the other methods for cost ratios from 2 to 30. Other better classifiers are MV and bag.CART for cost ratios from 1 to 3 and bag.linearSVM for cost ratios of 5, 25, and 30. However, the good performance of these approaches does not hold for a long range of cost ratios. The results indicate that the UV ensemble offers a more stable performance than the other classifiers, with the least misclassification cost over various cost ratios.

Figure 3 shows the ROC curves for the UV ensemble, linear.SVM (the best single classifier) and bag.CART (the best classifier ensemble) for further comparison. We can observe that the UV ensemble produces the smallest area under the curve. In other words, our UV ensemble approach offers the best performance.

Table 8 Prediction accuracy and Type I/II errors of different classifiers over the China dataset

China	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
UV										
Acc	91.27	90.83	**90.83**	**90.83**	**90.83**	**90.83**	**89.53**	**72.96**	**72.96**	**72.96**
Type I	4.66	3.50	**3.50**	**3.50**	**3.50**	**3.50**	**3.22**	**1.47**	**1.47**	**1.47**
Type II	12.79	14.83	**14.83**	**14.83**	**14.83**	**14.83**	**17.73**	**52.61**	**52.61**	**52.61**
Stacking										
Acc	91.57	91.57	91.57	91.57	91.57	91.57	91.57	49.85	49.85	49.85
Type I	6.12	6.12	6.12	6.12	6.12	6.12	6.12	0.29	0.29	0.29
Type II	10.74	10.74	10.74	10.74	10.74	10.74	10.74	100.00	100.00	100.00
MV										
Acc	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00	92.00
Type I	5.84	5.55	5.55	5.55	5.55	5.55	5.55	5.55	5.55	5.55
Type II	10.16	10.45	10.45	10.45	10.45	10.45	10.45	10.45	10.45	10.45
Single										
LinearSVM										
Acc	91.27	90.83	90.83	90.83	90.83	90.83	89.37	**80.80**	**80.80**	72.67
Type I	7.00	4.97	4.97	4.97	4.97	4.97	4.69	**3.52**	**3.52**	2.93
Type II	10.45	13.37	13.37	13.37	13.37	13.37	16.56	**34.88**	**34.88**	51.73
KNN										
Acc	90.39	90.39	88.94	88.94	88.94	88.94	88.94	80.34	80.34	80.34
Type I	5.55	5.55	4.09	4.09	4.09	4.09	4.09	2.93	2.93	2.93
Type II	13.67	13.67	18.03	18.03	18.03	18.03	18.03	36.39	36.39	36.39
MLP										
Acc	89.38	89.38	89.38	89.38	89.38	89.38	89.38	89.38	89.38	89.38
Type I	9.88	9.88	9.88	9.88	9.88	9.88	9.88	9.88	9.88	9.88
Type II	11.36	11.36	11.36	11.36	11.36	11.36	11.36	11.36	11.36	11.36
CART										
Acc	**93.00**	**93.00**	**93.00**	**92.71**	**92.71**	**92.71**	**92.71**	50.00	50.00	50.00
Type I	**5.25**	**4.96**	**4.96**	**4.67**	**4.67**	**4.67**	**4.67**	0.00	0.00	0.00
Type II	**8.75**	**9.04**	**9.04**	**9.90**	**9.90**	**9.90**	**9.90**	100.00	100.00	100.00
Bayes										
Acc	88.82	88.82	88.82	88.82	88.82	88.82	88.82	88.82	88.82	88.82
Type I	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39	6.39
Type II	15.97	15.97	15.97	15.97	15.97	15.97	15.97	15.97	15.97	15.97
Bag										
LinearSVM										
Acc	91.27	90.68	90.68	89.81	89.81	88.51	88.51	**88.51**	74.70	67.44
Type I	7.00	5.26	5.26	4.69	4.69	4.39	4.39	**4.39**	3.23	2.64
Type II	10.45	13.37	13.37	15.69	15.69	18.59	18.59	**18.59**	47.38	62.48

Table 8 continued

China	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
KNN										
Acc	90.24	90.24	89.08	89.08	89.08	89.08	89.08	89.08	89.08	82.09
Type I	5.56	5.56	4.09	4.09	4.09	4.09	4.09	4.09	4.09	3.51
Type II	13.96	13.96	17.74	17.74	17.74	17.74	17.74	17.74	17.74	32.31
MLP										
Acc	91.42	91.42	91.42	91.42	91.42	91.42	91.42	91.42	91.42	91.42
Type I	6.41	6.41	6.41	6.41	6.41	6.41	6.41	6.41	6.41	6.41
Type II	10.74	10.74	10.74	10.74	10.74	10.74	10.74	10.74	10.74	10.74
CART										
Acc	**92.86**	**92.86**	**92.86**	**92.86**	**92.86**	**90.70**	**90.70**	90.70	52.35	52.35
Type I	**4.96**	**4.96**	**4.96**	**4.96**	**4.96**	**4.38**	**4.38**	4.38	0.59	0.59
Type II	**9.33**	**9.33**	**9.33**	**9.33**	**9.33**	**14.23**	**14.23**	14.23	94.71	94.71
Bayes										
Acc	89.11	89.11	89.11	89.11	89.11	89.11	89.11	89.11	89.11	89.11
Type I	6.11	6.11	6.11	6.11	6.11	6.11	6.11	6.11	6.11	6.11
Type II	15.67	15.67	15.67	15.67	15.67	15.67	15.67	15.67	15.67	15.67
Boost										
LinearSVM										
Acc	91.13	91.13	91.13	89.24	86.77	86.77	86.77	86.77	71.94	71.94
Type I	5.83	5.83	5.83	4.69	3.82	3.82	3.82	3.82	2.34	2.34
Type II	11.92	11.92	11.92	16.82	22.65	22.65	22.65	22.65	53.77	53.77
KNN										
Acc	83.88	83.88	83.88	83.88	83.88	83.88	83.88	83.88	83.88	83.88
Type I	13.38	13.38	13.38	13.38	13.38	13.38	13.38	13.38	13.38	13.38
Type II	18.87	18.87	18.87	18.87	18.87	18.87	18.87	18.87	18.87	18.87
MLP										
Acc	85.60	85.60	85.60	85.60	85.60	85.60	85.60	85.60	85.60	85.60
Type I	8.44	8.44	8.44	8.44	8.44	8.44	8.44	8.44	8.44	8.44
Type II	20.36	20.36	20.36	20.36	20.36	20.36	20.36	20.36	20.36	20.36
CART										
Acc	**92.85**	**92.85**	**92.85**	92.85	92.85	92.85	92.85	92.85	92.85	92.85
Type I	**6.42**	**6.42**	**6.42**	6.42	6.42	6.42	6.42	6.42	6.42	6.42
Type II	**7.87**	**7.87**	**7.87**	7.87	7.87	7.87	7.87	7.87	7.87	7.87
Bayes										
Acc	83.10	83.10	83.10	83.10	83.10	83.10	83.10	83.10	83.10	83.10
Type I	17.52	17.52	17.52	17.52	17.52	17.52	17.52	17.52	17.52	17.52
Type II	16.28	16.28	16.28	16.28	16.28	16.28	16.28	16.28	16.28	16.28

The bold-faced numbers mean that they are significantly different from the others ($p < 0.01$)

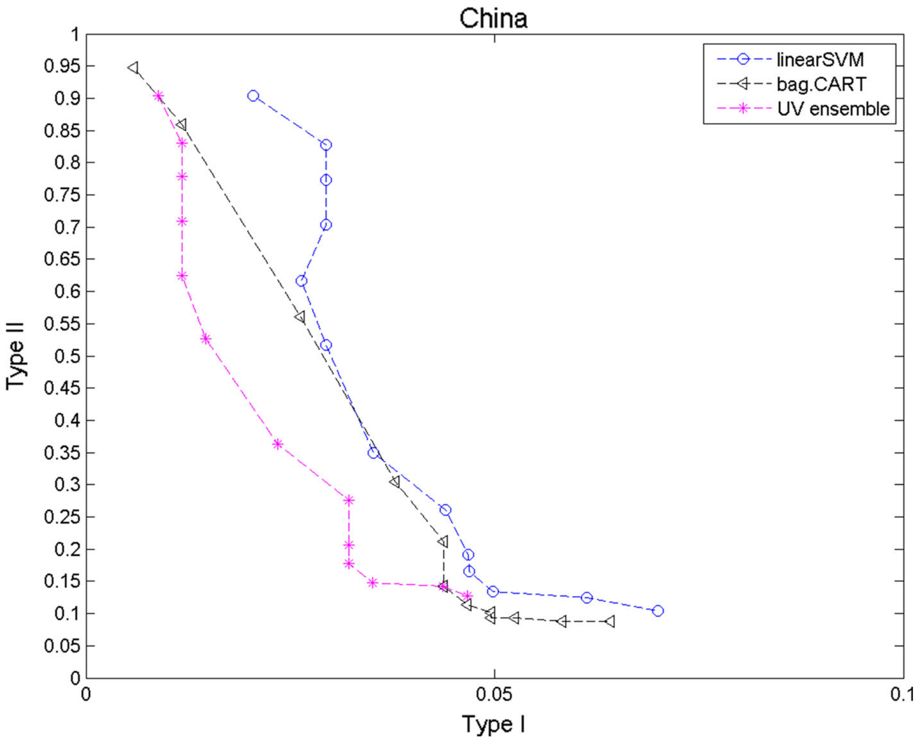


Fig. 4 ROC curves for UV ensemble, linear.SVM and bag.CART over the China dataset

4.2.2 Results for the China dataset

Table 8 illustrates the performance of single classifiers and classifier ensembles over the China dataset. As we can see, the UV ensemble, CART, and bag.CART perform better than the other methods. In particular, the UV ensemble performs very well for a range of cost ratios from 3 to 30, while CART and bag.CART perform better for cost ratios from 1 to 15.

Figure 4 shows the ROC curves for the UV ensemble, linearSVM, and bag.CART methods. Further examination shows that the UV ensemble performs better than the other methods for various misclassification costs. The only exception is when the Type I error is 0.05 or less, in which case bag.CART performs better than the UV ensemble method. However, the differences in performance are very small. Therefore, generally speaking, the UV ensemble, CART, and bag.CART are all good choices for the China dataset.

4.2.3 Results for the Australian dataset

Table 9 shows the performance results for different classifiers over the Australian dataset. Like the results for the Taiwan and China datasets, the UV ensemble method outperforms most of the other classifiers for cost ratios in the range of 2–30, while linearSVM provides good performance when the cost ratios range from 10 to 30. The performance of bag.linearSVM and boost.linearSVM is similar to that of linearSVM. Another classifier that performs relatively well is bag.CART when the cost ratios are 3 and from 15 to 30.

Table 9 Prediction accuracy and Type I/II errors of different classifiers over the Australian dataset

Australian	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
UV										
Acc	86.37	**84.77**	**82.31**	**82.31**	**82.31**	**79.41**	**79.41**	**56.38**	**56.38**	**56.38**
Type I	9.61	**6.21**	**3.11**	**3.11**	**3.11**	**2.33**	**2.33**	**0.00**	**0.00**	**0.00**
Type II	18.61	**26.43**	**35.83**	**35.83**	**35.83**	**43.31**	**43.31**	**98.03**	**98.03**	**98.03**
Stacking										
Acc	85.67	85.67	85.67	65.24	65.24	65.24	65.24	65.24	65.24	65.24
Type I	14.55	14.55	14.55	3.62	3.62	3.62	3.62	3.62	3.62	3.62
Type II	14.02	14.02	14.02	73.56	73.56	73.56	73.56	73.56	73.56	73.56
MV										
Acc	86.52	**86.52**	86.52	86.52	86.52	86.52	86.52	86.52	86.52	86.52
Type I	9.87	**9.61**	9.61	9.61	9.61	9.61	9.61	9.61	9.61	9.61
Type II	17.95	**18.28**	18.28	18.28	18.28	18.28	18.28	18.28	18.28	18.28
Single										
LinearSVM										
Acc	85.54	82.45	79.55	79.55	79.55	**79.55**	**79.55**	**56.53**	**56.53**	**56.53**
Type I	20.01	5.72	2.59	2.59	2.59	**2.59**	**2.59**	**0.00**	**0.00**	**0.00**
Type II	7.53	32.25	42.66	42.66	42.66	**42.66**	**42.66**	**97.70**	**97.70**	**97.70**
KNN										
Acc	84.94	82.17	82.17	82.17	82.17	68.96	68.96	68.96	68.96	68.96
Type I	14.80	5.99	5.99	5.99	5.99	3.11	3.11	3.11	3.11	3.11
Type II	15.34	32.53	32.53	32.53	32.53	65.85	65.85	65.85	65.85	65.85
MLP										
Acc	83.05	83.05	83.05	83.05	83.05	83.05	83.05	83.05	83.05	83.05
Type I	16.66	16.66	16.66	16.66	16.66	16.66	16.66	16.66	16.66	16.66
Type II	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26	17.26
CART										
Acc	84.96	81.91	81.91	75.37	75.37	75.37	**70.69**	55.51	55.51	55.51
Type I	18.18	6.51	6.51	2.33	2.33	2.33	**1.55**	0.00	0.00	0.00
Type II	11.08	32.49	32.49	52.45	52.45	52.45	**63.98**	100.00	100.00	100.00
Bayes										
Acc	68.24	68.24	68.24	68.24	68.24	68.24	68.24	68.24	68.24	68.24
Type I	22.73	22.73	22.73	22.73	22.73	22.73	22.73	22.73	22.73	22.73
Type II	42.98	42.98	42.98	42.98	42.98	42.98	42.98	42.98	42.98	42.98
Bag										
LinearSVM										
Acc	85.54	**83.89**	83.89	79.85	79.85	75.65	75.65	**56.82**	**56.82**	**56.82**
Type I	20.01	**5.46**	5.46	3.12	3.12	2.07	2.07	**0.00**	**0.00**	**0.00**
Type II	7.53	**29.31**	29.31	41.34	41.34	52.08	52.08	**97.04**	**97.04**	**97.04**

Table 9 continued

Australian	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
KNN										
Acc	84.80	82.89	82.89	82.89	82.89	72.01	72.01	72.01	72.01	72.01
Type I	15.06	6.50	6.50	6.50	6.50	3.63	3.63	3.63	3.63	3.63
Type II	15.34	30.28	30.28	30.28	30.28	58.35	58.35	58.35	58.35	58.35
MLP										
Acc	85.67	85.67	85.67	85.67	85.67	85.67	85.67	85.67	85.67	85.67
Type I	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55	14.55
Type II	14.02	14.02	14.02	14.02	14.02	14.02	14.02	14.02	14.02	14.02
CART										
Acc	86.39	85.35	**83.18**	78.12	78.12	78.12	**78.12**	**56.97**	**56.97**	**56.97**
Type I	11.45	7.81	**5.46**	2.60	2.60	2.60	**2.60**	**0.00**	**0.00**	**0.00**
Type II	16.29	23.16	**30.97**	45.91	45.91	45.91	**45.91**	**96.73**	**96.73**	**96.73**
Bayes										
Acc	68.97	68.97	68.97	68.97	68.97	68.97	68.97	68.97	68.97	68.97
Type I	23.77	23.77	23.77	23.77	23.77	23.77	23.77	23.77	23.77	23.77
Type II	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03	40.03
Boost										
LinearSVM										
Acc	85.38	83.90	**83.90**	81.44	76.94	76.94	76.94	**67.68**	**67.68**	67.68
Type I	13.25	5.47	**5.47**	4.14	2.34	2.34	2.34	**1.30**	**1.30**	1.30
Type II	16.32	29.35	**29.35**	36.48	48.85	48.85	48.85	**71.01**	**71.01**	71.01
KNN										
Acc	78.27	78.27	78.27	78.27	78.27	78.27	78.27	78.27	78.27	78.27
Type I	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33	15.33
Type II	29.68	29.68	29.68	29.68	29.68	29.68	29.68	29.68	29.68	29.68
MLP										
Acc	79.72	79.72	79.72	79.72	79.72	79.72	79.72	79.72	79.72	79.72
Type I	15.84	15.84	15.84	15.84	15.84	15.84	15.84	15.84	15.84	15.84
Type II	25.77	25.77	25.77	25.77	25.77	25.77	25.77	25.77	25.77	25.77
CART										
Acc	**87.97**	**87.97**	87.97	87.97	87.97	57.22	57.22	57.22	57.22	57.22
Type I	**11.44**	**11.44**	11.44	11.44	11.44	4.62	4.62	4.62	4.62	4.62
Type II	**12.75**	**12.75**	12.75	12.75	12.75	90.32	90.32	90.32	90.32	90.32
Bayes										
Acc	69.12	69.12	69.12	69.12	69.12	69.12	69.12	69.12	69.12	69.12
Type I	22.21	22.21	22.21	22.21	22.21	22.21	22.21	22.21	22.21	22.21
Type II	41.66	41.66	41.66	41.66	41.66	41.66	41.66	41.66	41.66	41.66

The bold-faced numbers mean that they are significantly different from the others ($p < 0.01$)

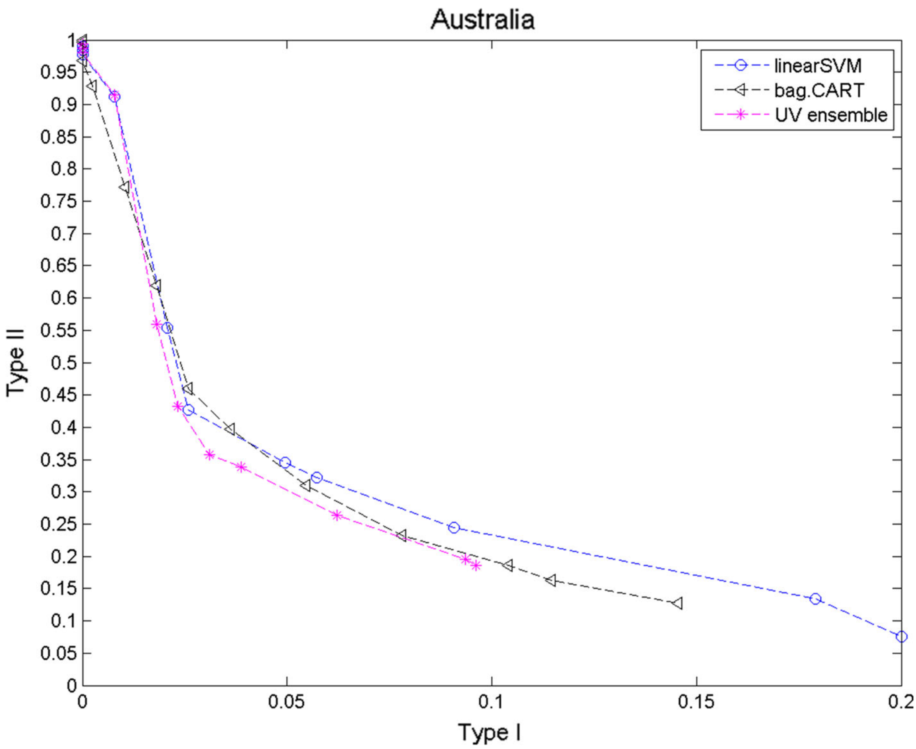


Fig. 5 ROC curves for UV ensemble, linear.SVM, and bag.CART over the Australian dataset

The results indicate that only the UV ensemble provides a more stable performance for cost ratio from 2 to 30 because the other classifiers, such as linearSVM, bag.linearSVM, boost.linearSVM, and bag.CART, only perform well over some cost ratios.

Figure 5 shows the ROC curves for the UV ensemble, linear.SVM and bag.CART methods for comparison. Although the curves for all three classifiers are similar, on average, the UV ensemble has the least area under the curve, which means that the UV ensemble is the better choice for the Australian dataset.

4.2.4 Results on the German dataset

Table 10 shows the prediction accuracy and Type I/II errors obtained for different classifiers based on different cost ratios obtained using the German dataset. As we can see, the UV ensemble performs significantly better than the other classifiers. Specifically, the UV ensemble provides better performances when the cost ratios range from 2 to 30, which is almost the same as the results obtained using the other three datasets. On the other hand, linearSVM and bag.linearSVM can provide good performance for some cost ratios.

Figure 6 shows the ROC curves for the UV ensemble, linear.SVM, and bag.CART approaches. Although the UV ensemble and linearSVM curves are very close, in most cases (i.e., the Type I error is smaller than 0.3) the UV ensemble approach produces the curve with the least area.

Table 10 Prediction accuracy and Type I/II errors of different classifiers over the German dataset

German	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
UV										
Acc	75.50	**72.50**	**72.50**	**63.50**	**53.40**	**46.10**	**46.10**	**46.10**	**46.10**	**30.00**
Type I	36.00	**23.67**	**23.67**	**11.67**	**5.67**	**2.00**	**2.00**	**2.00**	**2.00**	**0.00**
Type II	19.57	**29.14**	**29.14**	**47.14**	**64.14**	**76.14**	**76.14**	**76.14**	**76.14**	**100.00**
Stacking										
Acc	76.50	76.50	76.50	30.00	30.00	30.00	30.00	30.00	**30.00**	**30.00**
Type I	59.00	59.00	59.00	0.00	0.00	0.00	0.00	0.00	**0.00**	**0.00**
Type II	8.29	8.29	8.29	100.00	100.00	100.00	100.00	100.00	**100.00**	**100.00**
MV										
Acc	**77.60**	75.20	75.20	75.20	75.20	75.20	75.20	75.20	75.20	75.20
Type I	**53.00**	41.00	41.00	41.00	41.00	41.00	41.00	41.00	41.00	41.00
Type II	**9.29**	17.86	17.86	17.86	17.86	17.86	17.86	17.86	17.86	17.86
Single										
LinearSVM										
Acc	**77.60**	**73.80**	73.80	63.60	53.20	**46.00**	**46.00**	**46.00**	**30.00**	**30.00**
Type I	**49.67**	**28.67**	28.67	13.33	6.33	**2.33**	**2.33**	**2.33**	**0.00**	**0.00**
Type II	**10.71**	**25.14**	25.14	46.29	64.14	**76.14**	**76.14**	**76.14**	**100.00**	**100.00**
KNN										
Acc	69.40	68.70	62.30	46.50	46.50	46.50	46.50	46.50	46.50	46.50
Type I	66.33	38.00	20.33	5.00	5.00	5.00	5.00	5.00	5.00	5.00
Type II	15.29	28.43	45.14	74.29	74.29	74.29	74.29	74.29	74.29	74.29
MLP										
Acc	71.40	71.40	71.40	71.40	71.40	71.40	71.40	71.40	71.40	71.40
Type I	61.00	61.00	61.00	61.00	61.00	61.00	61.00	61.00	61.00	61.00
Type II	14.71	14.71	14.71	14.71	14.71	14.71	14.71	14.71	14.71	14.71
CART										
Acc	73.20	67.60	64.20	59.20	45.30	30.00	30.00	30.00	**30.00**	**30.00**
Type I	63.00	28.67	20.00	15.33	7.33	0.00	0.00	0.00	**0.00**	**0.00**
Type II	11.29	34.00	42.57	51.71	75.00	100.00	100.00	100.00	**100.00**	**100.00**
Bayes										
Acc	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70
Type I	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00	96.00
Type II	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71
Bag										
LinearSVM										
Acc	**77.20**	75.60	**70.30**	**64.70**	**50.60**	50.60	**39.80**	**39.80**	**39.80**	**39.80**
Type I	**50.33**	37.33	**21.67**	**13.00**	**4.67**	4.67	**1.00**	**1.00**	**1.00**	**1.00**
Type II	**11.00**	18.86	**33.14**	**44.86**	**68.57**	68.57	**85.57**	**85.57**	**85.57**	**85.57**

Table 10 continued

German	Cost ratios/penalties									
	1%	2%	3%	5%	7%	10%	15%	20%	25%	30%
KNN										
Acc	70.40	64.30	64.30	49.50	49.50	49.50	49.50	49.50	49.50	49.50
Type I	65.33	19.33	19.33	6.00	6.00	6.00	6.00	6.00	6.00	6.00
Type II	14.29	42.71	42.71	69.57	69.57	69.57	69.57	69.57	69.57	69.57
MLP										
Acc	76.50	76.50	76.50	76.50	76.50	76.50	76.50	76.50	76.50	76.50
Type I	59.00	59.00	59.00	59.00	59.00	59.00	59.00	59.00	59.00	59.00
Type II	8.29	8.29	8.29	8.29	8.29	8.29	8.29	8.29	8.29	8.29
CART										
Acc	76.00	76.00	72.10	59.70	**59.70**	42.30	32.10	32.10	**30.00**	**30.00**
Type I	43.33	43.33	31.67	11.33	**11.33**	3.33	0.33	0.33	**0.00**	**0.00**
Type II	15.71	15.71	26.29	52.71	**52.71**	81.00	96.86	96.86	**100.00**	**100.00**
Bayes										
Acc	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70	70.70
Type I	95.67	95.67	95.67	95.67	95.67	95.67	95.67	95.67	95.67	95.67
Type II	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
Boost										
LinearSVM										
Acc	76.60	**74.40**	**65.60**	65.60	65.60	58.70	58.70	58.70	58.70	58.70
Type I	50.00	**32.33**	**15.33**	15.33	15.33	11.67	11.67	11.67	11.67	11.67
Type II	12.00	**22.71**	**42.57**	42.57	42.57	54.00	54.00	54.00	54.00	54.00
KNN										
Acc	66.20	65.10	53.70	52.70	52.70	52.70	52.70	52.70	52.70	52.70
Type I	54.00	36.33	10.33	9.33	9.33	9.33	9.33	9.33	9.33	9.33
Type II	25.14	34.29	61.71	63.57	63.57	63.57	63.57	63.57	63.57	63.57
MLP										
Acc	72.50	72.50	72.50	72.50	72.50	72.50	72.50	72.50	72.50	72.50
Type I	42.33	42.33	42.33	42.33	42.33	42.33	42.33	42.33	42.33	42.33
Type II	21.14	21.14	21.14	21.14	21.14	21.14	21.14	21.14	21.14	21.14
CART										
Acc	75.90	75.90	75.90	75.90	75.90	75.90	75.90	75.90	75.90	75.90
Type I	49.00	49.00	49.00	49.00	49.00	49.00	49.00	49.00	49.00	49.00
Type II	13.43	13.43	13.43	13.43	13.43	13.43	13.43	13.43	13.43	13.43
Bayes										
Acc	73.60	73.60	73.60	73.60	73.60	73.60	73.60	73.60	73.60	73.60
Type I	50.67	50.67	50.67	50.67	50.67	50.67	50.67	50.67	50.67	50.67
Type II	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00	16.00

The bold-faced numbers mean that they are significantly different from the others ($p < 0.01$)

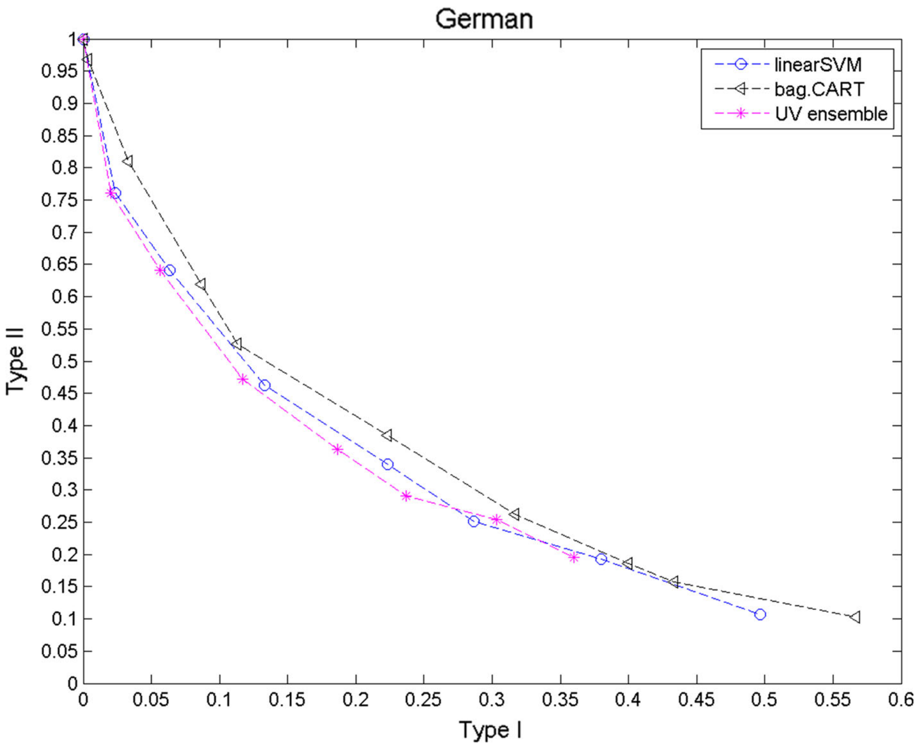


Fig. 6 ROC curves of UV ensemble, linear.SVM, and bag.CART over the German dataset

Table 11 Recommended classifiers for the different four datasets

	Taiwanese	Chinese	Australian	German
1	UV	UV	UV	UV
2		CART		linearSVM
3		bag.CART		bag.linearSVM

4.2.5 Discussion

Table 11 summarizes the best (or better) prediction models over the four datasets. Note that only the UV ensemble approach is recommended for the Taiwan and Australian datasets because none of the other classifiers offers similar performance (i.e., better performances over many cost ratios).

Specifically, when the cost ratio is larger than 2, the UV ensemble performs the best over the Taiwanese and Australian datasets. On the other hand, for the Chinese and German datasets, the UV ensemble outperforms the other classifiers, when the cost ratio is larger than 3. In other words, which single classifier or classifier ensemble performs that best differs when used over different datasets. However, the UV ensemble performs consistently better than the other classifiers over all four datasets. This suggests that the UV ensemble has better potential for use on different financial distress datasets.

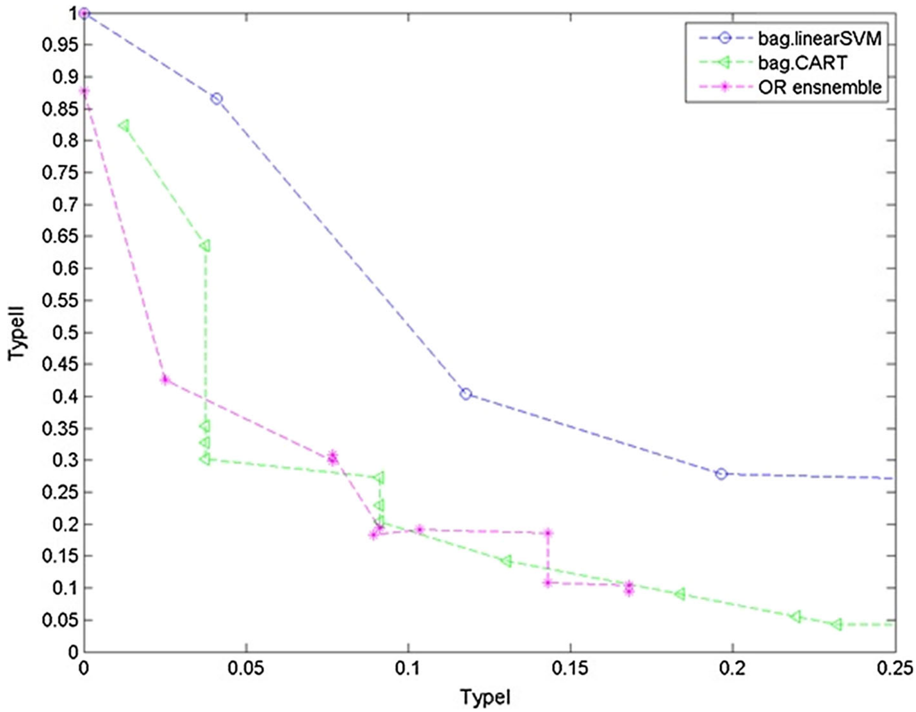


Fig. 7 ROC curves of UV ensemble, linear.SVM, and bag.CART

4.2.6 Practical example

In order to obtain a better understanding of the effectiveness of using this method to calculate the ratio between bankruptcy and non-bankruptcy in a realistic example, another dataset is collected, using data from 2005 to 2015. The dataset contains data for 77 high bankruptcy risk and 231 normal companies. Figure 7 shows the ROC curves for the UV ensemble, linear.SVM, and bag.CART methods. It can be seen that the UV ensemble method still performs reasonably well.

Table 12 shows the cost ratios for the different classifiers. As we can see, the cost of the UV ensemble increases very slowly, which is similar to the bag.CART and boost.CART results. According to the Wilcoxon test, the costs of these three classifiers are not significantly different.

5 Conclusion

Although improving on the prediction accuracy of financial distress prediction models is usually the most important issue addressed in research studies, lowering the Type I error is also important in practice, because a prediction model that provides a lower Type I error rate can reduce the risk to the financial institution.

In order to obtain good prediction accuracy as well as the lower Type I error, we propose a classifier ensemble approach based on the unanimous voting (UV) method, which com-

Table 12 Cost ratios of different classifiers

	1	1.2	1.5	2	3	5	7	10	15	20	25	30
UV	34.79	37.37	41.25	46.89	57.89	76.59	90.34	110.96	127.19	136.81	146.44	156.06
Stacking	38.78	42.76	48.74	58.71	78.65	118.53	158.4	212.44	231	231	231	231
MV	27.78	30.97	35.75	43.73	59.68	86.49	112.34	151.11	215.74	280.36	344.99	409.61
Single												
LinearSVM	32.73	37.87	45.58	58.44	84.15	135.58	156.89	184.11	229.49	231	231	231
KNN	69.99	81.65	97.49	123.89	159.09	182.74	206.39	241.86	300.99	360.11	419.24	478.36
MLP	48.54	53.74	61.53	74.53	100.51	152.49	204.46	282.43	412.36	542.3	672.24	802.18
CART	31.76	36.93	44.69	57.61	70.26	82.36	94.46	112.61	142.86	173.11	203.36	231
Bayes	68.89	70.81	73.7	78.51	88.14	107.39	126.64	155.51	203.64	251.76	299.89	348.01
Bag												
LinearSVM	32.04	36.63	43.52	55	77.96	122.79	152.49	197.04	231	231	231	231
KNN	66	76.75	92.88	119.76	154.83	186.45	218.08	265.51	344.58	423.64	502.7	581.76
MLP	38.78	42.76	48.74	58.71	78.65	118.53	158.4	218.21	317.9	417.59	517.28	616.96
CART	27.78	31.35	36.71	45.65	62.98	82.23	90.06	98.73	113.16	127.6	142.04	156.48
Bayes	71.09	73.23	76.45	81.81	92.54	113.99	135.44	167.61	221.24	274.86	328.49	382.11
Boost												
LinearSVM	36.99	41.03	46.06	54.45	71.23	104.78	138.33	188.65	272.53	356.4	440.28	524.15
KNN	70.95	77.47	87.24	103.54	131.45	172.29	211.61	270.6	368.91	467.23	565.54	663.85
MLP	52.94	56.1	60.84	68.75	74.56	116.19	147.81	195.25	274.31	353.38	432.44	511.5
CART	26.68	30.2	34.44	41.53	55.69	84.01	112.34	154.83	225.64	296.45	367.26	438.08
Bayes	64.76	69.01	77.62	90.48	116.19	167.61	219.04	296.18	424.74	553.3	681.86	810.43

bines multiple outputs produced by different classifiers. This new UV ensemble approach is different from classic classifier ensemble approaches that use the majority voting method. The final prediction result of our UV ensemble is based on choosing the class for which all classifiers agree, whereas the majority voting method is based on choosing the class which receives the largest number of votes.

Using four relevant datasets, two bankruptcy prediction datasets and two credit scoring datasets, the results obtained with our proposed UV ensemble approach are compared with those obtained with five well-known single classifiers including SVM, CART, KNN, MLP, and Bayes and classifier ensembles using bagging and boosting algorithms. The results show that the UV ensemble outperforms the other classifiers over all four datasets. Specifically, the UV ensemble can provide the least misclassification cost.

References

1. Altman EI (1968) Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J Finance 23:589–609
2. Balcaen S, Ooghe H (2006) 35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems. Br Account Rev 38:63–93
3. Beaver WH (1966) Financial ratios as predictors of failure. J Account Res 4:71–111
4. Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
5. Boritz JE, Kennedy DB (1995) Effectiveness of neural network types for prediction of business failure. Expert Syst Appl 9:503–512
6. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
7. Clyde MA, Lee HK (2001) Bagging and the Bayesian bootstrap. In: International conference on artificial intelligence and statistics, pp 169–174
8. Crook JN, Edelman DB, Thomas LC (2007) Recent developments in consumer credit risk assessment. Eur J Oper Res 183:1447–1465
9. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
10. Dietterich TG (1997) Machine-learning research. AI Mag 18:97
11. Ethem A (2004) Introduction to machine learning. MIT Press, Cambridge
12. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874
13. Fitzpartrick PJ (1932) A comparison of the ratios of successful industrial enterprises with those of failed companies. The Accountants Publishing Company.
14. Freund Y, Schapire RE (July 1996) Experiments with a new boosting algorithm. In: Proceedings of the international conference on machine learning, Bari, Italy, pp 148–156
15. Geng R, Bose I, Chen X (2015) Prediction of financial distress: an empirical study of listed Chinese companies using data mining. Eur J Oper Res 241:236–247
16. Heo J, Yang JY (2014) AdaBoost based bankruptcy forecasting of Korean construction companies. Appl Soft Comput 24:494–499
17. John GH Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In: International conference on uncertainty in artificial intelligence, pp 338–345
18. Kim H, Kim H, Moon H, Ahn H (2011) A weight-adjusted voting algorithm for ensembles of classifiers. J Korean Stat Soc 40:437–449
19. Kittler J, Hatef M, Duin RPW, Matas J (1998) On combining classifiers. IEEE Trans Pattern Anal Mach Intell 20(3):226–239
20. Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence, pp 1137–1143
21. Kumar PR, Ravi V (2007) Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review. Eur J Oper Res 180:1–28
22. Lei S, Xinming M, Lei X, Xiaohong H (2010) Financial data mining based on support vector machines and ensemble learning. In: International conference on intelligent computation technology and automation, pp 313–314
23. Li H, Sun J, Wu J (2010) Predicting business failure using classification and regression tree: an empirical comparison with popular classical statistical methods and top classification mining methods. Expert Syst Appl 37:5895–5904

24. Lin W-Y, Hu Y-H, Tsai C-F (2012) Machine learning in financial crisis prediction: a survey. IEEE Trans Syst Man Cybern Part C Appl Rev 42(4):421–436
25. Ohlson JA (1980) Financial ratios and the probabilistic prediction of bankruptcy. J Account Res 18:109–131
26. Ozkan-Gunay EN, Ozkan M (2007) Prediction of bank failures in emerging financial markets: an ANN approach. J Risk Finance 8:465–480
27. Paleologo G, Elisseeff A, Antonini G (2010) Subagging for credit scoring models. Eur J Oper Res 201:490–499
28. Sexton RS, Sriram RS, Etheridge H (2003) Improving decision effectiveness of artificial neural networks: a modified genetic algorithm approach. Decis Sci 34:421–442
29. Shi L, Xi L, Ma X, Hu X (2009) Bagging of artificial neural networks for bankruptcy prediction. In: International conference on information and financial engineering, pp 154–156
30. Shin K-S, Lee TS, Kim H-J (2005) An application of support vector machines in bankruptcy prediction model. Expert Syst Appl 28:127–135
31. Tam KY, Kiang MY (1992) Managerial applications of neural networks: the case of bank failure predictions. Manag Sci 38:926–947
32. Tsai C-F (2009) Feature selection in bankruptcy prediction. Knowl Based Syst 22:120–127
33. Verikas A, Kalsyte Z, Bacauskiene M, Gelzinis A (2010) Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey. Soft Comput 14:995–1010
34. Wang G, Hao J, Ma J, Jiang H (2011) A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl 38:223–230
35. Wang G, Ma J (2012) A hybrid ensemble approach for enterprise credit risk assessment based on support vector machine. Expert Syst Appl 39(5):5325–5331
36. Wang S-J, Mathew A, Chen Y, Xi L-F, Ma L, Lee J (2009) Empirical analysis of support vector machine ensemble classifiers. Expert Syst Appl 36:6466–6476
37. West D (2000) Neural network credit scoring models. Comput Oper Res 27:1131–1152
38. West D, Dellana S, Qian J (2005) Neural network ensemble strategies for financial decision applications. Comput Oper Res 32:2543–2559
39. Wolpert DH (1992) Stacked generalization. Neural Netw 5(2):241–259
40. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14:1–37
41. Xu X Frank E (2004) Logistic regression and boosting for labeled bags of instances. In: Dai H, Srikant R, Zhang C (eds) Proceedings 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, 26–28 May 2004. Springer, Berlin, pp. 272–281
42. Yao P (2009) Credit scoring using ensemble machine learning. In: International conference on hybrid intelligent systems, pp 244–246
43. Zhang D, Zhou X, Leung SCH, Zheng J (2010) Vertical bagging decision trees model for credit scoring. Expert Syst Appl 37:7838–7843

Deron Liang is a professor at the Department of Computer Science and Information Engineering, National Central University. His research interests cover business intelligence and distributed systems.

Chih-Fong Tsai is a professor at the Department of Information Management, National Central University, Taiwan. He has published more than 60 technical publications in journals, book chapters, and international conference proceedings. His current research focuses on multimedia information retrieval and data mining.

An-Jie Dai Chen received a Master's degree from the Department of Computer Science and Information Engineering, National Central University. His research interests cover machine learning and data mining applications.

William Eberle is currently an associate professor in the Department of Computer Science at Tennessee Technological University. His research areas of interest include data mining, artificial intelligence, machine learning, graph theory and language processing.