Contents lists available at ScienceDirect

Journal of Business Research

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

Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction

Chih-Fong Tsai, Kuen-Liang Sue, Ya-Han Hu [*], Andy Chiu

Department of Information Management, National Central University, Taiwan, ROC

ARTICLE INFO

Keywords:
Data mining
Ensemble learning
Feature selection
Financial distress prediction
Instance selection

ABSTRACT

Bankruptcy prediction and credit scoring are major problems in financial distress prediction. Studies have shown that prediction models can be made more effective by performing data preprocessing procedures. Moreover, classifier ensembles are likely to outperform single classifiers. Although feature selection, instance selection, and classifier ensembles are known to affect final prediction results, their combined effects on bankruptcy prediction and credit scoring problems have not been fully explored. This study compares the performance of three feature selection algorithms, three instance selection algorithms, four classification algorithms, and two ensemble learning techniques. The results obtained using five bankruptcy prediction and five credit scoring datasets indicate that by carefully considering the combination of these three factors, better prediction models can be developed than by considering only one related factor.

1 Introduction

Financial distress prediction has long been regarded as a critical research problem for financial institutions, and various statistical and machine learning techniques have been employed to construct prediction models, the output of which can be used for loan decision-making (Alaka et al., 2018; Chen, Ribeiro, & Chen, 2016; Kumar & Ravi, 2007; Liang, Tsai, Lu, & Chang, 2020; Lin, Hu, & Tsai, 2012; Tang, Li, Tan, & Shi, 2020). In general, financial distress prediction covers two related subtopics: bankruptcy prediction and credit scoring. Bankruptcy prediction models usually focus on identifying the likelihood that the company applying for the loan will go bankrupt. Credit scoring models are developed to rate the credit score (high, medium, or low) of the customer (either an individual or a company) applying for the loan for later decision-making processes (Climent, Momparler, & Carmona, 2019; Lin et al., 2012; Pérez-Martín, Pérez-Torregrosa, & Vaca, 2018; Thomas, 2000).

Prediction models must be made as effective as possible, irrespective of the techniques used to develop them. Therefore, studies have mainly aimed to determine the best model that can provide the highest prediction accuracy or the lowest prediction error. Several key factors affect the final performance of prediction models, as discussed further on.

One crucial factor is feature selection. Although the types of related features (i.e., indicators) that can provide the highest discriminative power to distinguish between bankrupt or nonbankrupt classes and high and low credit classes remain unclear, performing feature selection to analyze the level of feature representativeness from the collected datasets is usually helpful to improve the prediction performance of models. For example, Liang, Tsai, and Wu (2015) examined the effect of applying filter- and wrapper-based feature selection methods to financial distress prediction performance. Lin, Lu, and Tsai (2019) compared various combinations of feature selection methods and ensemble classification techniques for bankruptcy prediction. Jadhav, He, and Jenkins (2018) proposed a novel feature selection approach that combines information gain and a genetic algorithm (GA) for enhanced credit rating.

The effectiveness of prediction models can alternatively be improved by filtering out some noisy data (or outliers) from the collected training datasets. Specifically, instance selection or related data sampling methods can be used to allow the two-class data distribution of the processed training datasets to be more easily separated in the feature space than would be possible with the original training dataset. For example, Ahn and Kim (2009) and Liu and Pan (2018) used a GA and fuzzy-rough set for instance selection to improve the performance of bankruptcy prediction and credit scoring models, respectively. Tsai and

[*] Corresponding author at: Department of Information Management, National Central University, No. 300, Zhongda Rd., Zhongli District, Taoyuan City 320317, Taiwan, ROC.

E-mail address: yhhu@mgt.ncu.edu.tw (Y.-H. Hu).

https://doi.org/10.1016/j.jbusres.2021.03.018
Received 22 February 2020; Received in revised form 10 March 2021; Accepted 12 March 2021
Available online 30 March 2021

Cheng (2012) introduced a simple distance-based clustering outlier detection method for improved bankruptcy prediction. Tsai (2014) also comprehensively compared clustering techniques for instance selection and various classification techniques for financial distress prediction models.

Unlike the previous two factors—feature and instance selection—-that are related to data preprocessing issues, ensemble learning techniques that work by combining multiple classifiers (or prediction models) have been shown to outperform single classifiers in terms of bankruptcy prediction and credit scoring (Sun, Li, Fujita, Fu, & Ai, 2020). For example, Choi, Son, and Kim (2018) proposed a voting-based ensemble model for financial distress prediction. Feng, Xiao, Zhong, Qiu, and Dong (2018) constructed dynamic ensemble classifiers based on soft probability for credit scoring. Sun, Fujita, Chen, and Li (2017) constructed support vector machine ensembles based on the time weight of data samples. Garcia, Marques, and Sanchez (2019) investigated the training set size effect on various classifier ensembles for bankruptcy prediction.

The aforementioned studies have demonstrated the importance of considering feature selection, instance selection, and ensemble learning techniques for improved bankruptcy prediction and credit scoring. However, most of them only focused on one of these three approaches; thus, no study has employed all three factors together for bankruptcy prediction or credit scoring problems. In other words, the research motivation of this paper is to answer the question: whether performing both feature and instance selection to preprocess the data and constructing the prediction models by ensemble learning techniques can outperform the single usage of these approaches or combining any two of these three approaches? To the best of our knowledge, this research question has never been answered before in the domain of financial distress prediction related literatures.

In this light, the present study aimed to fill the gap in the literature concerning the joint use of feature selection, instance selection, and classifier ensembles for bankruptcy prediction and credit scoring. To answer the research question, there is a novel technical issue to combine feature and instance selection before the construction of classifier ensembles. Since feature and instance selection are two individual data preprocessing steps, different orders of execution can result in different reduced datasets, which can make classifier ensembles perform differently. As a result, a further research objective is to assess which order of combining feature selection and instance selection can allow classifier ensembles to perform better.

The contribution of this paper is threefold. First, this study is the first to examine all three factors together affecting the financial distress prediction. From the technical viewpoint, a very large number of combinations (i.e. prediction models) are implemented for performance comparison. In particular, three feature selection algorithms (i.e., GA, principal component analysis [PCA], and *t*-test), three instance selection algorithms (i.e., affinity propagation [AP], GA, and self-organizing map [SOM]), and four classification algorithms (i.e., artificial neural network [ANN], decision tree [DT], logistic regression [LR], and support vector machine [SVM]) as well as two ensemble learning techniques (i.e., bagging and boosting) are employed, which result in 288 combinations.

Second, from a practical point of view, although there are no general rules for different data sets (i.e., no free lunch theorem), the ten data sets used in our study cover both small and large bankruptcy prediction and credit scoring problems. The findings can be used as the representative baselines for future researches. When a new dataset is collected, and feature selection, instance selection, and classifier ensembles are required to be employed, it is infeasible to test all possible combinations of these algorithms. Therefore, to reduce the computational load and meet the practical need in a time-efficient manner, the best combination of algorithms determined by this research can be used as a starting point.

Third, from an academic point of view, when future studies propose novel algorithms related to feature selection, instance selection, and classifier ensembles, the results can be compared with our recommended approaches instead of conducting experimental evaluations from scratch (i.e., all possible combinations).

The rest of this paper is organized as follows. Section 2 provides an overview of the feature selection, instance selection, and classifier ensembles and discusses the limitations of related works. Section 3 describes the research methodology, including the experimental setup and procedure. Section 4 presents the experiment results. Finally, Section 5 presents the conclusions of this study.

2 Literature review

2.1 Feature selection

Feature selection or dimensionality reduction aims at selecting a number of representative features from a given training dataset, where the dimension-reduced training dataset can provide more discriminative power to distinguish between classes than can the original one. As a result, a model constructed using a dimension-reduced training dataset is likely to outperform one obtained using the original training dataset (Guyon & Elisseeff, 2003).

In general, the feature selection process consists of four steps: subset generation, subset evaluation, stopping criterion, and result validation. Subset generation is based on the use of some search strategy to produce candidate feature subsets for later evaluation steps. In subset evaluation, each candidate subset is compared with the previous best one based on some evaluation criterion. The subset generation and subset evaluation steps are repeated until a given stopping criterion is satisfied. The final step of result validation is based on using synthetic and/or real-world datasets to validate the selected best subset.

Previously reported feature selection algorithms can be divided into three categories: those using filter, wrapper, and embedded methods (Bolon-Canedo, Sanchez-Marono, & Alonso-Betanzos, 2013; 2004;; Li, Li, & Liu, 2017). Filter methods use some feature ranking criteria for identifying features that have higher discriminative power. By contrast, wrapper methods use a chosen predictor or classifier, the performance of which is used as the objective function to evaluate the feature subset. Embedded methods incorporate the feature selection process as part of the model training procedure, meaning that feature selection is embedded in the construction stage of the prediction model; consequently, the representative features are selected as the prediction model is developed.

2.2 Instance selection

Instance selection is used to not only reduce the size of the training dataset but also filter out noisy data or outliers. The instance selection process is outlined in the following. Given a training set *T*, an instance selection algorithm aims at selecting a more representative subset *S* of *T*, where *S* does not contain redundant instances. The models trained by *S* are likely to outperform or at least perform similarly to the models trained by *T* (Garcia, Derrac, Cano, & Herrera, 2012; Olvera-Lopez, Carrasco-Ochoa, Martinez-Trinidad, & Kittler, 2010).

As with feature selection algorithms, instance selection algorithms can be classified into those using filter and wrapper methods, where filter methods use a selection function instead of a classifier function and wrapper methods are based on the accuracy obtained by using a classifier to select the best subset of a given training dataset.

Clustering algorithms are one type of filter-based instance selection method. According to the clustering results, where each data sample is grouped into a specific cluster, the data samples that are far away from their cluster centers can be removed, meaning that the data samples that are closer to their cluster centers should be more representative than those that are further from their cluster centers (Tsai & Chen, 2014).

A GA is a wrapper-based instance selection method. It is based on using a fitness function as a classifier to iteratively train and test subsets to identify the optimal subset that allows the classifier to provide the highest classification accuracy (Tsai & Chen, 2014).

2.3 Classifier ensembles

Classifier ensembles or multiple classifiers are based on ensemble learning techniques. In this approach, multiple classifiers are constructed, and their classification results are combined for the final output (Masoudnia & Ebrahimpour, 2014; Rokach, 2009; Sun et al., 2020; Wozniak, Grana, & Corchado, 2014). Two widely used ensemble learning techniques are bagging and boosting.

The bagging ensemble method or bootstrap aggregation method samples the original training dataset to produce a number of training subsets. Then, various models are constructed based on these training subsets, with the final result obtained by combining the outputs produced by the constructed models. Averaging the multiple outputs and voting are the most commonly used combination methods for regression and classification problems, respectively (Breiman, 1994).

In the boosting method, weak classifiers are trained iteratively and finally converted to strong classifiers by considering the data weights or reweighting, which relate to the accuracy of weak classifiers. Thus, misclassified input data are assigned a higher weight, whereas correctly classified data are assigned a lower weight. As a result, future weak classifiers focus more on the data that are misclassified by previous weak classifiers (Schapire, 1990).

2.4 Comparison of related works

Table 1 summarizes related works in terms of the techniques employed (including feature and instance selection and ensemble learning), domain datasets used (including bankruptcy prediction and credit scoring), and evaluation metrics considered.

Some limitations of these related works are clear from Table 1. First, some considered either feature or instance selection when performing data preprocessing step, except for Ahn and Kim (2009) and Kim, Lee, and Ahn (2018), both of which used only one algorithm (i.e., GA) for both feature and instance selection tasks over Korean datasets. In particular, Kim et al. (2018) demonstrated that performing both feature and instance selection can enable an SVM classifier to provide higher prediction accuracy (i.e., 80.3%) than those obtained by performing feature and instance selection individually (i.e., 77.8% and 80%, respectively). Similarly, Ahn and Kim (2009) found that a case-based reasoning model based on the combination of both feature and instance selection provided a prediction accuracy of 87.08%, which is higher than that obtained by combining feature and instance selection alone (i.e., 83.71% and 84.64%, respectively).

Second, some studies focused on constructing classifier ensembles and demonstrated how they outperform single classifiers. However, only some of them performed feature or instance tasks. Moreover, studies that considered feature or instance selection and classifier ensembles only used either bankruptcy prediction or credit scoring datasets, except for Lin et al. (2019) and Tsai (2014).

Third, because datasets for financial distress prediction contain a class imbalance and because the number of bankruptcy (or bad credit) cases is much smaller than the number of nonbankruptcy (or good credit) cases, it is misleading to only assess prediction models based on the average prediction accuracy. The AUC and type I/II errors should be examined to provide a reliable conclusion. However, few related works have considered these evaluation metrics. If the collected datasets were preprocessed by some data-sampling techniques or systematically chosen to become class-balanced datasets for the experiments, examining AUC may not be critical. However, among the eight related works listed in Table 1 that examined the rate of prediction accuracy alone, only four studies used class-balanced datasets, which are Kim et al. (2018), Sun et al. (2017), Finlay (2011), and Ahn and Kim (2009).

By contrast, for type I/II errors depending on the definition, the error of misclassifying bankruptcy or bad credit cases into the nonbankruptcy or good credit class, respectively, is critical for financial institutions to reduce the risk of bad debts. However, only six related works out of the 18 examined the type I/II errors of prediction models.

In summary, both feature and instance selection tasks are crucial in the data preprocessing step and can directly affect the performance of classifier ensembles. This study explored the effect of combining feature selection and instance selection on the prediction performance by ensemble classification techniques. Thus, various combinations of these three technical factors—feature selection, instance selection, and ensemble classification techniques—were investigated.

3. Research methodology

3.1 Data preprocessing procedures

In this study, 10 related datasets—five bankruptcy prediction and five credit scoring datasets—containing various numbers of features and instances were chosen for the experiments. Table 2 summarizes the information on these 10 datasets.

Fig. 1 shows the four data preprocessing procedures—feature selection, instance selection, performing feature selection first and instance selection second, and performing instance selection first and feature selection second—that result in four processed datasets that are respectively denoted as FS Dataset, IS Dataset, FS-IS Dataset, and IS-FS Dataset.

For the baseline approach, the original dataset D is divided into training and testing sets that are respectively denoted as $D_{training}$ and $D_{testing}$. The four data preprocessing procedures are described below.

- Feature selection (FS): $D_{training}$ is input into the chosen FS algorithm to identify representative features. Suppose that D contains m feature dimensions and n features are finally selected where $n < m$. The resulting training set containing n features, denoted as $D_{training\_reduced\_features}$, is used to train a classifier. For classifier testing, the features of $D_{testing}$ should be selected to be the same as the ones in $D_{training\_reduced\_features}$, leading to a new testing set that contains n features, denoted as $D_{testing\_reduced\_features}$. This new testing set is used to assess the performance of the classifier.
- Instance selection (IS): $D_{training}$ is input into the chosen IS algorithm to select representative instances. Suppose $D_{training}$ contains o instances and p instances are finally selected where $p < o$. As a result, a subset of $D_{training}$ is produced, denoted as $D_{training\_reduced\_instances}$, and is then used to train a classifier. For classifier testing, $D_{testing}$ is used.
- FS-IS: $D_{training\_reduced\_features}$ is input into the chosen IS algorithm to select representative instances, where q instances are selected, denoted as $D_{training\_reduced\_features+instances}$. The selected q instances of $D_{training\_reduced\_features+instances}$ are not necessarily the same as the selected p instances of $D_{training\_reduced\_instances}$. Next, $D_{training\_reduced\_features+instances}$ is used to train the classifier, and $D_{testing\_reduced\_features}$ is used to test the classifier.
- IS-FS: $D_{training\_reduced\_instances}$ is input into the chosen FS algorithm, resulting in a new training set that contains l features, denoted as $D_{training\_reduced\_instances+features}$. The selected l features of $D_{training\_reduced\_instances+features}$ are not necessarily the same as the selected n features of $D_{training\_reduced\_features}$. Next, $D_{training\_reduced\_instances+features}$ is used for classifier training. To test the classifier, the features of $D_{testing}$ selected are the same as the ones in $D_{training\_reduced\_instances+features}$, which results in a new testing set that contains l features.

For the feature selection algorithms, PCA, *t*-test, and GA[1] methods are used, whereas the instance selection algorithms are based on SOM[2], AP (Frey & Dueck, 2007), and GA. Consequently, three sets of results are

[1] The parameters of GA are based on Tsai and Chen (2014).

[2] The map size of SOM is based on Tsai (2014).

Table 1
Overview of related research.

Works	Techniques			Domain datasets		Evaluation metrics
	Feature selection	Instance selection	Ensemble techniques	Bankruptcy prediction	Credit scoring	
Garcia et al. (2019)			Bagging & boosting ANN[1]/C4.5/kNN[2]	Australian/Finland/German/Japanese/Polish/SabiSPQ/Taiwan/Thomas		AUC/Type I/II errors
Lin et al. (2019)	GA/IG[3]		Bagging & boosting LR/NB[4]/ANN/DT[5]/SVM/kNN	Australian/German/Taiwan		Type I error
Choi et al. (2018)			Voting of SVM[6], ANN, DT, NB, LR[7]		Korea	AUC
Du Jardin (2018)	Mann-Whitney test		Bagging & Boosting DA[8]/LR/DT/SVM/ANN/ELM[9]	France		Accuracy
Feng et al. (2018)			Bagging DT/ANN/SVM; boosting DT; voting of DT, ANN/SVM	AER credit card/Australian/Chinese/German/Japanese/Kaggle/PAKDD2010/Taiwan		Accuracy/AUC/Type I/II errors
Jadhav et al. (2018)	IGDFS [10]/GA [11]				Australian/German/Taiwan	Accuracy/AUC
Kim et al. (2018)	GA	GA		Korea		Accuracy
Liu and Pan (2018)		FRIS [12]			Australian/German	Accuracy
Maldonado et al. (2017)	CBFS [13]				New/returning customer datasets	AUC
Sun et al. (2017)			Boosting SVM	Chinese		Accuracy
Zhou and Lai (2017)			Boosting DT/ANN	Japan/USA		Accuracy/AUC
Liang et al. (2015)	GA/DA/LR/PSO [14]/*t*-test			Australian/Chinese/German/Taiwan		Accuracy/Type I/II errors
Tsai (2014)		k-means/SOM [15]	Voting of CART [16], ANN, LR; bagging ANN	Australian/German/Japanese/UC competition		Accuracy/Type I/II errors
Tsai et al. (2014)			Bagging & boosting ANN/DT/SVM	Australian/German/Japanese/Taiwan		Accuracy
Tsai and Cheng (2012)		k-means		Australian/German/Japanese/UC competition		Accuracy/Type I/II errors
Finlay (2011)			Bagging & boosting LDA/CAT/ANN/kNN		UK	Accuracy
Sun et al. (2011)			Boosting DT	Chinese		Accuracy
Ahn and Kim (2009)	GA	GA		Korea		Accuracy

[1] ANN: artificial neural network.
[2] kNN: k-nearest neighbor.
[3] IG: information gain.
[4] NB: naïve Bayes.
[5] DT: C4.5 decision tree.
[6] SVM: support vector machine.
[7] LR: logistic regression.
[8] DA: discriminant analysis.
[9] ELM: extreme learning machine.
[10] IGDFS: information gain directed feature selection.
[11] GA: genetic algorithm.
[12] FRIS: fuzzy-rough instance selection.
[13] CBFS: cost-based feature selection.
[14] PSO: particle swarm optimization.
[15] SOM: self-organizing maps.
[16] CART: classification and regression tree.

produced by the data preprocessing procedures for FS Dataset and IS Dataset, and nine sets of results are produced for FS-IS Dataset and IS-FS Dataset.

3.2 Construction of classifier ensembles

In this study, four classification techniques—LR, SVM, ANN, and DT—were used to construct classifier ensembles. For the ensemble techniques, the bagging and boosting methods were employed, resulting in both bagging- and boosting-based LRs, SVMs, ANNs, and DTs.

The classifier ensembles were constructed using the Weka machine learning software[3] with the related parameters set to the default values. In addition, as in Tsai, Hsu, and Yen (2014), the number of multiple classifiers was set to 100.

Moreover, each dataset was divided into 90% training and 10% testing subsets based on the 10-fold cross validation method. Many studies listed in Table 1 also considered 10-fold cross validation, including Garcia et al. (2019), Choi et al. (2018), Jadhav et al. (2018), Liu and Pan (2018), Maldonado, Perez, and Bravo (2017), Liang et al. (2015), Tsai and Cheng (2012), Sun, Jia, and Li (2011). The average of 10 results is reported for performance comparison between data preprocessing approaches and classifier ensembles. Furthermore, the training and testing subsets of each fold were controlled to contain almost the same class ratio of the original dataset.

In terms of evaluation metrics, the AUC and type II error were examined. AUC is a suitable measurement for class-imbalanced datasets (Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2012), and the type II error focuses on the misclassification error of bankruptcy or bad credit cases into the nonbankruptcy or good credit class, which can leave

[3] https://www.cs.waikato.ac.nz/ml/weka/

Table 2
Basic information for the investigated datasets.

Dataset	No. of features	No. of instances	No. of bankruptcy/ bad credit cases
Bankruptcy prediction			
Bankruptcy (Olson, Delen, & Meng, 2012)	16	1321	697
JPNBDS (Zhou & Lai, 2017)	11	152	76
TEJ-China (Liang et al., 2016)	101	3058	69
TEJ-Taiwan (Liang et al., 2016)	95	6819	220
USABDS (Zhou & Lai, 2017)	11	2336	1168
Credit scoring			
Australian[1]	14	690	383
German[2]	20	1000	300
Japanese[3]	15	690	383
Kaggle[4]	10	150,000	10,026
PAKDD[5]	37	50,000	13,041

[1]https://archive.ics.uci.edu/ml/datasets/Statlog+(Australian + Credit + Approval).
[2]https://archive.ics.uci.edu/ml/datasets/Statlog+(German + Credit + Data).
[3]https://archive.ics.uci.edu/ml/datasets/Credit + Approval.
[4]http://www.kaggle.com/c/GiveMeSomeCredit.
[5]http://sede.neurotech.com.br/PAKDD2010/.

financial institutions vulnerable to bad debts.

4 Experimental results

4.1 Feature and instance selection factors

The first experiment was designed to examine the effect of data preprocessing procedures, namely, FS, IS, FS-IS, and IS-FS, on the prediction performance of classifier ensembles. To facilitate understanding and illustrate the results obtained using the various data preprocessing procedures, the classifier ensemble factor was fixed and based on boosting SVM. This was because SVM is one of the most widely used classifiers and has been demonstrated to be superior to many other well-known classifiers such as LR, ANN, and DT (Lin et al., 2012; Liang, Lu, Tsai, & Shih, 2016; Maldonado et al., 2017; Sun et al., 2017). Moreover, ensemble classifiers are known to usually outperform single classifiers (Tsai, 2014), and the boosting algorithm has the potential to achieve a lower error rate than other related algorithms (Freund & Schapire, 1997).

4.1.1 Results for bankruptcy prediction datasets

Tables 3 and 4 show the AUC and Type II errors from boosting SVM for various preprocessing procedures over the five bankruptcy prediction datasets. The best result for each dataset is underlined. With most of the preprocessing approaches, the boosting SVM classifier provided a higher AUC and lower Type II error than the baseline approach did over these datasets. However, the algorithm(s) for the data preprocessing step should be chosen carefully to obtain the best result.

Based on the evaluation metrics (i.e., AUC and type II error), the best approach for each dataset differed somewhat. According to the AUC and type II error results, the best approach was the same only for the TEJ-Chain dataset, namely, GA-SOM. The relationship between data characteristics and prediction performance is noteworthy. For the JPNBDS and USABDS datasets, which were not class imbalanced (i.e., class imbalance ratio is 1:1) and contained smaller numbers of features (i.e., 11) than the Bankruptcy, TEJ-China, and TEJ-Taiwan datasets, performing AP-PCA provided the highest AUC rate ($p < 0.05$)[4]. However, no direct relationship was identified between the data characteristics and the prediction performances for type II error.

Fig. 2 shows the average AUC and type II errors obtained with the 24 data preprocessing approaches as well as the baseline method over the five bankruptcy prediction datasets. In terms of the average AUC, all data preprocessing approaches outperformed the baseline approach, except for GA-GA, for which PCA-AP was the best approach (AUC of 0.879, $p < 0.05$).

For the average type II error, only the GA and PCA approaches produced a poorer result than the baseline. Among the 24 data preprocessing approaches, GA-SOM exhibited the best performance ($p < 0.05$).

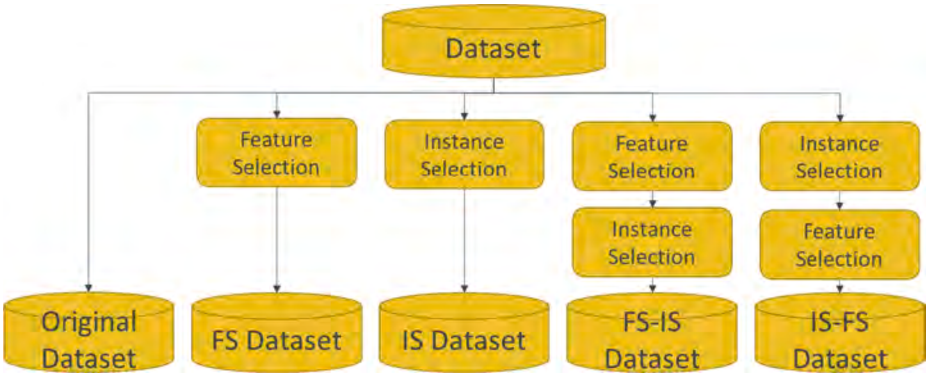

Fig. 1. . The data pre-processing procedures.

4.1.2 Results for credit scoring datasets

Tables 5 and 6 present the AUC and Type II errors obtained using boosting SVM with various preprocessing procedures over the five credit scoring datasets. Similar to the results of the bankruptcy prediction datasets, not all data preprocessing approaches outperformed the baseline approach. Moreover, the best approaches differed based on the AUC and type II error results.

These results indicate that for most credit scoring datasets, except Australian, PCA-SOM and *t*-test-SOM were the superior choices depending on the dataset size. For small-scale datasets containing small numbers of instances, such as German (1000) and Japanese (690), PCA-SOM provided the highest AUC rate ($p < 0.05$). For datasets containing large numbers of instances, such as Kaggle (15000) and PAKDD (50000), *t*-test-SOM is recommended ($p < 0.05$)[5]. For the type II error, PCA-AP

[4] The statistical analysis is based on Welch's T-test (Derrick, Toher, & White, 2016).

[5] For the Kaggle dataset, there is no a significant difference between SOM, *t*-test-SOM, and SOM-*t*-test.

Table 3
AUC under different pre-processing procedures (bankruptcy prediction).

	Bankruptcy	JPNBDS	TEJ-China	TEJ-Taiwan	USABDS
Baseline					
	0.723	0.797	0.846	0.905	0.418
FS					
GA	0.741	0.851	0.862	0.903	0.507
PCA	0.641	0.8	0.845	0.908	0.525
t-test	0.737	0.824	0.868	0.9	0.514
IS					
AP	0.802	0.925	0.844	0.798	0.871
GA	0.704	0.919	0.821	0.884	0.513
SOM	0.824	0.9	0.851	0.896	0.801
FS-IS					
GA-AP	0.768	0.712	0.876	0.888	0.787
GA-GA	0.66	0.702	0.819	0.902	0.508
GA-SOM	0.828	0.95	0.958	0.885	0.672
PCA-AP	0.844	0.9	0.955	0.969	0.725
PCA-GA	0.667	0.793	0.838	0.9	0.521
PCA-SOM	0.848	0.861	0.878	0.878	0.705
t-test-AP	0.807	0.81	0.894	0.826	0.705
t-test-GA	0.702	0.822	0.86	0.902	0.491
t-test-SOM	0.853	1	0.933	0.909	0.544
IS-FS					
AP-GA	0.803	0.833	0.876	0.914	0.819
AP-PCA	0.787	0.967	0.853	0.785	0.877
AP-*t*-test	0.808	0.942	0.889	0.842	0.862
GA-GA	0.623	0.925	0.859	0.897	0.497
GA-PCA	0.641	0.879	0.825	0.89	0.517
GA-*t*-test	0.671	0.879	0.876	0.902	0.5
SOM-GA	0.808	0.375	0.893	0.922	0.799
SOM-PCA	0.831	0.95	0.887	0.83	0.779
SOM-*t*-test	0.837	0.9	0.919	0.882	0.775

Table 4
Type II error under different pre-processing procedures (bankruptcy prediction).

	Bankruptcy	JPNBDS	TEJ-China	TEJ-Taiwan	USABDS
Baseline					
	0.304	0.225	0.9	0.877	0.516
FS					
GA	0.219	0.259	1	0.968	0.45
PCA	0.266	0.209	0.867	0.886	0.642
t-test	0.257	0.105	0.926	0.918	0.504
IS					
AP	0.531	0.2	0.717	0.633	0.21
GA	0.275	0.208	0.927	0.607	0.237
SOM	0.418	0.1	0.558	0.59	0.33
FS-IS					
GA-AP	0.538	0.367	0.833	0.55	0.275
GA-GA	0.335	0.115	0.967	0.912	0.054
GA-SOM	0.382	0	0.417	0.7	0.038
PCA-AP	0.42	0.2	0.5	0.55	0.205
PCA-GA	0.288	0.192	0.947	0.795	0.221
PCA-SOM	0.257	0.3	0.78	0.907	0.372
t-test-AP	0.464	0.195	0.717	0.8	0.42
t-test-GA	0.291	0.392	0.933	0.757	0.426
t-test-SOM	0.376	0	0.675	0.643	0.652
IS-FS					
AP-GA	0.539	0.2	0.633	0.567	0.274
AP-PCA	0.524	0.1	0.7	0.633	0.265
AP-*t*-test	0.517	0.3	0.608	0.55	0.186
GA-GA	0.179	0.108	0.967	0.874	0.002
GA-PCA	0.348	0.133	0.857	0.725	0.223
GA-*t*-test	0.365	0.158	0.94	0.655	0
SOM-GA	0.473	0.75	0.492	0.4	0.325
SOM-PCA	0.394	0.25	0.608	0.665	0.357
SOM-*t*-test	0.448	0.2	0.575	0.6	0.362

outperformed most credit scoring datasets except Australian ($p < 0.05$).

Fig. 3 displays the average AUC and type II errors obtained using the 24 data preprocessing approaches as well as the baseline method over the five credit scoring datasets. In terms of AUC, SOM (0.786), GA-SOM (0.779), PCA-AP (0.78), PCA-SOM (0.827), *t*-test-SOM (0.829), AP-*t*-test (0.774), SOM-GA (0.819), SOM-PCA (0.787), and SOM-*t*-test (0.801) outperformed the baseline approach (0.772), with *t*-test-SOM performing the best ($p < 0.05$).

For the type II error, many data preprocessing approaches produced better results than the baseline approach, except for PCA, *t*-test, GA-AP, PCA-GA, *t*-test-GA, and GA-PCA. Among the 24 data preprocessing approaches, PCA-AP significantly outperformed the others ($p < 0.05$).

4.2 Classifier ensemble factors

The second experiment was designed to fix the data preprocessing procedure and to examine the prediction performance of the various classifier ensembles combined with the best data preprocessing approach, meaning that the effect of combining the identified best data preprocessing approach and various classifier ensembles on the prediction performance was examined. The GA-SOM approach was chosen for the bankruptcy prediction datasets because it provides the lowest type II error rate and an AUC similar to that obtained with the best data preprocessing approach (i.e., PCA-AP). The *t*-test-SOM approach was chosen for the credit scoring datasets because it achieved the best and second-best performance in terms of the AUC and type II error, respectively.

4.2.1 Results for bankruptcy prediction datasets

Tables 7 and 8 present the AUC and type II errors obtained with various classifier ensembles based on the GA-SOM data preprocessing approach. The best result for each classification technique for each dataset is underlined. The results indicate that classifier ensembles do not always outperform single classifiers. However, upon examining the average AUC and type II errors of various classifiers over the five bankruptcy prediction datasets, much clearer conclusions can be made.

Figs. 4 and 5 show that, on average, ensemble techniques can improve the prediction performance of single classifiers. However, the ensemble method should be chosen carefully to make the various classifiers more effective. For example, in terms of the AUC and type II error, bagging DT outperforms single DT and boosting DT, whereas boosting SVM outperforms single SVM and bagging SVM.

By contrast, bagging ANN and bagging LR provide better AUC results than single ANN and LR and their ensembles by the boosting method. Boosting ANN and boosting LR provide lower type II errors.

The best and second-best AUC results were obtained by using bagging DT (i.e., 0.903) and bagging ANN (i.e., 0.895), respectively. Although the performance of these two ensemble classifiers did not differ significantly, they significantly outperformed the other classifiers ($p < 0.05$). In terms of the type II error, boosting SVM significantly outperformed the other classifiers ($p < 0.05$).

4.2.2 Results for credit scoring datasets

Tables 9 and 10 exhibit the AUC and type II errors obtained with various classifiers based on the GA-SOM data preprocessing approach. Similar to the previous results, classifier ensembles did not necessarily outperform single classifiers every time. Figs. 6 and 7 present the average AUC and type II error obtained with various classifiers over the five credit scoring datasets.

Figs. 6 and 7 indicate that bagging ANN and boosting SVM outperform their corresponding classifiers in terms of the AUC and type II error. Among DT classifier ensembles, bagging DT and boosting DT provided the best results in terms of the AUC and type II errors, respectively. Among LR ensembles, bagging LR exhibited a better AUC, and both the single LR classifier and boosting LR exhibited the lowest type II error.

Similar to the results for bankruptcy prediction datasets, bagging DT and bagging ANN were the best and second-best classifiers, respectively, in terms of AUC (i.e., 0.852 and 0.846, respectively). No significant

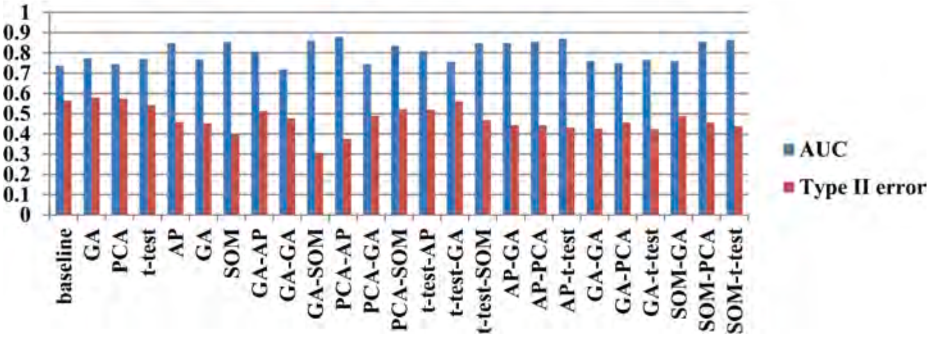

Fig. 2. . Average AUC and type II errors obtained with the different approaches.

Table 5
AUC under different pre-processing procedures (credit scoring).

	Australian	German	Japanese	Kaggle	PAKDD
Baseline					
	0.903	0.762	0.907	0.674	0.612
FS					
GA	0.914	0.738	0.912	0.617	0.607
PCA	0.924	0.722	0.911	0.645	0.610
t-test	0.920	0.725	0.909	0.674	0.613
IS					
AP	0.846	0.725	0.821	0.730	0.616
GA	0.903	0.697	0.903	0.673	0.609
SOM	0.898	0.737	0.848	0.826	0.619
FS-IS					
GA-AP	0.907	0.565	0.916	0.754	0.625
GA-GA	0.900	0.708	0.917	0.618	0.604
GA-SOM	0.918	0.795	0.924	0.626	0.631
PCA-AP	0.894	0.758	0.899	0.793	0.558
PCA-GA	0.872	0.688	0.857	0.618	0.611
PCA-SOM	0.950	0.796	0.982	0.805	0.604
t-test-AP	0.885	0.722	0.873	0.736	0.619
t-test-GA	0.916	0.681	0.894	0.670	0.610
t-test-SOM	0.944	0.778	0.959	0.826	0.638
IS-FS					
AP-GA	0.844	0.769	0.843	0.707	0.611
AP-PCA	0.881	0.757	0.857	0.712	0.619
AP-*t*-test	0.880	0.786	0.867	0.723	0.615
GA-GA	0.902	0.724	0.913	0.604	0.602
GA-PCA	0.891	0.710	0.903	0.681	0.607
GA-*t*-test	0.896	0.735	0.914	0.669	0.606
SOM-GA	0.980	0.788	0.885	0.820	0.623
SOM-PCA	0.953	0.646	0.901	0.824	0.610
SOM-*t*-test	0.922	0.732	0.911	0.826	0.615

Table 6
Type II error under different pre-processing procedures (credit scoring).

	Australian	German	Japanese	Kaggle	PAKDD
Baseline					
	0.160	0.527	0.167	0.985	1
FS					
GA	0.154	0.52	0.168	0.967	1
PCA	0.133	0.74	0.173	0.989	1
t-test	0.172	0.667	0.162	0.985	1
IS					
AP	0.194	0.407	0.144	0.96	1
GA	0.173	0.458	0.158	0.983	1
SOM	0.125	0.457	0.186	0.885	1
FS-IS					
GA-AP	0.155	0.597	0.127	1	1
GA-GA	0.193	0.482	0.156	0.973	1
GA-SOM	0.146	0.420	0.135	1	1
PCA-AP	0.1	0.35	0.05	0.68	0.984
PCA-GA	0.197	0.698	0.213	0.985	1
PCA-SOM	0.04	0.467	0.158	0.722	1
t-test-AP	0.129	0.525	0.138	0.958	1
t-test-GA	0.178	0.622	0.197	0.983	1
t-test-SOM	0.1	0.383	0.046	0.861	1
IS-FS					
AP-GA	0.141	0.359	0.123	0.96	1
AP-PCA	0.155	0.488	0.145	0.979	1
AP-*t*-test	0.156	0.37	0.159	0.961	1
GA-GA	0.173	0.437	0.186	0.97	1
GA-PCA	0.173	0.496	0.208	0.993	1
GA-*t*-test	0.173	0.447	0.208	0.983	1
SOM-GA	0.113	0.443	0.152	0.898	1
SOM-PCA	0.05	0.580	0.177	0.879	1
SOM-*t*-test	0.063	0.580	0.095	0.861	1

difference was identified in the performance of these two classifiers, and both significantly outperformed the other classifiers ($p < 0.05$). In terms of the type II error, boosting DT significantly outperformed the other classifiers ($p < 0.05$).

4.3 Further comparisons

To arrive at a more reliable conclusion about the effects of the data preprocessing and ensemble classification technique on the final prediction result, nine data preprocessing approaches combined with twelve different classification techniques for bankruptcy prediction and credit scoring datasets were further compared, resulting in 288 combinations for each dataset. For convenience, only the top five combinations in terms of the AUC and type II error are listed; Tables 11 and 12 show the results for the bankruptcy prediction and credit scoring datasets, respectively.

These results indicate that the three factors—feature selection, instance selection, and ensemble classification techniques—affect each other. Careful consideration of the combination of the data preprocessing algorithm and the classification technique is needed to develop superior prediction models for bankruptcy prediction and credit scoring. Furthermore, using the original dataset without performing feature and instance selection and constructing classifier ensembles is unlikely to produce the best prediction performance.

For bankruptcy prediction problems, no significant difference was identified in the performance of the top three combined approaches in terms of the AUC, namely, AP + bagging DT, AP-GA + bagging ANN, and AP-*t*-test + bagging ANN. Further, they significantly outperformed the fourth- and fifth-best approaches, namely, GA + bagging DT and original + bagging DT ($p < 0.05$). To realize the highest AUC, AP + bagging DT is recommended because it does not require the additional feature selection step.

Furthermore, no significant difference was identified in the performance of the top three combined approaches in terms of the type II error, namely, PCA-AP + LR, PCA-AP + boosting LR, and GA-SOM + boosting SVM. To realize the best type II error, PCA-AP + LR is recommended because it requires the lowest computational effort during the feature selection step—because it uses PCA instead of GA—and the classifier construction step—because it uses LR instead of boosting LR or boosting SVM.

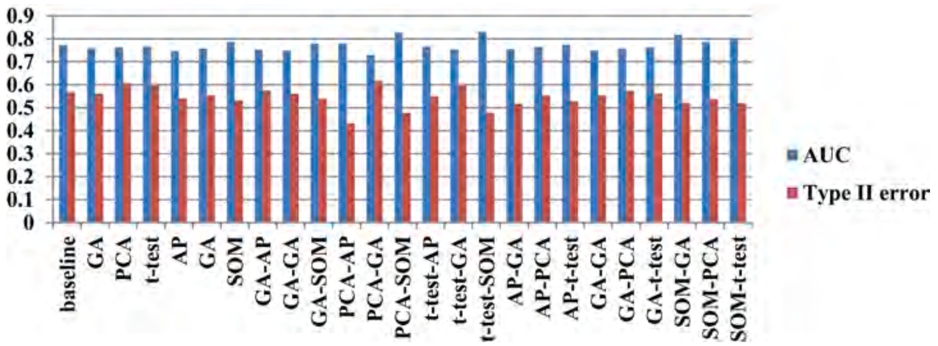

Fig. 3. . Average AUC and type II errors for the different approaches.

Table 7
AUC under different classifier ensembles.

	Bankruptcy	JPNBDS	TEJ-China	TEJ-Taiwan	USABDS
ANN					
Single	0.846	1.000	0.961	0.880	0.730
Bagging	0.863	1.000	0.962	0.893	0.755
Boosting	0.844	1.000	0.959	0.849	0.715
DT					
Single	0.862	0.950	0.897	0.648	0.712
Bagging	0.916	1.000	0.973	0.877	0.747
Boosting	0.884	0.950	0.972	0.869	0.696
LR					
Single	0.857	1.000	0.959	0.893	0.700
Bagging	0.863	1.000	0.959	0.889	0.700
Boosting	0.806	1.000	0.940	0.814	0.634
SVM					
Single	0.500	0.950	0.784	0.548	0.499
Bagging	0.820	1.000	0.920	0.619	0.589
Boosting	0.828	0.950	0.958	0.885	0.672

Table 8
Type II errors under different classifier ensembles.

	Bankruptcy	JPNBDS	TEJ-China	TEJ-Taiwan	USABDS
ANN					
Single	0.344	0.000	0.467	0.633	0.253
Bagging	0.344	0.000	0.517	0.700	0.193
Boosting	0.319	0.000	0.467	0.633	0.219
DT					
Single	0.308	0.100	0.383	0.767	0.247
Bagging	0.259	0.100	0.383	0.700	0.199
Boosting	0.288	0.100	0.367	0.700	0.247
LR					
Single	0.392	0.050	0.500	0.717	0.097
Bagging	0.382	0.050	0.517	0.667	0.101
Boosting	0.392	0.050	0.417	0.717	0.097
SVM					
Single	0.368	0.000	0.900	1.000	0.000
Bagging	0.368	0.000	0.900	0.983	0.003
Boosting	0.382	0.000	0.417	0.700	0.038

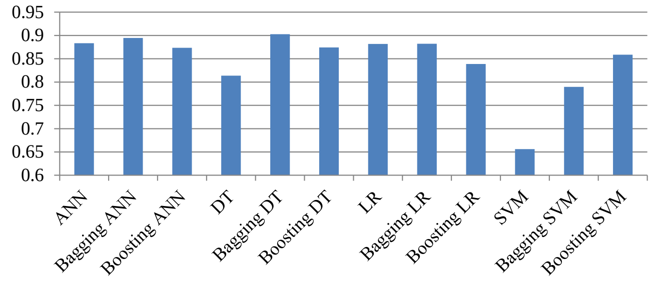

Fig. 4. . Average AUC obtained with the different classifiers.

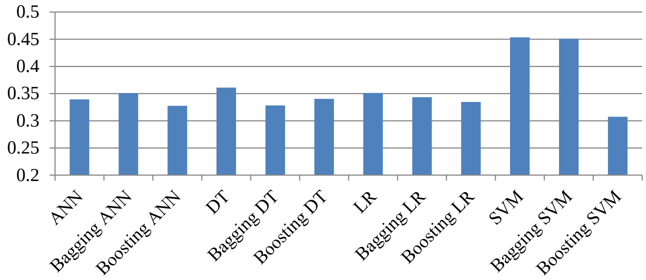

Fig. 5. . Average type II errors obtained with the different classifiers.

Table 9
AUC under different classifier ensembles (credit scoring).

	Australian	German	Japanese	Kaggle	PAKDD
ANN					
Single	0.940	0.950	0.786	0.620	0.837
Bagging	0.959	0.818	0.963	0.850	0.642
Boosting	0.944	0.737	0.942	0.793	0.603
DT					
Single	0.954	0.922	0.697	0.577	0.679
Bagging	0.970	0.843	0.968	0.844	0.634
Boosting	0.958	0.809	0.952	0.781	0.594
LR					
Single	0.940	0.969	0.775	0.655	0.831
Bagging	0.944	0.799	0.961	0.832	0.655
Boosting	0.927	0.741	0.953	0.736	0.622
SVM					
Single	0.949	0.940	0.777	0.500	0.500
Bagging	0.949	0.821	0.946	0.500	0.501
Boosting	0.944	0.778	0.959	0.826	0.638

Table 10
Type II error under different classifier ensembles (credit scoring).

	Australian	German	Japanese	Kaggle	PAKDD
ANN					
Single	0.100	0.093	0.333	0.907	0.871
Bagging	0.086	0.300	0.046	0.874	0.957
Boosting	0.100	0.358	0.068	0.871	0.906
DT					
Single	0.086	0.093	0.375	0.865	0.838
Bagging	0.071	0.325	0.046	0.839	0.897
Boosting	0.086	0.375	0.056	0.790	0.813
LR					
Single	0.114	0.046	0.433	0.981	0.876
Bagging	0.114	0.458	0.046	0.877	0.980
Boosting	0.114	0.433	0.046	0.876	0.981
SVM					
Single	0.071	0.046	0.275	1.000	1.000
Bagging	0.071	0.300	0.046	1.000	1.000
Boosting	0.100	0.383	0.046	0.861	1.000

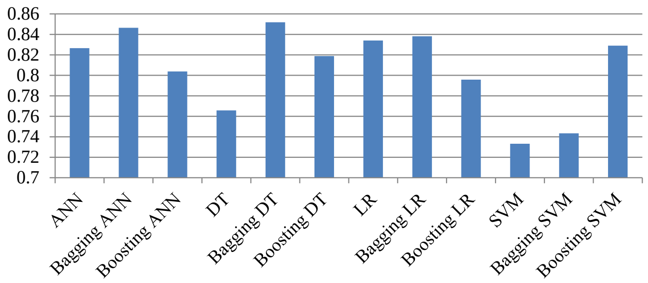

Fig. 6. . Average AUC obtained with different classifiers.

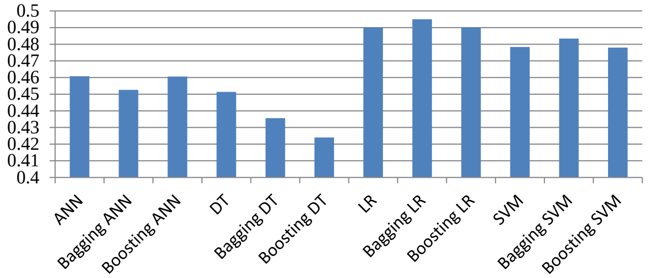

Fig. 7. . Average type II errors obtained with different classifiers.

Table 11
The top five combinations for the bankruptcy prediction datasets.

Top five combinations	AUC	Top five combinations	Type II error
AP + bagging DT	0.934	PCA-AP + LR	0.28
AP-GA + bagging ANN	0.931	PCA-AP + boosting LR	0.28
AP-*t*-test + bagging ANN	0.929	GA-SOM + boosting SVM	0.289
GA + bagging DT	0.914	GA-SOM + boosting LR	0.32
Original + bagging DT	0.913	GA-SOM + boosting ANN	0.33

Table 12
The top five combinations for the credit scoring datasets.

Top five combinations	AUC	Top five combinations	Type II error
t-test-SOM + bagging DT	0.852	PCA-AP + ANN	0.414
t-test-SOM + bagging ANN	0.846	*t*-test-SOM + boosting DT	0.424
t-test-SOM + bagging LR	0.838	PCA-AP + boosting SVM	0.433
t-test-SOM + LR	0.834	AP + boosting DT	0.44
SOM-GA + ANN	0.833	AP-*t*-test + boosting DT	0.44
SOM-GA + bagging ANN			

For the credit scoring problem, no significant difference was evident in the top two combined approaches in terms of the AUC and type II error. However, in light of the computational time, *t*-test-SOM + bagging DT and PCA-AP + ANN are recommended to realize the best AUC and type II error, respectively.

5 Conclusion

This study examined three important factors—feature selection, instance selection, and classifier ensembles—that affect the prediction performance of bankruptcy prediction models and credit scoring models. Most related studies that have used one of these three types of techniques have reported their positive impact on the final prediction performance. However, the present study is the first to consider all three types of techniques together for financial distress prediction.

To demonstrate the combinations of these three factors and determine superior bankruptcy prediction and credit scoring models, three feature selection algorithms (i.e., GA, PCA, and t-testing), three instance selection algorithms (i.e., AP, GA, and SOM), and four classification techniques (i.e., ANN, DT, LR, and SVM) as well as the bagging and boosting ensemble algorithms were individually combined to construct classifier ensembles, leading to 288 combinations for performance comparison.

Studies have found that performing data preprocessing by using either feature or instance selection and constructing the prediction model by using ensemble learning techniques can improve the prediction performance. Our experiment results indicate that these three types of techniques should be carefully combined to develop more effective bankruptcy prediction and credit scoring models. In particular, the order in which feature and instance selection and related algorithms used for feature selection, instance selection, and model construction steps are performed greatly affects the final prediction performance.

For bankruptcy prediction, the top three combinations exhibited insignificant performance differences from each other and significantly outperformed the other combinations in terms of the AUC and type II error. However, considering the computational cost for data preprocessing and the classifier training, AP + bagging DT is the best choice to achieve the highest AUC, and PCA-AP + LR is the best choice to achieve the lowest type II error rate.

For credit scoring, some combinations exhibited insignificant performance differences from each other and yielded a superior AUC and type II error than did other combinations. The approaches of *t*-test-SOM + bagging DT and PCA-AP + ANN are the superior choices for realizing the best AUC and type II error, respectively, in consideration of lower computational costs.

The results of this study have practical value for both general and institutional investors. When faced with a high-dimensional and large amount of corporate data, it is difficult for general investors to make reasonable investment decisions. The proposed approach can help determine the key factors of financial distress and develop a decision support system for investment recommendations, thereby enabling investors to avoid investing in high-risk companies. With respect to institutional investors, as they may suffer severe losses from the bankruptcy of investment companies, it is a crucial task to detect financial distress as early as possible. A real-time financial distress prediction model with high precision and low tolerance of error is required. Our study gives sufficient evidence of the effectiveness of combining a number of data preprocessing and classifier ensemble techniques. Institutional investors can develop early warning systems based on the proposed approaches to provide more optimal performance of financial distress prediction.

This study has some limitations that can be improved upon in future studies. First, other more sophisticated feature and instance algorithms can be used for performance comparison, such as ensemble feature selection (Pes, 2020; Tsai & Sung, 2020), constrained nearest neighbor–based instance selection (Yang et al., 2019), and natural ensemble margin instance selection (Saidi, Bechar, & A., Settouti, N., and Chikh, M.A. , 2018). Second, because many financial distress prediction datasets are class imbalanced, some data sampling approaches such as over- and under-sampling can be performed to balance the datasets before or after the feature and instance selection steps to examine the performance improvement (Galar et al., 2012). Third, future studies can employ related deep learning techniques such as deep belief networks, deep neural networks, and convolutional neural networks (Luo, Wu, & Wu, 2019; Qu, Quan, Lei, & Shi, 2019).

Acknowledgements

This research was supported in part by the Ministry of Science and Technology of the Republic of China (grant number MOST 109-2410-H-008-059-MY2).

References

Ahn, H., & Kim, K.-J. (2009). Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing, 9*, 599–607.

Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications, 94*, 164–184.

Bolon-Canedo, V., Sanchez-Marono, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems, 34*, 483–519.

Breiman, L. (1994). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Chandrashekar, G., & Sahin, F. (2004). A survey on feature selection methods. *Computers and Electrical Engineering, 40*, 16–28.

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: A recent review. *Artificial Intelligence Review, 45*, 1–23.

Choi, H., Son, H., & Kim, C. (2018). Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications, 110*, 1–10.

Climent, F., Momparler, A., & Carmona, P. (2019). Anticipating bank distress in the Eurozone: An extreme gradient boosting approach. *Journal of Business Research, 101*, 885–896.

Derrick, B., Toher, D., & White, P. (2016). Why Welch's test is type I error robust? *The Quantitative Methods for Psychology, 12*(1), 30–38.

Du Jardin, P. (2018). Failure pattern-based ensembles applied to bankruptcy forecasting. *Decision Support Systems, 107*, 64–77.

Feng, X., Xiao, Z., Zhong, B., Qiu, J., & Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability. *Applied Soft Computing, 65*, 139–151.

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research, 210*, 368–378.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and application to boosting. *Journal of Computer and System Sciences, 55*, 119–139.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972–976.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 42*(4), 463–484.

Garcia, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(3), 417–435.

Garcia, V., Marques, A. I., & Sanchez, J. S. (2019). Exploring the synergetic effects of sample types on the performance of ensembles for credit risk and corporate bankruptcy prediction. *Information Fusion, 47*, 8–101.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Jadhav, S., He, H., & Jenkins, K. (2018). Information gain directed genetic algorithm wrapper feature selection for credit rating. *Applied Soft Computing, 69*, 541–553.

Kim, K. J., Lee, K., & Ahn, H. (2018). Predicting corporate financial sustainability using novel business analytics. *Sustainability, 11*(1), 64–80.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – a review. *European Journal of Operational Research, 180*(1), 1–28.

Li, Y., Li, T., & Liu, H. (2017). Recent advances in feature selection and its applications. *Knowledge and Information Systems, 53*(3), 551–577.

Liang, D., Lu, C.-C., Tsai, C.-F., & Shih, G.-A. (2016). Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. *European Journal of Operational Research, 252*(2), 561–572.

Liang, D., Tsai, C.-F., & Wu, H.-T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems, 73*, 289–297.

Liang, D., Tsai, C. F., Lu, H. Y. R., & Chang, L. S. (2020). Combining corporate governance indicators with stacking ensembles for financial distress prediction. *Journal of Business Research, 120*, 137–146.

Lin, W.-C., Lu, Y.-H., & Tsai, C.-F. (2019). Feature selection in single and ensemble learning-based bankruptcy prediction models. *Expert Systems, 36*(1), 1–8.

Lin, W.-Y., Hu, Y.-H., & Tsai, C.-F. (2012). Machine learning in financial crisis prediction: A survey. *IEEE Transactions on Systems, Man and Cybernetics – Part C: Applications and Reviews, 42*(4), 421–436.

Liu, Z. F., & Pan, S. (2018). Fuzzy-rough instance selection combined with effective classifiers in credit scoring. *Neural Processing Letters, 47*(1), 193–202.

Luo, C., Wu, D., & Wu, D. (2019). A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence, 65*, 465–470.

Maldonado, S., Perez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research, 261*, 656–665.

Masoudnia, S., & Ebrahimpour, R. (2014). Mixture of experts: A literature review. *Artificial Intelligence Review, 42*, 275–293.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems, 52*(2), 464–473.

Olvera-Lopez, J. A., Carrasco-Ochoa, J. A., Martinez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review, 34*, 133–143.

Pérez-Martín, A., Pérez-Torregrosa, A., & Vaca, M. (2018). Big Data techniques to measure credit banking risk in home equity loans. *Journal of Business Research, 89*, 448–454.

Pes, B. (2020). Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. *Neural Computing and Applications, 32*, 5951–5973.

Qu, Y., Quan, P., Lei, M., & Shi, Y. (2019). Review of bankruptcy prediction using machine learning and deep learning techniques. *Procedia Computer Science, 162*, 895–899.

Rokach, L. (2009). Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics and Data Analysis, 53*, 4046–4072.

Saidi, M., Bechar, M. El A., Settouti, N., and Chikh, M.A. (2018) Instance selection algorithm by ensemble margin. Journal of Experimental& Theoretical Artificial Intelligence, vol. 30, no. 3, pp. 457-478.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning, 5*(2), 197–227.

Sun, J., Jia, M.-Y., & Li, H. (2011). AdaBoost ensemble for financial distress prediction: An empirical comparison with data from Chinese listed companies. *Expert Systems with Applications, 38*, 9305–9312.

Sun, J., Fujita, H., Chen, P., & Li, H. (2017). Dynamic financial distress prediction with concept drift based on time weighting combined with Adaboost support vector machine ensemble. *Knowledge-Based Systems, 120*, 4–14.

Sun, J., Li, H., Fujita, H., Fu, B., & Ai, W. (2020). Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Information Fusion, 54*, 128–144.

Tang, X., Li, S., Tan, M., & Shi, W. (2020). Incorporating textual and management factors into financial distress prediction: A comparative study of machine learning methods. *Journal of Forecasting, 39*(5), 769–787.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149–172.

Tsai, C.-F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion, 16*, 46–58.

Tsai, C.-F., & Cheng, K.-C. (2012). Simple instance selection for bankruptcy prediction. *Knowledge-Based Systems, 27*, 333–342.

Tsai, C.-F., & Chen, Z.-Y. (2014). Towards high dimensional instance selection: An evolutionary approach. *Decision Support Systems, 61*, 79–92.

Tsai, C.-F., & Sung, Y.-T. (2020). Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. *Knowledge-Based Systems, 203*, Article 106097.

Tsai, C.-F., Hsu, Y.-F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing, 24*, 977–984.

Wozniak, M., Grana, M., & Corchado, E. (2014). A survey of multiple classifier systems as hybrid systems. *Information Fusion, 16*, 3–17.

Yang, L., Zhu, Q., Huang, J., Wu, Q., Cheng, D., & Hong, X. (2019). Constraint nearest neighbor for instance selection. *Soft Computing, 23*, 13235–13245.

Zhou, L., & Lai, K. K. (2017). AdaBoost Models for corporate bankruptcy prediction with missing data. *Computational Economics, 50*(1), 69–94.

Chih-Fong Tsai is now a professor at the Department of Information Management, National Central University, Taiwan. He received a PhD degree at School of Computing and Technology from the University of Sunderland, UK in 2005. He has published more than 30 refereed journal papers including *ACM Transactions on Information Systems*, *Decision Support Systems*, *Pattern Recognition*, *Information Processing & Management*, *Applied Soft Computing*, *Neurocomputing*, *Knowledge-Based Systems*, *Expert Systems with Applications*, *Expert Systems*, *Online Information Review*, *International Journal on Artificial Intelligence Tools*, *Journal of Systems and Software*, etc. He received the Distinguished New Faculty Award from National Central University in 2010 and the Highly Commended Award (Emerald Literati Network 2008 Awards for Excellence) for a paper published in *Online Information Review* ("A Review of Image Retrieval Methods for Digital Cultural Heritage Resources"). His current research focuses on multimedia information retrieval and data mining applications.

Kuen-Liang Sue is now an assistant professor at the Department of Information Management, National Central University, Taiwan. He received a PhD degree in Information Management from National Chiao Tung University of Taiwan in 2003. His current research focuses on data mining, machine learning, FinTech, and information security.

Ya-Han Hu is currently a Professor of Department of Information Management at National Central University, Taiwan. He received a PhD degree in Information Management from National Central University of Taiwan in 2007. His current research interests include text mining and information retrieval, clinical decision support systems, and recommender systems. His research has appeared in *Information & Management*, *Decision Support Systems*, *Journal of the American Society for Information Science and Technology*, *IEEE Transactions on Systems, Man, and Cybernetics*, *International Journal of Information Management*, *Artificial Intelligence in Medicine*, *Applied Soft Computing*, *Computers in Human Behavior*, *Data & Knowledge Engineering*, *Expert Systems*, *Knowledge-Based Systems*, *Information Systems and e-Business Management*, *Journal of Information Science*, *Journal of Clinical Epidemiology*, *Methods of Information in Medicine*, *Online Information Review*, and *Journal of Systems and Software*.

Andy Chiu received his MS degree in Information Management from National Central University of Taiwan in 2018. His research interests include data mining, information retrieval and EC technologies.