Information Fusion 16 (2014) 46–58

Contents lists available at SciVerse ScienceDirect

Information Fusion

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

Combining cluster analysis with classifier ensembles to predict financial distress

Chih-Fong Tsai *

Department of Information Management, National Central University, Taiwan

ARTICLE INFO

Article history:
Received 6 October 2010
Received in revised form 24 March 2011
Accepted 7 December 2011
Available online 21 December 2011

Keywords:
Financial distress
Machine learning
Classifier ensembles
Hybrid classifiers
Bankruptcy prediction
Credit scoring

ABSTRACT

The ability to accurately predict business failure is a very important issue in financial decision-making. Incorrect decision-making in financial institutions is very likely to cause financial crises and distress. Bankruptcy prediction and credit scoring are two important problems facing financial decision support. As many related studies develop financial distress models by some machine learning techniques, more advanced machine learning techniques, such as classifier ensembles and hybrid classifiers, have not been fully assessed. The aim of this paper is to develop a novel hybrid financial distress model based on combining the clustering technique and classifier ensembles. In addition, single baseline classifiers, hybrid classifiers, and classifier ensembles are developed for comparisons. In particular, two clustering techniques, Self-Organizing Maps (SOMs) and *k*-means and three classification techniques, logistic regression, multilayer-perceptron (MLP) neural network, and decision trees, are used to develop these four different types of bankruptcy prediction models. As a result, 21 different models are compared in terms of average prediction accuracy and Type I & II errors. By using five related datasets, combining Self-Organizing Maps (SOMs) with MLP classifier ensembles performs the best, which provides higher predication accuracy and lower Type I & II errors.

1. Introduction

Financial crises and distress can cause many social costs affecting owners or shareholders, managers, workers, lenders, suppliers, clients, the community and government, etc. Therefore, it is very important to predict business failure. Consequently, to develop a financial distress prediction model has long been regarded as an important and widely studied issue in the academic and business community [1,2]. Bankruptcy prediction, for example, has a significant impact on lending decisions and the profitability of financial institutions [3]. Before extending a loan, banks need to predict the possibility of failure of the potential counterparty.

In addition, with the rapid growth in credit industry and the management of large loan portfolios, credit scoring models have been extensively used for the credit admission evaluation. Financial institutions also need to determine if loan customers belong to either a good applicant group or a bad applicant group. That is, credit scoring models help to decide whether to grant credit to new applicants using customer's characteristics such as age, income and martial status [4].

In the past, the research problem of bankruptcy prediction and credit scoring is usually solved by statistical approaches, such as logistic regression and regression analysis. However, many recent studies have considered machine learning techniques and show that machine learning techniques are superior to that of (traditional) statistical methods, e.g. [5–9]. A detailed review can be refereed to Kumar and Ravi [1].

In machine learning, combining multiple classifiers (i.e. classifier ensembles) has recently been an active research area. They aim at obtaining highly accurate classifiers by combining less accurate ones. That is, the combination is able to complement the errors made by the individual classifiers on different parts of the input space [10]. In the literature, classifier ensembles have shown better prediction performances than many single classifiers, e.g. [11].

Another advanced machine learning technique focuses on combining clustering and classification techniques as the hybrid models. That is, clustering is used for pre-classification which is to arrange a given collection of input patterns into natural groupings or meaningful clusters based on a measure of similarity [12]. Then, the clustering results as the representative data groups are used to construct a classifier. In Hsieh [13], the cluster analysis is used to filter out unrepresentative data (i.e. outliers) to reduce prediction errors during training a classifier.

As related work has shown that classifier ensembles and hybrid classifiers outperform single classifiers, such as [11] and [13], they have not been compared each other, especially in the bankruptcy prediction domain. In addition, there is no related study combining

* Tel.: +886 3 422 7151; fax: +886 3 4254604.
E-mail address: cftsai@mgt.ncu.edu.tw

1566-2535/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.inffus.2011.12.001

the clustering method and classifier ensembles as a more sophisticated model to predict bankruptcy.

Therefore, the aim of this paper is to examine the prediction performance by comparing single and advanced machine learning techniques. In particular, the combination of the clustering method and classifier ensembles is compared with classifier ensembles, hybrid classifiers, and three well-known single classifiers, i.e. logistic regression, the multilayer-perceptron (MLP) neural network, and decision trees, as the baseline classifiers. Totally, 21 different prediction models are examined over five public bankruptcy prediction and credit scoring datasets. In addition, two evaluation measures are used for the assessment task over the testing and validation sets. They are the rate of prediction accuracy, Type I & II errors. For model validation, it is based on the outlier data obtained from each of the five datasets by the clustering method (c.f. Section 3.3).

Consequently, the contribution of this paper is three-fold. First, besides single machine learning techniques, we further compare three advanced machine learning techniques, which are (a) the combination of the clustering method and classifier ensembles, (b) classifier ensembles, and (c) hybrid ensembles. Second, the evaluation strategy using the noisy data as the validation set is considered to more rigorously examine the stableness and robustness of these techniques. Third, instead of artificial neural networks and other single machine learning techniques, the finding of this paper can allow us to identify the best prediction model, which can be regarded as the reliable baseline for future research.

This paper is organized as follows. Section 2 reviews literatures which include the description of financial distress, the introduction of related machine learning techniques, and the comparison and analysis of related work. In addition, the general evaluation method considered in the literature is overviewed. Section 3 describes the research methodology including the datasets used, construction of different models, and the evaluation strategies. Section 4 presents the experimental results. Conclusion and future work are provided in Section 5.

2. Literature review

2.1. Financial distress

A firm sometimes can become distressed if it takes on a higher level of debts and continues to operate in that condition for many years. That is, the company cannot meet or has difficulty to payoff the financial obligation. Therefore, financial distress sometimes can lead to bankruptcy. Some firms enter bankruptcy immediately after a highly distressing event, such as a major fraud. There are a number of factors affecting bankruptcy potentially. They are audit, financial ratios, fraud indicators, start-up and stress which are measured by qualitative or quantitative variables [14].

As a result, for financial institutions to accurately predict whether the loan customers belong to a good applicant group or not, i.e. to grant credit to new applicants, is necessary.

2.2. Machine learning techniques

2.2.1. Classification

Classification belongs to supervised machine learning. It can be regarded as learning by examples or learning with a teacher [15]. The teacher has knowledge of the environment which is represented by a set of input–output examples. In order to classify unknown patterns, a certain number of training samples are available for each class, and they are used to train the classifier [16].

The learning task is to compute a classifier or model that approximates the mapping between the input–output examples and correctly labels the training set with some level of accuracy. This can be called as the *training* or *model generation* stage. After the model is generated or trained, it is able to classify an unknown instance, into one of the learned class labels in the training set.

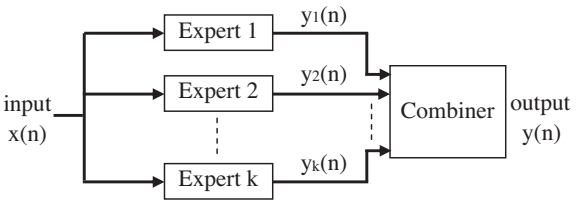

Fig. 1. Architecture of a classifier ensemble.

In this paper, three mostly used classification techniques for bankruptcy prediction are considered, which are logistic regression, multilayer-perceptron (MLP) neural network, and decision trees [1].

2.2.2. Clustering

Clustering, as opposed to classification, is to arrange a given collection of unlabelled input patterns into natural groupings or meaningful clusters based on a measure of similarity [12]. That is, it seeks to extract information from unlabelled samples [17]. Pattern clustering results in a number of well-separated clusters in the feature space which provide a summarization and visualisation of data in the given collection.

Two well-known clustering techniques are used in this paper, which are Self-Organizing Maps (SOMs) and *k*-means [17].

2.2.3. Classifier ensembles

Classifier ensembles are based on combining multiple classifiers. They aim at obtaining highly accurate classifiers by combining less accurate ones. They are proposed to improve the classification performance of a single classifier [18]. That is, the combination is able to complement the errors made by the individual classifiers on different parts of the input space. Therefore, the performance of classifier ensembles is likely better than the one of the best single classifier used in isolation [10].

The simplest method to combine classifiers is majority voting. The outputs of different individual classifiers are pooled together. Then, the output class which receives the largest number of votes is selected as the final classification decision. On the other hand, the weighted voting approach considers some output result(s) of classifiers which have more weights than others for the final classification output [18].

Fig. 1 shows the general architecture of a classifier ensemble [16]. A number of differently trained neural networks (i.e. experts) share the input and whose outputs are combined to produce an overall output. More specifically, the experts can be trained by different examples (or different features) of a given training set or different learning models trained by the same training set.

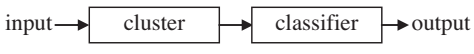

Fig. 2. Architecture of a hybrid classifier.

Table 1
The confusion matrix for prediction accuracy and Type I/II errors.

↓ actual/predicted →	Bad credit/bankruptcy	Good credit/non-bankruptcy
Bad credit/bankruptcy	(a)	II (b)
Good credit/non-bankruptcy	I (c)	(d)