Expert Systems with Applications 39 (2012) 9308–9314

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction

Myoung-Jong Kim [a], Dae-Ki Kang [b,*]

[a] School of Business, Pusan National University, South Korea
[b] Division of Computer & Information Engineering, Dongseo University, South Korea

ARTICLE INFO

Keywords:
Ensemble learning
Genetic algorithm
Coverage optimization
Bankruptcy prediction

ABSTRACT

Ensemble learning is a method to improve the performance of classification and prediction algorithms. Many studies have demonstrated that ensemble learning can decrease the generalization error and improve the performance of individual classifiers and predictors. However, its performance can be degraded due to multicollinearity problem where multiple classifiers of an ensemble are highly correlated with. This paper proposes a genetic algorithm-based coverage optimization technique in the purpose of resolving multicollinearity problem. Empirical results with bankruptcy prediction on Korea firms indicate that the proposed coverage optimization algorithm can help to design a diverse and highly accurate classification system.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Since bankruptcy is a critical event that could inflict a great loss to management, stockholders, employees, customers and nation, the development of bankruptcy prediction models has been one of important issues in accounting and finance research fields.

The widely used methods for developing bankruptcy prediction models are statistics and machine learning. The statistical techniques, including multiple regression, discriminant analysis, logistic models, and probit, have been traditionally used in forecasting business failures (Altman, 1968; Altman, Edward, Haldeman, & Narayanan, 1977; Dimitras, Zanakis, & Zopounidis, 1996; Meyer & Pifer, 1970; Ohlson, 1980; Pantalone & Platt, 1987; Zmijewski, 1984). However, one major drawback is that it should be based on strict assumptions. Such strict assumptions include linearity, normality, independence among predictor variables and pre-existing functional forms relating the criterion variables and the predictor variables. Those strict assumptions of traditional statistics have limited their application to the real world.

Machine learning techniques also used in bankruptcy prediction models include decision trees (DT), neural networks (NN), and Support Vector Machine (SVM) (Bryant, 1997; Buta, 1994; Han, Chandler, & Liang, 1996; Laitinen & Kankaanpaa, 1999; Min, Lee, & Han, 2006; Odom & Sharda, 1990; Ravi & Ravi, 2007; Shaw & Gentry, 1998; Shin, Lee, & Kim, 2005).

One of the recent techniques applied in bankruptcy prediction is ensemble learning (Alfaro, García, Gámez, & Elizondo, 2008; Alfaro, Gámez, & García, 2007; Kim & Kang, 2010). Ensemble learning is a machine learning technique for improving the performance of individual classifiers and predictors. Basically, ensemble learning constructs a highly accurate classifier (a single strong classifier) on the training set by combining an ensemble of weak classifiers, each of which needs only to be moderately accurate on the training set. Many studies on ensemble learning have shown an experimental confirmation and a theoretical explanation that combination of diverse hypotheses can produced a strong ensemble, whose error is reduced with respect to the average error of members. In the last decade, many studies have applied ensemble learning for designing high performance classification systems, mainly in terms of classification accuracy, in several pattern recognition tasks such as alphanumeric character recognition and face recognition (Czyz, Sadeghi, Kittler, & Vandendorpe, 2004; Lemieux & Parizeau, 2003; Zhou & Zhang, 2002). Recently, empirical studies on bankruptcy prediction have also demonstrated the reduction in generalization error and the prominent performance improvement (Alfaro et al., 2007, 2008; Kim & Kang, 2010).

However, some studies have reported the performance degradation problem of ensemble learning caused by the multicollinearity among classifiers. (Buciu, Kotrooulos, & Pitas, 2001; Dong & Han, 2004; Eom, Kim, & Zhang, 2008; Valentini, Muselli, & Ruffino, 2003). Several studies have proposed coverage optimization to

* Corresponding author. Address: Division of Computer & Information Engineering, Dongseo University, 47, Churye-Ro, Sasang-Gu, Busan, 617-716, South Korea. Tel.: +82 51 320 1724; fax: +82 51 327 8955.
E-mail address: dkkang@dongseo.ac.kr (D.-K. Kang).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.072

cope with such problem (Banfield, Hall, Bowyer, & Kegelmeyer, 2003; Giacinto & Roli, 2001; Valentino, 2005). Coverage optimization, also known as diversity-based classifier selection, is a method for selecting classifiers in order to decrease the number of ensemble members and keeping the diversity among the selected members as well (Santana, Soares, Canuto, & Soouto, 2006). Those experimental studies have reported that the optimized ensembles have fewer classifiers than the original ensembles, but their accuracies are higher than the original ensembles.

This paper proposes a genetic algorithms-based coverage optimization system for ensemble learning. The optimal (or near optimal) classifiers subset is selected based on prediction accuracy and diversity measurement represented as statistical value of variance influence factor (VIF). The proposed coverage optimization is applied to a company failure prediction task to validate the effect on the performance improvement. Experimental results with the bankruptcy prediction on Korean firms indicate that the proposed genetic algorithms-based coverage optimization can help to design a diverse and highly accurate classification system.

The remainder of this paper is organized as follows: The next section describes two popular ensemble algorithms Bagging and Boosting, and the diversity problem in ensemble learning. Section 3 explains the algorithm of the proposed coverage optimization. Section 4 presents data descriptions and experimental design process. Section 5 discusses experimental results. The final section presents several concluding remarks and future research issues.

2. The diversity problem in ensemble learning

Several ensemble methods for constructing and combining a collection of classifiers have been proposed. Two main methods which have been widely used are Bagging (Breiman, 1994) and Boosting (Freund, 1995; Schapire, 1990).

Bagging is a method that creates and combines multiple classifiers, each of which is trained on a bootstrap replicate of the original training set. The bootstrap data is created by resampling examples uniformly with replacement from the original training set. Each classifier is created by training on corresponding bootstrap replicate. The classifiers could be trained in parallel and the final classifier is generated by combining ensemble of classifier. Bagging has been considered as a variance reduction technique for a given classifier. Bagging is known to be particularly effective when the classifiers are unstable, that is, when perturbing the learning set can cause significant changes in the classification behavior, because Bagging improves generalization performance due to a reduction in variance while maintaining or only slightly increasing bias.

Boosting constructs a composite classifier by sequentially training classifiers while increasing weight on the misclassified observations through iterations. The observations that are incorrectly predicted by previous classifiers are chosen more often than examples that are correctly predicted. Thus Boosting attempts to produce new classifiers that are better able to predict examples for which the current ensemble's performance is poor. Boosting combines predictions of ensemble of classifiers with weighted majority voting by giving more weights on more accurate predictions.

In the last decade, many studies have applied ensemble learning to designing high performance classification systems. Particularly, many empirical studies using DT as a base classifier have been shown that ensemble learning can enhance the prediction performance of DT classification algorithms such as Classification and Regression Tree (CART) and C4.5 (Banfield, Hall, Bowyer, & Kegelmeyer, 2007; Bauer & Kohavi, 1999; Drucker & Cortes, 1996; Quinlan, 1996). Recently, several studies have applied ensemble learning to bankruptcy classification trees. They have shown that ensemble learning decreases the generalization error and improve the accuracy (Alfaro et al., 2007).

On the other hand, many studies on NN/SVM ensemble have also reported that ensemble learning can improve individual classifier's accuracy. However, some studies have indicated that the ensemble combination with NN/SVM is less effective than DT ensemble in the respect of performance improvement and that ensemble's performance is often even worse than that of a single classifier (Buciu et al., 2001; Dong & Han, 2004; Eom et al., 2008; Valentini et al., 2003). Several works have investigated the cause of performance degradation and insisted that the performance of ensemble can be degraded where multiple classifiers of an ensemble are highly correlated with, and thereby result in multicollinearity problem, which leads to performance degradation of the ensemble (Banfield et al., 2003; Breiman, 1994; Giacinto & Roli, 2001; Hansen & Salamon, 1990; Valentino, 2005).

Hansen and Salamon (1990) insisted that it is necessary and sufficient for the performance enhancement of an ensemble that the ensemble should contain diverse classifiers and each classifier in the ensemble needs to be more accurate than random guess. This means that the accuracy of each classifier in the ensemble should be over 50% when there are two class labels, and the classifiers in the ensemble should be diverse to minimize mis-classification rate. Therefore, the key to successful ensemble methods is to construct individual classifiers with error rates below 0.5 whose errors are at least somewhat uncorrelated.

Breiman's work (1994) reported that bagging (and to a lesser extent, boosting) can increase the performance of unstable learning algorithms, but does not show remarkable performance improvement on stable learning algorithms. Ensemble learning applies various sampling techniques such as bagging, boosting, etc. to guarantee the diversity in a classifier pool. Unstable learning algorithms such as DT learners are sensitive to the change of the training data, and thus small changes in the training data can yield large changes in the generated classifiers. Therefore, ensemble with unstable learning algorithms can guarantee some diversity among the classifiers. To the contrary, stable learning algorithms such as NN/SVM generate similar classifiers in spite of the changes of the training data, and thus the correlation among the resulting classifiers is very high. This high correlation results in multicollinearity problem, which leads to performance degradation of the ensemble.

The concept of the coverage optimization is introduced to cope with performance degradation due to multicollinearity problem. Coverage optimization is a method for selecting classifiers in order to decrease the number of ensemble members and, at the same time, keeping the diversity among the selected members. It arises from the intuition that a set of dissimilar classifiers would perform better than a single good decision maker because its error is compensated by the decisions of the others. For example, there is clearly no accuracy gain in an ensemble that is composed of a set of identical classifiers. Thus, if there are many different classifiers to be combined, one would expect an increase in the overall accuracy when combining them, as long as they are divers (Santana et al., 2006).

In the previous literature, several studies have proposed the methods for the diversity-based classifier selection problem (Banfield et al., 2003; Giacinto & Roli, 2001; Valentino, 2005). For instance, classifiers could be clustered based on the diversity they produce. In prediction task, one classifier of each group is selected to be a member of the ensemble to avoid multicollinearity problem because classifiers that belong to the same group tend to make correlated errors (Giacinto & Roli, 2001). Banfield et al. (2003) proposed an ensemble diversity procedure based on uncertain points (patterns). These uncertain points are considered to deliver diversity to the ensemble, since there is no general agreement among

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات