Expert Systems with Applications 38 (2011) 223–230

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A comparative assessment of ensemble learning for credit scoring

Gang Wang [a,b,*], Jinxing Hao [b,c], Jian Ma [b], Hongbing Jiang [b]

[a] *School of Management, Hefei University of Technology, Hefei, Anhui 230009, PR China*
[b] *Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong*
[c] *School of Economics and Management, BeiHang University, Beijing 100083, PR China*

ARTICLE INFO

Keywords:
Credit scoring
Ensemble learning
Bagging
Boosting
Stacking

ABSTRACT

Both statistical techniques and Artificial Intelligence (AI) techniques have been explored for credit scoring, an important finance activity. Although there are no consistent conclusions on which ones are better, recent studies suggest combining multiple classifiers, i.e., ensemble learning, may have a better performance. In this study, we conduct a comparative assessment of the performance of three popular ensemble methods, i.e., Bagging, Boosting, and Stacking, based on four base learners, i.e., Logistic Regression Analysis (LRA), Decision Tree (DT), Artificial Neural Network (ANN) and Support Vector Machine (SVM). Experimental results reveal that the three ensemble methods can substantially improve individual base learners. In particular, Bagging performs better than Boosting across all credit datasets. Stacking and Bagging DT in our experiments, get the best performance in terms of average accuracy, type I error and type II error.

1. Introduction

The recent world financial tsunami arouses unprecedented attention of financial institutions on credit risk. Credit scoring has become one of the primary ways for financial institutions to assess credit risk, improve cash flow, reduce possible risks and make managerial decisions (Huang, Chen, & Wang, 2007).

The purpose of credit scoring is to classify the applicants into two types: applicants with good credit and applicants with bad credit. Applicants with good credit have great possibility to repay financial obligation. Applicants with bad credit have high possibility of defaulting. The accuracy of credit scoring is critical to financial institutions' profitability. Even 1% of improvement on the accuracy of credit scoring of applicants with bad credit will decreases a great loss for financial institutions (Hand & Henley, 1997).

Credit scoring was originally evaluated subjectively according to personal experiences, and later it was based on 5Cs: the character of the consumer, the capital, the collateral, the capacity and the economic conditions. But with the tremendous increase of applicants, it is impossible to conduct the work manually. Two categories of automatic credit scoring techniques, i.e., statistical techniques and Artificial Intelligence (AI) techniques, have been studied by prior researches (e.g., Huang, Chen, Hsu, Chen, & Wu, 2004).

Some statistical techniques have been widely applied to build the credit scoring models, such as Linear Discriminant Analysis (LDA) (Karels & Prakash, 1987; Reichert, Cho, & Wagner, 1983), Logistic Regression Analysis (LRA) (Thomas, 2000; West, 2000), Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991). However, the problem with applying these statistical techniques to credit scoring is that some assumptions, such as the multivariate normality assumptions for independent variables, are frequently violated in the practice of credit scoring, which makes these techniques theoretically invalid for finite samples (Huang et al., 2004).

In recent years, many studies have demonstrated that AI techniques such as Artificial Neural Networks (ANN) (Desai, Crook, & Overstreet, 1996; West, 2000), Decision Tree (DT) (Hung & Chen, 2009; Makowski, 1985), Case-Based Reasoning (CBR) (Buta, 1994; Shin & Han, 2001), and Support Vector Machine (SVM) (Baesens et al., 2003; Huang et al., 2007; Schebesch & Stecking, 2005) can be used as alternative methods for credit scoring. In contrast with statistical techniques, AI techniques do not assume certain data distributions. These techniques automatically extract knowledge from training samples. According to previous studies, AI techniques are superior to statistical techniques in dealing with credit scoring problems, especially for nonlinear pattern classification (Huang et al., 2004).

However, there is no overall best AI techniques used in building credit scoring models, for what is best depends on the details of the problem, the data structure, the characteristics used, the extent to which it is possible to segregate the classes by using those characteristics, and the objective of the classification (Hand & Henley, 1997; Yu, Wang, & Lai, 2008). Recently, there is a growing interest

* Corresponding author at: Department of Information Systems, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong. Tel.: +852 9799 0955; fax: +852 2788 8694.

E-mail address: wgedison@gmail.com (G. Wang).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.06.048

that existing applications of single AI technique can be further improved by ensemble methods. Latest researches (Hung & Chen, 2009; Yu et al., 2008) have shown that such ensemble methods have performed better than single AI technique for credit scoring. However, the application of ensemble methods in credit scoring is a relatively new and untried area. To the best of our knowledge, this may be the first attempt to systematically compare of classical ensemble methods for credit scoring.

Base on these considerations, we conduct a comparative assessment of the performance of three popular ensemble methods—Bagging, Boosting, and Stacking—on credit scoring problems. The aim of this study is to examine the performance of different ensemble methods for the field of credit scoring in terms of average accuracy, type I error and type II error. Besides two common used datasets, i.e., Australian and German credit datasets, which are from UCI machine learning repository (Asuncion & Newman, 2007), our studies use a new credit dataset from China, collected mainly by the Industrial and Commercial Bank of China. In experiments we choose four popular methods in the literature, i.e., LRA, DT, ANN and SVM, as base learner. The results reveal that the application of ensemble learning can bring substantial improvement for individual base learner. Especially in our experiments, Bagging performs better than Boosting across all datasets. In addition, Stacking, and Bagging DT get best results in terms of three performance indicators, i.e., average accuracy, type I error and type II error. And among four base learners, DT gets best improvement in terms of three performance indicators after the application of ensemble learning.

The remainder of the paper is organized as follows. In Section 2, the details of three different types of ensemble methods for credit scoring are presented. In Section 3, we present the details of experimental design. Section 4 reports the experimental results. Based on the observations and results of these experiments, Section 5 draws conclusions and future research directions.

2. Ensemble learning for credit scoring

2.1. Overviews of ensemble learning

Ensemble learning is a machine learning paradigm where multiple learners are trained to solve the same problem (Polikar, 2006). In contrast to ordinary machine learning approaches that try to learn one hypothesis from the training data, ensemble methods try to construct a set of hypotheses and combine them to use (Zhou, 2009). Learners composed of an ensemble are usually called base learners.

One of the earliest studies on ensemble learning is Dasarathy and Sheela's research (1979), which discusses partitioning the feature space using two or more classifiers. In 1990, Hansen and Salamon showed that the generalization performance of an ANN can be improved using an ensemble of similarly configured ANNs (Hansen & Salamon, 1990). While Schapire proved that a strong classifier in probably approximately correct (PAC) sense can be generated by combining weak classifiers through Boosting (Schapire, 1990), the predecessor of the suite of AdaBoost algorithms. Since these seminal works, studies in ensemble learning have expanded rapidly, appearing often in the literature under many creative names and ideas (Polikar, 2006).

The generalization ability of an ensemble is usually much stronger than that of a single learner, which makes ensemble methods very attractive (Dietterich, 1997). In practice, to achieve a good ensemble, two necessary conditions should be satisfied: accuracy and diversity (Windeatt & Ardeshir, 2004). In the next three subsections, we will introduce three popular ensemble methods, i.e., Bagging, Boosting, and Stacking, respectively.

2.2. Bagging

Bagging (short for bootstrap aggregating) is one of the earliest ensemble learning algorithms (Breiman, 1996). It is also one of the most intuitive and simplest to implement, with a surprisingly good performance. Diversity in Bagging is obtained by using bootstrapped replicas of the training data: different training data subsets are randomly drawn—with replacement—from the entire training data. Each training data subset is used to train a different base learner of the same type.

The base learners' combination strategy for Bagging is majority vote. Simple as it is, this strategy can reduce variance when combined with the base learner generation strategies. The pseudo-code of Bagging algorithm is shown in Fig. 1.

Bagging is particularly appealing when the available data is of limited size. To ensure that there are sufficient training samples in each subset, relatively large portions of the samples (75–100%) are drawn into each subset. This causes individual training subsets to overlap significantly, with many of the same instances appearing in most subsets, and some instances appearing multiple times in a given subset. In order to ensure diversity under this scenario, a relatively unstable base learner is used so that sufficiently different decision boundaries can be obtained for small perturbations in different training datasets.

2.3. Boosting

Boosting (Freund & Schapire, 1996; Schapire, 1990) encompasses a family of methods. Unlike Bagging, Boosting creates different base learners by sequentially reweighting the instances in the training dataset. Each instance misclassified by the previous base learner will get a larger weight in the next round of training.

```
Input: Data set $D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;
       Base learning algorithm $L$;
       Number of learning rounds $T$.
Process:
  For $t=1,2,\cdots,T$:
    $D_t = Bootstrap(D)$;     % Generate a bootstrap sample from $D$
    $h_t = L(D_t)$     % Train a base learner $h_t$ from the bootstrap sample
  end.
Output: $H(x)=\arg\max_{y\in Y}\sum_{t=1}^{T} 1(y=h_t(x))$     % the value of $1(\alpha)$ is 1 if $\alpha$ is true
                                                      % and 0 otherwise
```

Fig. 1. The Bagging algorithm.