A COMPARISON OF BOOSTING CLASSIFIERS FOR USE IN THE MORTGAGE LOAN GRANTING PROBLEM

Louis W. Glorfeld, Department of ISYS, Walton College of Business, University of Arkansas, Fayetteville, AR 72701, 479-575-6121, Lglorfeld@cox.net
Doug White, Department of CIS, Gabelli School of Business Administration, Roger Williams University, Bristol, RI 02809, 401-254-3165, doug.white@acm.org

ABSTRACT

One of the most basic problems that may arise in the data mining (DM) context is the two group classification problem. This type of problem is the cornerstone of such DM applications as customer relations management, credit granting, and fraud detection. Because of the importance of two group classification, a very large number of methodologies have been developed for dealing with this problem. Several methods for enhancing classifier performance have been developed. Of these, one of the most interesting is boosting. Using a set of mortgage loan data, the performance improvement of boosting is tested using three common classifier methodologies. It is demonstrated that on the mortgage loan data, boosting of a common decision tree classifier is most effective.

INTRODUCTION

The rapid development of database and data storage technology has led to the common place deployment of data warehousing and data mart systems in many businesses that store and allow easy retrieval of massive amounts of persistent historical data reflecting all aspects of a business's operations. The availability of such massive amounts of data has spurred the development of a complete process for the extraction, preprocessing, and analysis of huge data sets commonly called data mining (DM) in order to turn raw data into business intelligence to promote superior decision making that results in competitive advantage. In this paper the emphasis will be on a very simple but common problem which is frequently encountered in the modeling step of the DM process. The simple qualitative two group classification problem is frequently encountered in such diverse areas as credit granting, customer relations management, and fraud detection. As Giudici [1] points out, "the building block of most qualitative response models is the logistic regression model (LRM), one of the most important predictive data mining methods." Because of the importance of the logistic model, it has been included as our baseline two group classifier. Another important type of classifier is the decision tree classifier. As Hastie, Tibshirani, and Friedman [2] pont out, "Of all the well-known learning methods, decision trees come closest to meeting the requirements for serving as an off-theshelf procedure for data mining." One of the most widely employed tree classifiers is based on Quinlan's [3] C4.5 algorithm. In addition, since boosting will be employed, the classic tree stump (TS) will be used since boosting has been shown to be most effective when applied to simple classifiers.

TRAINING, VALIDATION, AND BOOSTING

Sample

The sample for this study consists of a random sample of 750 mortgage loan applications from the Columbia, South Carolina SMSA, 500 accepted loan applications and 250 rejected loan applications. Given the data proportions, the best base classification rate would be 67 percent that could be achieved

by assigning all observations to the loan acceptance group. The sample consisted of 17 independent variables that described a number of personal characteristics of the loan applicants and characteristics of the property. In this study, the dependent variable is the status of the mortgage loan, categorized as either accepted or rejected.

Training and Validation Process

It is well known that assessing a model's performance using the same data used to build the model will lead to optimistically biased results of the model's performance. To produce a less biased assessment of how a model's performance will generalize to new data not originally used to build the model, some form of model cross validation is used. A frequently recommended procedure used in this study for reducing the bias of the classification accuracy based on the training sample is 10-fold cross validation[2].

Boosting

Boosting is a method for improving classifier performance of classification models. The basic idea is that by combing the predictions of multiple complimentary models and then using a voting scheme, the combined classification performance of the boosted models will show improved performance over the original model run alone. The primary sacrifice made is that the ability to interpret the model may be lost. In boosting all models are of the same type. The form of each new model built is influenced by the performance of the previously built models. By the use of weighting the classification variable, new models are encouraged to specialize on classification instances incorrectly classified by previous models. Furthermore, boosting weighs a models contribution to the overall classification scheme by its performance based on an exponential loss function. Details of boosting theory and algorithms may be found in Witton & Frank [4] and Hastie et al., [2]. For the C4.5 and tree stump algorithms, the widely used AdaBoost.M1 was used. For the LRM model an additive logistic regression was used which can be viewed from a practical perspective as a boosting method adapted to regression models [4].

RESULTS

Normal Model Results

The results of the training sample demonstrate the three models' ability to capture the loan granting decision making process. The C4.5 tree had a just slightly better overall classification performance rate with an overall correct classification rate of 86.3 percent while the LRM had a rate of 85.2 percent. The TS had the worst overall rate of correct classification getting only 75.2 percent correct. This result for the TS is not a surprising since it split on the single best classification variable which was applicants credit rating (ACR). Essentially, the TS predicted anyone with a good credit rating as having the loan approved and anyone with a bad credit rating as having the loan approval rejected.

The validation results are of primary interest. The results indicate that the LRM model performed just slightly better than the C4.5 tree using 10-fold cross validation. The LRM had an overall correct classification rate 84.9 percent compared to the 84.1 percent rate of the C4.5 tree. Such a small difference in performance is of no real significance. As with the training data, the TS had the poorest cross validation results giving an overall correct classification rate of 75.2 percent. The fact that the cross validation result is identical to the training sample result is not surprising. Each of the 10-fold cross validation models split on the same ACR variable used in the full training sample, thus simply replicating the training sample result.

Boosted Model Results

The training sample results indicate the best model in terms of overall classification performance was the C4.5 decision tree with a perfect 100 percent classification rate. The boosted TS model is second with an overall correct classification rate of 90.8 percent followed by the LRM with a correct classification rate of 85.6 percent. The boosted performance represent a radical improvement on the training data for the C4.5 and TS models with a negligible improvement in performance for the LRM. Once again, the validation results are of primary interest. The best overall classifier is the boosted C4.5 decision tree with an overall correct classification rate of 88.4 percent. The boosted TS model is second with a correct classification rate of 86.3 percent followed by the boosted LRM with a correct classification rate of 85.3 percent. The most dramatic increase in classification performance was achieved by the simple TS model with over a 10 percent improvement in its cross validation performance. The simple TS is now actually competitive with the more complex LRM and C4.5 classifiers. This substantial improvement in classification performance serves to demonstrate the fact that boosting typically gives the greatest performance enhancement to very simple classifiers like the TS. The boosted C4.5 decision tree also showed reasonable improvement with just over a 4 percent gain in cross validated classification performance, turning it into the best performing classifier. Although the LRM did improve, the improvement was negligible.

CONCLUSION

This paper has evaluated three different inductive decision models in their ability to predict if a mortgage loan should be granted or rejected in the models' normal form and after the models were boosted. The basic decision models used in this study included logistic regression, the C4.5 decision tree and a simple tree stump. Evaluation was primarily based on the use of 10-fold cross validation, It was demonstrated that for the mortgage loan data, boosting provided a substantial performance increase for the simple TS classifier, allowing it to actual perform slightly better than the more complex LRM model. The performance increase in classification accuracy for the C4.5 decision tree placed it as the top performing classifier. These results indicate that if classification accuracy is a primary goal of using a classification model, then boosting is a worthwhile way to provide a possible performance increase. Although no attempt was made to interpret the classification models, there are methods available to allow interpretation of boosted tree classifiers, such as option trees and the boosted LRM model can be interpreted directly since the usual logistic regression weights are available for the boosted model.

(*A complete version of this paper may be obtained from the first author*)

REFERENCES

[1] Giudici, Paolo (2003). *Applied Data Mining: Statistical Methods for Business and Industry*. West Sussex, England: John Wiley & Sons Ltd.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer-Verlag.

[3] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann. Ridgeway, Greg (2003). Strategies and methods for prediction. In Ye, Nong (Ed.), *The Handbook of Data Mining*. Mahwah, NJ: Lawrence Erlbaum, 159-191.

[4] Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. New York, NY: Morgan Kaufmann.