CART의 예측 성능:은행 및 보험 회사 데이터 사용

박 정 선†

요 약

본 연구에서는 CART(Classification and Regression Tree)가 예측을 함에있어 통계적인 기법인 discriminant analysis와 비교된다. 은행 데이터를 사용하는 경우 discriminant analysis가 더 나은 성능을 보여줬으며, 보험 회사 데이터를 사용한 경우 CART가 더 나은 성능을 보여줬다. 이러한 모순된 결과가 데이터의 성격을 분석함으로 해석된다. 본 연구에서는 두가지 모델 모두 사용된 매개변수들인 사전 확률, 데이터, 타입 I/II 오류 코스트, 검증 방법에 의해 성능의 차이를 보여줬다.

The Prediction Performance of the CART Using Bank and Insurance Company Data

Jeong Sun Park†

ABSTRACT

In this study, the performance of the CART(Classification and Regression Tree) is compared with that of the discriminant analysis method. In most experiments using bank data, discriminant analysis shows better performance in terms of the total cost. In contrast, most experiments using insurance data show that the CART is better than discriminant analysis in terms of the total cost. The contradictory results are analyzed by using the characteristics of the data sets. The performances of both the Classification and Regression Tree and discriminant analysis depend on the parameters: failure prior probability, data used, type I error cost, type II error cost, and validation method.

1. Introduction

In order to identify banks/insurance companies at risk of failure[1], there have been numerous studies on bank/insurance company bankruptcy predictions. Most studies [1, 6] advocate the use of statistical methods called discriminant analysis. Typically, discriminant analysis methods assume that all variables are normally distributed. In the case of linear classifier, it also requires identical covariance matrices. In reality, however, these assumptions are rarely satisfied.

In view of these limitations, the Classification and Regression Tree [2, 4] has been introduced as an alternative method for bankruptcy prediction. The Classification and Regression Tree (CART) requires no distributional assumptions and no functional forms. Thus, it is argued as a more robust method than general statistical methods. In addition, the CART takes into account economic concepts of prior probabilities and misclassification costs like discriminant

† 정 회 원:명지대학교 산업공학과 조교수
논문접수:1996년 3월 28일, 심사완료:1996년 7월 3일

1. The bankruptcy of banks and insurance companies is made in America.

analysis. To use the class proportions in the sample as the prior probabilities may generate biased results. The cost of classifying as viable a bank/insurance company which subsequently fails should be much higher than the cost of classifying as failing a bank/insurance company which is actually viable. Thus, the concepts of priors and misclassification costs are important in the prediction of bank/insurance company bankruptcy.

Almost classifier systems show their performances differently depending on the data used in experiments [3, 9]. This paper explores comparing the Classification and Regression Tree with discriminant analysis for bank/insurance company failure prediction and analyzes the results of them in various experiments. The rest of this paper is organized as follows: In section 2, the CART is reviewed. The data sample, and the results of the classification are presented in section 3. Section 4 discusses the characteristics of the data and section 5 concludes the paper.

2. The CART

The CART classifier procedure requires four elements to create a tree: a set of questions involving a predicate of the form {Is x < A?} where x is a variable and A is a given threshold value, a goodness of split criterion that can be evaluated for any split condition at a node, stopping rules to terminate splitting, and a rule to assign a class to a terminal node.

The procedure employs a recursive splitting strategy that repeatly splits a node into 2 disjoint nodes (children). Observations contained in the original node are divided into 2 groups with each stored in one child. When a node decides to split, a set of questions is made by using the threshold values of each variable at the node. Each question is evaluated by the split criterion which yields an evaluation value. If all values obtained from evaluating the questions are less than some given minimum value for splitting, then the node stops splitting. Otherwise, the question form that yields the maximal evaluation value is selected as the splitting predicate and two new children nodes are created. Also, if all observations at a node belong to the same class, the node stops splitting. Such a node becomes a terminal or leaf node of the tree. The class assignment rule is then applied to determine the class tag of the node.

If the observations are exact and free of error, the largest tree which contains the most information will have the best classificatory accuracy. In reality, observations are recorded with error, important variables may be missed, and the chosen variables may be inappropriate for prediction. It is therefore necessary to prune a tree so constructed to make it more robust as reported in [8] and to improve the generality of the classifier. The minimal cost-complexity pruning method is directed towards this end. The minimal cost-complexity pruning method considers both misclassification costs and penalizes for additional complexity of the tree. In selecting a pruning node, the method finds a node with the weakest links to its branches. The pruning continues to the point where only one node remains. The pruning process creates a sequence of trees with decreasing order of tree size.

3. Experiments and Results

3.1 Experimental Design

We have constructed two prediction models, one-year ahead prediction (one-year period) and two-year ahead prediction (two-year period). The methods for comparing the trees with DA(discriminant analysis)s are the independent test sample, and the cross-validation methods. The independent test sample method uses a separate test sample which is different from the training sample. This method is useful when the available data sample is large. The parameters chosen for our experiments are: There are two prior probabilities of failure, 0.01 and 0.02. For each probability, type I misclassification cost takes on one value from [1, 5, 15, 25, 50, 60, 75, and 100], while type II cost remains

at 1. These sixteen tests are carried out using the one-year data, and are repeated with the two-year data. In each experiment that uses the one-year bank (insurance company) data, 59 (56) failed and 59 (56) non-failed banks (insurance companies) are used for training, with 22 (36) failed and 22 (36) non-failed for testing. In the case of the two-year data, 59 (56) failed and 59 (56) non-failed banks (insurance companies) are used for training, and 20 (36) failed and 20 (36) non-failed banks (insurance companies) are used for testing.

The second method, the cross-validation method, divides the entire sample into ten subsets randomly. Selecting ten as the V value is generally accepted [2]. Nine of these subsets are used for training and the remaining one for validation. Thus, there are ten training samples and ten test samples correspondingly. Experiments using the cross-validation method have the same parameter setups as those using the test sample method. The cross-validation method, however, requires ten times as many experiments as the test sample method. In each experiment using the one-year data, 146 (166) failed or non-failed data are used for training, and 16 (18) banks (insurance companies) are used for testing. In the case of two-year data, 142 (166) failed or non-failed bank (insurance companies) data are used for training in each experiment, and 16 (18) banks (insurance companies) are used for testing.

In both methods, the expected misclassification cost is used as the performance measure. Type I and type II error rates are calculated first and the rates are weighted by their priors and costs. Namely, expected misclassification cost = type I error rate * prior probability of class 1 * type I error cost + type II error rate * prior probability of class 0 * type II error cost.

3.2 Empirical Results

To validate trees using the test sample and the cross-validation methods, 32 and 320 experimentations are performed respectively using two bank (insurance company) failure prior probabilities, 0.01 and 0.02, type I error costs (i.e. 1, 5, 15, 25, 50, 60, 75, and 100), in two prediction periods, the one-year period and two-year period. The type I, type II error rates, and the total cost for the cross-validation method are the averages of 10 independent tests. The same test procedures are used for discriminant analysis validations.

As type I cost increases, the type I error rate using either classification method generally decreases. The type I error rate using the CART method decreases more drastically than that of discriminant analysis method. In contrast, as type I error cost increases, the type II error rate using the CART and discriminant analysis method generally increases. The type II error rate using the CART method increases more speedily than when using discriminant analysis method. The total cost (the sum of two expected misclassification costs) for both methods increases, as type I cost increases. The total cost using the binary tree method is generally higher than that by discriminant analysis method in bank data (58 cases out of 64) and is generally lower in insurance data (52 cases out of 64). The results show that the performance of a model may change depending on the characteristics of data. These results will be analyzed by the characteristics of data sets.

4. Task Characteristics

4.1 Task Characteristics of Insurance

When we consider task characteristics, we assume that sample data reflect the characteristics of the task. Insurance is one of the industries[2] controlled strictly by federal or state agencies [5]. As the task characteristics[3], we use dependence among the independent variables, linearity between the independent variables and the dependent variable, covariance equality of two classes, and normality of the independent variables.

2. Insurance, utilities, and banking are required to report their financial status in more detail than other industries.
3. To test the task characteristics, SAS package is used.

To test the dependence, correlation coefficients are used and P values show the significance of the coefficients. To test the linearity, linear regression is used and the F value shows the significance. Covariance equality and normality are tested by using an option of discriminant analysis and the Kolomogorov test respectively. The results are summarized in Table 1.

〈Table 1〉 Characteristics of Insurance Data Sets

characteristics	training (one-year)	test (one-year)	training (two-year)	test (two-year)
dependence	44%	17%	44%	19%
linearity	yes	no	yes	yes
normality	0%	0%	0%	0%
equal covariance	no	–	no	–

From the above results, we can see that dependence is low and linearity is high. Two assumptions of DA (normality and equal covariance) are violated, resulting the data have characteristics of non-normality and non-equal covariance. These characteristics may explain the weak performance of DA.

4.2 Relative Task Characteristics

To see the performance of a model changes depending on a task, we introduce bank industry. We extract the task characteristics from these bank data sets. Finally, we compare the characteristics of insurance with those of banking. We show that the performance of a model changes depending on the change of data characteristics. For example, DA shows a better performance in a domain which fits the assumptions of DA.

We use the same statistical methods. Table 2 summarizes these results.

From the above results, we can see that the dependence is medium and linearity is very high. In the insurance domain, the dependence is high. The normality assumption of DA is violated more strongly in insurance than in banking. This implies that DA may perform better in the banking domain.

〈Table 2〉 Characteristics of Bank Data Sets

characteristics	training (one-year)	test (one-year)	training (two-year)	test (two-year)
dependence	58%	50%	53%	46%
linearity	yes	yes	yes	yes
normality	11%	42%	5%	26%
equal covariance	no	–	no	–

5. Conclusions

In most experiments using bank data, discriminant analysis method shows better performance in terms of the type I error rate and the total cost. Only six experiments out of sixty four show less total cost of the binary tree than discriminant analysis. In contrast, most experiments using insurance data show that the CART method performs better than discriminant analysis. Only 12 cases out of 64 show less total cost of discriminant analysis than the binary tree.

The performances of both the CART and discriminant analysis depend on the parameters : failure prior probability, data used, type I error cost, type II error cost, and validation method. In this study, discriminant analysis is generally better than the CART in the classification of bank defaults ; the CART is generally better than discriminant analysis in the classification of insurance company defaults.

Default predictions of insurance companies show contradictory results to those of banks, confirming the previous results [4] that the CART shows better performances than discriminant analysis. This situation can be explained that the performance of a classifier depends on the characteristics of the data. If the data are dispersed appropriately for the classifier, the classifier will show a good performance. Otherwise, it may show a poor performance.

The two data sets (bank and insurance) are analyzed using four characteristics : dependence among

the independent variables, linearity between the independent variables and the dependent variable, covariance equality of two classes, and normality of the independent variables. The results show that insurance data have more independence, more non-linearity, and less normality compared with bank data. These results imply that the CART performs better and discriminant analysis worse in insurance data. As explained before, the CART method builds a tree recursively and terminal nodes show classes. The terminal nodes are independent and they have non-linear relationships each other. Thus, the CART may show better results in data with more independence and more non-linearity. One of the two assumptions of discriminant analysis is normality. Discriminant analysis may perform worse in data with less normality. Therefore, the characteristics of the two data sets explain the better performance of the CART in insurance and the worse performance in bank; the better performance of discriminant analysis in bank and the worse performance in insurance. These results indicate that the performance of a classifier should be considered from the view of data characteristics. Namely, the performance of a classifier changes depending on test situations such as data characteristics.

References

[1] E. I. Altman, "Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy," Journal of Finance, 23, 1968, 589-609.

[2] L. Breiman, J. Frieman, J. Olshen, and C. Stone, "Classification and Regression Trees," California, Wadsworth & Brooks, 1984.

[3] H. M. Chung and M. S. Silver, "Rule-Based Expert Systems and Linear Models: An Empirical Comparison of Learning-By-Examples Method," Decision Sciences, 23, 3, 1992, 687-707.

[4] H. Frydman, E. I. Altman, and D. Kao, "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress," Journal of Finance, 11, 1, March, 1985, 269-291.

[5] B. Koo, "Solvency Surveillance in the Property/Casualty Insurance Industry: A Financial Analysis and Statistical Evaluation of Sampling Biases," Unpublished Ph. D. Dissertation, The University of Texas at Austin, 1992.

[6] J. Lann-Tennant, L. Starks, and L. Stokes, "Solvency Surveillance: An Empirical Evaluation of the Property-Liability Insurance Industry," Proc. Third Int. Conf. on Ins. Fin. and Solvency, 1990.

[7] T. Liang, "A Composite Approach to Inducing Knowledge for Expert Systems Design," Management Science, 38, 1992, 1-17.

[8] J. Mingers, "An Empirical Comparison of Pruning Methods for Decision Tree Induction," Machine Learning, 4, 1989, 227-243.

[9] R. Mooney, J. Shavlik, G. Towell, and A. Gove, "An Experimental Comparison of Symbolic and Connectionist Learning Algorithms," Proceedings of Eleventh IJCAI, 1989.

박 정 선

1983년 서울대학교 산업공학과 졸업(학사)
1985년 한국과학기술원 경영과학과 졸업(석사)
1993년 미국 Univ. of Texas at Austin 경영정보시스템 전공(박사)
1985년~1988년 금성소프트웨어 근무
1988년~1989년 한국유니시스 근무
1993년~1995년 한국전산원 근무
1995년~현재 명지대학교 산업공학과 조교수
관심분야: AI/Expert System, CALS, WWW 응용