The Bump Hunting Using the Decision Tree Combined with the Genetic Algorithm: Extreme-value Statistics Aspect

Hideo Hirose Takahiro Yukizane *

Abstract—**In difficult classification problems of the z-dimensional points into two groups giving 0-1 responses due to the messy data structure, it is more favorable to search for the denser regions for the response 1 points than to find the boundaries to separate the two groups; this is called the bump hunting. In a series of previous studies, we have shown that a bump hunting method using the decision tree combined with the genetic algorithm is useful for certain customer database, where we have developed a trade-off curve between the pureness rate and the capture rate. This paper deals with the behavior of the trade-off curve from a statistical viewpoint.**

Keywords: ***data mining, ROC curve, recall precision curve, genetic algorithm, extreme-value statistics, trade-off curve, decision tree, bootstrapped hold-out.***

1 Introduction

Suppose that we are interested in classifying n points in a z-dimensional space into two groups according to their responses, where each point is assigned response 1 or response 0 as its target variable. For example, if a customer makes a decision to act a certain way, then we assign response 1 to this customer, and assign response 0 to the customer that does not. We want to know the customers' preferences presenting response 1. We assume that their personal features, such as gender, age, living district, education, family profile, etc., are already obtained.

Many classification problems have been dealt with elsewhere to rather simpler cases using the methods of the linear discrimination analysis, the nearest neighbor, logistic regression, decision tree, neural networks, support vector machine, boosting, etc. (see [9], e.g.) as fundamental classification problems. In some real data cases in customer classification, it is difficult to find the favorable customers, because many response 1 points and 0 points are closely located, resulting response 1 points are hardly separable from response 0 points [10, 11]. In such a case, to find the denser regions to the favorable customers is considered to be an alternative. Such regions are called the bumps, and finding them is called the bump hunting. The bump hunting has been studied in the fields of statistics, data mining, and machine learning [1, 2, 7, 8].

*Manuscript received July 22, 2007. The authors would like to thank Dr. Miyano for his cooperation and valuable comments. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research on Priority Areas, A16092223, Grants by Ohkawa Foundation for Information and Telecommunications 06-26, and Grants by Kayamori Foundation for Informational Science Advancement, K18-XI230.
Correspondence: Department of Systems Innovation and Informatics, Kyushu Institute of Technology, Fukuoka 820-8502, Japan Tel/Fax: +81(948)29-7711/7709, Email: hirose@ces.kyutech.ac.jp

2 Progress of Our Research

By specifying the pureness rate p, the ratio of the number of points of response 1 to the total number of points in some region R in advance, we may obtain the maximum capture rate c_R for the region R, the maximum ratio of the number of response 1 points in R to the whole number of ones. Then, a trade-off relationship, $T(p, c_R)$, between p and c_R can be constructed; see Figure1.

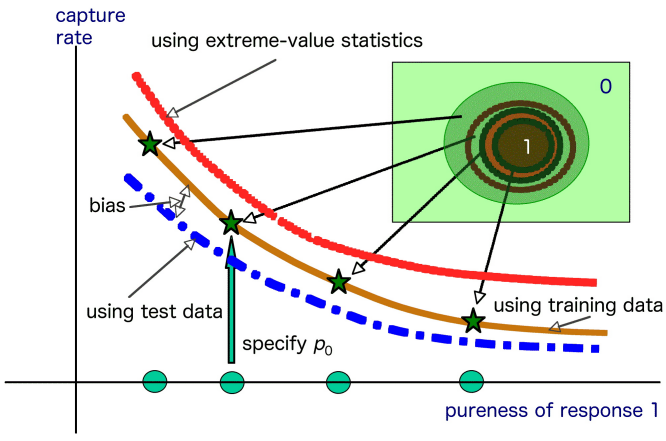

Figure 1: Three trade-off curves between the pureness rate and the capture rate: 1) using the training data, 2) using the extreme-value statistics, 3) using the test data.

You may remind similar curves, i.e., the ROC (Receiver Operator Characteristic) curve and the Recall-Precision curve in machine learning and medical fields [4, 6]. Let TP be true positive, TN be true negative, FP be false positive, and FN be false negative. Then, recall is de-

fined by TP/(TP+FN); precision by TP/(TP+FP); true positive rate by TP/(TP+FN); false positive rate by FP/(FP+TN); see [4, 6], e.g. Since a response 1 point in or outside the bump region is considered to be TP or FN, respectively, and a response 0 in or outside the bump is FP or TN, the pureness rate can be defined by TP/(TP+FP) which is identical to precision; the capture rate can also be defined by TP/(TP+FN) which is identical to true positive rate and to recall; see Figure 2. So, a Recall-Precision curve and a capture-rate pureness-rate curve seem to be equivalent superficially. However, we should note that these two are totally different from each other. As is seen in Figure 3, it can be considered that the capture-rate pureness-rate curve is constructed by collecting the skyline points consisting of many trade-off curves where each curve is corresponding to one classifier. In Figure 2, two kinds of error, α and β, in statistical hypothesis tests are depicted for comparison; the misclassification rate and the accuracy are also shown.

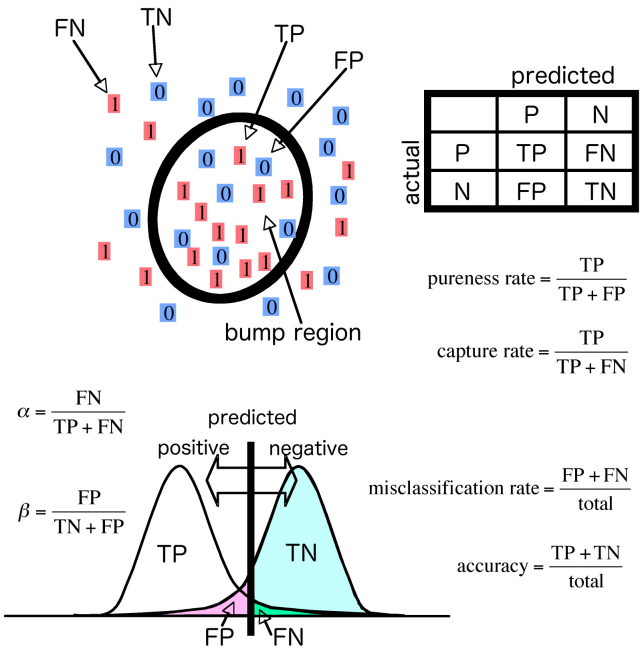

Figure 2: Confusion matrix in the bump hunting.

In order to make future actions easier, we adopt simpler boundary shapes such as the union of z-dimensional boxes located parallel to some explanation variable axes for the bumps; that is, we use the binary decision tree. In decision trees, by selecting optimal explanation variables and splitting points to split the z-dimensional explanation variable subspaces into two regions from the top node to downward using the Gini's index as in the conventional method, we may obtain the number of response 1 points by collecting nodes where the pureness rates are satisfying to be larger than the pre-specified pureness rate p_0. However, much response 1 points could be obtained if we

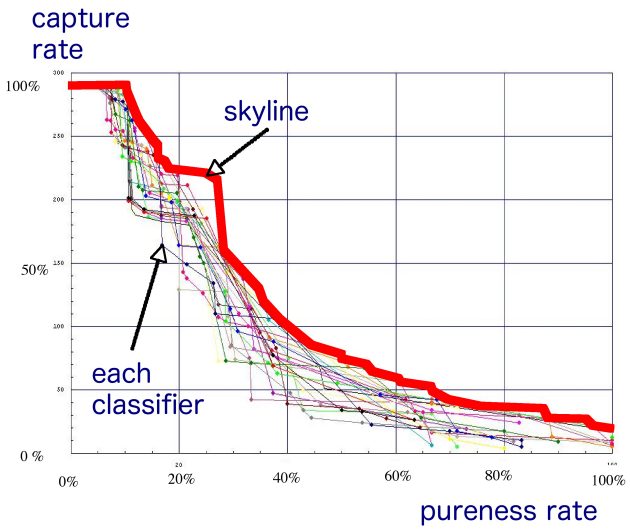

Figure 3: Trade-off curve as a skyline curve consisting of many classifiers.

locate adequate explanation variables to each branching knot. This is because the conventional algorithm has a property of the local optimizer. Thus, we have developed a new decision tree method with the assistance of the random search methods such as the genetic algorithm (GA) specified to the tree structure, where the most adequate explanation variables are selected by the GA, but the best splitting points are obtained by using the Gini's index [16]; see Figure 4.

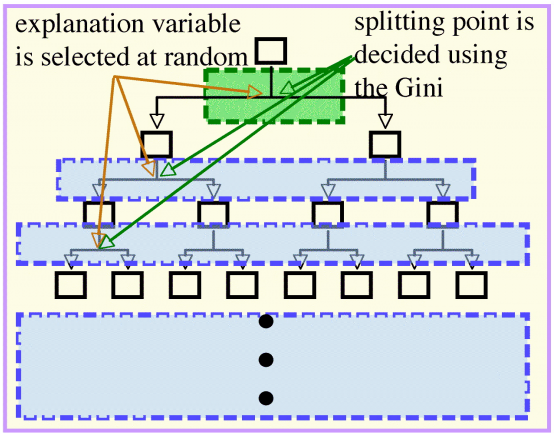

Figure 4: The bump hunting procedure using the decision tree with the genetic algorithm.

Solutions obtained by the GA primarily are not global optimal; this is a drawback of the algorithm. However, we have observed the existence of many multiple local maxima with each starting point in our GA procedure. This turns out to become a merit; the use of the extreme-value statistics [3] is used to estimate the return cap-

ture rate (expected global maximum capture rate), and the method did work successfully when the shape of the marginal density function of an explanation variable is simple, such as monotonic or unimodal [12, 16]. This property is also observed in a real customer database [13].

This kind of solution is, however, the best fitted solution [9]; that is, the rule is constructed using the training data and the evaluation is performed using the same training data [5, 15]; so, the solution should have a bias (see Figure 1, e.g.). If we apply the rule obtained by the training data to a new test data case having the same data structure, we may no longer expect the same performance in the new data case. We have noticed that we should pay much attention to this problem even though the size of the explanation variable is small. The bootstrapped hold-out proposed in [14, 17] is effective in this case.

3 Objective of the Research

Summarizing the above, the trade-off curve we are dealing with have three aspects. The first is the curve obtained by using the training data; we can apply the if-then-rules to the future data only to this curve. The second is the return curve obtained by using the extreme-value statistics; by estimating the ceiling for the capture rate, we can know where we are. The last is the curve obtained by using the test data; we can expect the actual capture rate for response 1. Figure 1 shows the relationships among these three curves. These three are indispensable like the Trinity to comprehend the whole figure of the trade-off curve between the pureness rate and the capture rate.

In our previous studies, however, we have not paid attention much to the statistical consideration to the GA outcomes. In this paper, we deal with this point.

4 Extreme-Value Statistics Approach

As shown [16], for the ten cases of the iteration procedures with ten different initial conditions in an simulated example, the converged solutions differ from each other when the initial value is set to different values. Our genetic algorithm have a strong inclination of searching for the local maxima because we are using the tree structure in evolution procedure. So far, we have been using the following evolution procedure:
1) the number of initial seeds is set to 30, and the successive number of seeds is 20,
2) the maximum number of evolution procedure is set to 20,
3) the crossover is done using the left wing of a tree and the right wing of another tree,
4) the mutation rate is set to around 5%

If the mother distribution function is a normal, exponential, log-normal, gamma, gumbel, or Rayleigh type distribution, then the limiting distribution of the maximum values from the mother distribution follows the gumbel distribution (see [3] e.g.). For example in a data case where samples are drawn at random with 1/100 probability from a real customer data case, we have 20 local maxima of 48, 45, 48, 39, 56, 44, 32, 41, 56, 70, 40, 49, 42, 52, 38, 53, 47, 55, 34, 45, when we specify the pureness rate of 50%. If we fit the gumbel distribution to the data, we can estimate the shape and scale parameters as 7.38 and 42.6. Then, the return capture rate for 500 trials is estimated to be 88.5. Here, the number of samples is 1,635 samples; the number of response 1 is 290; the number of variables is 44. The frequency distribution and the fitted gumbel density function to this 20 data are seen in Figure 5. However, the results by applying the test data to the rules obtained by the training data was very pessimistic. The bias between the training data trade-off and test data one shown in Figure 6 becomes very large because of large number of explanation variables, resulting that the rules obtained by the training data are not applicable.

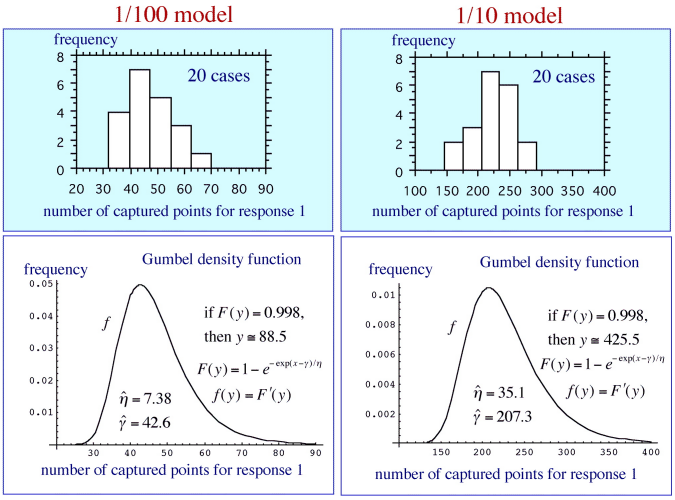

Figure 5: Gumbel distribution fit to the 20 local maxima using the genetic algorithm.

A much larger case is also investigated, where 15,870 samples, 2,863 response 1 points, and 41 variables are treated. The 20 local maxima are 207, 230, 251, 258, 255, 238, 170, 229, 204, 292, 247, 218, 281, 237, 230, 206, 195, 208, 193, 147, by the half training data, and the return capture is estimated as 425.5. The results in this case are found to be useful because of small biases. See Figures 5 and 6.

However, all the initial seeds have been set to 30 so far. Considering that the property of the local convergence of the GA procedure, it would be better to provide much larger number of seeds to verify if the extreme-value statistics works well. Figure 7 shows the results of the number of captures in a data case resampled from the

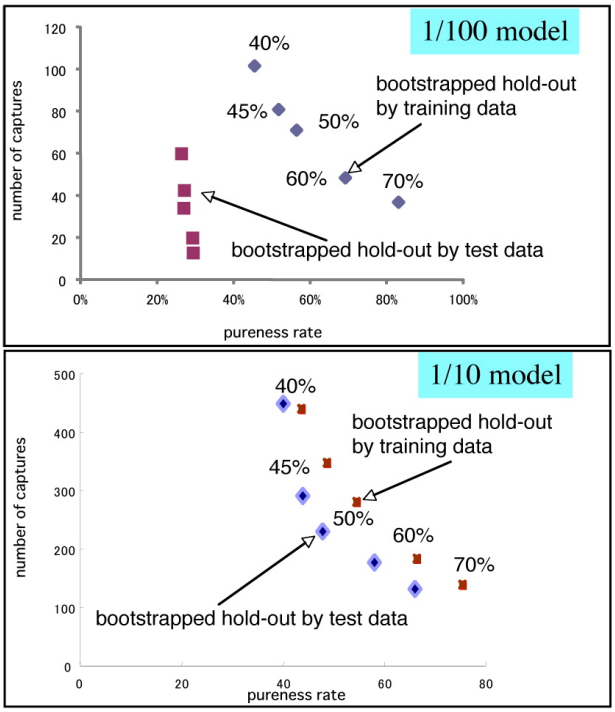

Figure 6: Gumbel distribution fit to the 20 local maxima using the genetic algorithm.

real customer data case when the number of seeds and the successive number of iterations are set to larger values. Here, the pre-specified pureness rate is 50% and the model is 1/100 scale. We can see that the extreme-value statistics work very well, and we can use this method even if the number of samples to the gumbel distribution is small such as 20; we can see that the predicted captures by the extreme-value statistics preserve almost a constant value even though the converged local maxima are gradually becoming large as the number of seeds becomes large.

5 Concluding Remarks

In difficult classification problems, the bump hunting method using the decision tree combined with the genetic algorithm is useful. In this paper, we have shown that the trade-off curve between the pureness rate and the capture rate can be characterized into three categories; 1) using the training data, 2) using the extreme-value statistics, and 3) using the test data. Of these, the behavior of the results using the extreme-value statistics is mainly investigated. To comprehend the whole figure of the trade-off curves between the pureness rate and the capture rate, it is recommended to take into account these three categories. If the number of explanation variables is large to some extent and the number of samples is small, the bias between the result using the training data and that using the test data may become large even if the relaxation

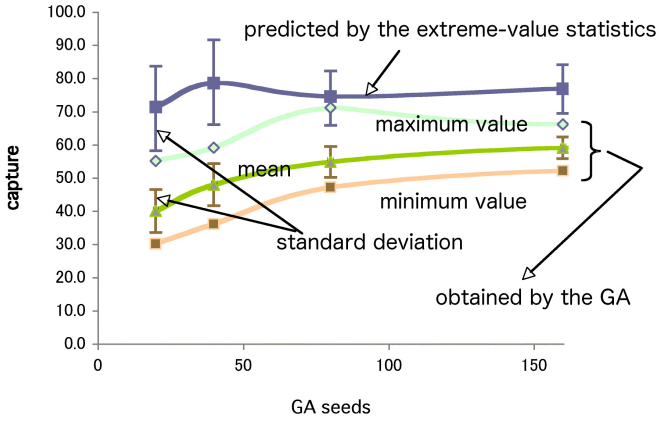

Figure 7: Number of captures of response 1 versus number of the seeds in the genetic algorithm.

method is used, showing the pessimistic results. Reduction of the explanation variables and the increase of the number of samples will solve this difficulty.

The ROC curve and precision recall curve are compared to the proposed trade-off curve, showing that the trade-off curve is different from the precision recall curve even if they are superficially equivalent. The misclassification rate, the accuracy, the first kind error and the second kind error in statistical hypothesis tests are also explained comparing to the confusion matrix.

References

[1] Agarwal, D., Phillips, J.M., and Venkatasubramanian, S. :The hunting of the bump: On maximizing statistical discrepancy, SODA'06. (2006) 1137–1146

[2] Becker, U and Fahrmeir, L. :Bump hunting for risk: a new data mining tool and its applications, Computational Statistics, 16. (2001) 373–386

[3] Castillo, E.: Extreme Value Theory in Engineering. Academic Press. 1988

[4] Davis, J., and Goadrich, M.: The relationship between precision-recall and ROC Curves, Proceedings of the 23 International Conference on Machine Learning, 2006

[5] Efron, B.: Estimating the error rate of a prediction rule: improvements in cross-validation. JASA. 78 (1983) 316-331

[6] Fawcett, T. : An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861-874.

[7] Friedman, J.H. and Fisher, N.I.: Bump hunting in high-dimensional data. Statistics and Computing. **9** (1999) 123–143.

[8] Gray, J.B. and Fan, G: Target: Tree analysis with randomly generated and evolved trees. Technical report. The University of Alabama (2003).

[9] Hastie, T., Tibshirani, R. and Friedman, J.H.: Elements of Statistical Learning. Springer (2001).

[10] Hirose, H.: A method to discriminate the minor groups from the major groups. Hawaii International Conference on Statistics, Mathematics, and Related Fields, (2005).

[11] Hirose, H.: Optimal boundary finding method for the bumpy regions. IFORS2005.

[12] Hirose, H., Yukizane,T. and Miyano, E.: Boundary detection for bumps using the Gini's index in messy classification problems. CITSA2006 pp.293-298.

[13] Hirose, H., Yukizane,T. and Deguchi T.: The bump hunting method and its accuracy using the genetic algorithm with application to real customer data. accepted, CIT2007.

[14] Hirose, H., Ohi, S. and Yukizane, T.: Assessment of the prediction accuracy in the bump hunting procedure. Hawaii International Conference on Statistics, Mathematics, and Related Fields, (2007).

[15] Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, 1995.

[16] Yukizane, T., Ohi, S., Miyano, E. and Hirose, H.: The bump hunting method using the genetic algorithm with the extreme-value statistics. IEICE Trans. Inf. Syst., E89-D. (2006) 2332–2339.

[17] Yukizane T., Hirose, H, Ohi, S., and Miyano, E.: Accuracy of the solution in the bump hunting. IPSJ MPS SIG report, MPS06-62-04 (2006) 13–16.