# Bump hunting とその顧客データへの応用

廣瀬英雄, 行實隆広　(九州工業大学)

**abstract**

Suppose that we are interested in classifying $n$ points in a $z$-dimensional feature variable space into two groups according to their responses, where each point is assigned response 1 or response 0 as its target variable. We assume that due to the messy data structure such that many response 1 points and 0 points are closely located in the feature variable space, response 1 points are hardly separable from response 0 points. In such a case, to find the denser regions for response 1 points could be an alternative to the usual classification problems. Such regions are called the bumps, and finding them is called the bump hunting. We have developed a bump hunting method using probabilistic and statistical methods. By specifying a pureness rate in advance, a maximum capture rate will be obtained. The pureness rate and the capture rate are illustrated in Fig.1 in terms of true positive, true negative, false positive, and false negative. Since the smaller the regions to capture the response 1 dense points, the higher the ratio of the response 1 points to the total in the regions. Thus, a trade-off curve between the pureness rate and the capture rate can be constructed; see Fig.1.

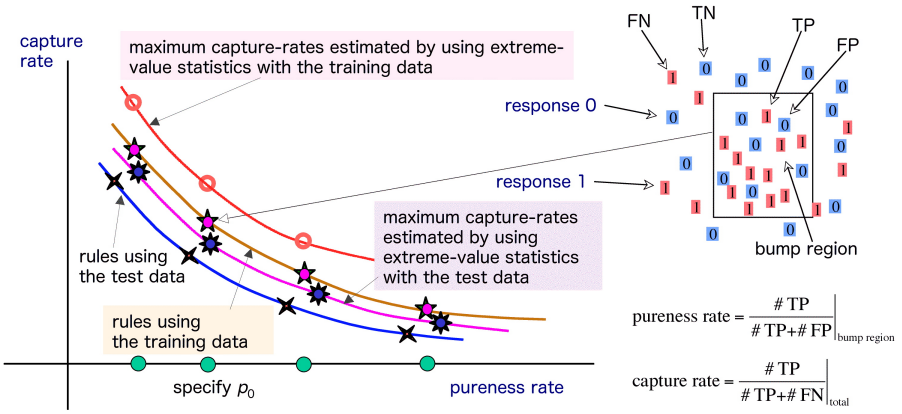


図 1: The trade-off curve between the pureness rate and the capture rate in the bump hunting.

To obtain the optimal trade-off curve, we have developed a new decision tree method which includes the genetic algorithm; we call this the GA tree. In the

GA tree, the explanation variables to each branching knot are randomly selected, but the splitting points of the feature variables are determined by using the Gini's index. This method intuitively seems to work in simple data structures; however, it works well also in real complex customer data.

To preserve a good inheritance property in evolution procedure, we have designed our own crossover method in the GA tree. This has caused the existence of many local maxima for the capture rates. This drawback, however, turns out to be a nice property to estimate the upper bound for the trade-off curve. We have used the extreme-value statistics to estimate this upper bound.

As is well known, to assess the the accuracy for the trade-off curve, it is recommended to apply the test data to the optimally obtained rules by using the training data. When the number of feature variables is large comparing to the sample size, the bias between the results using the training data and those using the test data may become large. To assess the bias efficiently, we have proposed to use the boostrapped hold-out method in case that the cross validation cannot be applicable due to the small sample size.

To assess the accuracy for the upper bound of the trade-off curve using the extreme-value statistics, the results using the test data at the final stage of the evolution procedure are supposed to be local maxima, as similarly observed using the training data only. This requires us to apply the test data to the rules obtained by using the training data at every stage of the evolution procedure. However, such the test data is no longer responsible for the role of testing; we have to provide another test data set. Thus, we next proposed a new GA tree; we first classify the original data into three subsets, the first is for training, the second is for evaluation to each evolution stage, and the third is for test at the final stage. Using the new GA tree proposed, we can obtain the upper bound accuracy for the trade-off curve. Then, we may expect the actually attainable trade-off curve upper bound with its accuracy. Using this, we will make future decisions by applying the rules obtained by the training data with the knowledge of how far the rules we are using are located from the optimal points. In Fig.1, two important curves are shown; one is the rules using the training data because we have to use these rules in actual cases, and the other is the estimated curve using the extreme-value statistics with the test data to know the position we stand.