Abstract

In difficult classification problems of the $z$-dimensional points into two groups having 0-1 responses due to the messy data structure, it is more favorable to search for the denser regions for the response 1 assigned points than to find the boundaries to separate the two groups. To such problems often seen in customer databases, 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. Then, a trade-off curve between the pureness rate and the capture rate can be constructed. In finding the maximum capture rate, we have used the decision tree method combined with the genetic algorithm. We first explain a brief introduction of our research: what the bump hunting is, the trade-off curve between the pureness rate and the capture rate, the bump hunting using the tree genetic algorithm, the upper bounds for the trade-off curve using the extreme-value statistics. Then, the assessment for the accuracy of the trade-off curve is tackled from the genetic algorithm procedure viewpoint. Using the new genetic algorithm procedure 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. The bootstrapped hold-out method is used in assessing the accuracy of the trade-off curve, as well as the cross validation method.