INFORMATION, Vol.14, No.10, October 2011

Bump Hunting using the Tree-GA

Hideo Hirose

International Information Institute

INFORMATION
Volume 14, Number 10, pp.3409-3424

ISSN 1343-4500
©2011 International Information Institute

Bump Hunting using the Tree-GA

Hideo Hirose

Kyushu Institute of Technology, Department of Systems Design and Informatics,
Iizuka, Fukuoka, 820-8502 Japan
hirose@ces.kyutech.ac.jp

Abstract

The bump hunting is to find the regions where points we are interested in are located more densely than elsewhere and are hardly separable from other points. By specifying a pureness rate p for the points, a maximum capture rate c of the points could be obtained. Then, a trade-off curve between p and c can be constructed. Thus, to find the bump regions is equivalent to construct the trade-off curve. We adopt simpler boundary shapes for the bumps such as the box-shaped regions located parallel to variable axes for convenience. We use the genetic algorithm, specified to the tree structure, called the tree-GA, to obtain the maximum capture rates, because the conventional binary decision tree will not provide the maximum capture rates. Using the tree-GA tendency providing many local maxima for the capture rates, we can estimate the return period for the trade-off curve by using the extreme-value statistics. We have assessed the accuracy for the trade-off curve in typical fundamental cases that may be observed in real customer data cases, and found that the proposed tree-GA can construct the effective trade-off curve which is close to the optimal one.

Key words: data mining, decision tree, genetic algorithm, bump hunting, extreme-value statistics, trade-off curve, accuracy, return period, evaluation.

1 Introduction

The bump hunting is to find the regions where points we are interested in are located more densely than elsewhere and are hardly separable from other points; see Figure1. Such a problem is seen everywhere in common. However, it has been rather neglected because of the difficulty to solve the problem. In finding such regions, the patient rule induction method (PRIM) is sometimes referred to because PRIM finds boxes in which the response average is high in the feature space; PRIM differs from tree-based partitioning methods (see [7, 9]). In contrast, we seek the bump regions using a newly proposed method, the tree-GA. The tree-GA is basically a kind of tree-based methods, but finds the better trees with the assistance of the genetic algorithm; in addition, we use the extreme-value statistics (e.g., [3]) for obtaining the optimal tree (see [23]). The bump hunting has been studied in the fields of statistics, data mining, and machine learning [1, 2, 8], as well.

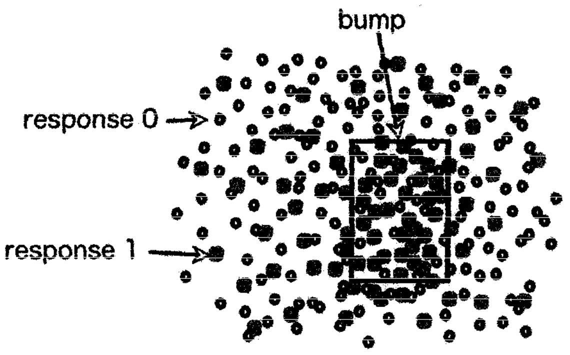


Fig. 1. The bump hunting for the denser regions to response 1 points which are hardly separable from response 0 points.

1.1 *Trade-off curve*

Suppose that n points are located in a z-dimensional feature variable space, where each point is assigned response 1 or response 0 as its target variable. By specifying a pureness rate p_0 in advance, where the pureness rate p is the ratio of the number of points of assigned response 1 to the total number of points assigned responses 0 and 1 in the target region, a maximum capture rate c_m will be obtained, where the capture rate c is the ratio of the number of points assigned response 1 to the number of points assigned responses 0 and 1 in the total regions. Then a trade-off curve between the pre-specified pureness rate p_0 and the maximum capture rate c_m can be constructed; see Figure 2.

Now, we let TP be true positive, TN be true negative, FP be false positive, and FN be false negative. Since a response 1 point in or outside the bump regions is considered to be TP or FN, respectively, and a response 0 in or outside the bumps is FP or TN, the pureness rate p can be defined by

$$p = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FP}}$$

in the bump regions; the capture rate c can also be defined by

$$c = \frac{\#\mathrm{TP}}{\#\mathrm{TP} + \#\mathrm{FN}}$$

in the total region, where "#" expresses the size of the samples. In a recall-precision curve, recall is defined by #TP/(#TP+#FN) which is identical to

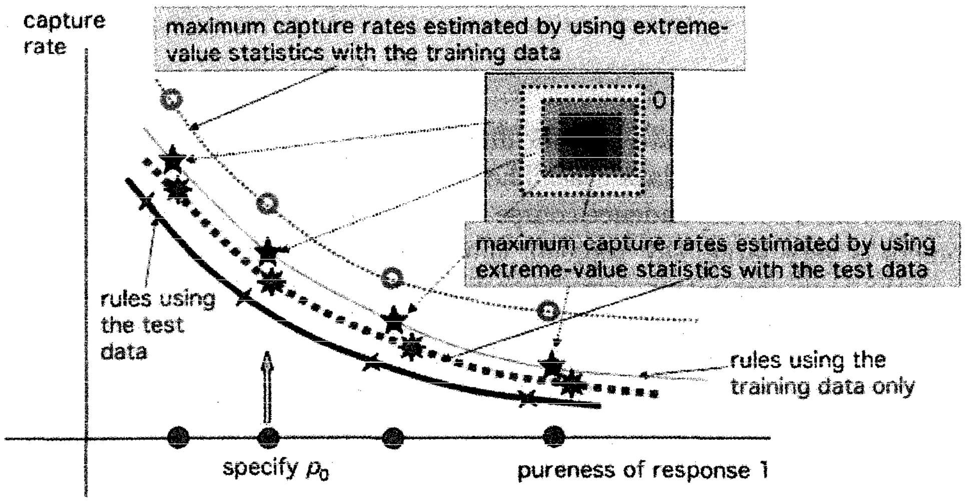

capture rate
maximum capture rates estimated by using extreme-value statistics with the training data
0
maximum capture rates estimated by using extreme-value statistics with the test data
rules using the test data
rules using the training data only
specify p_0
pureness of response 1

Fig. 2. Tradcoff-curve between the pureness rate and the capture rate.

the capture rate, and precision is defined by #TP/(#TP+#FP) which is identical to the pureness rate; thus, a trade-off curve between the capture rate and the pureness rate seems to be equivalent to a recall-precision curve superficially (see [4, 6], e.g). However, we should note that these two are totally different from each other. As is seen in Figure 3, it can be considered that our trade-off curve is constructed by collecting the skyline points of many recall-precision curves where each curve is corresponding to one classifier. Thus, the proposed trade-off curve is new. In classification problems, the misclassification rate is used as a criterion for efficiency. However, in bump hunting problems, the trade-off curve becomes a criterion instead.

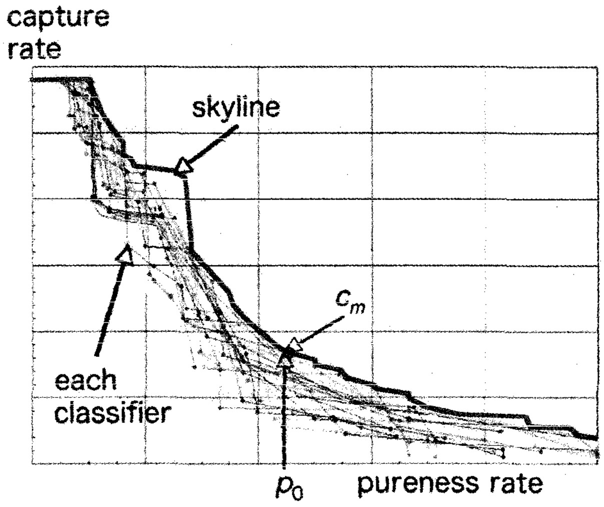


Fig. 3. Trade-off curve as a skyline curve constructed by many recall-precision curves where each curve is corresponding to one classifier.

1.2 Customer data

In this paper, we use a real data case for illustration. Figure 3 is the collection of many trees using the real data case. The customer data case we are dealing with is taken from a correspondence course in Japan (see [16, 18]). The number of customers is very large, say 160,000; thus, we will not use all these data because of the high computing cost. Therefore, we will treat 15,870 samples, randomly selected from the original database, where the number of response 1 (the customers, we are interested in) is 2,863; thus the mean pureness rate becomes 18.0%. The number of features of the customers is more than 60, but we will use 41 variables; the variables are both continuous and discrete. We call this 1/10 model here. A much smaller case consisting of 1,635 samples was also investigated, where the number of response 1 is 290; the mean pureness rate is 17.7%. The number of variables is 44. We call this 1/100 model. Our primary objective is how many response 1 samples can be captured if we require at least 40-50% pureness rate, p_0, from a practical viewpoint using these two smaller models.

1.3 Using the tree-based method

To make future actions easier, for example in the customer database, adopting the tree-based method is considered to be natural for explicit decision making. Thus, we adopt simpler boundary shapes such as the union of z-dimensional boxes located parallel to some explanation variable axes for the bumps as shown in Figure 1. However, the tree obtained by the conventional algorithm from the top node to downwards using the Gini's index will not provide the optimal solution. A tree in which randomly selected feature variables at each node may give a much nicer solution than that in which feature variables are determined by the conventional algorithm. The decision tree primarily tries to make some region classify into much purer subregions. Usually, the purer regions are concerned with as the target point regions (the response 1 points), and the decision tree works in such a situation. However, if we are not interested in response 0 point regions where the decision tree intended to find the purer regions, we may discard such regions and expect much denser regions for response 1 to the rest of the regions. In a messy data case as shown in Figure 1, the decision tree also can do this; consequently, it can find the boundaries for the bumps indirectly. Thus, we have proposed to use the random search for the feature variables preserving that the splitting point is determined by using the Gini's index (see [15, 23]). This is called the tree-GA.

2 The tree-GA with the extreme-value statistics

2.1 *Tree-GA procedure applied to the training data*

In the decision tree, by selecting optimal explanation variables and splitting points to split 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 p are satisfying to be larger than the pre-specified pureness rate p_0. However, much response 1 points could be obtained if we locate appropriate 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 using the GA, but the best splitting points are determined by using the Gini's index. The mutation can be done in the same manner to the standard genetic algorithms. However, the crossover should be different from those used in common because we are dealing with the tree structures. To preserve good inheritance in the tree structures, we have designed our crossover method as shown in Figure 4; we will know later that this causes many local maxima for the capture rates.

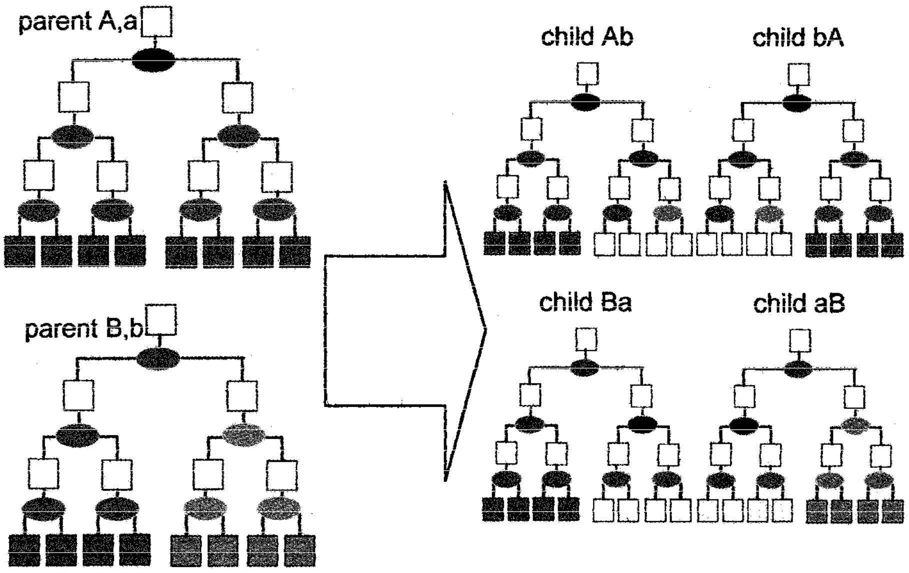
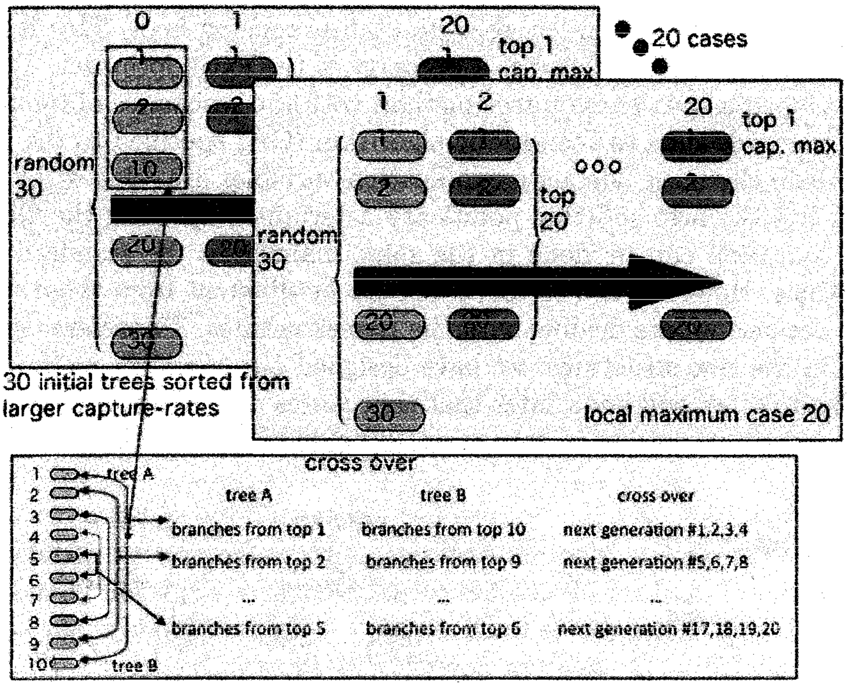


Fig. 4. Crossover in the tree-GA.

So far, we have been using the following evolution procedure in the tree-GA: 1) the number of initial seeds is set to 30; here, the initial seeds mean the trees where the explanation variables to be allocated to each branch are randomly

selected,
2) obtain the capture rate to each seed tree, and select the top 20 best trees,
3) in the next generation, divide each tree to the left wing with or without the top node and the right wing with or without the top node, and combine the left wing and right wing trees of different parents to produce children trees (see Figure 5); why we adopt this crossover method is to preserve a good inheritance in evolution procedure; the mutation rate is set to around 5%; then, 40 children are then generated, and select the top 20 best trees,
4) this evolution procedure is repeated by the 20th generation,
5) at the final stage, select the best rule.
6) we do procedures 1) - 5) for 20 cases, and select the best one rule.



tree A	tree B	cross over
branches from top 1	branches from top 10	next generation #1,2,3,4
branches from top 2	branches from top 9	next generation #5,6,7,8
...	...	...
branches from top 5	branches from top 6	next generation #17,18,19,20

Fig. 5. A typical tree-GA procedure applied to the training data.

The tree-GA algorithm has a strong inclination of searching for the local maxima because we are intended to preserve a good inheritance in evolution procedure. Solutions obtained by the tree-GA primarily are not the global optimal; this is a drawback of the algorithm. However, we have observed the existence of many local maxima with each starting point in the tree-GA procedure. This turns out to become an advantage; the use of the extreme-value statistics (e.g., [3]) can then be used to estimate the return period (a maximum capture rate with many starting points, e.g., $c_{\mathrm{RP}_{1000}}$ for 1000 points), 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. This property is also observed in a real customer database [16]. Thus, we add a function of
7) estimating the return period capture rate by using the best 20 trees in each

final stage of the evolution to our tree-GA procedure; that is, we do procedures 1) - 5) for 20 cases, and estimate the return period using these 20 local maxima. The underlying probability distribution is assumed to be the gumbel distribution, and the estimated return period (e.g., $c_{\mathrm{RP}_{1000}}$) is found to be consistent to the actual value (the maximum value using 1000 starting points) in the simulation study. In Figure 1, how we have obtained the trade-off curve for the return period is shown. The procedures explained above is applied only to the training data. To assess the estimation accuracy, the tree-GA procedure applied to the evaluation and test data is required (see [19, 12, 13, 14, 20]). Whether the assumption that the extreme-value statistics is valid to the test data should also be investigated.

2.2 Tree-GA procedure applied to the evaluation and test data

In the GA procedures explained above, the optimal tree is constructed by using the training data only; that is, 20 rules after 20 evolution generations to each procedure are obtained by using the training data. Although the trade-off curve using the training data is actually applicable rules, it is well-known that the result by this method is optimistic, and the estimated return period will also be optimistic, in other words, be conservative. We want to know a much more accurate return period for the trade-off curve, even though the return period rule shall be unknown. We should use the test data for accurate evaluation. This is performed by the assessment methods such as the cross-validation [5] and the bootstrapped hold-out method, the BHO method [17]. One way to do this is to use the very last generation rule for the test. However, someone may suspect that the accuracy evaluation by using the test data in evolution procedure in the tree-GA would still be optimistic. The test data are always treated like the training data. In addition, such a method cannot be applicable to assess the accuracy for the return period obtained by using the extreme-value statistics because the test data results do not necessarily follow the extreme-value distributions even though the training data results do. The return period capture rates using the extreme-value statistics would no longer be obtained because the results using the test data would not necessarily have the property of local maximum.

To overcome this problem, we modify the scheme of the tree-GA procedure adaptable to the test data assessment. First, we classify the original data into three subsets: to 1/3, 1/3, 1/3, or to 50, 25, 25%; the first, the second, and the third are for the training, the evaluation, and test data, respectively. Selection of the seeds, crossover method, and the mutation rate, and etc. are almost the same as mentioned in 2.1. The difference is the following. At each evolution generation stage, we produce the trees by using the training data, and select the best trees using the evaluation data. Then, we can expect that the final stage results could be the local maxima for the evaluation data, and we may apply the extreme-value statistics to these final results, just as we applied the extreme-value statistics to the optimal results obtained by using the training data. Finally, to assess the accuracy, we apply the optimal rule to the test data that are refereed to nowhere else and provided in advance. Figure 6 shows one of the tree-GA procedure results using the BHO method. On the upper right in the figure, although the capture rates for the training data and the test data are not necessarily monotonically increasing because the optimality in the training data case is not succeeded to the next generation, we can see that the capture rates are stabilized within 10 generations, and this tendency is observed in simulation data and real data in common. Using this method, 20 local maxima could be obtained (on the lower left in the figure), and we can estimate the parameters in the underlying probability distribution (on the lower right in the figure). Then, we may apply the extreme-value statistics to these results to estimate a much more accurate trade-off curve. In addition, we could assess the accuracy for this trade-off curve.

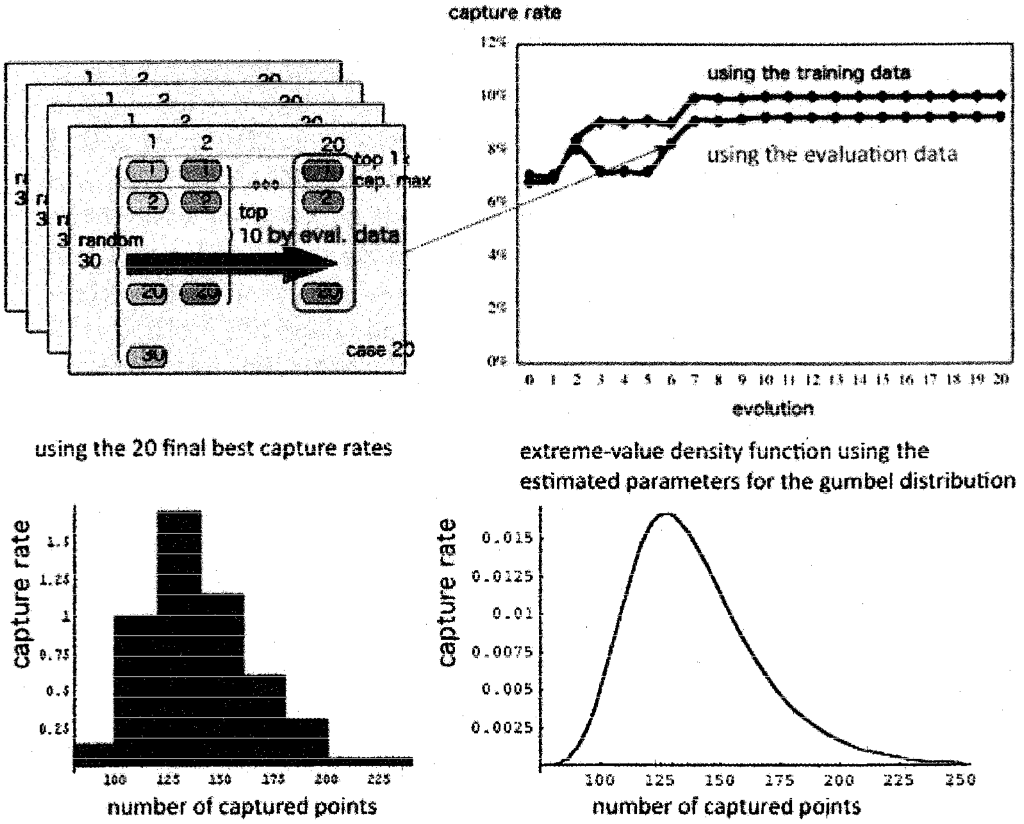


Fig. 6. Applying the evaluation and test data to the tree-GA procedure.

Figure 7 shows the distribution similarity between the evaluation and the test data. We may assume that the evaluation data results follow the extreme-value statistics by the property of the local maxima. Using 200 initial cases, we have checked if the similarity holds between the evaluation data results and the test data results. It seems that the test data results look like the evaluation data results. Therefore, we may estimate the return period for the trade-off curve by the test data results. In Figure 1, we can see that the dotted curve is representing this.

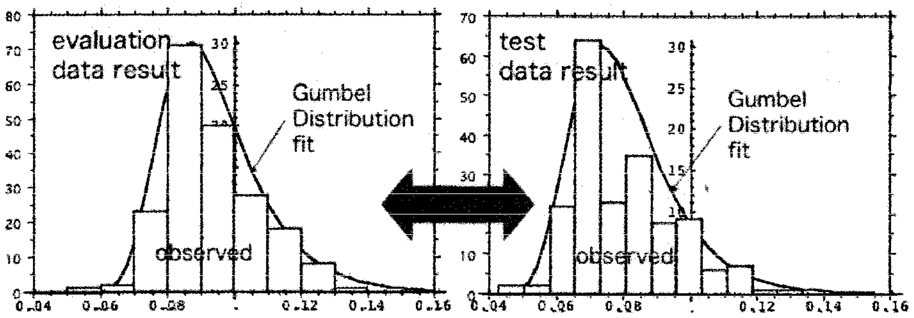


Fig. 7. The relation of the capture rates between by the evaluation data results and the test data results.

3 Accuracy assessment of the trade-off curve between the simulation results and the theoretical results

3.1 Simulation model

In this paper, we investigate the trade-off curve accuracy by applying our proposed tree-GA bump hunting method to three typical simulation data cases, which are mimicked by real data; the cases are shown in Figure 8. Response 1 points are embedded with Gaussian distribution in uniformly distributed response 0 point area. The case (1) is that: on uniformly distributed response 0 points, response 1 points of Gaussian distribution are located in the center. The cases (2,3) are that: on uniformly distributed response 0 points, response 1 points are distributed according to the mixture of the uniform distribution and the Gausissian distribution. When we assume the numbers of feature variables are 1, 2, 3, 4, 8, 16, 32, 64, and that there are no correlations among the feature variables. We have generated 20,000 points random numbers by using the Mersenne twister [22].

3.2 Incompetence of finding the bump regions by the conventional classifiers

First, we used R package *ipred* for bagging (one of the ensemble methods) and *e1071* for SVM (support vector machine) to check if the well-known classifiers can find the bump regions efficiently. In bagging, we used CART as the base

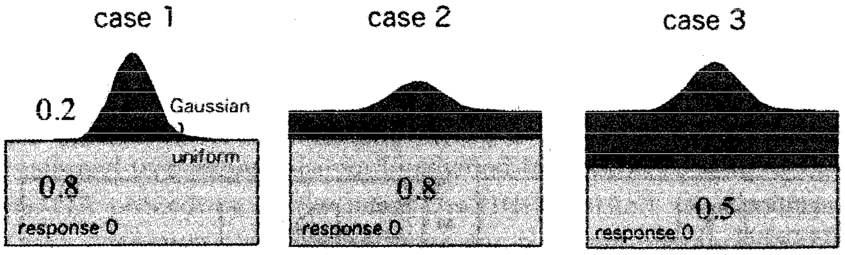


Fig. 8. Three kinds of data case for bump hunting accuracy assessment.

classifier. In the SVM, radial basis function kernel is used, the value of cost parameter is 8, and the value of gamma is 4. We can see that the single decision tree CART, bagging, and the SVM performed very poorly as shown in Table 1. The misclassification rates using these conventional classifiers are incompetent. For example, in 1-dimensional case, all the methods provide the trivial results: if a classifier determines that all the data are classified into response 0, then the misclassification rate becomes 0.2. This means that the classifier did nothing effective. We next apply the proposed tree-GA to the same simulation cases.

Table 1
Misclassification rates appled to the simulation cases using the conventional classifiers

	case 2			case 3		
dimension	CART	bagging	SVM	CART	bagging	SVM
1	0.199	0.198	0.198	0.368	0.497	0.497
2	0.170	0.168	0.258	0.340	0.343	0.501
4	0.130	0.125	0.268	0.338	0.318	0.497
8	0.127	0.111	0.320	0.343	0.309	0.498
16	0.127	0.115	0.322	0.335	0.310	0.501
32	0.131	0.108	0.319	0.336	0.303	0.499
64	0.132	0.102	0.321	0.342	0.305	0.500

3.3 Effectiveness of finding the bump regions by the tree-GA

We introduce the assessment results for the trade-off curve accuracy using the tree-GA bump hunting. The data cases are the same in Figure 8. Table 2 shows the capture rate results by the tree-GA bump hunting along with the theoretical value when the pureness rate is 0.6. We see that most of them show very good results, in contrast with misclassification rates obtained by using the conventional classifiers.

Table 2
Capture rates when pureness rate is 0.6.

	case 1		case 2		case 3	
dim.	theo.	tree-GA	theo.	tree-GA	theo.	tree-GA
1	0.0002	0.0043	0.0001	0.0065	0.6612	0.5399
2	0.9684	0.9069	0.4362	0.4307	0.6666	0.6447
3	0.9971	0.9266	0.5334	0.4798	0.6666	0.6732
4	0.9995	0.9245	0.5433	0.4840	0.6666	0.6740
8	0.9999	0.9317	0.5454	0.4777	0.6666	0.6633
16	0.9999	0.9308	0.5455	0.4807	0.6666	0.6677
32	0.9999	0.9310	0.5455	0.4074	0.6666	0.6686

Figure 9 shows other cases with various pre-specified pureness rates. In the figure, the simulation results of the trade-off points are superimposed on the theoretical (i.e., maximum obtainable capture rate vs. pre-specified pureness rate) curves. We can see that the capture rates using the tree-GA are closely to those of theoretical values. Although we can see that as the dimension becomes larger, the trade-off curve moves to the right, the theoretical values and the tree-GA results are consistent overall.

Table 3 shows the ratio of the capture rate results by the tree-GA to the theoretical values. Almost all the results show very good results. In case (1), these ratios are larger than 91%; in case (2), they are larger than 86%; in case (3), they are larger than 93%. Figure 10 shows a typical example case in the two-dimensional case in the case of (3). We can see that the tree-GA results are almost the same as the theoretical values.

4 Application to real customer data case

Since the real data is messy [10, 11], the trade-off curves obtained by using the proposed tree-GA method may also fluctuate so much. We here used the 1/10 model as a typical application case. Figure 11 shows the trade-off curve for the real customer data case using the BHO evaluation method. We can use one of the rules out of many seeds for a pre-specified pureness rate. In addition, we can know how far the capture rate using this rule is located below the return periods.

Why we adopt the BHO rather than the cross-validation method is that the computing time by the BHO is shorter than that by the cross-validation; in the typical real customer data case, the computing time becomes 80 hours by

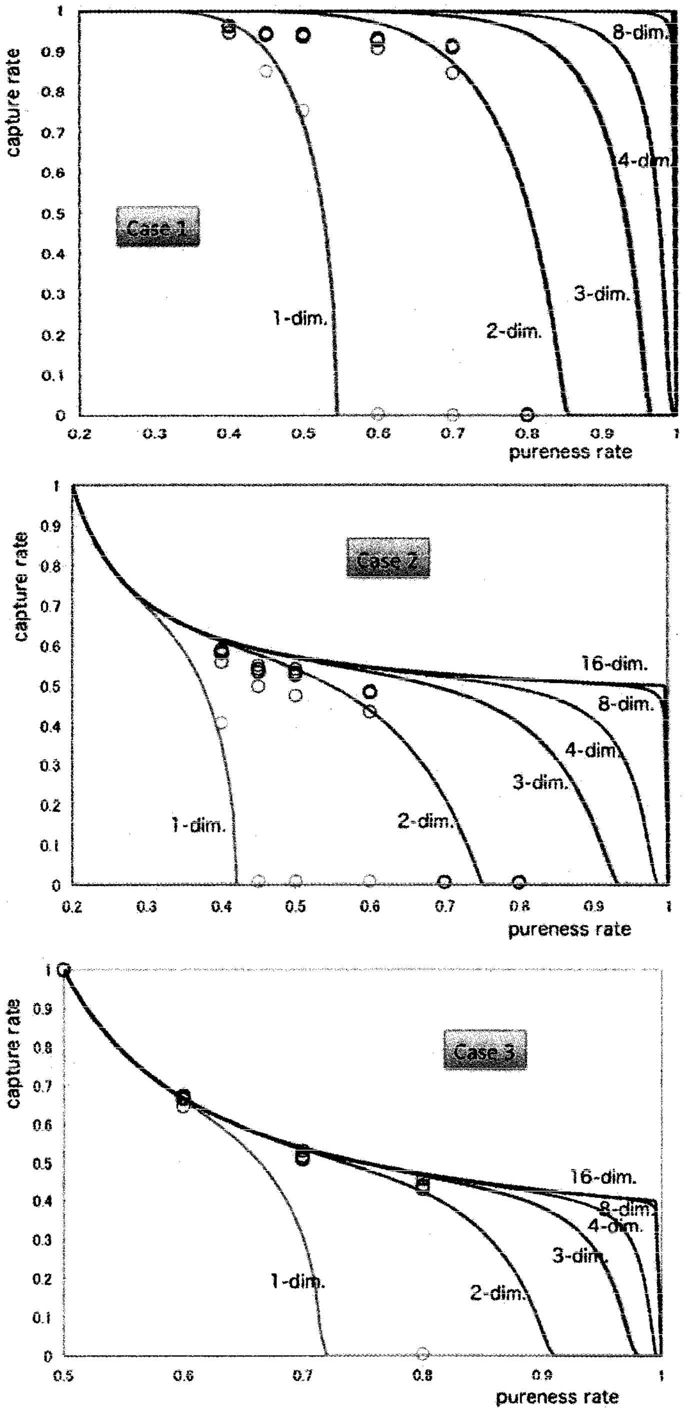


Fig. 9. Capture rates using the tree-GA and the theoretical values.

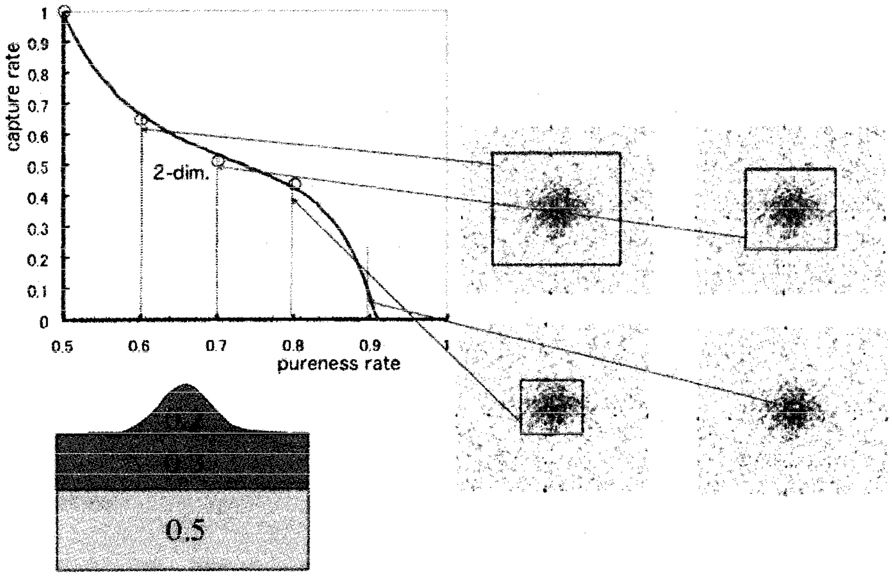


Fig. 10. Simulated bump regions using the tree-GA.

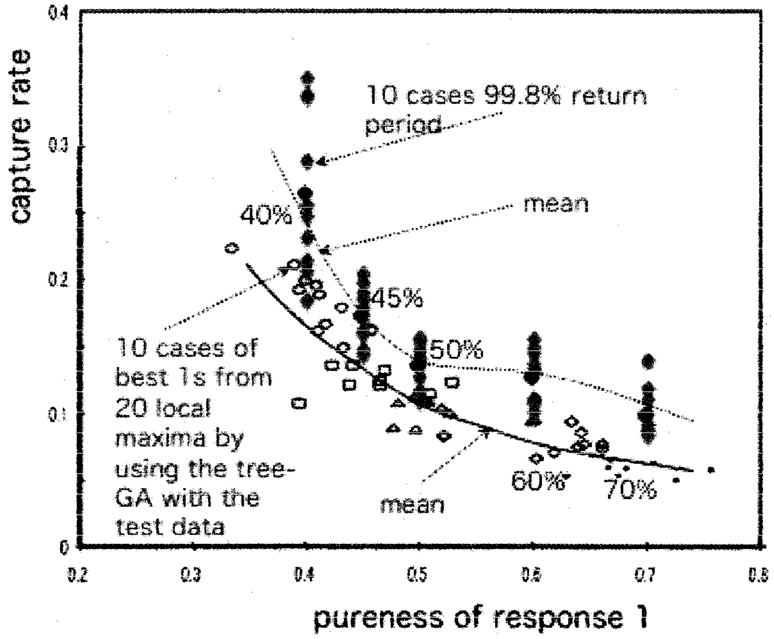


Fig. 11. Trade-off curve for the real customer data case.

a personal computer when we use the cross-validation, but 40 hours by the BHO method.

Table 3
Ratios of the capture rates between the tree-GA results and the theoretical values.

pureness	40%	45%	50%	60%	70%	80%
dimension			case 1			
2	0.95	0.94	0.94	0.94	0.97	-
3	0.96	0.94	0.94	0.93	0.92	-
4	0.96	0.94	0.94	0.92	0.91	-
8	0.96	0.95	0.94	0.93	0.91	-
16	0.96	0.94	0.94	0.93	0.91	-
dimension			case 2			
2	0.91	0.86	0.87	0.99	-	-
3	0.94	0.90	0.92	0.90	-	-
4	0.95	0.91	0.92	0.89	-	-
8	0.95	0.91	0.93	0.88	-	-
16	0.96	0.93	0.94	0.88	-	-
dimension			case 3			
2	-	-	1.00	0.97	0.95	1.01
3	-	-	1.00	1.01	0.94	0.93
4	-	-	1.00	1.01	0.95	0.93
8	-	-	1.00	0.99	0.96	0.93
16	-	-	1.00	1.00	0.98	0.97

5 Concluding Remarks

In difficult classification problems of z-dimensional points into two groups having 0-1 responses due to the messy data structure, we have shown to use the tree-based and genetic algorithm assisted bump hunting method, a newly proposed tree-GA procedure. We have also shown that the criterion in the bump hunting is not the misclassification rate but the trade-off curve between the pureness rate p and the capture rate c; this curve is not the recall-precision curve (i.e., not the ROC curve) because the trade-off curve consists of many trees. Due to the inclination to find the local maxima in the proposed tree-GA, we have proposed to use the extreme-value statistics to find the return period as a global maximum capture rate starting with many initial seeds. The gumbel distribution can be fitted to the real and simulated data well in our study. Knowing this, we will make future decisions by applying the rules obtained by using the training data with the knowledge of how far the rules we are using are located from the optimal points. To assess the accuracy of the trade-off curve, we have shown to use the training, evaluation and test data sets; the evaluation is done either by the cross-validation method or the bootstrapped hold-out method. We have also shown the simulation results that the conventional classification methods, such as the bagging and the SVM, are not efficient to the bump hunting. On the contrary, we have shown that the tree-GA worked very well, e.g., for typical cases which are mimicked by real case. In the simulation study, we have shown that the capture rates are obtained larger than about 90% of the theoretical values.

References

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

[2] Becker, U and Fahrmeir, L. , "Bump hunting for risk, a new data mining tool and its applications", *Computational Statistics*, 16, pp.373-386, 2001.

[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", *Journal of American Statistical Association*, 78, pp.316-331, 1983.

[6] Fawcett, T. , "An introduction to ROC analysis", *Pattern Recognition Letters*, 27, pp.861-874, 2006.

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

[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, 2009.

[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", *IFORS Triennial Conference*, 2005.

[12] Hirose, H., "The bump hunting by the decision tree with the genetic algorithm, *Advances in Computational Algorithms and Data Analysis*, pp.305-318, Springer, 2008.

[13] H. Hirose, "Evaluation of the trade-off curve in the bump hunting using the tree genetic algorithm", *1st IMS Asia Pacific Rim Meetings*, 2009.

[14] H. Hirose, "Assessment of the trade-off curve accuracy in the bump hunting using the tree-GA", *Third International Conference on Knowledge Discovery and Data Mining*, pp.597-600, 2010.

[15] Hirose, H., Yukizane,T. and Miyano, E., "Boundary detection for bumps using the Gini's index in messy classification problems", *The 3rd International Conference on Cybernetics and Information Technologies, Systems and Applications*, pp.293-298, 2006.

[16] Hirose, H., Yukizane,T. and Deguchi T., "The bump hunting method and its accuracy using the genetic algorithm with application to real customer data", pp.128-132, *IEEE 7th International Conference on Computer and Information Technology*, 2007.

[17] Hirose, H., Ohi, S. and Yukizane, T., "Assessment of the prediction accuracy in the bump hunting procedure", *The 6th Annual Hawaii International Conference on Statistics, Mathematics, and Related Fields*, 2007.

[18] Hirose, H. and Yukizane, T., "The bump hunting using the decision tree combined with the genetic algorithm, extreme-value statistics aspect", *International Conference on Machine Learning and Data Analysis*, 2007.

[19] Hirose, H. and Yukizane, T., "The accuracy of the trade-off curve in the bump hunting", *The 7th Annual Hawaii International Conference on Statistics, Mathematics, and Related Fields*, 2008.

[20] Hirose, H., Yukizane, T. and Zaman, F., "Accuracy assessment for the trade-off curve and its upper bound curve in the bump hunting using the new tree genetic algorithm", *7th World Congress in Probability and Statistics*, 2008.

[21] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", *International Joint Conferences on Artificial Intelligence*, 1995.

[22] Matsumoto, M. and Nishimura, T., Mersenne Twister, "A 623-dimensionally equidistributed uniform pseudorandom number generator", *ACM Transactions on Modeling and Computer Simulation*, 8, pp.pp.3-30, 1998.

[23] 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*. pp.2332-2339, 2006.