A Novel Method of Evaluating a Classification Tree: The Structural Evaluation Index

Jing Zhang, Rui Song, Shengping Xia, Weidong Hu, Wenxian Yu
(State Lab of Automatic Target Recognition, National University of Defense Technology, Changsha 410073, China)
Email: ben_bbzj@sina.com

Abstract: Popular analytic methods for classification trees mainly focus on the inner information contained in tree nodes, while ignoring the integrative information implied in the tree structure. In order to describe and utilize the structural attributes of classification trees effectively, with three types of the leaf nodes in a classification tree defined, generalized propositions on the structure of a classification tree have been constituted, which have led to the normative definitions of the structural information of leaf nodes with a universal form. Then a comprehensive structural evaluation index (SEI) of classification trees has been formulated. Finally, the hierarchical SOM tree algorithm and IRIS data have been used to test the validity and performance of the presented evaluation method. The results have shown that the method could depict the discrimination ability and quality effectively.

Key Words: Classification Tree; Propositions on Tree Structure; Structural Evaluation Index (SEI); Leaf Node Type; Structural Information

I. INTRODUCTION

The classification tree is a popular clustering method widely used in pattern recognition, knowledge discovery, decision support and machine learning, etc. It has shown satisfying performance in class discrimination, rule extraction, feature selection and syntax analysis. With a hierarchical structure, the classification tree can partition a complex pattern through the nodes and their connections into a series of simpler sub patterns, which is suitable for solving the problem of classifying complicated samples. [1, 2]In order to depict and compare different classification trees generated with various algorithms, multiple methods have been introduced to evaluate their overall performance, The methods basing on information entropy, mutual information, node purity, misclassification rate, or the ratio of within-cluster scatter to between-cluster separation[3, 4], get the most popular use.

Yet, most of the evaluation methods mentioned above only analyze the performance of each node itself, while ignoring the organization of the nodes in the target tree, i.e. the information hidden in the hierarchy of a tree. However, the inner information of nodes and the relationship between those nodes make up the two fundamental elements of a classification tree. So, the tree structure, viz. the relationship between nodes, should be granted sufficient importance as the local performance of nodes. The Number of Leaf Nodes [4] and the Minimal Description Length [5] can give an illustration of the tree structure, but the former is too simple while the latter is not explicit to embody.

To solve the problem, a novel method named **structural evaluation index (SEI)** that utilizes both the inner property of the individual nodes and the structural information of the whole tree, has been introduced in this paper. In section 2, after the definition of three types, the pure, the lopsided and the confused, of leaf nodes, generalized propositions on the structure of a classification tree have been constituted. Then normative definitions of the structural information of different node types have been given, and the overall structural evaluation index S_T has been constructed. Section 3 shows how to constitute S_T in detail. In section 4, hierarchical SOM trees have been used to analyze IRIS data sets, and the comparison of SEI with some other popular indices of these trees is given. Section 5 is the summary which has analyzed the extendibility and applicability of the novel evaluating method.

II. EVALUATING THE STRUCTURE

In substance, the classification tree is composed of nodes and their connections [1]. So, to take the inner status of nodes into account only, yet to ignore their hierarchical organization and global functionality, would neglect the other important essential attribute of classification trees. Meanwhile, most of the evaluation methods are embedded in the tree generating process, and different indices are always associated with different training algorithms respectively. Thus, there is an absence of a consentaneous post-processing analysis method. Therefore, it is necessary to catch the global property of tree structure so as to make more effective and more comprehensive evaluation of the classification tree.

2.1 Leaf Node Types

A classification tree consists of branching nodes, representing specific discrimination criteria, and leaf nodes, containing different training data. The nodes are organized in a specific hierarchy. According to the partition ability of leaf nodes, which is reflected by the training data contained in them, leaf nodes could be divided into three types. **The pure leaf node (PLN)** contains data from only one class, and samples falling into it are all recognized as this unique class. **The lopsided leaf node (LLN)** contains data from more than one class, but shows tendency to one class under specific criterion, and samples falling into it are recognized as this "distinctive class". **The confused leaf node (CLN)** contains data from several classes, yet the applied discrimination criterion cannot divide them apart, so at recognition time it cannot determine which class the sample falling in belongs to. Fig. 1 shows an illustration of the three types of leaf nodes in a classification tree.

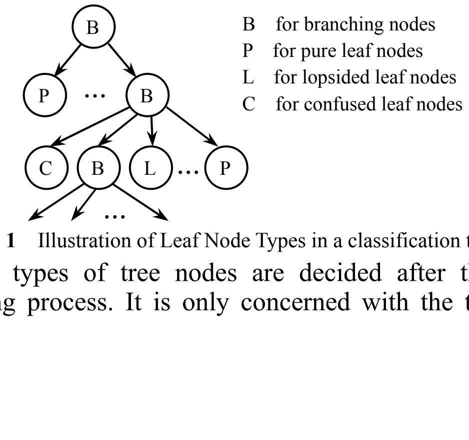


Fig. 1 Illustration of Leaf Node Types in a classification tree

The types of tree nodes are decided after the tree generating process. It is only concerned with the training result, and independent of the different tree construction algorithm. Then the evaluation of the performance of a classification tree is transformed into the analysis of the synthetic performance of each leaf nodes, which is defined as the **structural information** of leaf nodes.

2.2 Propositions on the Structure of Classification Trees

The propositions on tree structure below have been induced, basing on the analysis of the applications of classification trees.

Proposition 1: As to a classification tree, if the pure leaf nodes and the lopsided leaf nodes are at higher hierarchy, i.e. nearer to the root node, their portion in all leaf nodes is greater, and the absolute and relative sample ratios for them are larger, it is more suitable for the classification use. Here the absolute ratio is the number of samples contained in one leaf node to the number of total training samples, and the relative ratio is the number of samples of each class in a node to the number of total samples of that corresponding class.

Proposition 2: If the confused leaf nodes are at lower hierarchy, and their portion in all nodes is smaller and their sample ratios are lower, the tree is more suitable for classification.

2.3 Structural Information of Leaf Nodes

The structural information of a leaf node represents the leaf node's overall classification performance within a specific tree structure. According to the propositions in Section 2.2, the structural information should consist of three components: *the hierarchical level* a leaf node belongs to, *the ratio of the sample number* it contained to the total sample number, and *the functional measurement* of the node itself. This can be represented as $S = f\left(W(l), R(n), A_c(i)\right)$, where S is the structural information of leaf nodes, $W(l)$ is the level factor of the node, $R(n)$ is the sample ratio factor, and $A_c(i)$ represents the discrimination ability for a pure and a lopsided leaf nodes, or the confusion degree for a confused leaf node. The following form of $f(\cdot)$ is applied in this paper:

$$S = W(l)\times R(n)\times A_c(i) \qquad (1)$$

Define $S_{T,P}$ as the overall structural information of the pure leaf nodes in a classification tree, $S_{T,L}$ for the lopsided ones and $S_{T,C}$ for the confused ones.

2.4 Structural Evaluation Index S_T

Based on the definitions of the overall structural information above, **the structural evaluation index (SEI)**, S_T, of a classification tree can be constituted:

$$S_T = \left(e^{\alpha\cdot S_{T,P}} + e^{\beta\cdot S_{T,L}}\right)\Big/ e^{\gamma\cdot S_{T,C}} \qquad (2)$$

Here, α, β and γ are defined as the saliency factors for different nodes types.

III. STRUCTURALEVALUATION INDEX

3.1 Denotation

The denotations to be used are defined here.

N_T - Total number of samples in a classification tree;

C_T - Total number of sample classes;

$N_{CT,c}$ - Total number of samples in class c, $c = 1,\cdots,C_T$;

N_l - Total number of leaf nodes;

$N_{X,l}$ - Number of X type leaf nodes, X=*pure, lopsided, or confused*;

$L_{\max}$ - Max number of hierarchy level;

$LN_{P,k}$, $LN_{L,k}$ and $LN_{C,k}$ - Numbers of the three different types of leaf nodes in level k respectively;

L_i - Hierarchical level of leaf node i (root node at level 0);

C_i - Number of classes of samples in node i;

$N_{T,i}$ - Total number of samples in leaf node i;

$N_{i,k}$ - Number of samples of class k in node i, $i = 1,\cdots,N_l$ and $k = 1,\cdots,C_i$.

3.2 Structural Information of PLN

The effectual discrimination ability $A_c(i)$ of pure leaf nodes is defined as the number of samples contained in it to the total number of samples of that corresponding class, which represents its ability to classify samples of this class correctly. $R(n)$ is the number of samples in that node to the total sample number, and represents the node's importance in the whole sample space and the influence it has. The level factor,

$$W(l) = \frac{2\exp(-\kappa(l-\lambda))}{1+\exp(-\kappa(l-\lambda))} + 0.1, \quad l = 1,\cdots,L_{\max},$$

where κ is the scale factor and λ is the transition factor. We take $\kappa = 1$ and $\lambda = 1$ in this paper.

Suppose pure node i contains samples from class c only, then:

$$A_c(i) = \frac{N_{T,i}}{N_{CT,c}} \qquad (3)$$

$$R_{i,c}(n) = R_{i,c}(N_{T,i}) = \frac{N_{T,i}}{N_T} \qquad (4)$$

$$W(L_i) = \frac{2\exp(-(L_i-1))}{1+\exp(-(L_i-1))} + 0.1 \qquad (5)$$

The structural information of pure node i is:

$$s_{i,L,P} = W(L_i)\times R_{i,c}\left(N_{T,i}\right)\times A_c(i) \qquad (6)$$

Then the overall structural information of all pure leaf nodes is:

$$\begin{aligned} S_{T,P} &= \sum_{k=1}^{L_{\max}}\left(\sum_{i=1}^{LN_{P,k}} s_{i,L,P}\right) \\ &= \sum_{k=1}^{L_{\max}}\left(\left(\frac{2\exp(-(L_k-1))}{1+\exp(-(L_k-1))}+0.1\right)\times\sum_{i=1}^{LN_{P,k}}\left(\frac{N_{T,i}}{N_T}\times\frac{N_{T,i}}{N_{CT,c}}\right)\right) \end{aligned}$$

3.3 Structural Information of LLN

The lopsided leaf nodes are different under variant judging methods. For instance, the lopsided nodes confirmed by *the sample number ratio* is different from those confirmed by *the ratio of within-cluster scatter to between-cluster separation*. But once they are specified, the meaning and the effect of the lopsided nodes are the same. The sample number ratio has been used to illustrate how to define the overall structural information of the lopsided leaf nodes, $S_{T,L}$.

Firstly, a lopsided leaf node should been defined by the sample number ratio criterion. In a leaf node, once the number of samples of one class to the total number of samples in the node has reached the threshold R $(R > 0.5)$, it can be confirmed as a lopsided node, i.e. if $\max_k (N_{i,k}/N_{T,i}) \geq R$, it is a lopsided one.

Let $S_{i,L,L}$ denote the structural information of a lopsided leaf node. Because a lopsided node will judge all the samples falling in as one class, it will inevitably misclassify samples from other classes. The discrimination ability of that node and the overall performance of the tree are practically decreased.

So for leaf node i which is "lopsided to" class c, its effectual discrimination ability:

$$A_c(i) = \frac{N_{i,c}}{N_{CT,c}} \times \frac{N_{i,c}}{N_{T,i}} - \sum_{\substack{k=1 \\ k \neq c}}^{C_i} \left(\frac{N_{i,k}}{N_{CT,k}} \times \frac{N_{i,k}}{N_{T,i}} \right) \tag{7}$$

Note, when $N_{i,c} = N_{T,i}$, the lopsided node is just pure, and formula 7 can be transformed to formula 3.

The sample proportional factor is

$$R_{i,c}(n) = R_{i,c}(N_{i,c}) = \frac{N_{i,c}}{N_T} \tag{8}$$

The level factor is the same as formula 5.

Take formula 7, 8, and 5 into formula 1, and then the overall structural information of all the lopsided nodes will be gotten.

3.4 Structural Information of CLN

The confused leaf nodes cannot determine which class the samples falling in should belong to under their corresponding discrimination functions. As they do make negative influence to the classification tree, the structural information of confused nodes, $S_{i,L,C}$, should represent the uncertainty they bring up in the tree.

The information entropy can be used to show the uncertainty in a confused leaf node:

$$A_c(i) = -\sum_{k=1}^{C_i} \frac{N_{i,k}}{N_{T,i}} \left(\log_2 \frac{N_{i,k}}{N_{T,i}} \right) \tag{9}$$

The sample proportional factor is

$$R_{i,c}(n) = R_{i,c}(N_{T,i}) = \frac{N_{T,i}}{N_T} \tag{10}$$

The level factor is the same as formula 5.

Take formula 9, 10, and 5 into formula 1, and then the overall structural information of all the confused nodes will be gotten.

At last, set the saliency factors for the three types of leaf nodes:

$$\begin{cases} \alpha = N_{P,d} / N_T, & N_{P,d} = \sum_{i=1}^{LMax} N_{T,i},\ i \in PLN \\ \beta = N_{L,d} / N_T, & N_{L,d} = \sum_{i=1}^{LMax} N_{T,i},\ i \in LLN \\ \gamma = N_{C,d} / N_T, & N_{C,d} = \sum_{i=1}^{LMax} N_{T,i},\ i \in CLN \end{cases} \tag{11}$$

Bring $S_{T,P}$, $S_{T,L}$, $S_{T,C}$ and formula 11 into equation 2, then we could get the overall structural evaluation index of the classification tree, S_T.

According to the construction process above, we could draw the conclusion that S_T is an incremental function of the classification performance, i.e., if the tree is more suitable for classification, then its structural information will be more, and S_T will get the higher value.

IV. EXPERIMENTS

Different data sets and HSOM[6,7] (Hierarchical SOM) training methods of classification trees have been used to compare the SEI with other indices.

4.1. With IRIS Data

The IRIS plant data set has been used as a standard data set in pattern recognition field[8]. The data set is composed of 150 samples, 50 for each of the three types of plants: setosa, versicolor, and virginica. Each sample features four attributes (petal length, petal width, sepal length, and sepal width). Data of each feature have a different confusion degree, so the clustering performance of different classification trees based on diverse combinations of features would differ greatly.

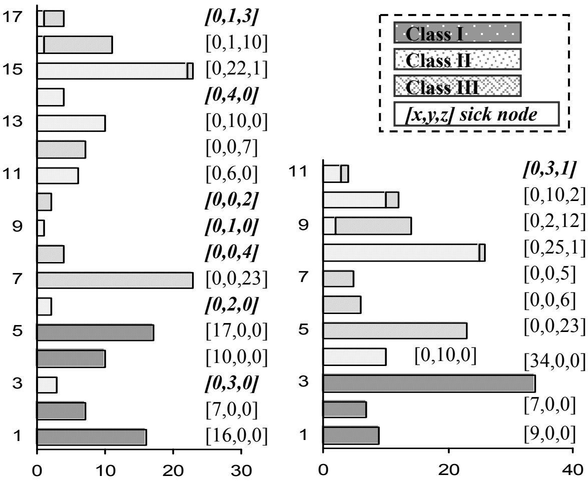
15 training sample sets have been obtained by selecting 4 single features, 6 combinations of 2 features, 4 combinations of 3 features, and all 4 features together. Training these 15 data sets with the HSOM has yielded 15 classification trees. Besides SEI, the average error rate and the information gain [9] have also been applied to assess their performance. Table 1 gives the comparison of the three indices.

Table 1. Comparison of Indices of Classification Trees

Evaluation Index / Feature Set	Leaf Node Quan.	Ave. Error Rate (%)	Infor-mation Gain	Structural Evaluation Index
1d	10	64.7	0.7451	1.0772
2d	9	70.7	0.4043	0.3800
3d	11	26	1.3434	2.8417
4d	11	6	1.4151	3.1669
1-2d	10	67.3	0.9507	1.0142
1-3d	15	10.7	1.3743	2.6538
1-4d	17	8.7	1.4294	3.1762
2-3d	18	10.7	1.3766	2.4794
2-4d	21	9.3	1.4237	2.5373
3-4d	**13**	**6**	**1.4301**	**4.6603**
1-2-3d	**17**	**4**	**1.3357**	**2.8356**
1-2-4d	17	10	1.4032	2.7644
1-3-4d	15	6.7	1.4527	3.1982
2-3-4d	**13**	**9.3**	**1.4819**	**4.3426**
1-2-3-4d	**18**	**6**	**1.4826**	**4.1732**

As shown in Table 1, the three indices can all represent the discrimination performance of the trees from various aspects. The average error rate is the simplest and the most intuitive, but is somehow unilateral. For example, feature combination [1-2-3d] has the lowest average error rate of 4%, but the classification tree is over-fitted [3] substantially. The tree is ill-formed with a lot of "sick nodes" that contains few samples. It is over branched with many redundant nodes, and its ability to generalize is poor.

The compositions of leaf nodes in HSOM trees for feature set [1-2-3d] and for feature set [4d], which has both fine classification performance and tidy structure, have been shown in Fig. 2. The vertical coordinate is the order of the leaf node, and the horizontal coordinate is the number of samples contained in each nodes. Each bar stands for a leaf node. The numbers at the right are the quantity of samples it contained, of the three classes respectively. A leaf node containing less than 5 samples is classified as the "sick" node, and its composition is marked by italic bold fonts. It can be found that the HSOM tree in figure 2a has 7 sick nodes, while the tree in figure 2b has only one.



(2a) On Feature Set [1-2-3d] (2b) On Feature Set [4d]

Fig. 2 Composition of Leaf Node in HSOM Trees

The information gain has integrated multiple factors, so its performance is better than the average error rate; but it does not take into account the information implicated in the tree structure, and cannot depict the over-fit phenomenon. For instance, combinations [2-3-4d] and [1-2-3-4d] have the greatest information gain, but combination [2-3-4d] has a high error rate and [1-2-3-4d] generates too many redundant "sick nodes".

The structural evaluation index has generally solved the aforementioned problem. The combination chosen by it, [3-4d], has relatively low error rate, high information gain and a condensed tree structure at the same time. This accords with the fact that the combination of feature [3-4d] has the highest separability. Meanwhile, combination [3-4d] is the minimal combination of the optimal features [2]. Though its information gain is a little lower than the top one, combination [1-2-3-4d], but it's still among the highest ones.

The IRIS data have also been divided into two parts for training and testing respectively. The results also show that the classification tree generated by feature combination [3-4d] has the best ability.

4.2. With Radar Target Data

Radar echoes from multiple vessel types have also been used to testify the proposed index. Table 2 has shown the sample sets used for training and test. Each sample consists of 23 features extracted from various aspects, including 1d-FFT and 2d-FFT, time-frequency transformation, radar waveform visual features, etc[11].

Table 2. Radar Target Data

Set	A	B	C	D	E	F	G	H	Total
Training	1162	2905	1015	1795	1070	696	409	510	9562
Test	150	476	361	527	317	136	102	94	2163

Two HSOM trees have been trained with the information gain and the structural evaluation index as their branching criteria respectively. The recognition rates of the training set and the test set have been shown in Table 3. It can be seen that the recognition rates in the HSOM with the SEI are significantly higher than the ones with the information gain, which has approved the performance of the SEI.

Table 3. Recognition Rate

	Information Gain		Structural Evaluation Index	
	Training Result(%)	Test Result(%)	Training Result(%)	Test Result(%)
A	81.3	73.2	90.3	87.3
B	84.2	60.4	92.4	88.6
C	63.5	37.9	97.7	92.5
D	33.2	20.5	88.6	85.8
E	20.8	12.8	100	92.1
F	13.4	0	93.7	89.2
G	27.5	10.2	95.0	90.4
H	2.5	0	96.3	94.7

V. SUMMARY

According to the theoretical analysis and the synthetic experiment, we could get the conclusion as follows:

1. The SEI method is a kind of result-oriented, post-process analysis of classification trees, so it could represent the overall performance of them effectively.

2. SEI could be directly extended and applied in fields of pattern recognition and artificial intelligence, such as classifier selection /combination [10], machine learning and decision support [9], etc. Especially, it owns excellent performance in feature selection, and it is hopeful to avoid redundant features with it.

3. SEI can be integrated with existing generation and evaluation indices to improve traditional generation algorithms for classification trees. The resulting trees will satisfy specific requirements and have fine overall structural performance as well.

Acknowledgements

The authors sincerely acknowledge Master Hua Yu for his help in experiments.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth, 1984.

[2] A. K. Jain, R. P. W. Duin and J. Mao , "Statistical Pattern Recognition: A Riview," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, no.1, pp.4-37, 2000.

[3] J. R. Quinlan, *C4.5: Programs for Machine Learning*. Wadsworth, 1993.

[4] T. S. Lim, W. Y. Loh, Y. S. Shih, *An Empirical Comparison of Decision Trees and Other Classification Method*. University of Wisconsin-Madison, Technique Report 979, Jun, 1997.

[5] A. Barron, J. Rissanen, B. Yu, "The Minimum Description Length Principle in Coding and Modeling," *IEEE Trans. Information Theory*, vol.44, no.6, pp.2743-2760, 1998.

[6] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.

[7] Z. J. Tu, G. S. Liu, "A Self-Organizing Neural Network Tree Based on Entropy," *Chinese J. Computers*, vol.23, no,11, pp.1226-1229, 2000.

[8] Http://www.ics.uci.edu/pub/machine-learning-databas.

[9] W. W. Chen, *Intelligent Decision Making*. Beijing: Publishing House of Electronics Industry, 1998.

[10] R. Song, J. Zhang, S. P. Xia, W. X. Yu, "An Adaptive Classification Method of BP-NN Group Based Classification System and its Application," *Electronica Sinica*, vol.29, no.12A, pp.1950-1953, 2001.

[11] J. Zhang, R. Song, W. X. Yu, etc. Visual Effects Based Feature Extraction for Dynamic Radar Target Echo Series. ICSP 04, 2004.