DECISION TREE CLASSIFIER WITH UNDETERMINED NODES

Masanobu Yoshikawa*, Sadao Fujimura**,
Shojiro Tanaka* * *, and Ryuei Nishii* * **

* Research Associate, Faculty of Engineering, Yamanashi University, Japan
** Professor, Faculty of Engineering, University of Tokyo, Japan
* * * Associate Professor, Faculty of Engineering, Yamanashi University, Japan
* * ** Associate Professor, Faculty of Integrated Arts and Sciences, Hiroshima University, Japan

ISPRS commission III

KEYWORDS: Land Cover, Classification, Design Algorithms, Multispectral Vector, Pattern Recognition

ABSTRACT

A new approach to preserve undetermined data for classification is proposed in this paper. The proposed classifier includes a mechanism to suspend classification for the indistinct data. The triplet decision tree has two 'determined nodes' based on binary splitting of categories and one additional 'undetermined node' for uncertain part of data. A design procedure for this type of triplet decision trees is proposed as an extension of the design procedure for binary decision trees. This method maintains advantages of general tree classifiers about computing efficiency. An effective and flexible classification is enabled by this decision tree by appling various data segmentation methods in the feature space to uncertain sample groups. Moreover, this classification tree has the effect to display hierarchical structure of similar categories and uncertainly-classified data groups.

1. INTRODUCTION

In general tree classifiers, samples in a category are processed in one group, i.e. one tree node. While classification is very effective in these usual methods, the following three major drawbacks are pointed out.

(P1) Decision trees have only one terminal node for one classification category. In these tree classifiers, samples mis-classified at one non-terminal node division in the tree have no chance to correctly classified by the succeeding steps.

(P2) Land cover categories possibly have variety of vagueness in actual data representabilities, such as indistinct distribution or existence of adjacent categories. It is true that usual decision trees make it possible to adopt the node division even with this ill conditioned data segmentation. If a multibranch tree structure is selected, decision trees may suit the nature of the data better, and classification accuracy may become better. However, the processing becomes very complicated for the design of general multibranch trees. It is one problem that a complex tree structure is required for accurate classification but is not desirable for efficient design method.

(P3) Data segmentation is executed by rigid boundaries at each tree node. In case rigid boundary is adopted, as for the data which is far from the boundary in the feature space, the node division is suitable. However, as for distributions overlap each other, the node division is less suitable and may include many mis-classifications.

A design method of decision trees taking these problems into account is useful for the processing of remotely sensed data. The following mechanisms are required for dealing with these problems:

- Samples with indefinite character can be detected and considered separately;
- Classification at each node is partially executed and decision for the indefinite samples are postponed to lower nodes;
- Each category is able to have plural terminal nodes depending on its nature in the hierarchical structure.

In this paper, a triplet tree structure is proposed to overcome the problems considering both classification accuracy and computing costs.

2. A TRIPLET TREE DIVISION–WAIT MECHANISM

The proposed methods is an extension of binary decision tree classifiers. The proposed triple tree classifier has two 'determined nodes' based on binary splitting of categories and one optional 'undetermined node' for uncertain subgroups.

Samples definitely classified by group distance criteria are classified to determined nodes. Determined nodes are labeled as definite categories and processed in a similar manner repeatedly. The ambiguous samples are classified to undetermined node, and undetermined node is labeled as the same categories as parent node. That is, undetermined node redundantly inherit categories from parent node. Samples, on the other hand, are divided into two determined nodes and one undetermined node with no redundancy. Classification of undetermined node is tried based on newly calculated group distances in the succeeding step.

The mechanisms introduced in this proposed method are as follows.

- Samples definitely classified by group distance criteria are assigned to determined nodes. Determined nodes are labeled as definite categories.

- The uncertain samples are assigned to undetermined node and undetermined node is labeled as the same categories as the parent node.

- Boundaries of data segmentation to three children nodes are determined based on error tolerance criterion of training samples. Namely, by two boundary lines approaching from the two terminals of a variable to the middle point, each determined nodes contain at most p percent of mis-classification, where p is control parameter.

For more detail, Fig. 1 shows a triplet tree and Fig. 2 illustrates a segmentation of samples to three child nodes. Fig. 3 illustrates a effective segmentation of feature space by two steps in a triplet tree, in case of two categories A, B. Based on binary splitting of categories and two boundaries in the feature space, a data histogram of parent node is calculated. Samples belong outside two boundaries are allotted respectively to determined nodes. Samples belong between boundaries are allotted to undetermined node. The blackly shown parts in the Fig. 1 are equivalent to the mis-classified samples in this segmentation, as a tolerated limit of error.

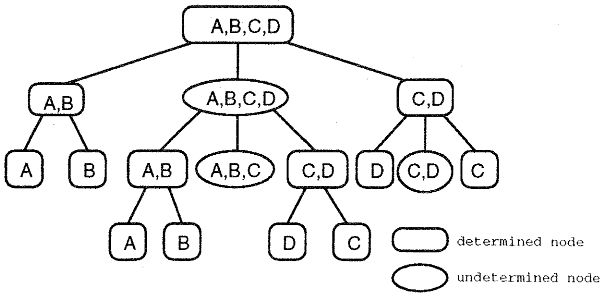

Fig.1 Triplet tree structure.

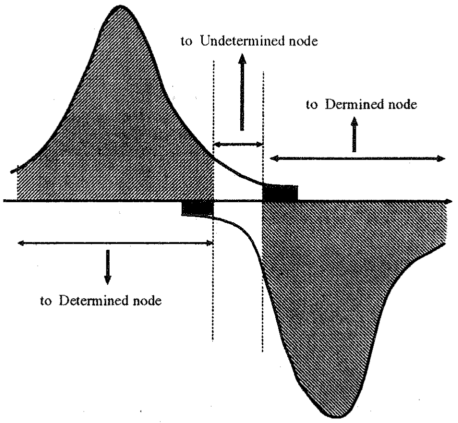

Fig.2 Segmentation to two determined nodes and one undetermined node:
sample histogram and two boundaries.

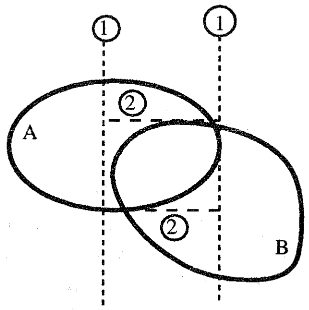

Fig.3 Segmentation of feature space by a triplet tree.

3. ADVANTAGES OF THE PROPOSAL

By this triplet tree approach, problems (P1),(P2) and (P3) with usual tree methods stated at the Section 1 can be solved. Solutions by this approach are following (S1),(S2) and (S3) respectively.

(S1) Plural terminal nodes could be appeared for each category in proposed tree classifier. Partial decision of classification is made at every division of node, and it restricts the influence of mis-classification at higher nodes.

(S2) Triplet tree is treated as en extension of binary tree. It is more adaptive than binary tree and has advantages in classification accuracy in terms of making the middle node hold uncertain groups of data. Based on binary trees, number of splitting patterns of categories at each node is the same as in the design of binary tree. As small additional costs are needed compared with binary tree, computations of design and classification are more effective than general multibranch trees.

(S3) As for data segmentation boundary in the feature space, 'undetermined node' containing samples nearby the boundary is introduced in triplet tree. The label of undetermined node is the same as its parent node. Decision of category is suspended about undetermined node and other divisions are tried to classify in the following steps.

As the decision tree approach considering ambiguity of categories, Wang proposed a binary tree design using rejection strategy(Wang,1986a).

1. Categories are divided in two subgroups using group distance.
2. Using Bhattacharyya distance, distance between each category and two subgroup is calculated. If the distance above is smaller than a threshold value D_0 decided in advance, this category is decided to be 'rejected' at this node.
3. Rejected categories are inherited to two children nodes.

Decision tree by Wang's method is binary tree, so computing efficiency is excellent. One defect of this method is the selection of threshold value D_0. A heuristic is used in design procedure.

General Bayesian classifiers with rejection area in discrimination is able to extended with 'determined area' and 'undetermined area'. Compared with these approach, proposed triplet tree classifier has advantages as a tree classifier such as efficiency in computation at multistep operation and ability to get hierarchical view of characteristics of both categories and uncertainly-classified data part. Moreover, proposed design method behaves well even when normality of data distribution is rejected.

4. ALGORITHMS

4.1 Node division constraints

An important point of the design algorithm of this triplet tree is how to control creation of undetermined nodes. Four constraints are applied in the proposed method, considering proper tree size and informative nodes according to training samples. The effect of constraints (a),(b), and (d) are illustrated in Fig.4.

(a) The number of division of undetermined node containing M categories are limited to $M-1$ times.

(b) At the division of an undetermined node, if one or two determined nodes with single category is created, its child undetermined node becomes terminal node.

(c) An undetermined node with less samples than 1% of the whole training samples become a terminal node.

(d) In case training samples can be partitioned by one boundary, only determined child nodes are created.

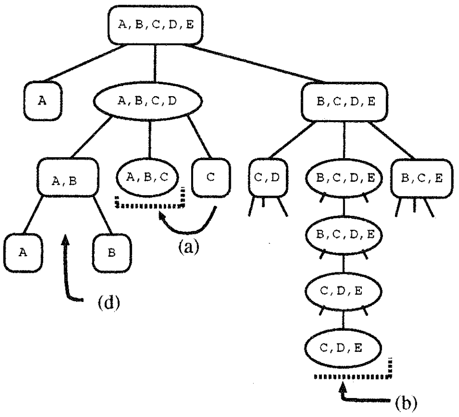

Fig.4 Example of the effect of constraints (a),(b), and (d)

Some undetermined nodes become terminal nodes in this method. These nodes mean indistinct part in the multidimensional data. The characteristic of each ambiguity is shown by the location in the decision tree.

4.2 Design procedure

STEP1: All types of binary division of categories are compared using group distance, and categories for two determined nodes and variables are selected.

STEP2: Two boundaries in selected variables are selected and one node is segmented two determined nodes and one determined node.

1. Error tolerance parameter $p(\%)$ in determined node is choiced at the beginning. When sample numbers for two groups of categories are N_1 and N_2 respectively, error classified samples for two determined nodes are limited to $\lfloor N1*p/100 \rfloor$ and $\lfloor N2*p/100 \rfloor$ respectively.
2. Two determined nodes can be made by one boundary, no undetermined node is created.
3. If two determined node cannot be created, one determined node and one undetermined node are tried to create.
4. If no determined node is created, this node becomes terminated as undetermined node.

STEP3: Node division through STEP1 and STEP2 is repeated while there are nonterminal nodes.

Error tolerance parameter $p(\%)$ ensures reliability about determined nodes with training data. When p is smaller, classification about determined node is better and size of determined node is smaller.

5. EXPERIMENT

Tests for the design of trees and classification of samples by this methods were executed. In the experiment, error tolerance parameter p was selected as 3%, 5%, and 7% making a compromise between classification accuracy and total proportion of determined data. Simulated artificial random data and real Landsat completely-enumerated data were used for the experiment. The relations between nature of categories and tree configuration were checked. Performance of triplet tree classifier was compared with usual binary decision tree and Bayesian classifiers.

5.1 Simulated artificial data

A set of five categories, two-feature data was generated by program. Each category had 100 training samples and 1000 test samples. The mean vectors of the five categories are as follows and illustrated in Fig .5:

$$\mu_i = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 10 \end{pmatrix}, \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

$$\sigma = \begin{pmatrix} 9 \\ 16 \end{pmatrix}$$

$$R = \begin{pmatrix} 1.0 & 0.1 \\ 0.1 & 1.0 \end{pmatrix}$$

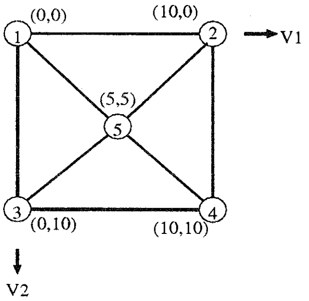

Fig.5 The mean vectors in simulated data

Example of designed triplet trees is shown in Fig.6. A variable was selected in tern at tree node hierarchy based on the design procedure.

Table 1,2,3 show classification results by triplet tree and Table 4 and 5 usual show classification results by binary decision tree (BDT) and two Bayesian method: linear discriminate function(LDF), and quadratic discriminate function(QDF). Results of triplet trees is only for samples in determined nodes, but result of other methods is for all samples. Triplet trees improve performance of BDT. Classification accuracy and classification reliability are found to be equivalent to one step Bayesian methods, by excluding indistinct part of data to undetermined nodes.

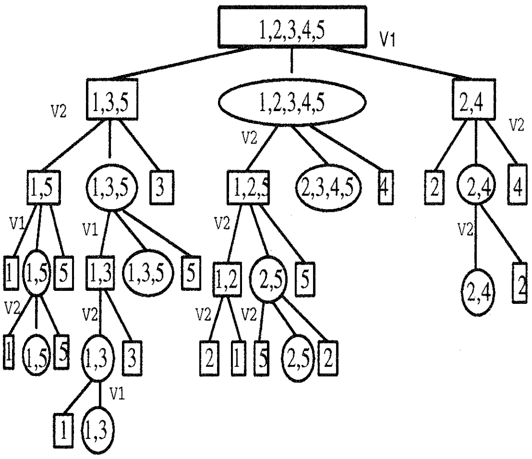

Fig.6 Triplet tree for $p = 5\%$

Table 1. Number of samples in determined node for simulation data

tolerance parameter p	3%	5%	7%
Number of samples	3106	3569	4014
ratio (%)	62.1	71.4	80.3

Table 2. Classification accuracy for determined nodes

p	CAT1	CAT2	CAT3	CAT4	CAT5	Total
3%	87.1	86.5	88.2	94.0	31.9	82.3
5%	89.4	82.1	85.4	82.9	42.6	79.3
7%	83.7	80.5	85.1	77.9	39.6	80.3

Table 3. Classification reliability ($= 100 - commission\_error$) for determined nodes

p	CAT1	CAT2	CAT3	CAT4	CAT5
3%	83.6	89.5	84.5	82.1	51.6
5%	76.5	86.7	81.9	87.5	52.6
7%	75.0	79.2	78.5	86.1	48.8

Table 4. Classification accuracy by BDT, LDF and QDF

Classifier	CAT1	CAT2	CAT3	CAT4	CAT5	Total
BDT	40.1	69.0	19.9	64.3	28.2	44.3
LDF	78.2	79.4	82.3	79.0	51.5	74.1
QDF	78.1	78.9	82.1	79.2	53.1	74.3

Table 5. Classification reliability by BDT, LDF and QDF

Classifier	CAT1	CAT2	CAT3	CAT4	CAT5
BDT	69.1	77.6	13.9	83.4	20.5
LDF	76.0	80.2	78.8	77.8	55.9
QDF	76.4	80.7	79.2	78.1	56.0

5.2 Real Landsat completely-enumerated data

A completely enumerated image(Tanaka,1992) was used in this experiment. Detailed digital land-use data were aggregated to $50m \times 50m$ cell size of seven land cover classes, and the synthesis image was built by matching these classes for geocoded four bands of Landsat MSS pixels in May 1984. Test field has $16km \times 12km$ area and the population is 14768 after excluding mixels. The test field contains seven categories with widely varied frequencies (Table 6): agricultural field(A),barren(B), developed area(D), forest(F), paddy field(P),residential area(R) and water surface(W).

Table 6 Pixels of each category (total 14768)

Category	A	B	D	F	P	R	W
Number	3398	175	1903	273	8084	566	369
Frequency (%)	23	1.2	13	1.8	55	1.2	2.5

Training samples were randomly selected in proportional to the frequency, sampling rates was 5 percent for the whole pixels. The rest 95 percent samples were used for classification test. Namely, number of training samples was 738 and number of classification samples was 14030.

Table 7 Classification accuracy about samples in determined nodes for error tolerance p

p	3%	5%	7%
Size of determined nodes (%)	42.8	62.5	72.4
Total accuracy(%)	91.5	86.7	84.4
Averaged accuracy(%)	42.4	36.7	36.5

The classification result about Landsat MSS'data are shown in Fig.7 and Table 7.

Fig. 7 shows the triplet tree produced by the proposal method. Categories expected have similar features such as {P, W} and {B,D,R} were in the near node and divided repeatedly in the tree. In this case, no determined node for category F was appeared. Characteristics of Forest area in this image is near other category, especially agricultural field, so classification with specified reliability could not be executed.

Table 7 shows the sizes of determined nodes and their classification accuracies. When parameter p becomes larger, size of determined nodes increases but classification accuracies decreases.

Table 8 and Table 9 show classification accuracy. Bayesian method is superior in averaged performance. Triplet trees is superior in total accuracy, by excluding indistinct part to undetermined node. Present design method calculate boundary with sample numbers, so in case categories with similar distribution have large difference in frequency, categories with small frequency is tend to be ignored.

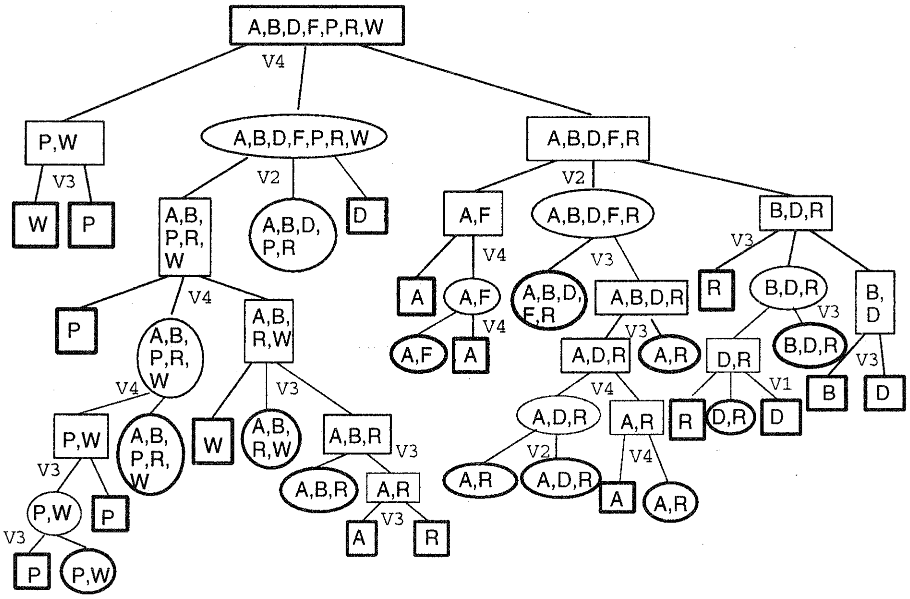

Fig.7 Tree for Landsat MSS data in case error tolerance $p = 5\%$.

Table 8 Classification accuracy of triplet tree ($p = 5\%$) using 95% of samples: size of determined node, accuracy and reliability ($= 100 - commission\_error$) for each category, averaged accuracy and total accuracy

Classifier	A	B	D	F	P	R	W	Averaged	Total
Size of determined nodes	1365	42	411	0	6773	140	41	(Total 62.5%,8772 pixels)	
Classification accuracy	75.9	8.1	53.9	0	98.3	11.8	9.1	36.7	86.7
Classification reliability	72.3	14.3	68.1	NaN	93.0	12.9	43.9	-	-

Table 9 Classification accuracy of Bayesian classifiers using 95% of samples: accuracies for each category, averaged accuracy, and total accuracy

Classifier	A	B	D	F	P	R	W	Averaged	Total
LDF	33.0	30.3	52.1	86.1	75.3	20.5	40.1	48.2	59.2
QDF	29.8	33.7	53.2	68.9	77.5	18.2	52.3	47.7	59.8

5.3 Summary of the results

Tree structures which support the assumption and sufficient classification accuracy were obtained in the tests. Proportion of sizes of determined terminal node and undetermined terminal nodes was dependent on classification accuracy about training samples and specified tolerance parameter p.

Proposed triplet tree classifier was demonstrated that it not only has advantages of general tree classifiers, but also enables to treat uncertainty of data. Samples were effectively classified by the decision tree. Moreover, both relations between categories and the uncertainty were shown in the hierarchical tree structure.

6. CONCLUSION

Undetermined nodes are introduced as an extension of binary decision trees. Data with indistinct feature at binary division are classified to the undetermined node at each node division.

Proposed triplet tree classifiers can be widely adapted even when adjacent pair of category exists or representabilities of training samples is relatively poor.

Proposed triplet tree was shown to enable classification with flexible and effective boundaries. Proposed design method ensures classification reliability in determined nodes about training samples. Classification accuracy for determined nodes is almost the same to the Bayesian classifiers.

It was also demonstrated that the hierarchical structure can represent the relations of categories and uncertainly-classified data part.

References

[1] Wang Ru-Ye,1986a. "An approach to tree-classifier design based on a splitting algorithm," Int. J. Remote Sensing, vol.7,89-104.

[2] S. Tanaka,1992. "A Comparison and Rating of Conditioned Bayesian Discriminant Classifiers by Quantitative Term of Training Representability," J. Remote Sensing Society of Japan, vol.12,3-21(In Japanese).

[3] Wang Ru-Ye,1986b. "An approach to tree-classifier design based on a splitting algorithm," Int. J. Remote Sensing, vol.7,89-104.

[4] B. Kim and D. A. Landgrebe,1991. "Hierarchical Classifier Design in High-Dimensional, Numerous Class Cases," IEEE Trans. on Geoscience and Remote Sensing,vol.29,518-529.

[5] R. Chin, P. Beaudet, and P. Argentiero.,1980. "An Automated Approach to the Design of of Decision Tree Classifiers," Proceedings of the 5th International Conference in Pattern Recognition, 660-665.