한국통계학회논문집 제2권 1호
1995년 4월 pp. 74-88

Tree-structured Classification based on Variable Splitting[1)]

Sung Jin Ahn[2)]

Abstract

This article introduces a unified method of choosing the most explanatory and significant multiway partitions for classification tree design and analysis. The method is derived based on the impurity reduction (IR) measure of divergence, which is proposed to extend the proportional-reduction-in-error (PRE) measure in the decision-theory context. For the method derivation, the IR measure is analyzed to characterize its statistical properties which are used to consistently handle the subjects of feature formation, feature selection, and feature deletion required in the associated classification tree construction. A numerical example is considered to illustrate the proposed approach.

1. Introduction

Classification trees have been successfully applied to such diverse areas as pattern recognition, remote sensing, medical diagnosis and prognosis, taxonomy, and machine learning. Such successful applications are due to the associated conceptual simplicity and classification accuracy. (See, for example, Brieman et al. (1984), Chou (1991), and Zhou and Dillon (1991).)

Classification trees are commonly constructed by recursively partitioning the feature space into mutually exclusive and exhaustive subsets in a systematic way of tree generation. Four essential elements in a classification tree design include 1) a data set consisting of feature vectors and their corresponding class labels, 2) a node-splitting criterion, 3) a rule for determining terminal nodes (leaves), and 4) a class-assigning rule.

The performance of a classification tree is generally evaluated in terms of the expected loss. Due to the inherent computational complexity in constructing trees with the minimum expected loss, heuristics are usually considered by employing steepest-descent greedy procedures. Each step of such a greedy procedure operates on a partially grown tree such that for splitting each node, a collection of splits permitted at that node are all evaluated to make the performance of the resulting tree improved most under the assumption that the branched nodes (children) will be terminal nodes. This implies that the resulting partitioning may become locally (one-step) optimal.

Kass (1980) and Biggs et al. (1991) have proposed algorithms (known as CHAID and KS, respectively) that were offshoots of AID (Automatic Interaction Detection) designed for a

1) This study was supported by Non-directed Research Fund, Korea Research Foundation, 1991.
2) Department of Statistics, Gyeongsang National University, Chinju, 660-701, KOREA

categorized dependent variable by Sonquist et al. (1973). In the algorithms, the best multiway partition has been chosen on the basis of statistical significance calculated by using the Pearson's chi-square statistic and the Bonferroni adjustment factor. Both the algorithms CHAID and KS have been organized to search the best partition (or clustering) in a stepwise (or agglomerative) manner. Although the agglomerative clustering algorithms may yield satisfactory results in practice, they do not guarantee that the optimal clustering, in the sense of maximizing statistical significance, will be found. Further, the chi-square criterion, while it provides a test of significance with regard to independence between variables, does not give a measure of association, which is a critical drawback in the classification (or prediction) purpose.

In this article we propose a unified method for optimal partitioning (or clustering) by use of a new criterion measure, called IR criterion, which practically seeks a partition characterized as being the most explanatory and the most significant.

The organization of this article is briefed as follows. Section 2 discusses the IR measure of divergence in the context of decision theory. Section 3 proposes the tree construction method for the optimal classification and gives numerical test results, and Section 4 gives conclusive remarks.

2. Impurity Reduction Measures of Divergence for Distributions

Consider a situation where a population is classified on the basis of two categorical variables, X and Y. In contingency table notation, let p_{ij} be the probability that an individual of the population is classified into category i of the row variable X and category j of the column variable Y, for $i \in R = \{1, \ldots, r\}$ and $j \in C = \{1, \ldots, c\}$. Then the ith row total p_{i+} is the marginal probability that an individual falls in category i of variable X. Similarly, p_{+j} is the marginal probability that an individual falls in category j of variable Y. Let q_{ij} denote the conditional probability that an individual of the population falls in category j of the column variable Y, given that the individual's X category is i: that is, $q_{ij} = p_{ij} / p_{i+}$.

Suppose that some decision rule is used to classify an individual randomly selected from a given population with respect to variable Y, first using only the population information at hand, and then using the information that the individual's X category is known. Both the associated class label y and its prediction $\hat{y}$ will be expressed in c-dimensional vectors. Let CP denote the set of all probability distributions defined on C, which, of course, contains the set CI of all degenerate probability distributions or elementary vectors e_j where components

of CP and CI will be called class probability vectors and class indicator vectors, respectively.

The loss, or cost, incurred due to representing the class label $y \in CI$ by the prediction $\hat{y} \in CP$ will be measured by the loss function $l(y, \hat{y})$. Typical examples of loss functions are given as

1) the misclassification error loss function

$$l(y, \hat{y}) = 1 - y^T \hat{y}, \tag{2.1}$$

2) the squared error loss function

$$l(y, \hat{y}) = (y - \hat{y})^T (y - \hat{y}), \tag{2.2}$$

3) the log likelihood loss function

$$l(y, \hat{y}) = -y^T \log \hat{y}, \tag{2.3}$$

where $\log \hat{y}$ denotes $(\log \hat{y}_1, \ldots, \log \hat{y}_J)$. (For other types of loss functions, referring to Chou (1991).)

As long as any loss is accepted as being the most suitable measure of prediction, we are entirely entitled to select the particular prediction which gives the minimum amount of the expected loss.

The minimum expected loss in predicting the Y category of a randomly selected individual by using only the population information at hand may be called the impurity of Y and denoted by $\phi(Y)$:

$$\phi(Y) = \min_{\hat{y} \in CP} E[\, l(Y, \hat{y})] \,. \tag{2.4}$$

The prediction $\hat{y}$ that minimizes the expected loss $E[\, l(Y, \hat{y})]$ may be called the centroid of Y, and denoted by $\mu(Y)$:

$$\mu(Y) = \arg\min_{\hat{y} \in CP} E[\, l(Y, \hat{y})] \,. \tag{2.5}$$

Moreover, $\phi(Y)$ and $\mu(Y)$ have the relation

$$\phi(Y) = E[\ l(Y, \mu(Y)]\ . \tag{2.6}$$

The conditional impurity of Y, given $X = i$, denoted by $\phi(Y|i)$, may be defined as the minimum expected loss in predicting the Y category by using the conditional probability distribution $q_i = (q_{i1}, \ldots, q_{ic})$:

$$\phi(Y|i) = \min_{\hat{y}(i) \in CP} E[\ l(Y, \hat{y}(i))|X = i]\ , \tag{2.7}$$

where $\hat{y}(i)$ represents the prediction of Y, given $X = i$.

The conditional centroid of Y, $\mu(Y(i))$, given $X = i$, is the prediction $\hat{y}(i)$ that minimizes the conditional expected loss $E[\ l(Y, \hat{y}(i))|X = i]$:

$$\mu(Y(i)) = \arg\min_{\hat{y}(i) \in CP} E[\ l(Y, \hat{y}(i))|X = i]\ , \tag{2.8}$$

where $Y(i)$ represents the conditional variation of Y, given $X = i$.

The conditional impurity $\phi(Y|i)$, given $X = i$, can be rewritten as

$$\phi(Y|i) = E[\ l(Y, \mu(Y(i)))|X = i]\ . \tag{2.9}$$

Thus the optimal decision rule b^* that minimizes the expected loss is just the decision rule that gives the (conditional) centroid as the prediction.

The impurity of Y conditioned by X, denoted by $\phi(Y|X)$, is defined by

$$\phi(Y|X) = \sum_{i \in R} p_{i+}\, \phi(Y|i). \tag{2.10}$$

Let's notice here that the conditional impurity $\phi(Y|i)$ is obviously a random variable defined in the space R. Its value is completely determined by the knowledge of which event i of the finite space R actually occurs. This implies that the conditional impurity $\phi(Y|X)$ is the mathematical expectation of the random variable.

The Impurity Reduction (IR) is defined as the reduction of the impurity in predicting the Y category obtained when the X category is known, as opposed to the situation when the X category is not known:

$$IR(Y;i) = \phi(Y) - \phi(Y|i)$$

and

$$IR(Y;X) = \phi(Y) - \phi(Y|X). \tag{2.11}$$

The Proportional-reduction-in impurity (PRI) is defined as the relative reduction of the impurity in predicting the Y category obtained when the X category is known, as opposed to the situation when the X category is not known:

$$PRI(Y;X) = \frac{IR(Y;X)}{\phi(Y)}. \tag{2.12}$$

The PRI measure of association may be thought of as a generalization of PRE measures. Each loss function may yield its corresponding PRI measure. The PRI measures will be discussed in another study.

3. Tree Construction based on IR Measure

Based on Jensen's inequality and the Shannon entropy, Lin (1991) proposed an information-theoretic divergence measure between two probability distributions, called Jensen-Shannon Divergence (JSD). One of the salient features of the JSD is that a different weight can be assigned to each probability distribution. Also, the JSD can be easily generalized to measure the overall difference of more than two distributions. These make it particularly suitable for the study of multiclass decision making problems.

Let Y be a categorical random variable with c categories and let $q_1, q_2, \ldots, q_r$ be r probability distributions of Y with weights $w_1, w_2, , , w_r$, respectively. The JSD between the distributions is defined as in Lin (1991)

$$JSD(q_1, q_2, \ldots, q_r; w_1, w_2, \ldots, w_r) = H\left(\sum_{i=1}^{r} w_i q_i\right) - \sum_{i=1}^{r} w_i H(q_i),$$

where H is the Shannon entropy function,

$$H(q_i) = -\sum_{j=1}^{c} q_{ij} \log q_{ij}.$$

In the classification tree construction we are concerned with the problem of optimally

clustering the r probability distributions (or categories or subpopulations) into k probability distributions $(2 \leq k < r)$. Later we will seek an 'optimal' k-clustering by the principle of JSD maximization.

Suppose that the features $X_1, X_2, \ldots, X_m$ are all categorical. Let $\boldsymbol{X}$ denote the m-dimensional categorical feature vector and F the feature space containing all possible values of the feature vector. Let C denote the set of all class labels $\{1, 2, \ldots, c\}$ and CP denote the set of all class probability vectors defined on C, which, of course, contains the set CI of all class indicator vectors defined on C. Let's take both the class label y and its prediction $\hat{y}$ to be c-dimensional vectors. Then, assuming that $(\boldsymbol{X}, Y)$ be jointly distributed random variables with $\boldsymbol{X}$ taking values in F, and Y taking values in CI, the problem is to, based on an observation $\boldsymbol{x}$ of the feature vector $\boldsymbol{X}$, make a prediction $\hat{y}$ of the associated class label y, given a training sample D of feature vectors and their associated class indicator vectors. We assume full multinomial, or independent 'row' mulitinomial sampling in which, for each feature variable, a random sample of size n is taken from the subpopulation in category i of the variable and in which all the other features are random.

From now on, all probabilities may represent sample proportions of training sequence: that is, for a feature variable X, $p_{ij} = n_{ij}/n$, $p_{i+} = n_{i+}/n$, $p_{+j} = n_{+j}/n$, $q_{ij} = n_{ij}/n_{i+}$. And thus impurities, IRs, and PRIs may be interpreted as sample statistics.

In standard statistical diagnostic techniques based on simple data, the type decision statement for a new case (newly appearing individual) is made in composite form. (See Aitchison and Begg (1976).) Accordingly, we do not assume that a class of cases is absolutely characterized by selected combinations of features. Rather, we assume that each combination of features is distributed among all the classes. Thereupon, our approach is to consider classification trees as probabilistic classifiers such that a random object Y is predicted by a decision rule $\hat{Y} = b(X)$.

A classification tree predicts a random object Y by a decision rule $\hat{Y} = b(X)$. Let T denote the set of nodes in the tree or the tree itself, and take each $t \in T$ to be an event in the feature space that an individual belongs to node t. Thus the root node t_o corresponds to the feature space F. Then $P(t)$ is the probability that node t is reached when an individual is classified. Let $\tilde{T}$ denote the set of terminal nodes, or leaves, of T. We can see that $\tilde{T}$ forms a partition of the feature space F. Considering the tree T as a compound feature, we can measure the performance of the tree by proportional reduction in impurity $PRI(Y; T)$, defined by

$$PRI(Y;T) = \frac{IR(Y;T)}{\phi(Y)} = \frac{\phi(Y) - \phi(Y|T)}{\phi(Y)}, \tag{3.1}$$

which is achieved under the optimal decision rule b^* giving the centroid $\mu_t(Y)$ as the output label for each node t, where

$$\phi(Y|T) = \sum_{t \in T} P(t)\ \phi_t(Y), \tag{3.2}$$

in which the subscript t indicates that the measure is associated with the node t.

Now the classification tree design problem can be restated; given the training sample D, the loss function l, and the decision rule b^*, the classification tree is to be designed so as to be most significant and also to be most explanatory, that is, to maximize $PRI(Y;T)$.

We will use an IR measure as the basis for constructing probabilistic classification trees. While any appropriate loss function can be utilized in tree construction, we will apply the log likelihood loss function based on the following statistical properties.

Proposition 3.1. The log likelihood loss function (2.3) yields

$$IR(Y;X) = H(Y) - H(Y|X), \tag{3.3}$$

which is the JSD between the conditional distributions $q_1, q_2, , , q_I$ with weights $p_{1+}, p_{2+}, \ldots, p_{I+}$, where $H(Y)$ is the Shannon entropy and $H(Y|X)$ is the noise or conditional entropy,

$$H(Y) = -\sum_{j=1}^{c} p_{+j} \log p_{+j},$$

$$H(Y|X) = -\sum_{i=1}^{r} p_{i+} H(Y|i)$$

$$= -\sum_{i=1}^{r} p_{i+} \sum_{j=1}^{c} q_{ij} \log q_{ij}.$$

That is, the JSD between the conditional distributions of Y, given $X = i\ (i = 1, 2, \ldots, r)$ with weights their marginal probabilities is the impurity reduction $IR(Y;X)$ under the log

likelihood loss function.

Proposition 3.2. The log likelihood loss function (2.3) yields

$$PRI(Y;X) = \frac{H(Y) - H(Y|X)}{H(Y)}, \tag{3.4}$$

which is the uncertainty coefficient $U_{Y|X}$ (see Särdnal (1974)).

Knowing these relationships, we will denote, for convenience and clarity, $IR(Y;X)$ and $PRI(Y;X)$ under the log likelihood loss function (2.3) by $JSD(Y;X)$ and $U(Y;X)$, respectively. Let IND (X, Y) denote that X and Y are independent. If IND (X, Y), the following lemma is well known from the likelihood-ratio theory.

Lemma 3.1. Under the log likelihood loss function (2.2), the information statistic

$$IS(Y;X) = 2nJSD(Y;X) \tag{3.5}$$

is approximately distributed as a χ^2 variable with $(r-1)(c-1)$ degrees of freedom under the null hypothesis of IND (X, Y).

This lemma, coupled with the above propositions, imply that the IR or JSD criterion provides both a test of significance with regard to independence between variables and a measure of the degree of association between variables, properties desired for a criterion to be used in classification tree construction. Let's denote a k-nary categorization (or k-clustering) of a feature X by $X'(k)$ $(2 \leq k < r)$. Then IND (X, Y) implies IND $(X'(k), Y)$, which leads to the next corollary.

Corollary 3.1. Under the log likelihood loss function (2.2), the information statistic

$$IS(Y;X'(k)) = 2nJSD(Y;X'(k)) \tag{3.6}$$

is approximately distributed as a χ^2 variable with $(k-1)(c-1)$ degrees of freedom under the null hypothesis of IND (X, Y).

Theorem 3.1. The k-clustering of a feature X, $X'(k)$, that maximizes the measure

$JSD(Y;X'(k))$ is the most significant k-clustering under the log likelihood loss function.

Proof. It can be seen from the results of Corollary 3.1 that for a given k, $X^*(k)$ maximizing the JSD measure also maximizes IS statistic. Thus $X^*(k)$ makes the significance probability (p-value) to be a minimum.

This completes the proof.

3.1 Outline of Tree-construction Algorithm

A greedy growing algorithm is proposed here for constructing classification trees, which searches at each node t for an optimal selection of feature variable with an optimal categorization.

(1) (Feature formation) Categorize optimally each in the set A of the available features. For each feature, we first find optimal k-clusterings on the basis of an IR measure, $k = r-1, \ldots, 2$, and then determine the 'best' clustering on the basis of significance test discussed below. The resulting set of optimally-formed features will be denoted by A^*.

(2) (Feature selection) We seek the 'best' feature among A^* for splitting on the basis of PRI measure and the significance test. Although we can find an effective set of features to split, we will restrict ourselves to searching for a single best feature to split.

(3) (Splitting) Split the node in accordance with the best feature selected in step (2).

3.2 Feature Formation

(1) Optimal k-clustering of a feature ($k = r-1, \ldots, 2$).

For each X in A, we seek the optimal k-clustering $X^*(k)$ that maximizes $JSD(Y;X'(k))$ and that, from Theorem 3.1, also minimizes p-value. Now that we decide to solve the clustering problem by an iterative method, we must ask ourselves the question: given any k -clustering $X'(k)$, how can we improve it upon? In the spirit of parallel gradient methods, we can let one iteration step process all the elements and update the composition of all the clusters. Let's denote by G_i the ith cluster (or compound category) of $X'(k)$. Let $p(i')$ denote the probability that an individual falls in cluster G_i of $X'(k)$, and $p(i|i')$ denote the conditional probability that an individual falls in category i of X, given that he is in cluster G_i of $X'(k)$. The contribution of cluster G_i to the total measure $\phi(Y|X'(k))$ can be written as

$$\begin{aligned}\delta(i') &= p(i')E[\ l(Y,\mu(Y(i')))|i'] \\ &= p(i')\sum_{i\in G_{i'}} p(i|i')E[\ l(Y,\mu(Y(i')))|i] \\ &= p(i')\sum_{i\in G_{i'}} p(i|i')d(i,i'),\end{aligned} \tag{3.7}$$

where the distance measure $d(i,i')$ representing the distance of category i to the centroid of cluster $G_{i'}$ is defined by

$$d(i,i') = E[\ l(Y,\mu(Y(i')))|i]\ . \tag{3.8}$$

Thus the total intracluster impurity $\delta(i')$ for cluster $G_{i'}$ turns out to be a minimal distance criterion. This formulation lends itself to iterative improvement: an obvious step is to minimize the distance for each cluster. In fact. an optimal k-clustering is a minimal distance clustering. Without proof we present the following theorem. (See Späth (1985).)

Theorem 3.2. A necessary condition for a k-clustering $X'(k)$ to maximize the $IR(Y;X'(k))$ (or to minimize $\phi(Y|X'(k))$) is that each category satisfy the minimal distance (or nearest neighbor) condition:

$$i \in G_{i'} \text{ only if } i' = \arg\min_{l} d(i,l'). \tag{3.9}$$

For the log likelihood loss function (2.3),

$$d(i,i') = -[\ \mu(Y(i))]^{T}\log\mu(Y(i')), \tag{3.10}$$

which is the cross entropy $H(\mu(Y(i));\mu(Y(i')))$, measuring the goodness of fit of $\mu(Y(i'))$ to $\mu(Y(i))$. (See Bozdogan (1987).)

The result of Theorem 3.2 can help to reduce greatly the number of clusterings to be considered in feature formation. Because the cluster centroids will vary as categories are redistributed among the clusters, we can let the process take place in the following two steps: Given a k-clustering $X'(k)$ with categories $\{G_{1'}, G_{2'}, \ldots, G_{k'}\}$;

(Representation step) Compute the centroid of each current cluster, $\mu(Y(i'))$.

(Assignment step) Assign the category i to the cluster $G_{i'}$ by the minimal distance or nearest neighbor condition:

$$i \in G_{i'} \text{ only if } i' = \arg\min_{l} d(i, l').$$

Iterate the above two steps to reduce the conditional impurity $\phi(Y|X'(k))$, until no further reduction occurs. In this procedure, the final clustering corresponds to a locally optimal clustering.

This procedure is actually a k-means clustering algorithm. The computational complexity at each iteration of the clustering algorithm is $O(rkc)$, where r is the total number of categories of the feature X, k is the target number of clusters, and c is the total number of classes of Y. The number of iterations is bounded above by the total number of k-clusterings $O(k^r)$ (referring to Feller (1968)), since each clustering cannot be tried twice, except in the last iteration. The average number of iterations is not known, but according to the empirical study of, for example, Jain (1988), the number of iterations in k-means clustering algorithms is small (less than 20), and does not seem to depend heavily on r, k, or, c. In contrast, the complexity of an exhaustive search is $O(ck^r)$. Thus our k-means clustering algorithm can find an optimal clustering practically in linear, rather than exponential or polynomial, time.

(2) Best clustering of a feature

We seek the most significant clustering X^* of each feature X among the optimal k-clusterings, $X^*(k)$, $k = 2, \ldots, r$. It is possible to calculate the significance probabilities (p-values) of each optimal k-clustering $X^*(k)$ under the null hypothesis of IND (X, Y) from the information statistics. Denoting the p-value of the clustering $X^*(k)$ by $p^*(k)$, the most significant clustering X^* is the clustering corresponding to the minimum of $p^*(k)$, $k = 2, \ldots, r$. This leads to the optimally formed features A^*.

Although the Bonferroni adjustment approaches or the multiple comparison approaches have been proposed by Biggs et al. (1991) and Kass (1980), they are appropriate to use only in a situation where all involved tests are performed independently. Therefore, it seems that employing the approaches in the repeated testing environment for the search procedures does not make any precise probabilistic interpretation possible. Thus, without being too concerned with the actual probability levels, we simply regard the search procedures as making a series of internal comparisons that will produce what appears to be the most useful or unuseful (set of) feature(s).

(3) Initial k-clustering

It is necessary to choose an initial clustering to start the algorithm. A promising initial clustering can be found by a k-means clustering method itself. Given the optimal $(k+1)$ -clustering, assign its $(k+1)$st cluster into one of the other clusters by the nearest-neighbor condition (3.10). Choose the clustering among the resulting (k+1) k -clusterings that has a minimum impurity. This way our initial clustering may be chosen.

3.4 A Numerical Example

This approach is illustrated through an artificial example of recognizing LCD digits, suggested by Brieman et al. (1984). A digit in a caculator display consists of seven lines, each of which may be on or off. Thus there are ten classes (one for each digit) and seven binary valued attributes (one for each line). See Figure 3.1 and Table 3.1.

Residual variation is introduced by assuming that a malfunction leads to a 10% chance of a line being incorrect. The JSD criterion is designed for constructing multiway classification trees, so we use the all possible Boolean combinations of attributes as features. Three hundred samples from this distribution constitute the training sample. Here a simplified notation was used: for example, "X_2X_5" denotes the feature generated from the Boolean combination of attributes X_2 and X_5, and "01", say, means "$X_2 = 0$ and $X_5 = 1$".

Table 3.1 Digit Patterns

Digit	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	0	0	1	0	0	1	0
2	1	0	1	1	1	0	1
3	1	0	1	1	0	1	1
4	0	1	1	1	0	1	0
5	1	1	0	1	0	1	1
6	1	1	0	1	1	1	1
7	1	0	1	0	0	1	0
8	1	1	1	1	1	1	1
9	1	1	1	1	0	1	1
0	1	1	1	0	1	1	1

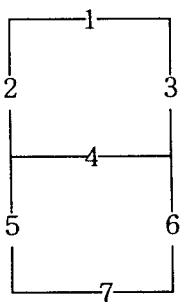

Figure 3.1 Digit Display

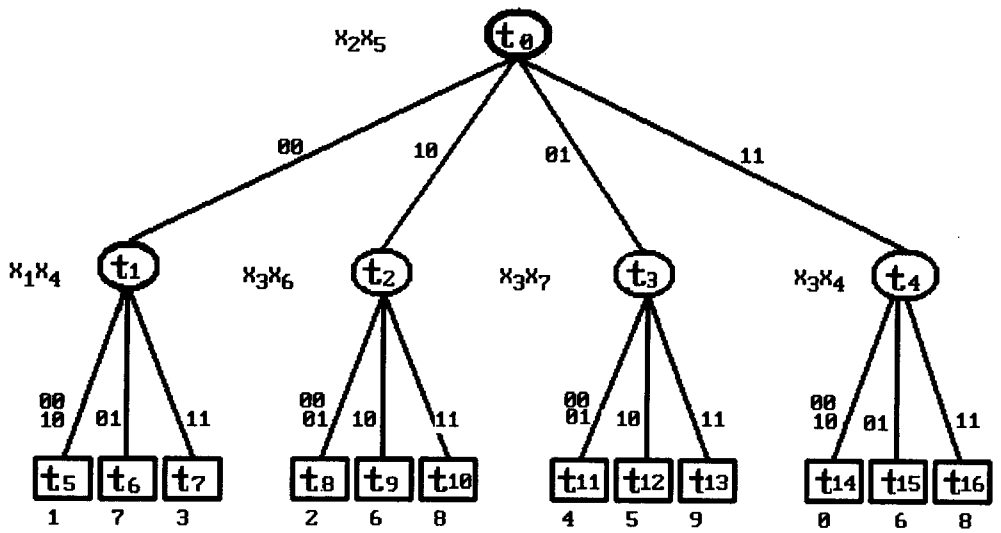

Figure 3.2 Classification Tree Generated from LCD Digit Data

Table 3.2 Impurities and JSDs of each node in the tree

node	size	$\phi_t(Y)$	$\phi_t(Y;X)$	$JSD_t(Y;X)$
t_0	300	2.3025		
t_1	88	1.7620		
t_2	50	1.3501	1.5389	0.7636
t_3	92	1.5907		
t_4	70	1.3253		
t_5	29	0.4770		
t_6	27	0.4192	0.5887	1.1733
t_7	32	0.8332		
t_8	26	0.3224		
t_9	5	0.0	0.7599	0.5902
t_{10}	19	1.5586		
t_{11}	44	1.3371		
t_{12}	25	0.9256	1.1146	0.4761
t_{13}	23	0.8946		
t_{14}	29	0.9215		
t_{15}	18	0.2158	0.8205	0.5048
t_{16}	23	1.1664		

In the Fig. 3.2, the digit attached to each terminal node represents the class label assignable

when the classification tree is used as a deterministic classifier. We set the significance level at 0.1%. From the Table 3.2, the impurity of Y given the tree T is calculated as $\phi(Y|T) = \sum_{t \in T} P(t)\ \phi_t(Y) = 0.8325$. Thus $JSD(Y; T) = 1.47$ and $PRI(Y; T) = 0.6384$. That is, the tree classifier is reducing the impurity of Y as much as 63.84%. The test sample estimate made from the independent sample of size 3000 is obtained as $PRI^{ts}(Y; T) = 0.5927$. Though validation, or the process of verifying the classification effectiveness on independent data is an important aspect of tree design, it is not our primary concern here.

4. Conclusive Remarks

The classification tree technique is a useful tool to analyze and construct a decision-making structure derived from large multivariate data. Based upon this property, we presented a unified method of feature formation and feature selection for constructing multiway classification trees. In exploiting the method, as a measure of divergence between categorical distributions, the IR measure was proposed, and as a measure of association (dependence) between categorical variables, the PRI measure was proposed as an extension of the PRE measure in the decision-theory context. The extension of the associated discussion to multivariate and conditional PRI measure was also made. These properties were all used to develop the tree construction algorithm based on IR criterion, which was also found suitable to extend the classification tree to the probabilistic classifier. For feature formation, a k-means clustering algorithm was presented for finding a locally optimal k-clustering, and then the p-value approach was proposed for finding the best clustering among all the optimal (r-1) k-clusterings of the feature. For feature selection, both the unconditional and conditional PRI measures or their corresponding p-values were considered. The classification algorithm was applied to the problem of recognizing LCD digits, and summarized test results.

References

[1] Aitchison, J. and Begg, C. B. (1976). Statistical diagnosis when basic cases are not classified with certainty, *Biometrika,* vol. 63, no. 1, pp. 1-12.

[2] Biggs, David, Ville, Barry de, and Suen, Ed (1991). A method of choosing multiway partitions for classification and decision trees, *J. of Appl. Statist.,* vol. 18, pp. 49-62.

[3] Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions, *Psychometrika,* vol. 52, no. 3, pp. 345-370.

[4] Brieman, L., Friedman, J. H., Olsen, R. A., and Stone, C. G. (1984) *Classification and regression trees,* Belmont, CA: Wadsworth.

[5] Chou, P. A. (1991). Optimal partitioning for classification and regression trees, *IEEE Trans. on Pattern Anal. and Machine Intell.,* vol. 13, pp. 340-354.

[6] Feller, W. (1968). *An Introduction to Probability Theory and its Applications,* New York, NY: John Wiley and Sons, vol. I, p. 101, ff.

[7] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data, Englewood Cliffs,* NJ: Prentice-Hall, p. 99, ff.

[8] Kass, G.A. (1980). An exploratory technique for investigating large quantities of categorical data, *Appl. Statist.,* vol. 29, no. 2, pp. 119-127, 1980.

[9] Lin, J. (1991). Divergence measures based on the Shannon entropy, *IEEE Trans. on Info. Theory,* vol. 37, no. 1, pp. 145-151.

[10] Sonquist, J. A., Baker, E., and Morgan, J. (1973). *Searching for Structure,* Ann Arbor, MI: University of Michigan.

[11] Späth, H. (1985). *Cluster Dissection and Analysis,* Chichester, England: Ellis Horwood.

[12] Särndal, C. E. (1974). A comparative study of association measures, *Psychometrika,* vol. 39, pp. 165-187.

[13] Zhou, X. J. and Dillon, T. S. (1991). A statistical-heuristic feature selection criterion for decision tree induction, *IEEE Trans. on Pattern Anal. and Mach. Intell.,* vol. 13, pp 834-841.