Statistics and Probability Letters, 2001, Vol 54, pp. 341-345

Selecting the best categorical split for classification trees

Yu-Shan Shih
Department of Mathematics
National Chung Cheng University
Minghsiung, Chiayi 62117, Taiwan

Abstract

Based on a family of splitting criteria for classification trees, methods of selecting the best categorical splits are studied. They are shown to be very useful in reducing the computational complexity of the exhaustive search method.

Keyword: classification tree; power divergence; splitting criteria

1 Introduction

When constructing a binary classification tree, a naive approach to choose the split based on a single variable is to search through all the possible points. A criterion is then used to select the best one. For example, the Gini criterion is used in Breiman, Friedman, Olshen and Stone (1984) and the entropy criterion is used in Ciampi, Chang, Hogg and McKinney (1987), Clark and Pregibon (1992) and Quinlan (1993). Taylor and Silverman (1993) suggest the mean posterior improvement criterion. Shih (1999) proposes a weighted sum method which creates a family of splitting criteria.

If the variable is numerical with M distinct values, the procedure has to check $M-1$ possible splits. If the variable is categorical with M elements, then the set of all the possible splits is of size $2^{M-1}-1$. When M becomes larger, the procedure will spend more time on finding the best split, especially for categorical variables. For two-class problem, the search for the best categorical split can be reduced to $M-1$ steps using the Gini criterion (Breiman et al., 1984). For three or more classes, Mola and Siciliano (1997; 1999) provide algorithms that could reduce the computational

complexity of searching through several categorical variables, when the Gini or entropy criterion is used.

In this paper, similar results are obtained for the family of splitting criteria proposed in Shih (1999). The family is introduced in Section 2. It is shown in Section 3 that the search can also be reduced to $M-1$ steps, if any member of the family is used for two-class problem. A divergence index is defined in Section 4 and its property is studied. In Section 5, an algorithm based on the index is given and an example is shown to demonstrate the computational efficiency of the algorithm. Conclusions are given in Section 6.

2 Family of splitting criteria

Suppose there are J classes in the current node. For every binary split, denote L and R to be its two subnodes. Let π_L and π_R be the proportions that are placed into L and R, respectively. The relative proportion of class j in the current node is defined as p_j while that in the subnode k is defined as $p_{jk}, k \in \{L, R\}$. We use $\mathbf{p} = (p_1, \ldots, p_J)$ as the proportion vector in the current node and $\mathbf{p}_k = (p_{1k}, \ldots, p_{Jk}), k \in \{L, R\}$ as the proportion vector in the subnodes.

Definition 1 (Read and Cressie (1988)) Let $\mathbf{u}$ and $\mathbf{v}$ be two discrete probability distributions defined on the $(J-1)$-dimensional simplex: $\{\boldsymbol{\pi} | \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J)$ with $\pi_j \geq 0$ and $\sum_j \pi_j = 1,$ where $1 \leq j \leq J\}$. The power divergence for $\mathbf{u}$ and $\mathbf{v}$ is

$$I^{\lambda}(\mathbf{u} : \mathbf{v}) = \{\lambda(\lambda+1)\}^{-1} \sum_{j=1}^{J} u_j \{(u_j/v_j)^{\lambda} - 1\}; \quad -1 < \lambda < \infty,$$

where the value at $\lambda = 0$ is taken to be the continuous limit as $\lambda \to 0$. Thus, $I^0(\mathbf{u} : \mathbf{v}) = \sum_{j=1}^{J} u_j \log(u_j/v_j)$. The value $u_j\{(u_j/v_j)^{\lambda} - 1\}/\lambda = 0$, if $u_j = v_j = 0$.

Based on the power divergence family, a family of splitting criteria defined in Shih (1999) is

$$C(\lambda) \equiv \pi_L I^{\lambda}(\mathbf{p}_L : \mathbf{p}) + \pi_R I^{\lambda}(\mathbf{p}_R : \mathbf{p}), \quad -1 < \lambda < \infty.$$

For a given λ value, the best split based on $C(\lambda)$ is the one that maximizes $C(\lambda)$. Shih (1999) shows that the chi-squared criterion ($\lambda = 1$) and the entropy criterion ($\lambda = 0$) belong to this family. It also contains the Freeman-Tukey ($\lambda = -1/2$) and the Crssie-Read ($\lambda = 2/3$) criteria.

3 Two-class problem

Breiman et al. (1984, Theorem 4.5) prove the following theorem showing that the search can be reduced to $M-1$ subsets, if there are only two classes.

Theorem 1 Suppose there are two classes, class 1 and class 2. Let X be a categorical variable taking values on $\{1, 2, \ldots, M\}$ where the categories are in increasing $p(1|X = i)$ values. If ϕ is a concave function, then one of the $M-1$ splits, $X \in \{1, \ldots, m\}$ where $1 \leq m < M$, minimizes $p_L\phi(p_{1L}) + p_R\phi(p_{1R})$.

Given λ, the best split based on $C(\lambda)$ is the one that maximizes $p_L I^\lambda(\mathbf{p}_L : \mathbf{p}) + p_R I^\lambda(\mathbf{p}_R : \mathbf{p})$ which is equivalent to the split that minimizes $-p_L I^\lambda(\mathbf{p}_L : \mathbf{p}) - p_R I^\lambda(\mathbf{p}_R : \mathbf{p})$. For fixed $\mathbf{p} = (p_1, 1-p_1)$ with $p_1 \neq 0$, $I^\lambda(\mathbf{p}_L : \mathbf{p})$ is a convex function in p_{1L}. Thus, $-I^\lambda(\mathbf{p}_L : \mathbf{p})$ is a concave function. Therefore, this computational advantage holds for the splitting criteria based on $C(\lambda)$, where λ is given.

For more than two classes, there is no simple extension of Theorem 1. Chou (1991) provides a partial extension of Theorem 1 to three or more classes. However, it is only locally optimal (Ripley, 1996, p. 238).

4 Divergence index and its property

Let X be a categorical random variable which takes values on $\{1, 2, \ldots, M\}$ and Y be the class variable which takes values on $\{1, 2, \ldots, J\}$. The proportion vector of Y given that $X = i$ is denoted by $\mathbf{p}_{Y|i}$ for $i = 1, 2, \ldots, M$. We define the divergence index based on Y and X as follows.

Definition 2 Given λ, the divergence index of Y given X is

$$\Delta(Y|X) = \sum_{i=1}^{m} P(X = i) I^\lambda(\mathbf{p}_{Y|i} : \mathbf{p}).$$

A nice property of the divergence index is shown in Theorem 2. It claims that any split based on X using $C(\lambda)$ has value at most equal to the divergence index $\Delta(Y|X)$.

Theorem 2 Given a nonempty subset $A \subset \{1, 2, \ldots, M\}$, split s based on variable X is created such that a case with $X \in A$ goes to the left node,

otherwise, it goes to the right node. Let C_s be the goodness of split value based on $C(\lambda)$, where λ is known. We have

$$\Delta(Y|X) \geq C_s.$$

Proof. It is observed that

$$\sum_{i\in A} P(X=i) = \pi_L, \quad \sum_{i\in A} P(X=i)\mathbf{p}_{Y|i} = \mathbf{p}_L\pi_L$$
$$\sum_{i\notin A} P(X=i) = \pi_R, \quad \sum_{i\notin A} P(X=i)\mathbf{p}_{Y|i} = \mathbf{p}_R\pi_R.$$

By the fact that $f(x) = (x^{\lambda+1} - x)/\lambda(\lambda+1)$ is a convex function and Jensen's inequality. We have

$$\begin{aligned}
\Delta(Y|X) &= \sum_{i\in A} P(X=i)I^\lambda(\mathbf{p}_{Y|i} : \mathbf{p}) + \sum_{i\notin A} P(X=i)I^\lambda(\mathbf{p}_{Y|i} : \mathbf{p}) \\
&= \pi_L \sum_{i\in A} P(X=i)/\pi_L I^\lambda(\mathbf{p}_{Y|i} : \mathbf{p}) + \pi_R \sum_{i\notin A} P(X=i)/\pi_R I^\lambda(\mathbf{p}_{Y|i} : \mathbf{p}) \\
&\geq \pi_L I^\lambda(\mathbf{p}_L : \mathbf{p}) + \pi_R I^\lambda(\mathbf{p}_R : \mathbf{p}) \\
&= C_s.
\end{aligned}$$

Suppose S is the set of all the possible splits based on X, we conclude that

$$\Delta(Y|X) \geq \max_{s\in S} C_s.$$

Thus, the best split based on X has value less than or equal to $\Delta(Y|X)$.

5 Algorithm and example

Based on Theorem 2, the fast algorithm of Mola and Siciliano (1997) can be extended to the family of divergence measures. Let $X_1, X_2, \ldots, X_n, n \geq 2$ be categorical variables . Based on the family of splitting criteria $C(\lambda)$, the best split s^* can be selected by the following steps.

1. Compute the divergence index for each variable and order the variables based on their index values. Denote the ordered variables by $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ such that $\Delta(Y|X_{(1)}) \geq \Delta(Y|X_{(2)}) \geq \ldots \geq \Delta(Y|X_{(n)})$.
2. Initialize $i = 1$.

3. Find all the possible splits based on $X_{(i)}$ and let the best split be $s_{(i)}$ with value $C_{s_{(i)}}$.
4. If $i = 1$, set $s^* = s_{(1)}$.
5. If $i > 1$ and $C_{s_{(i)}} > C_{s^*}$, set $s^* = s_{(i)}$.
6. If $i = n$, exit, otherwise, continue the procedure.
7. If $C_{s_{(i)}} \geq \Delta(Y|X_{(i+1)})$, set $s^* = s_{(i)}$ and exit, otherwise let $i = i + 1$ and go back to step 3.

The algorithm could reduce the number of splits needed to be checked at each node. We study the effect of the algorithm using the following example. A random sample of 300 cases is generated. The class variable Y is equally distributed on $\{1, 2, 3\}$. The categorical predictors X_1, X_2 and X_3 which take values on $\{1, 2, \ldots, M\}$ are created such that when $Y = 1$, $X_1 = 1$ and when $Y = 2$, $X_1 = 2$. When $Y = 3$, X_1 is equally distributed on $\{3, \ldots, M\}$. X_2 and X_3 are equally distributed on $\{1, 2, \ldots, M\}$ and are independent of Y.

Based on this generating scheme, we know that the best split at the root node is $X_1 \in \{1, 2\}$. The following table shows the results with $M = 3$, 6 and 9 based on the chi-squared criterion: $C(1)$. The proposed method chooses the same best split in all three cases. The CPU seconds for the proposed algorithm are 1.2, 17.8 and 804.1 respectively. The computational speed of the algorithm relative to the totally exhaustive search algorithm is also reported. All the results are obtained by using S-PLUS on a Sun Ultra 2 workstation.

M	$\Delta(Y\|X_i)$	C_{s^*}	Speed
3	1,.004,.007	.500	2.7
6	1,.020,.011	.169	3.0
9	1,.016,.032	.159	3.6

It is found that the C_{s^*} value is greater than $\Delta(Y|X_{(2)})$ in all cases. As a result, the proposed algorithm only check about 1/3 of all the possible splits at the root node to find the best split. Thus, it reduces the computational time. The saving will be more significant as the number of categorical predictors increasing in these cases.

6 Conclusions

In this paper, the best categorical splits for binary classification trees based on a family of splitting criteria are studied. We find that, for two-class problem, the theorem of Breiman et al. (1984) can be applied to the family. For three or more classes with many categorical variables, the fast algorithm of Mola and Siciliano (1997) can also be extended via the divergence index between the class variable and the categorical predictor. In the tree growing process, the maximal tree is usually built and then pruned upward through cross validation (Breiman et al., 1984; Ripley, 1996). This gain over the usual exhaustive search method becomes even more profounding in the process when many categorical predictors are present.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*, Wadsworth International Group.

Chou, P. A. (1991). Optimal partitioning for classification trees, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**: 340–354.

Ciampi, A., Chang, C.-H., Hogg, S. and McKinney, S. (1987). Recursive partitioning: a versatile method for exploratory data analysis in biostatistics, *in* I. McNeil and G. Umphrey (eds), *Biostatistics*, D. Reidel, New York, pp. 23–50.

Clark, L. A. and Pregibon, D. (1992). Tree-based models, *in* J. M. Chambers and T. J. Hastie (eds), *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, CA.

Mola, F. and Siciliano, R. (1997). A fast splitting procedure for classification trees, *Statistics and Computing* **7**: 209–216.

Mola, F. and Siciliano, R. (1999). A general splitting criterion for classification trees, *Metron* **57**: 155–171.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Shih, Y.-S. (1999). Families of splitting criteria for classification trees, *Statistics and Computing* **9**: 309–315.

Taylor, P. C. and Silverman, B. W. (1993). Block diagrams and splitting criteria for classification trees, *Statistics and Computing* **3**: 147–161.