

# Families of splitting criteria for classification trees

Yu-Shan Shih

*Department of Mathematics  
National Chung Cheng University  
Minghsiung, Chiayi 62117, Taiwan*

## Abstract

Several splitting criteria for binary classification trees are shown to be written as weighted sums of two values of divergence measures. This weighted sum approach is then used to form two families of splitting criteria. One of them contains the chi-squared and entropy criterion, the other contains the mean posterior improvement criterion. Both family members are shown to have the property of exclusive preference. Furthermore, the optimal splits based on the proposed families are studied. We find that the best splits depend on the parameters in the families. The results reveal interesting differences among various criteria. Examples are given to demonstrate the usefulness of both families.

*Keywords:* Classification tree, divergence measure, exclusive preference, splitting criteria

## 1 Introduction

Various criteria have been proposed for split selection of growing classification trees. Kass (1980) uses a testing procedure based on Pearson's chi-squared statistic to choose the best multi-way splits. Breiman, Friedman, Olshen and Stone (1984) introduced CART which provides the Gini and twoing criterion to choose. Likelihood is used to form a criterion in Ciampi, Chang, Hogg and McKinney (1987), Clark and Pregibon (1992), and Quinlan (1993). Taylor and Silverman (1993) proposed the mean posterior improvement (MPI) criterion as an alternative to the Gini rule. Loh and Vanichsetakul (1988) and Loh and Shih (1997) employ statistical tests to select splits. Some splitting rules are compared in Fayyad and Irani (1992), Buntine and Niblett (1992), and Loh and Shih (1997).

A weighted sum approach is considered in this paper. It is shown in Section 2 that several aforementioned criteria can be written as weighted sums of two divergence values. As results, this unified approach can be used to form families of splitting criteria for classification trees. Some properties of these families are studied in Section 3 and 4. One simulation and two real data sets are given as examples in Section 5 to demonstrate the advantage of the proposed families. Conclusions are given in Section 6.

## 2 Weighted sum of two divergence values

In this paper, only binary splits are considered. Suppose there are totally  $N$  objects in a node and there are  $1, 2, \dots, J$  classes. Let  $N_j$  be the number of class  $j$  objects. For every binary split, it creates two subnodes  $L$  and  $R$  with the numbers of objects  $N_L$  and  $N_R$ , respectively. Let  $\pi_L$  and  $\pi_R$  be the proportion that are placed into  $L$  and  $R$ , respectively. Denote  $N_{jk}$

be the number of class  $j$  objects that are in node  $k \in \{L, R\}$ . The relative proportion of class  $j$  in the root node is defined as  $p_j$  while that in the subnode  $k$  is defined as  $p_{jk}$ . We use  $\mathbf{p} = (p_1, \dots, p_J)$  as the proportion vectors in the root node and  $\mathbf{p}_k = (p_{1k}, \dots, p_{Jk}), k \in \{L, R\}$  as the proportion vector in the subnodes.

After the set of candidate splits is decided, a goodness of split measure is performed to select the best split. Note that the more heterogeneous a split makes two candidate children nodes, the more it should be rewarded. Moreover, the compositions of the two children nodes can be treated as a  $J \times 2$  contingency table. Thus, any statistic for testing homogeneity of a  $J \times 2$  contingency table will be a suitable choice as measure. To incorporate priors, a generalized version of Pearson's chi-squared statistic is discussed in the following.

First, let us assume the priors are estimated. The expected value of  $N_{jk}$  is given by  $m_{jk} = N_j N_k / N$  for  $k \in \{L, R\}$ . We adopt the convention  $(N_{jk} - m_{jk})^2 / m_{jk} = 0$ , when  $N_j = 0$ . Then Pearson's chi-squared statistic can be rewritten as

$$X^2/N = \pi_L \sum_{j=1}^J p_{jL}(p_{jL}/p_j - 1) + \pi_R \sum_{j=1}^J p_{jR}(p_{jR}/p_j - 1),$$

where  $\pi_L = N_L/N, \pi_R = N_R/N, p_j = N_j/N$ , and  $p_{jk} = N_{jk}/N_k, k \in \{L, R\}$ .

Let  $\mathbf{u} = (u_1, u_2, \dots, u_J)$  and  $\mathbf{v} = (v_1, v_2, \dots, v_J)$  be two discrete probability distributions. Define the divergence measure for  $\mathbf{u}$  and  $\mathbf{v}$  to be

$$d(\mathbf{u} : \mathbf{v}) = \sum_{j=1}^J u_j(u_j/v_j - 1).$$

We adopt the convention  $u_j(u_j/v_j - 1) = 0$  when  $u_j = v_j = 0$ . The generalized version of Pearson's chi-squared statistic is defined as

$$X^2 = N\{\pi_L d(\mathbf{p}_L : \mathbf{p}) + \pi_R d(\mathbf{p}_R : \mathbf{p})\}.$$

That is,  $X^2$  is proportional to a weighted sum of two values of the divergence measure. Similarly, the generalized version of the likelihood ratio statistic  $G^2$  is defined as

$$G^2 = N\{\pi_L d(\mathbf{p}_L : \mathbf{p}) + \pi_R d(\mathbf{p}_R : \mathbf{p})\},$$

with the divergence measure

$$d(\mathbf{u} : \mathbf{v}) = 2 \sum_{j=1}^J u_j \log(u_j/v_j),$$

where  $u_j \log(u_j/v_j)$  is defined to be 0 when  $u_j = v_j = 0$ . This statistic is also known as the entropy or deviance criterion (Clark and Pregibon, 1992).

In CART, the Gini index is

$$i(\mathbf{p}) = \sum_{j \neq l} p_j p_l = 1 - \sum_{j=1}^J p_j^2.$$

Thus the Gini criterion function

$$i(\mathbf{p}) - \pi_L i(\mathbf{p}_L) - \pi_R i(\mathbf{p}_R)$$

can be written as

$$\pi_L \sum_{j=1}^J p_{jL}^2 + \pi_R \sum_{j=1}^J p_{jR}^2 - \sum_{j=1}^J p_j^2$$

which can also be written in terms of a weighted sum of two values as

$$\pi_L d(\mathbf{p}_L : \mathbf{p}) + \pi_R d(\mathbf{p}_R : \mathbf{p}),$$

where  $d(\mathbf{u} : \mathbf{v}) = \sum_j (u_j^2 - v_j^2)$ .

### 3 Families of splitting criteria

Basically, any divergence measure can be used to form a proper measure of goodness of split. The theoretical properties of various divergence measures of two discrete probability distributions are discussed in depth in Read and Cressie (1988, Section 7.4). In this section, two families of splitting criteria are constructed.

**Definition 1** Let  $\mathbf{u}$  and  $\mathbf{v}$  be two discrete probability distributions defined on the  $(J - 1)$ -dimensional simplex  $\Delta_J = \{\boldsymbol{\pi} | \boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_J)$  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

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

where the value at  $\lambda = 0$  is taken to be the continuous limit as  $\lambda \rightarrow 0$ . 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  $u_j = v_j = 0$ .

The original power-divergence family is defined on the real line (Read and Cressie, 1988). We only consider  $\lambda > -1$ , because the value of power divergence is infinity when  $\lambda \leq -1$  with  $u_i = 0$  and  $v_i \neq 0$  for some  $i$ . We have shown that  $X^2$  ( $\lambda = 1$ ) and  $G^2$  ( $\lambda = 0$ ) belong to this family. Various other goodness-of-fit statistics are also in this family. For example, Freeman-Tukey statistic  $F^2$  ( $\lambda = -1/2$ ) and Cressie-Read statistic ( $\lambda = 2/3$ ) (see Read and Cressie (1988) and references therein).

A split shall be rewarded more if it leads to two children nodes that are mutually exclusive. Based on this idea, Taylor and Silverman (1993) give the following definition of the *exclusivity preference* property.

**Definition 2** A splitting criterion has the exclusivity preference property, if the following two conditions are satisfied.

1. Given  $\pi_L \pi_R$ , it takes its maximum value when  $\sum_j p_{jL} p_{jR} = 0$ .
2. It takes its minimum value when  $p_{jL} = p_{jR} = p_j, \forall j$ .

Based on the power-divergence family, we can define a family of splitting criteria via weighted sums. This family is defined as

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

**Theorem 1**  $C(\lambda)$  has the exclusivity preference property.

**Proof.** Rewrite  $C(\lambda)$  as

$$C(\lambda) = \frac{1}{\lambda(\lambda+1)} \sum_{j=1}^J \{ \pi_L p_{jL} [(p_{jL}/p_j)^\lambda - 1] + \pi_R p_{jR} [(p_{jR}/p_j)^\lambda - 1] \}.$$

Without lost of generality, we can just consider the first term in the summation. For fixed  $\pi_L = 1 - \pi_R \neq 0$ , let  $x = p_{1L}, y = p_{1R}, \alpha = \pi_L$ , and  $\beta = \pi_R$ . Thus  $p_1 = \alpha x + \beta y$  which is a nonzero constant. Therefore, the first term in the summation becomes

$$\frac{1}{\lambda(\lambda+1)} \{ \alpha x [(x/p_1)^\lambda - 1] + \beta y [(y/p_1)^\lambda - 1] \}.$$

Since  $p_1$  is a constant, we could just consider the following function.

$$f(x) = \begin{cases} \frac{1}{\lambda(\lambda+1)} \{ \alpha x^{(\lambda+1)} + (1/\beta)^\lambda (p_1 - \alpha x)^{(\lambda+1)} \} & \text{if } \lambda \neq 0 \\ \alpha x \log(x/p_1) + \beta (p_1 - \alpha x) \log\{(p_1 - \alpha x)/(p_1 \beta)\} & \text{if } \lambda = 0 \end{cases}$$

It is straightforward to see that  $f(x)$  has a minimum at  $x = p_1$  and a maximum at  $x = 0$  or  $x = p_1/\pi_L$  which is equivalent to  $y = 0$ .

The second condition of Definition 2 is also proved to be satisfied for the Gini criterion (Breiman et al., 1984, Proposition 4.4). However, an example in Section 6 shows that the Gini criterion does not satisfy the first condition of Definition 2.

The MPI criterion also has the exclusivity preference property (Taylor and Silverman, 1993). We now show that it is actually a special member of a family of splitting criteria, too. The MPI splitting criterion is defined as

$$M = \pi_L \pi_R \left( 1 - \sum_{j=1}^J p_{Lj} p_{Rj} / p_j \right).$$

Substituting  $(p_j - \pi_L p_{Lj})/\pi_R$  for  $p_{Rj}$ , the summation term becomes

$$1 - \pi_L / \pi_R \sum_{j=1}^J \{ p_{Lj} (p_{Lj} / p_j - 1) \}.$$

Therefore, the MPI criterion is  $M = 2\pi_L^2 I^1(\mathbf{p}_L : \mathbf{p})$ . By replacing  $p_{Lj} = (p_j - \pi_R p_{Rj})/\pi_L$  instead, we also obtain that  $M = 2\pi_R^2 I^1(\mathbf{p}_R : \mathbf{p})$ . Therefore, the MPI criterion can be rewritten as a weighted sum of two values. That is

$$\begin{aligned} M &= 2\pi_L^2 I^1(\mathbf{p}_L : \mathbf{p}) \\ &= 2\pi_R^2 I^1(\mathbf{p}_R : \mathbf{p}) \\ &= \pi_L \{ 2\pi_R^2 I^1(\mathbf{p}_R : \mathbf{p}) \} + \pi_R \{ 2\pi_L^2 I^1(\mathbf{p}_L : \mathbf{p}) \}. \end{aligned}$$

A new family of splitting criteria can therefore be defined as

$$D(\lambda) \equiv \pi_L \{ \pi_R^2 I^\lambda(\mathbf{p}_R : \mathbf{p}) \} + \pi_R \{ \pi_L^2 I^\lambda(\mathbf{p}_L : \mathbf{p}) \} = \pi_L \pi_R C(\lambda), \quad -1 < \lambda < \infty.$$

**Theorem 2**  $D(\lambda)$  has the exclusivity preference property.

**Proof.** By similar argument of Theorem 1 with  $f(x)$  being replaced by

$$\frac{1}{\lambda(\lambda + 1)} \{ \beta \alpha^2 x^{(\lambda+1)} + \alpha \beta^2 [(p_1 - \alpha x) / \beta]^{\lambda+1} \}.$$

## 4 Optimal splits

In real world, the set of candidate splits are restricted, for example, univariate or linear. If all possible splits are allowed, it is also of interest to know what split corresponding to the maximum value of  $C(\lambda)$  or  $D(\lambda)$ . It is shown in this section that the behavior of optimal splits of these families is dependent on the value of parameter  $\lambda$ . Based on  $C(\lambda)$ , the criteria favor splits that balance the sizes of two subnodes, if  $-1 < \lambda < 1$  and prefer splits that channel the smallest class into one pure node and all the other into the other node, with  $1 < \lambda$ . If  $\lambda = 1$ , the criterion becomes the chi-squared criterion which favors splits that send the classes into two disjoint subnodes. Similarly, the criterion based on  $D(\lambda)$  favors splits that balancing the sizes, if  $-1 < \lambda \leq 1$  and prefers splits that lead the smallest class into one pure node if  $\lambda \geq 2$ .

From Theorem 1, it is known that the optimal splits do not split classes. Let  $\mathcal{L}$  be the set of classes that is channeled into the left node and  $\alpha = \pi_{\mathcal{L}}$ . Then

$$\begin{aligned} \max_{0 < \alpha < 1} C(\lambda) &= \max_{0 < \alpha < 1} \frac{1}{\lambda(\lambda + 1)} \left\{ \sum_{j \in \mathcal{L}} p_j (\alpha^{-\lambda} - 1) + \sum_{j \in \mathcal{L}^c} p_j [(1 - \alpha)^{-\lambda} - 1] \right\} \\ &= \max_{0 < \alpha < 1} \frac{1}{\lambda(\lambda + 1)} \{ \alpha (\alpha^{-\lambda} - 1) + (1 - \alpha) [(1 - \alpha)^{-\lambda} - 1] \}. \end{aligned}$$

Denote

$$g(\alpha) = \begin{cases} \frac{1}{\lambda(\lambda+1)} \{ \alpha^{1-\lambda} + (1-\alpha)^{1-\lambda} - 1 \} & \text{if } \lambda \neq 0 \\ \frac{1}{2} \{ -\alpha \ln \alpha - (1-\alpha) \ln(1-\alpha) \} & \text{if } \lambda = 0. \end{cases} \quad (1)$$

**Lemma 1**  $g(\alpha)$  attains its maximum values at the following points. (1)  $\alpha = 1/2$ , if  $-1 < \lambda < 1$ . (2)  $\alpha = \min_j p_j$ , if  $1 < \lambda < \infty$ . Moreover,  $g(\alpha)$  is a constant function, if  $\lambda = 1$ .

**Proof.** The second derivative of function  $g$  follows

$$g''(\alpha) = \frac{\lambda - 1}{\lambda + 1} \{ \alpha^{-\lambda-1} + (1 - \alpha)^{-\lambda-1} \}. \quad (2)$$

For  $-1 < \lambda < 1$ , observe that  $g(\alpha)$  is symmetric with respect to  $1/2$  and is a concave function. Therefore  $g(\alpha)$  has a maximum at  $\alpha = 1/2$ . For the case  $1 < \lambda < \infty$ , by equation (2),  $g$  is a convex function which indicates  $g(\alpha)$  has a maximum at the boundaries  $\alpha = \min_j p_j$  or  $\alpha = 1 - \min_j p_j$ . By the symmetric property of  $g(\alpha)$  with respect to  $1/2$ ,  $g(\alpha)$  has a maximum at  $\alpha = \min_j p_j$ . Furthermore, when  $\lambda = 1$ ,  $g(\alpha)$  is a constant function.

**Theorem 3** The optimal split based on  $C(\lambda)$  has the following properties.

1. It sends the classes to two disjoint subsets  $\mathcal{L}, \mathcal{L}^c \subset \{1, 2, \dots, J\}$  such that  $\mathcal{L}$  minimizes  $|\sum_{j \in \mathcal{L}} p_j - 1/2|$ , if  $-1 < \lambda < 1$ .
2. It sends the classes to two disjoint subsets, if  $\lambda = 1$ .

3. It sends the class  $i$  such that  $p_i = \min_j p_j$  to one subnode and all the other classes into another subnode, if  $1 < \lambda < \infty$ .

**Proof.** By Lemma 1 and the fact that  $g(\alpha)$  is concave for  $-1 < \lambda < 1$ .

Breiman (1996) shows that part (1) is also true for the twoing criterion and part (3) is true for the Gini criterion by replacing  $\min_j p_j$  with  $\max_j p_j$ .

**Theorem 4** The optimal split based on  $D(\lambda)$  has the following properties.

1. It sends the classes to two disjoint subsets  $\mathcal{L}, \mathcal{L}^c \subset \{1, 2, \dots, J\}$  such that  $\mathcal{L}$  minimizes  $|\sum_{j \in \mathcal{L}} p_j - 1/2|$ , if  $-1 < \lambda \leq 1$ .
2. It sends the class  $i$  such that  $p_i = \min_j p_j$  to one subnode and all the other classes into another subnode, if  $\lambda \geq 2$ .

**Proof.** By equation (1), we have

$$\max_{0 < \alpha < 1} D(\lambda) = \max_{0 < \alpha < 1} \alpha(1 - \alpha)g(\alpha).$$

Let  $h(\alpha) = \alpha(1 - \alpha)g(\alpha)$ . By Lemma 1, we know that both  $\alpha(1 - \alpha)$  and  $g(\alpha)$  attain their maximum values at  $\alpha = 1/2$  if  $-1 < \lambda \leq 1$ . Thus, we only need to study the case for  $\lambda > 1$ .

The second derivative of  $h$  follows

$$\lambda(\lambda + 1)h''(\alpha) = (\lambda - 2)\{\lambda - 1 - \alpha(\lambda - 3)\}\alpha^{-\lambda} + (\lambda - 2)\{\alpha(\lambda - 3) + 2\}(1 - \alpha)^{-\lambda} + 2.$$

We have  $h''(\alpha) > 0$  for all  $\alpha \in (0, 1)$  and  $\lambda \geq 2$ . Thus  $h$  is a convex function for  $\lambda \geq 2$ . Since  $h(\alpha)$  is symmetric with respect to  $1/2$ ,  $h(\alpha)$  has a maximum at  $\alpha = \min_j p_j$ .

**Remark.** As for  $1 < \lambda < 2$ , let

$$\begin{aligned} \eta(\lambda) &= h''(1/2) = 2^\lambda(1 - 2/\lambda) + 2/\{\lambda(\lambda + 1)\} \\ &= \frac{2^\lambda(\lambda - 2)(\lambda + 1) + 2}{\lambda(\lambda + 1)}. \end{aligned}$$

Since  $\eta(1) < 0, \eta(2) > 0$  and  $\eta'(\lambda) > 0$ ,  $\eta(\lambda)$  has only one root between 1 and 2. Let  $\lambda_0 \in (1, 2)$  be the solution of  $\eta(\lambda) = 0$ . We then have,  $\eta(\lambda) = h''(1/2) < 0$ , if  $\lambda < \lambda_0$ .

Moreover,  $\lambda - 1 - \alpha(\lambda - 3) \geq \frac{1}{2}(\lambda + 1)$  and  $\alpha(\lambda - 3) + 2 \geq \frac{1}{2}(\lambda + 1)$  for all  $\alpha \in (0, 1/2]$ . Since  $h''(\alpha) = h''(1 - \alpha)$  and  $\alpha^{-\lambda} + (1 - \alpha)^{-\lambda} \geq 2^\lambda$ ,

$$\begin{aligned} \frac{1}{\lambda - 2}h''(\alpha) &\geq \frac{1}{2\lambda}\alpha^{-\lambda} + \frac{1}{2\lambda}(1 - \alpha)^{-\lambda} + \frac{2}{\lambda(\lambda + 1)(\lambda - 2)} \\ &\geq \frac{h''(1/2)}{\lambda - 2}. \end{aligned}$$

Hence,  $h''(\alpha) \leq h''(1/2) < 0$  for all  $\lambda \in (1, \lambda_0)$  and  $\alpha \in (0, 1)$ . That is,  $h$  is a concave function for  $\lambda \in (1, \lambda_0)$ . We conclude that part (1) of Theorem 4 still holds for  $-1 < \lambda < \lambda_0 \approx 1.79346$ .

## 5 Case studies

Three examples are given in this section to compare the performance of the Gini, MPI, chi-squared, and entropy criterion. The first one is a simulation. The other two consist of real data studies. For real data sets, all the results are produced by using univariate exhaustive search, ten-fold CV pruning, and 1-SE rule as described in Breiman et al. (1984). Priors are estimated from the data. It is shown that, as far as simplifying splits are concerned, the chi-squared or entropy criterion performs better than the other two criteria.

### 5.1 Simple example

The following example given in Taylor and Silverman (1993) shows that the Gini criterion does not have the exclusivity preference property. Suppose we gather 80 observations, 40 of class 1, 20 of class 2 and 10 each of classes 3 and 4. Two splits are compared each with  $\pi_R = 0.25$ . Split A channels all the class 3 and 4 into the right node while Split B places 17 class 2 and 3 class 4 into it. The following summary is obtained for various splitting criteria.

	Gini	MPI	$G^2$	$X^2$
Split A	0.198	0.188	1.000	0.562
Split B	0.209	0.129	0.690	0.380

The values are computed up to proportional constants. In practice, Split A is preferred to Split B. However, the Gini value is higher for Split B.

### 5.2 Genus *Chaetocnema* data

This data set is from Lubischew (1962) and is used in Taylor and Silverman (1993). The data set consists of three species of male flea-beetles of the genus *Chaetocnema*. There are 21 *Chaetocnema concinna*, 31 *Chaetocnema heikertingeri*, and 22 *Chaetocnema heptapotamica* beetles. Six variables are selected to discriminant between the three species. The classification tree obtained by using the Gini, entropy, or MPI criterion is shown in Figure 1. The classification tree obtained by using the chi-squared criterion is shown in Figure 2.

The root split in Figure 2 reduces the problem from a three class problem to a two class problem. It is more simple than that in Figure 1, although the ten-fold CV error rates of the two trees are not different significantly.

*Figure 1 and 2 about here*

### 5.3 Wine recognition data

This data set consists of three types of wines grown in the same region of Italy. For each type, 13 constituent variables are measured. The data set is obtained from UCI repository (Merz and Murphy, 1996) and is used in Aeberhard, Coomans and de Vel (1993). There are 59 type 1, 71 type 2 and 48 type 3 observations. The trees obtained by using the Gini, MPI, chi-squared, entropy criterion are presented in Figure 3, Figure 4, Figure 5, and Figure 6, respectively. The results show that the root splits of Figure 5 and 6 are again more simple than those of Figure 3 and Figure 4. Furthermore, in Figure 3, the split chosen for the right children node of the root ( $X_{11} \leq 0.64$ ) may not be a good choice. Based on simplicity and error rate, the tree based on the entropy criterion is clearly the winner among them.

*Figure 3, 4, 5 and 6 about here*

## 6 Conclusions

Two families of splitting criteria for classification trees have been proposed. They contain the chi-squared, entropy, and mean posterior improvement criterion as special cases. Both families possess the property of exclusive preference which is not owned by the popular Gini criterion. Examples are given to show that some proposed criteria can improve the interpretation of classification trees. In one case, the accuracy of classification trees produced by the family members are even better than that obtained by the Gini criterion.

The main strength of tree-structured classification is that it provides understanding and insight of the data (Breiman et al., 1984; Hand, 1997). More insight can be obtained by using several different splitting criteria. The proposed families can certainly serve this need.

### Acknowledgments

I am very grateful to the referee for the valuable comments and suggestions. I would also like to thank W.-Y. Loh for his interests.

## References

- Aeberhard, S., Coomans, D. and de Vel, O. (1993). Improvements in the performance of regularized discriminant analysis, *Journal of Chemometrics* **7**: 99–115.
- Breiman, L. (1996). Some properties of splitting criteria, *Machine Learning* **24**: 41–47.
- Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont, California.
- Buntine, W. and Niblett, T. (1992). A further comparison of splitting rules for decision tree induction, *Machine Learning* **8**: 75–85.
- Ciampi, A., Chang, C.-H., Hogg, S. and McKinney, S. (1987). Recursive partitioning: a versatile method for exploratory data analysis in biostatistics, in M. I. B. and G. J. 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.
- Fayyad, U. M. and Irani, R. B. (1992). The attribute selection problem in decision tree generation, *10th National Conference on AI, AAAI-92*, MIT Press, pp. 104–110.
- Hand, D. J. (1997). *Construction and Assessment of Classification Rules*, John Wiley, Chichester, England.
- Kass, G. V. (1980). An exploratory technique for investigation large quantities of categorical data, *Applied Statistics* **29**: 119–127.
- Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees, *Statistica Sinica* **7**. To appear.

- Loh, W.-Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion), *Journal of the American Statistical Association* **83**: 715–728.
- Lubischew, A. A. (1962). On the use of discriminant functions in taxonomy, *Biometrics* **18**: 455–477.
- Merz, C. J. and Murphy, P. M. (1996). *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, CA.
- 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.
- Taylor, P. C. and Silverman, B. W. (1993). Block diagrams and splitting criteria for classification trees, *Statistics and Computing* **3**: 147–161.

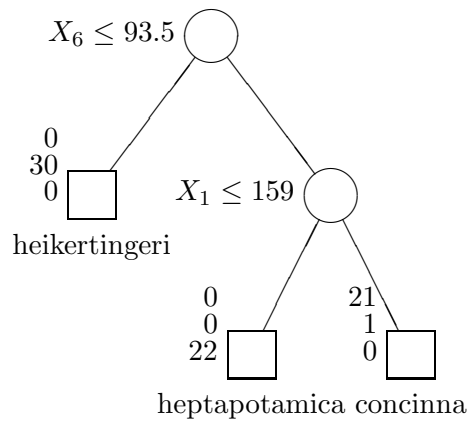


Figure 1:

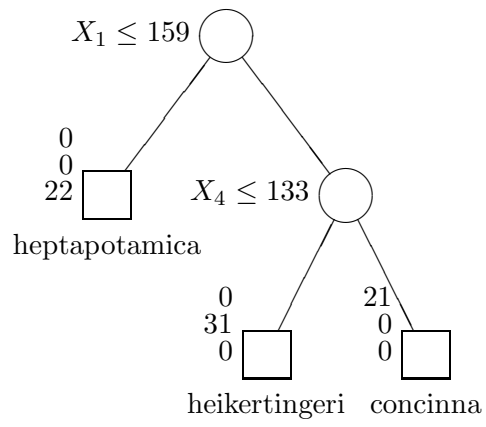


Figure 2:

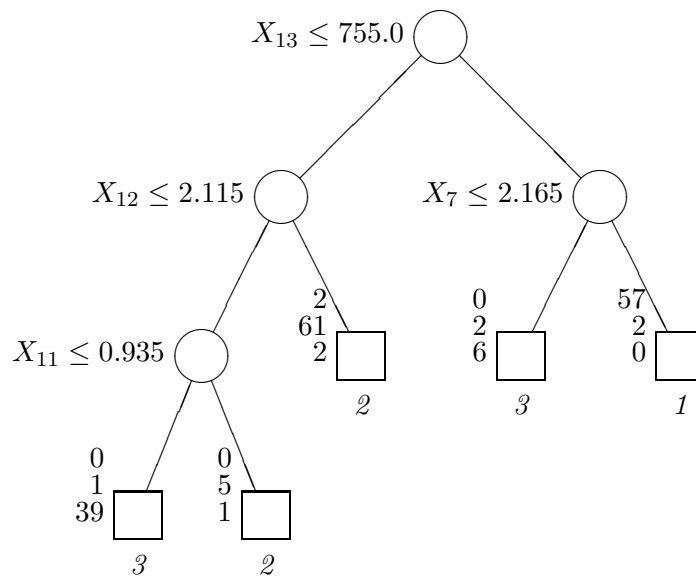


Figure 3:

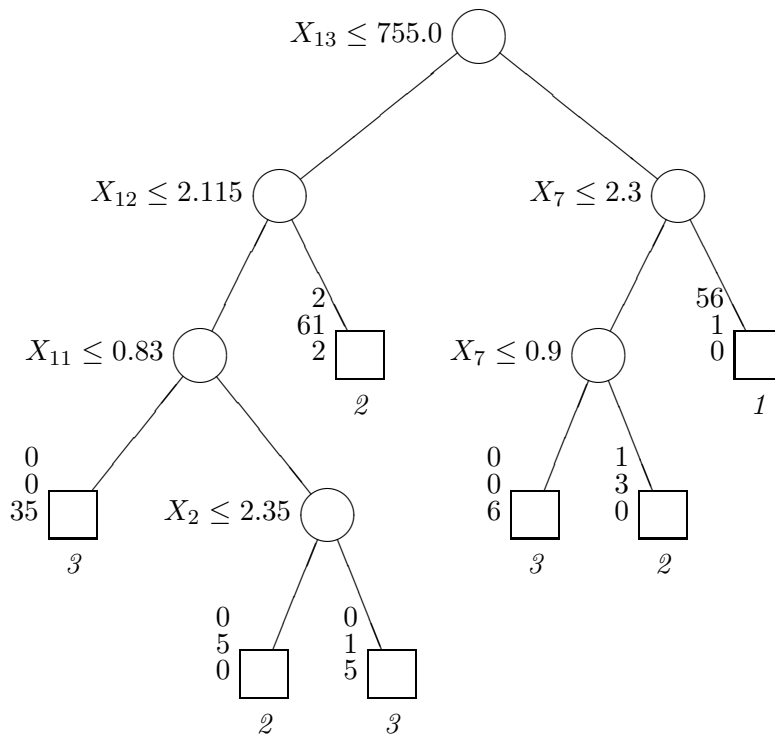


Figure 4:

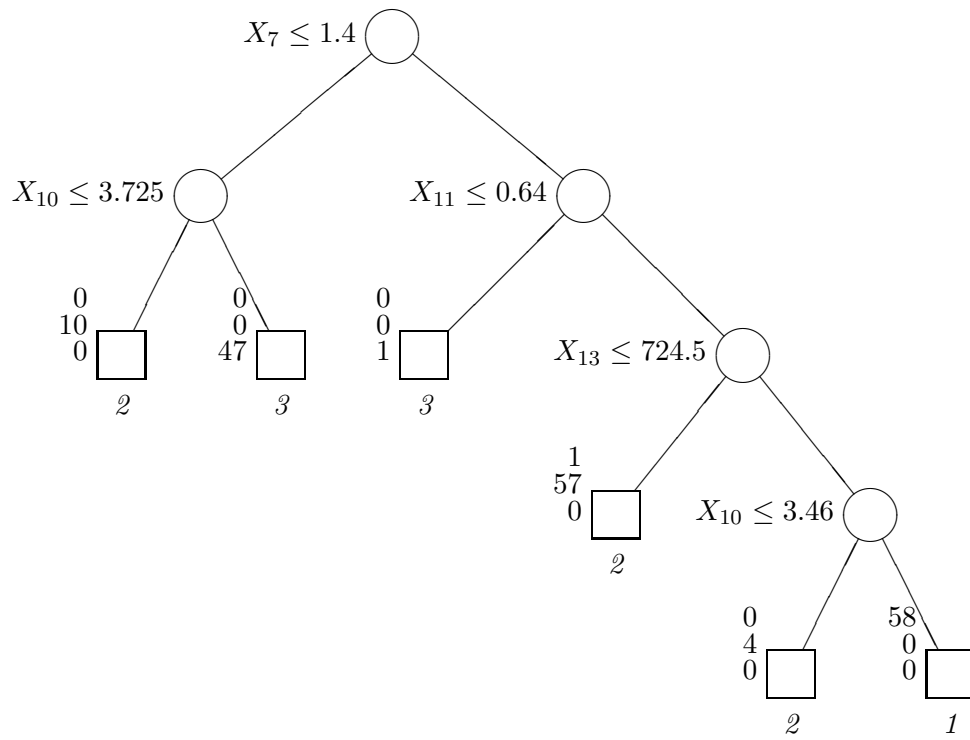


Figure 5:

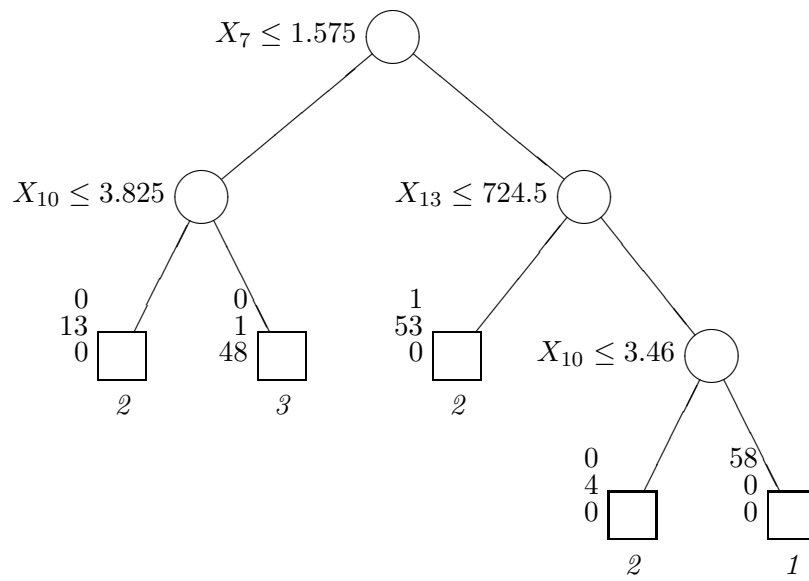


Figure 6:

Figure 1. Chaetocnema tree produced using the Gini, entropy, or MPI splitting criterion. The triple beside each terminal node gives the number of concinna, heikertingeri, and heptapotamica respectively, in the node. Its ten-fold CV error rate is  $0.068 \pm 0.029$ .

Figure 2. Chaetocnema tree produced using the chi-squared splitting criterion. The triple beside each terminal node gives the number of concinna, heikertingeri, and heptapotamica respectively, in the node. Its ten-fold CV error rate is  $0.041 \pm 0.023$ .

Figure 3. Wine tree produced using the Gini splitting criterion. The triple beside each terminal node gives the number of type 1, 2, and 3 respectively, in the node. Its ten-fold CV error rate is  $0.127 \pm 0.025$ .

Figure 4. Wine tree produced using the MPI splitting criterion. The triple beside each terminal node gives the number of type 1, 2, and 3 respectively, in the node. Its ten-fold CV error rate is  $0.101 \pm 0.023$ .

Figure 5. Wine tree produced using the chi-squared splitting criterion. The triple beside each terminal node gives the number of type 1, 2, and 3 respectively, in the node. Its ten-fold CV error rate is  $0.073 \pm 0.020$ .

Figure 6. Wine tree produced using the entropy splitting criterion. The triple beside each terminal node gives the number of type 1, 2, and 3 respectively, in the node. Its ten-fold CV error rate  $0.084 \pm 0.021$ .