CHAID and Exhaustive CHAID Algorithms

This document describes the tree growing process of CHAID and Exhaustive CHAID algorithms. The CHAID algorithm is originally proposed by Kass (1980) and the Exhaustive CHAID is by Biggs et al (1991). Algorithm CHAID and Exhaustive CHAID allow multiple splits of a node.

Both CHAID and exhaustive CHAID algorithms consist of three steps: merging, splitting and stopping. A tree is grown by repeatedly using these three steps on each node starting form the root node.

Notations

Y	The dependent variable, or target variable. It can be ordinal categorical, nominal categorical or continuous. If Y is categorical with *J* classes, its class takes values in C = {1, …, *J*}.
X_m, m = 1, …, M	The set of all predictor variables. A predictor can be ordinal categorical, nominal categorical or continuous.
$\hbar = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$	The whole learning sample.
w_n	The case weight associated with case *n*.
f_n	The frequency weight associated with case *n*. Non-integral positive value is rounded to its nearest integer.

The CHAID Algorithm

The following algorithm only accepts nominal or ordinal categorical predictors. When predictors are continuous, they are transformed into ordinal predictors before using the following algorithm.

Merging

For each predictor variable *X*, merge non-significant categories. Each final category of *X* will result in one child node if *X* is used to split the node. The merging step also calculates the adjusted *p*-value that is to be used in the splitting step.

1. If *X* has 1 category only, stop and set the adjusted *p*-value to be 1.
2. If *X* has 2 categories, go to step 8.
3. Else, find the allowable pair of categories of *X* (an allowable pair of categories for ordinal predictor is two adjacent categories, and for nominal predictor is any two categories) that is least significantly different (i.e., most similar). The most similar pair is the pair whose test statistic gives the largest *p*-value with respect to the dependent variable *Y*. How to calculate *p*-value under various situations will be described in later sections.

4. For the pair having the largest *p*-value, check if its *p*-value is larger than a user-specified alpha-level α_{merge} (*alpha_merge*). If it does, this pair is merged into a single compound category. Then a new set of categories of *X* is formed. If it does not, then go to step 7.
5. (*Optional*) If the newly formed compound category consists of three or more original categories, then find the best binary split within the compound category which *p*-value is the smallest. Perform this binary split if its *p*-value is not larger than an alpha-level $\alpha_{\text{split-merge}}$ (*alpha_spli-merge*).
6. Go to step 2.
7. (*Optional*) Any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest of the *p*-values.
8. The adjusted *p*-value is computed for the merged categories by applying Bonferroni adjustments that are to be discussed later.

Splitting

The "best" split for each predictor is found in the merging step. The splitting step selects which predictor to be used to best split the node. Selection is accomplished by comparing the adjusted *p*-value associated with each predictor. The adjusted *p*-value is obtained in the merging step.

1. Select the predictor that has the smallest adjusted *p*-value (i.e., most significant).
2. If this adjusted *p*-value is less than or equal to a user-specified alpha-level α_{split} (*alpha_split*), split the node using this predictor. Else, do not split and the node is considered as a terminal node.

Stopping

The stopping step checks if the tree growing process should be stopped according to the following stopping rules.

1. If a node becomes pure; that is, all cases in a node have identical values of the dependent variable, the node will not be split.
2. If all cases in a node have identical values for each predictor, the node will not be split.
3. If the current tree depth reaches the user specified maximum tree depth limit value, the tree growing process will stop.
4. If the size of a node is less than the user-specified minimum node size value, the node will not be split.
5. If the split of a node results in a child node whose node size is less than the user-specified minimum child node size value, child nodes that have too few cases (as compared with this minimum) will merge with the most similar child node as measured by the largest of the *p*-values. However, if the resulting number of child nodes is 1, the node will not be split.

The Exhaustive CHAID Algorithm

Splitting and stopping steps in Exhaustive CHAID algorithm are the same as those in CHAID. Merging step uses an exhaustive search procedure to merge any similar pair until only a single pair remains.

Also like CHAID, only nominal or ordinal categorical predictors are allowed, continuous predictors are first transformed into ordinal predictors before using the following algorithm.

Merging

1. If *X* has 1 category only, then set the adjusted *p*-value be 1.
2. Set *index* = 0. Calculate the *p*-value based on the set of categories of *X* at this time. Call the *p*-value *p*(*index*) = *p*(0).
3. Else, find the allowable pair of categories of *X* that is least significantly different (i.e., most similar). This can be determined by the pair whose test statistic gives the largest *p*-value with respect to the dependent variable *Y*. How to calculate *p*-value under various situations will be described in a later section.
4. Merge the pair that gives the largest *p*-value into a compound category.
5. (*Optional*) If the compound category just formed contains three or more original categories, search for a binary split of this compound category that gives the smallest *p*-value. If this *p*-value is larger than the one in forming the compound category by merging in the previous step, perform the binary split on that compound category.
6. Update the *index* = *index* + 1, calculate the *p*-value based on the set of categories of *X* at this time. Denote *p*(*index*) as the *p*-value.
7. Repeat 3 to 6 until only two categories remain. Then among all the indices, find the set of categories such that *p*(*index*) is the smallest.
8. (*Optional*) Any category having too few observations (as compared with a user-specified minimum segment size) is merged with the most similar other category as measured by the largest *p*-value.
9. The adjusted *p*-value is computed by applying Bonferroni adjustments which are to be discussed in a later section.

Unlike CHAID algorithm, no user-specified alpha-level (*alpha_split-merge* or *alpha_merge*) is needed. Only the alpha-level α_{split} (*alpha_split-node*) is needed in the splitting step.

The *p*-Value Calculations

Calculations of (unadjusted) *p*-values in the above algorithms depend on the type of dependent variable.

The merging step of both CHAID and Exhaustive CHAID sometimes needs the p-value for a pair of *X* categories, and sometimes needs the p-value for all the categories of *X*. When p-value for a pair of *X* categories is needed, only part of data in the current node is relevant. Let *D* denotes the relevant data. Suppose in *D* there are *I* categories of *X*, and *J* categories of *Y* (if *Y* is categorical). The p-value calculation using data in *D* is given below.

Continuous dependent variable

If the dependent variable *Y* is continuous, perform an ANOVA F test that tests if the means of *Y* for different categories of *X* are the same. This ANOVA F test calculates the *F*-statistic and hence derives the *p*-value as

$$F = \frac{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(\bar{y}_i - \bar{y})^2/(I-1)}{\sum_{i=1}^{I}\sum_{n\in D} w_n f_n I(x_n = i)(y_n - \bar{y}_i)^2/(N_f - I)},$$

$$p = \Pr\left(F(I-1, N_f - I) > F\right),$$

where

$$\bar{y}_i = \frac{\sum_{n\in D} w_n f_n y_n I(x_n = i)}{\sum_{n\in D} w_n f_n I(x_n = i)},\ \bar{y} = \frac{\sum_{n\in D} w_n f_n y_n}{\sum_{n\in D} w_n f_n},\ N_f = \sum_{n\in D} f_n,$$

and $F(I-1, N_f - I)$) is a random variable following a *F*-distribution with degrees of freedom I and $N_f - I$.

Nominal dependent variable

If the dependent variable *Y* is nominal categorical, the null hypothesis of independence of *X* and *Y* is tested. To do the test, a contingency (or count) table is formed using classes of *Y* as columns and categories of the predictor *X* as rows. The expected cell frequencies under the null hypothesis are estimated. The observed cell frequencies and the expected cell frequencies are used to calculate Pearson chi-squared statistic or likelihood ratio statistic. The *p*-value is computed based on either one of these two statistics.

The Pearson's Chi-square statistic and likelihood ratio statistic are respectively,

$$X^2 = \sum_{j=1}^{J}\sum_{i=1}^{I} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

$$G^2 = 2\sum_{j}^{J}\sum_{i=1}^{I} n_{ij} \ln(n_{ij}/\hat{m}_{ij})$$

where $n_{ij} = \sum_{n \in D} f_n \, I(x_n = i \wedge y_n = j)$ is the observed cell frequency and $\hat{m}_{ij}$ is the estimated expected cell frequency for cell $(x_n = i, y_n = j)$ from independence model as following. The corresponding *p*-value is given by $p = \Pr\left(\chi_d^2 > X^2\right)$ for Pearson's Chi-square test or $p = \Pr\left(\chi_d^2 > G^2\right)$ for likelihood ratio test, where χ_d^2 follows a chi-squared distribution with degrees of freedom $d = (J - 1)(I - 1)$.

Estimation of Expected Cell Frequencies without case Weights

$$\hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

where

$$n_{i.} = \sum_{j=1}^{J_t} n_{ij} \ , \quad n_{.j} = \sum_{i=1}^{I_t} n_{ij} \ , \quad n_{..} = \sum_{j=1}^{J_t} \sum_{i=1}^{I_t} n_{ij} \ .$$

Estimation of Expected Cell Frequencies with Case Weights

If case weights are specified, the expected cell frequency under the null hypothesis of independence is of the form

$$m_{ij} = \overline{w}_{ij}^{-1} \alpha_i \beta_j$$

where α_i and β_j are parameters to be estimated, and

$$\overline{w}_{ij} = \frac{w_{ij}}{n_{ij}}, \quad w_{ij} = \sum_{n \in D} w_n \, f_n I(x = i \wedge y_n = j) \, .$$

Parameters estimates $\hat{\alpha}_i$, $\hat{\beta}_j$, and hence $\hat{m}_{ij}$, are resulted from the following iterative procedure.

1. $k = 0$, $\alpha_i^{(0)} = \beta_j^{(0)} = 1$, $m_{ij}^{(0)} = \overline{w}_{ij}^{-1}$.
2. $\alpha_i^{(k+1)} = \dfrac{n_{i.}}{\sum_j \overline{w}_{ij}^{-1} \beta_j^{(k)}} = \alpha_i^{(k)} \dfrac{n_{i.}}{\sum_j m_{ij}^{(k)}}$.
3. $\beta_j^{(k+1)} = \dfrac{n_{.j}}{\sum_i \overline{w}_{ij}^{-1} \alpha_i^{(k+1)}}$.
4. $m_{ij}^{(k+1)} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)}$.

5. If $\max_{i,j}\left|m_{ij}^{(k+1)} - m_{ij}^{(k)}\right| < \varepsilon$, stop and output $\alpha_i^{(k+1)}$, $\beta_j^{(k+1)}$ and $m_{ij}^{(k+1)}$ as the final estimates $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{m}_{ij}$. Otherwise, $k = k+1$, go to 2.

Ordinal dependent variable

If the dependent variable *Y* is categorical ordinal, the null hypothesis of independence of *X* and *Y* is tested against the row effects model (with the rows being the categories of *X* and columns the classes of *Y*) proposed by Goodman (1979). Two sets of expected cell frequencies, $\hat{m}_{ij}$ (under the hypothesis of independence) and $\hat{\hat{m}}_{ij}$ (under the hypothesis that the data follow a row effects model), are both estimated. The likelihood ratio statistic and the *p*-value are

$$H^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\hat{\hat{m}}_{ij}\ln(\hat{\hat{m}}_{ij}/\hat{m}_{ij}),$$

$$p = \Pr(\chi^2_{I-1} > H^2).$$

Estimation of Expected Cell Frequencies under Row Effects Model

In the row effects model, scores for classes of *Y* are needed. By default, the order of a class of *Y* is used as the class score. Users can specify their own set of scores. Scores are set at the beginning of the tree and kept unchanged afterward. Let s_j be the score for class *j* of *Y*, *j* = 1, …, *J*. The expected cell frequency under the row effects model is given by

$$m_{ij} = \overline{w}_{ij}^{-1}\alpha_i\beta_j\gamma_i^{(s_j-\bar{s})}$$

where

$$\bar{s} = \sum_{j=1}^{J} w_{\cdot j}s_j \Big/ \sum_{j=1}^{J} w_{\cdot j},$$

in which $w_{\cdot j} = \sum_i w_{ij}$, α_i, β_j and γ_i are unknown parameters to be estimated.

Parameters estimates $\hat{\hat{\alpha}}_i$, $\hat{\hat{\beta}}_j$, $\hat{\hat{\gamma}}_i$ and hence $\hat{\hat{m}}_{ij}$ are resulted from the following iterative procedure.

1. $k = 0$, $\alpha_i^{(0)} = \beta_j^{(0)} = \gamma_i^{(0)} = 1$, $m_{ij}^{(0)} = \overline{w}_{ij}^{-1}$.

2. $\alpha_i^{(k+1)} = \dfrac{n_{.j}}{\sum_j \overline{w}_{ij}^{-1} \beta_j^{(k)} \left(\gamma_i^{(k)}\right)^{(s_j - \bar{s})}} = \alpha_i^{(k)} \dfrac{n_{i.}}{\sum_j m_{ij}^{(k)}}$.

3. $\beta_j^{(k+1)} = \dfrac{n_{.j}}{\sum_i \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \left(\gamma_i^{(k)}\right)^{(s_j - \bar{s})}}$.

4. $m_{ij}^{*} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)} \left(\gamma_i^{(k)}\right)^{(s_j - \bar{s})}$, $G_i = 1 + \dfrac{\sum_j (s_j - \bar{s})(n_{ij} - m_{ij}^{*})}{\sum_j (s_j - \bar{s})^2 m_{ij}^{*}}$.

5. $\gamma_i^{(k+1)} = \begin{cases} \gamma_i^{(k)} G_i & G_i > 0 \\ \gamma_i^{(k)} & \text{otherwise} \end{cases}$.

6. $m_{ij}^{(k+1)} = \overline{w}_{ij}^{-1} \alpha_i^{(k+1)} \beta_j^{(k+1)} \left(\gamma_i^{(k+1)}\right)^{(s_j - \bar{s})}$.

7. If $\max_{i,j} \left| m_{ij}^{(k+1)} - m_{ij}^{(k)} \right| < \varepsilon$, stop and output $\alpha_i^{(k+1)}$, $\beta_j^{(k+1)}$, $\gamma_i^{(k+1)}$ and $m_{ij}^{(k+1)}$ as the final estimates $\hat{\alpha}_i$, $\hat{\beta}_j$, $\hat{\gamma}_i$, $\hat{m}_{ij}$. Otherwise, $k = k + 1$, go to 2.

The Bonferroni Adjustments

The adjusted *p*-value is calculated as the *p*-value times a Bonferroni multiplier. The Bonferroni multiplier adjusts for multiple tests.

CHAID

Suppose that a predictor variable originally has *I* categories, and it is reduced to *r* categories after the merging step. The Bonferroni multiplier *B* is the number of possible ways that *I* categories can be merged into *r* categories. For $r = I$, $B = 1$. For $2 \le r < I$, use the following equation.

$$B = \begin{cases} \dbinom{I-1}{r-1} & \text{Ordinal predictor} \\ \sum_{v=0}^{r-1} (-1)^v \dfrac{(r-v)^I}{v!(r-v)!} & \text{Nominal predictor} \\ \dbinom{I-2}{r-2} + r\dbinom{I-2}{r-1} & \text{Ordinal with a missing category} \end{cases}.$$

Exhaustive CHAID

Exhaustive CHAID merges two categories iteratively until only two categories left. The Bonferroni multiplier *B* is the sum of number of possible ways of merging two categories at each iteration.

$$B = \begin{cases} \dfrac{I(I-1)}{2} & \text{Ordinal predictor} \\ \dfrac{I(I^2-1)}{2} & \text{Nominal predictor} \\ \dfrac{I(I-1)}{2} & \text{Ordinal with a missing category} \end{cases} .$$

Missing Values

If the dependent variable of a case is missing, it will not be used in the analysis. If all predictor variables of a case are missing, this case is ignored. If the case weight is missing, zero, or negative, the case is ignored. If the frequency weight is missing, zero, or negative, the case is ignored.

Otherwise, missing values will be treated as a predictor category. For ordinal predictors, the algorithm first generates the "best" set of categories using all non-missing information from the data. Next the algorithm identifies the category that is most similar to the missing category. Finally, the algorithm decides whether to merge the missing category with its most similar category or to keep the missing category as a separate category. Two p-values are calculated, one for the set of categories formed by merging the missing category with its most similar category, and the other for the set of categories formed by adding the missing category as a separate category. Take the action that gives the smallest *p*-value.

For nominal predictors, the missing category is treated the same as other categories in the analysis.

References

Bigss, D., Ville, B., and Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. *Journal of Applied Statistics*, 18, 1, 49-62.

Goodman, L. A. (1979). Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories. *Journal of the American Statistical Association*, 74, 537-552.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 20, 2, 119-127.