QUEST Algorithm

This document describes the tree growing process of the QUEST algorithm. QUEST is proposed by Loh and Shih (1997), and stands for Quick, Unbiased, Efficient, Statistical Tree. It is a tree-structured classification algorithm that yields a binary decision tree. A comparison study of QUEST and other algorithms was conducted by Lim et al (2000).

Notations

Y	The dependent variable, or target variable. It has to be nominal categorical. 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 continuous (including ordinal categorical) or nominal categorical.
$\hbar = \{\mathbf{x}_n, y_n\}_{n=1}^{N}$	The whole learning sample.
$\hbar(t)$	The learning samples that fall in node *t*.
f_n	The frequency weight associated with case *n*. Non-integral positive value is rounded to its nearest integer.
N_f	Total number of learning cases, $N_f = \sum_{n \in \hbar} f_n$
$N_{f,j}$	Total number of class *j* learning cases, $N_{f,j} = \sum_{n \in \hbar} f_n I(y_n = j)$
$N_f(t)$	Total number of learning cases in node *t*, $N_f(t) = \sum_{n \in \hbar(t)} f_n$
$N_{f,j}(t)$	Total number of class *j* learning cases in node *t*, $N_{f,j}(t) = \sum_{n \in \hbar(t)} f_n I(y_n = j)$.
$\pi(j)$, j = 1, …, J	Prior probability of $Y = j$, j = 1, …, J.
$p(j,t)$, j = 1, …, J	The probability of a case in class *j* and node *t*.
$p(j \mid t)$, j = 1, …, J	The probability of a case in class *j* given that it falls into node *t*.
$C(i \mid j)$	The cost of miss-classifying a class *j* case as a class *i* case. Clearly $C(j \mid j) = 0$.

QUEST Tree Growing Process

The QUEST tree growing process consists of the selection of a split predictor, selection of a split point for the selected predictor, and stopping. In this algorithm, only univariate splits are considered.

Selection of split predictor

1. For each continuous predictor *X*, perform an ANOVA F test that tests if all the different classes of the dependent variable *Y* have the same mean of *X*, and calculate the *p*-value according to the F statistics. For each categorical predictor, perform a Pearson's chi-square test of *Y* and *X*'s independence, and calculate the *p*-value according to the chi-square statistics.
2. Find the predictor with the smallest *p*-value and denote it X^*.
3. If this smallest *p*-value is less than α / M, where $\alpha \in (0,1)$ is a user specified level of significance and *M* is the total number of predictor variables, predictor X^* is selected as the split predictor for the node. If not, go to 4.
4. If this smallest *p*-value is greater than or equal to α / M, then
   - For each continuous predictor *X*, compute a Levene's *F*-statistic based on the absolute deviation of *X* from its class mean to test if the variances of *X* for different classes of *Y* are the same, and calculate the *p*-value for the test.
   - Find the predictor with the smallest *p*-value and denote it as X^{**}.
   - If this smallest *p*-value is less than $\alpha /(M + M_1)$, where M_1 is the number of continuous predictors, X^{**} is selected as the split predictor for the node. Otherwise, this node is not split.

ANOVA F test

Suppose, for node *t*, there are J_t classes of dependent variable *Y*. The *F*-statistic for a continuous predictor *X* is given by

$$F_X = \frac{\sum_{j=1}^{J_t} N_{f,j}(t)(\bar{x}^{(j)}(t) - \bar{x}(t))^2 / (J_t - 1)}{\sum_{n \in \hbar(t)} f_n \left(x_n - \bar{x}^{(y_n)}(t)\right)^2 \Big/ (N_f(t) - J_t)}$$

where

$$\bar{x}^{(j)}(t) = \frac{\sum_{n \in \hbar(t)} f_n x_n I(y_n = j)}{N_{f,j}(t)}, \quad \bar{x}(t) = \frac{\sum_{n \in \hbar(t)} f_n x_n}{N_f(t)}.$$

Its corresponding *p*-value is given by

$$p_X = \Pr\left(F(J_t - 1, N_f(t) - J_t) > F_X\right)$$

where $F(J_t - 1,\ N_f(t) - J_t)$ follows a F-distribution with degrees of freedom J_t - 1 and $N_f(t) - J_t$.

Pearson’s chi-square test

Suppose, for node *t*, there are J_t classes of dependent variable *Y*. The Pearson’s Chi-square statistic, for categorical predictor *X* with I_t categories, is by given

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

where

$$n_{ij} = \sum_{n \in \hbar(t)} f_n I(y_n = j \wedge x_n = i),\ \ \hat{m}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

with

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

where $I(y_n = j \wedge x_n = i)$ = 1 if case *n* has $y_n = j$ and $x_n = i$; 0 otherwise.

The corresponding *p*-value is given by $p_X = \Pr\left(\chi_d^2 > X^2\right)$ where χ_d^2 follows a chi-squared distribution with degrees of freedom $d = (J_t - 1)(I_t - 1)$.

Levene’s *F* test

For continuous predictor *X*, calculate $z_n = \left| x_n - \bar{x}_{\cdot}^{(y_n)}(t) \right|$. The Levene’s F statistics for predictor *X* is the ANOVA *F*-statistic for z_n.

Selection of split point

At a node, suppose that a predictor variable *X* has been selected for splitting. The next step is to determine the split point. If *X* is a continuous predictor variable, a split point *d* in the split $X \le d$ is to be determined. If *X* is a nominal categorical predictor variable, a subset *K* of the set of all values taken by *X* in the split $X \in K$ is to be determined. The algorithm is as follows.

Continuous splitting predictor

If the selected predictor variable *X* is continuous:

1. Group classes of dependent variable *Y* into two super-classes. If there are only two classes of *Y*, go to step 2. Otherwise, calculate the sample mean of *X* for each class of *Y*. If all class means are identical, the class with the most cases is gathered as super-class *A* and the other classes as super-class *B*. If there are two or more classes with the same maximum number of cases, the one with the smallest class index *j* is chosen to form *A* and the rest to *B*. If not all the class means are identical, a *k*-means clustering method, with the initial cluster centers set at the two most extreme class means, is applied to class means to divide classes of Y into two super-classes: *A* and *B*. Let $\bar{x}_A$ and s_A^2 denote the sample mean and variance for super-class *A*, $\bar{x}_B$ and s_B^2 the sample mean and variance for super-class *B*.
2. If $\min(s_A^2, s_B^2) = 0$, order the two super-classes by their variance in increasing order and denote the variances by $s_1^2 \le s_2^2$, and the corresponding means by $\bar{x}_1, \bar{x}_2$. Let ε be a very small positive number, say $\varepsilon = 10^{-12}$.

   If $\bar{x}_1 < \bar{x}_2$, $d = \bar{x}_1(1+\varepsilon)$. Else, $d = \bar{x}_1(1-\varepsilon)$.
3. If $\min(s_A^2, s_B^2) \neq 0$, quadratic discriminant analysis (QDA) is applied to determine the split point *d*. QDA assumes that *X* follows a normal distributions in each super-class with the calculated sample mean and variance. The split point is among the roots that make probability $\Pr(x, A \mid t) = \Pr(x, B \mid t)$ for node *t*, where

   $$\Pr(x, A \mid t) = P(x \mid A, t) P(A \mid t) = P(A \mid t) \frac{1}{\sqrt{2\pi s_A^2}} \exp\left\{ -\frac{(x - \bar{x}_A)^2}{2 s_A^2} \right\},$$

   with

   $$p(A \mid t) = \sum_{j \in A} p(j \mid t) = \sum_{j \in A} \frac{p(j,t)}{\sum_j p(j,t)}, \quad p(j,t) = \frac{\pi(j) N_{f,j}(t)}{N_{f,j}}.$$

   Solving $P(X, A \mid t) = P(X, B \mid t)$ is equivalent to solving the following quadratic equation

   $$ax^2 + bx + c = 0,$$

   where

   $$a = s_A^2 - s_B^2, \quad b = 2\left(\bar{x}_A s_B^2 - \bar{x}_B s_A^2\right),$$

   $$c = \bar{x}_B{}^2 s_A^2 - \bar{x}_A{}^2 s_B^2 + 2 s_A^2 s_B^2 \log \frac{p(A \mid t) s_B}{p(B \mid t) s_A}.$$

If there is only one real root, it is chosen to be the split point, provided this yields two non-empty nodes. If there are two real roots, choose the one that is closer to $\bar{x}_A$, provided this yields two non-empty nodes. Otherwise use the mean $(\bar{x}_A + \bar{x}_B)/2$ as split point.

Note: In step 3, the prior probability distribution for the dependent variable is needed. When user specified costs are involved, the altered priors can be used to replace the priors (optional). The altered prior is defined as $\pi'(j) = \dfrac{C(j)\pi(j)}{\sum_j C(j)\pi(j)}$, where $C(j) = \sum_i C(i|j)$.

Nominal splitting predictor

If the selected predictor variable X is nominal and with more than two categories (if X is binary, the split point is clear), QUEST first transforms it into a continuous variable (call it ξ) by assigning the largest discriminant coordinates to categories of the predictor. QUEST then applies the split point selection algorithm for continuous predictor on ξ to determine the split point.

Transform a categorical predictor into a continuous predictor

Let X be a nominal categorical predictor taking values in the set $\{b_1, \ldots, b_I\}$. Transform X into a continuous variable ξ such that the ratio of between-classes to within-classes sum squares of ξ is maximized (the classes here refer to the classes of dependent variable). The details are as following.

- Transforms *each* value x of X in $\hbar$ into an I-dimensional dummy vector $\boldsymbol{v} = (v_1, \ldots, v_I)'$, where $v_i = \begin{cases} 1 & x = b_i \\ 0 & \text{otherwise} \end{cases}$.

- Calculate the overall and class j mean of $\mathbf{v}$.

$$\bar{\boldsymbol{v}} = \frac{\sum_{n \in \hbar} f_n \boldsymbol{v}_n}{N_f}, \quad \bar{\boldsymbol{v}}^{(j)} = \frac{\sum_{n \in \hbar} f_n \boldsymbol{v}_n I(y_n = j)}{N_{f,j}}.$$

- Calculate the following $I \times I$ matrices.

$$\mathbf{B} = \sum_{j=1}^{J} N_{f,j} (\bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}})(\bar{\boldsymbol{v}}^{(j)} - \bar{\boldsymbol{v}})'$$

$$\mathbf{T} = \sum_{n \in \hbar} f_n (\boldsymbol{v}_n - \bar{\boldsymbol{v}})(\boldsymbol{v}_n - \bar{\boldsymbol{v}})'$$

- Perform single value decomposition on $\boldsymbol{T}$ to obtain $\boldsymbol{T} = \boldsymbol{QDQ'}$, where $\boldsymbol{Q}$ is an $I \times I$ orthogonal matrix, $\boldsymbol{D}$ = diag(d_1, …, d_I) such that $d_1 \geq \ldots \geq d_I \geq 0$. Let $\boldsymbol{D}^{-\frac{1}{2}}$ = diag(d_1^*, …, d_I^*) where $d_i^* = d_i^{-1/2}$ if $d_i > 0$, 0 otherwise. Perform single value decomposition on $D^{-\frac{1}{2}}Q'BQD^{-\frac{1}{2}}$ to obtain its eigenvector $\boldsymbol{a}$ which is associated with its largest eigenvalue.

- The largest discriminant coordinate of $\boldsymbol{v}$ is the projection

$$\xi = \boldsymbol{a}'D^{-\frac{1}{2}}Q'\boldsymbol{v}\ .$$

Note: The original QUEST by Loh and Shih (1997) transforms a categorical predictor into a continuous predictor at a considered node based on the data in the node. SPSS implementation of QUEST does the transformation only once at the very beginning based on the whole learning sample.

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 belong to the same dependent variable class at the node, 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, the node will not be split.

Missing Values

If the dependent variable of a case is missing, this case will be ignored in the analysis. If all predictor variables of a case are missing, this case will be ignored. If the frequency weight is missing, zero or negative, the case will be ignored.

Otherwise, the surrogate split method will be used to deal with missing data in predictor variables. If a case has a missing value at the selected predictor, the assignment will be done based on surrogate split. Method of defining and calculating surrogate splits is the same as that in CART (see CART algorithm for details).

Reference

Lim, T. S., Loh, W. Y. and Shih, Y. S., 2000. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms. *Machine Learning*, 40, 2000.

Loh, W. Y. and Shih, Y. S., 1997. Split selection methods for classification trees. *Statistica Sinica*, Vol. 7, p. 815 - 840.