Machine Learning Prediction Techniques by Classification and Regression

Shamy S
Ph.D Research Scholar,
Noorul Islam University,
Tamil Nadu,India

Dr. J Dheeba,
Associate Professor
Noorul Islam University, Tamil Nadu, India.

ABSTRACT

One of the machine-learning method for constructing prediction models from data is Classification and Regression. By partitioning the data space recursively these models are configuring and in each prediction model are fitting with a simple predictions. Finally, the partitioning can be represented pictorially as a decision tree. Finite number of unordered values are taken for the designing the classification trees and are designed for independent variables. And the prediction error are measured in terms of misclassification cost. Squared difference between the predicted and observed values are measured in regression trees, which are dependent variables that have ordered discrete values or continuous values. Here in this article reviewing and comparing some of the widely acceptable algorithms such as QUEST, GUIDE, CRUISE, C4.5 and RPART with their strengths, weakness and capabilities.

Keywords

Classification and regression tree algorithm, cross-validation, discriminant, linear model, prediction accuracy, recursive partitioning, selection *bias, unbiased, QUEST, GUIDE, CRUISE, C4.5, and RPART*

INTRODUCTION

The classic Classification and regression tree algorithm was introduced by Breiman et al. (Breiman, Friedman, Olshen, & Stone, 1984; see also Ripley, 1996). In statistics classification and regression takes a vital role. A recursive partitioning methods are used in classification and regression tree for predicting regression variables (continues dependent variable) and classification variables (categorical predictor variable). There are various algorithms are available for continuous variables or categorical variables from the set of categorical factor effects or continuous predictors. Some of these types of classification algorithms are QUEST (Quick, Unbiased, and Efficient Statistical Trees) algorithm. It is working in the principle of the Classification Trees Analysis. CHAID (Chi-square Automatic Interaction Detector; see Kass, 1980) is another similar type of algorithm.

As the name implies **Classificationtrees** are used to separate the dataset into classes belonging to the response variable. Usually it has two classes for the response variable that is Yes or No (1 or 0). The algorithm C4.5 are used if there are more than 2 categories. However the standard CART procedure is used for binary splits. The classification tree are used depends on the response or target variable is categorical in nature. The response of the target variable is continues or numeric in that cases regression trees are used. For example the count of new cancer patient in a year. Regression trees are suitable for the prediction type of problems that are opposed to classification. Anyhow the type of target variable which is predictor or independent variable like categorical or numeric, is the one determine which type decision tree needed.

CART and RPART

CART is the classification with the Gini index as node impurity criterion, $i(x) = 1 - \sum_{k=1}^{K} S^2 \left(\frac{k}{x}\right)$. If the split divides the data in x into a right node x_R and left node x_L.. sL and SR be the propotions of data in x_L and x_R respectively. CART select the splits that decreases the impurity maximum $i(x)$-$s_L i(x_L)$-$sRi(x_R)$. CART generate a sequence of subtrees by growing large tree instead of employing stopping rule. And pruning it until the root node. RPART is recursive partitioning and regression trees. Continuous variables or categorical can be used depends on target data whether wants regression trees or classification trees.

QUEST and CRUISE

The two later classification algorithms CART and C4.5 that follows this approach. CART uses the Gini index, a generalization of the binomial variance, whereas Entropy for impurity function uses in C4.5. To minimize an estimate of the misclassification error, unlike THAID, first they grow an overly large tree and later prune it into smaller size. To estimate error rates, heuristic formula uses C4.5, but cross validation employs in default (10 fold) CART. In the examples below CART is implemented in R system as RPART[3][4].

The exhaustive search approach has an undesirable property, despite its elegance and simplicity.Note that an ordered variable with n distinct values has (n− 1) splits of the form X ≤ c, and an unordered variable with n distinct unordered values has (2n−1 − 1) splits of the form X ∈ S. Therefore, variable which have more distinct values have a greater chance to be selected if everything equal. This selection bias effect the tree structure's integrity of inference drawn.

C4.5 and GUIDE

Another classification algorithm is C4.5. It divided the node into two split of usual form, if the variable a_i chosen to split non categorical node, it split in the form a_i<=c. If there are n values in the categorical variable, the node split into n branches with one branch for each categorical value. Regardless the size of n, there is no difficult for dealing with categorical variable to split in C4.5. The variable selection in GUIDE is unbiased same like CRUISE and QUEST. It construct a simple polynomial tree model for least squares. It constructs multiple linear piecewiseconstant. GUIDE is quantile, Poisson and proportional hazards regression method.

STATISTICAL METHODS

If we consider a classification problem, we have a training sample of mobservations on a class variable Zthat takes values 1, 2, . . . , n, and ipredictor variables, X_1, . . . , X_i. Our goal is to find a model for predicting the values of Z from new X values. In theory, the solution is simply a partition of the X space into k disjoint sets, A_1, A_2, . . . , A_n, such that the predicted value of Z is j if X belongs to A_k, for j = 1, 2, . . . , n. If the X are theordered values of variables , two classical solutions are linear discriminant analysis1 and nearest neighbor classification.2 These methods yield sets $A_k$with piecewise linear and nonlinear, respectively, boundaries that are not easy to interpret if iis large.

For easy to interpret classification tree methods yield rectangular sets $A_k$by partitioning recursively the data set one X variable at a time. For example, Figure 1 illustrates wherein there are two two X variables in three classes. The data points are plotted in left panel and the corresponding decision tree structure shows in right panel. The plot on its left is limited to at most two whereas a key advantage of the tree structure is its applicability to any number of variables.

The first published the classification tree algorithm is THAID [1] [2]. By the means of node impurity measures the distribution of the observed Z values in the node. To minimizes the total impurity of its two child nodes by searching over all X and S, THAID splits a node exhaustively, the split $\{X \in S\}$. The set S is an interval of the form $(-\infty, c]$, if X takes ordered values. Otherwise, S is a subset of the values taken by X. The data on in each child node is applied recursively on this process. If the relative decrease in impurity is below a prespecified threshold, splitting stops. The pseudocode for the basic steps gives in Algorithm 1.

Algorithm 1

Pseudocode for tree construction - exhaustive search
Start at the root node

Repeat

For each X,

Find the set S that minimizes the sum of the node impurities in the two child nodes and choose the split {X∗∈S∗} that gives the minimum overall X and S.
If a stopping criterion is reached, exit

End

Otherwise, apply step 2 to each child node in turn.

The two later classification algorithms CART and C4.5 that follows this approach. CART uses the Gini index, a generalization of the binomial variance, whereas Entropy for impurity function uses in C4.5. To minimize an estimate of the misclassification error, unlike THAID, first they grow an overly large tree and later prune it into smaller size. To estimate error rates, heuristic formula uses C4.5, but cross validation employs in default (10 fold) CART. In the examples below CART is implemented in R system as RPART[3][4].

The exhaustive search approach has an undesirable property, despite its elegance and simplicity.Note that an ordered variable with n distinct values has (n− 1) splits of the form X ≤ c, and an unordered variable with n distinct unordered values has (2n−1 − 1) splits of the form X ∈ S. Therefore, variable which have more distinct values have a greater chance to be selected if everything equal. This selection bias effect the tree structure's integrity of inference drawn.

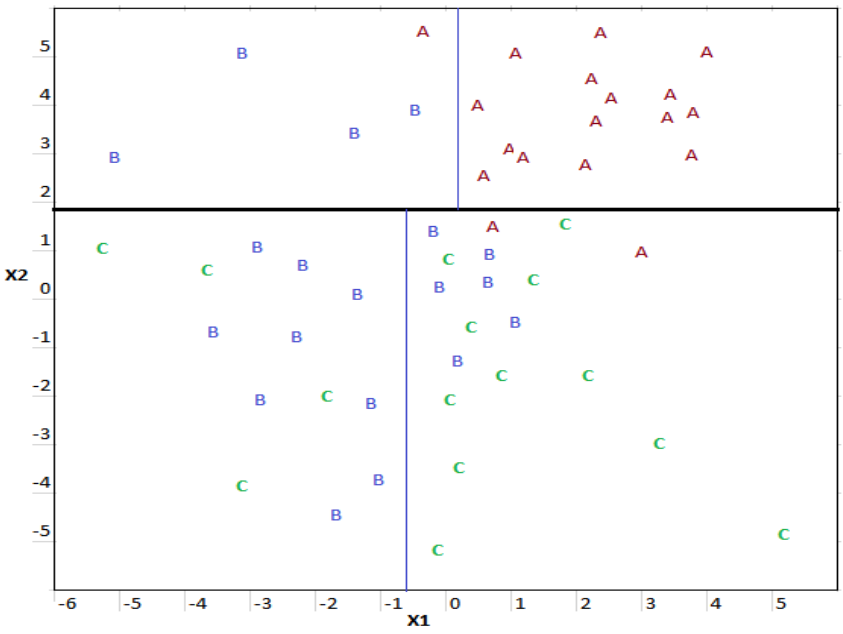

Figure 1.a

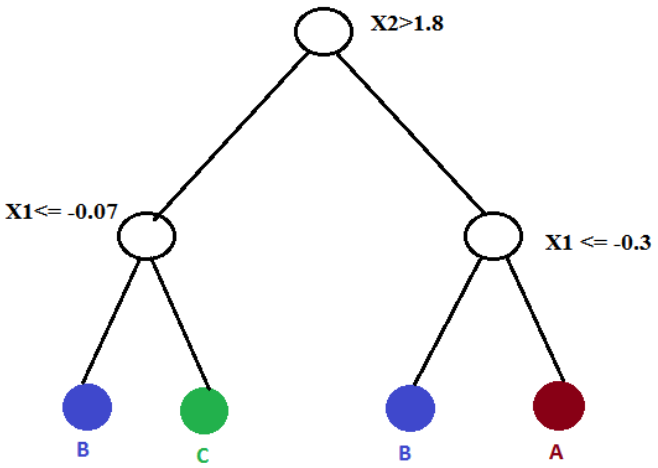

Figure 1.b

Figure 1.a partions and decision tree structure in figure 1.b for a classification of tree model with three classes A, B and C. At each intermediate mode, a case goes to the left child node if and only if the condition is satisfied. The predicted class is given beneath each leaf node.

The idea originated from the FACT[5] algorithm, on significant test to split each node, CRUISE[6][7], GUIDE[8], and QUEST[9] use a double step approach. First for selecting the most significant variable each P is tested association with Q. Then for set S an exhaustive search is performed. An effectively free selection bias approach is carried on if each P is independent of Q, every P has the same chance for selection. By the search for S is carried out only on the selected X variable, most computation is also saved. Chi squared tests are used in CRUISE and GUIDE. And in the case QUEST, analysis of variable (ANOVA) tests are used for ordered variables and chi squared test for unordered variables. Permutation test are used in CTree like an unbiased method[10]. Algorithm 2 gives pseudocode for the GUIDE. As the same way as CART, trees are pruned in CRUISE, GUIDE, and QUEST classifications.

Algorithm 2

GUIDE classification tree construction pseudocode

1. *Start at the root node.*
2. *For each ordered variable P, convert it to an unordered variable P' by grouping its values in the node into a small number of intervals. If P is unordered, set P' = P.*
3. *Perform a chi squared test of independence of each P' variable versus Q on the data in the node and compute its significance probability.*
4. *Choose the variable P* associated with the P' that has the smallest significance probability.*
5. *Find the split set {P* ∈ S*} that minimizes the sum of Gini indexes and use it to split the node into two child nodes.*
6. *If the termination criteria is reached, exit. Otherwise, repeat steps 2–5 to each child node.*
7. *Prune the tree with the CART method.*

CHAID functions in another strategy [11]. In the case of an ordered variable P, it is splitting into 10 intervals its data values and to each interval one child node is also assigned. One child node is assigned to each value of P, in the case of P as an unordered variable. To iteratively merge pairs of child nodes, CHAID uses significant tests and Bonferroni corrections. There are two consequences for this approach. The first one is there is chance for some nodes to split into more than two child nodes. The second thing is because of this method is biased towards few distinct values for the selecting variables, it will affect the correctness of the corrections and the sequential nature of the tests.

Splits in linear combinations of all the ordered variables are allows in CART, CRUISE, and QUEST, but combinations of two variables can only split in GUIDE at a time. CRUISE and CART are using alternate splits on other variables if there is a missing values.C4.5 send each observations are for missing values in split for every branching as a probability weighting scheme. Missing values are locally imputes by QUEST. GUIDE treating as separate category for missing values. User specified misclassification costs are accepted by all other than C4.5. User specified class prior probabilities are accepted by CHAID. For predicting Q to be the class with the lowest misclassification cost, all algorithms fit a constant model to each node by default. GUIDE is fit for nearest Neighbor model and bivariate kernel density models. Bivariate linear discriminant model can optionally fit for CRUISE. Ensemble models using bagging[12] and random forest[13] techniques are also can use GUIDE techniques. All the features of algorithms are summarized in table. **Table Summarized features of algorithms**

Methods	**GUIDE**	**CRUISE**	**QUEST**	**CHAID**	**CART**	**C4.5**
Split Types	*yes*	*yes*	*yes*	*No*	*no*	*no*
Unbiased Splits	*u,l*	*u,l*	*u,l*	*U*	*u,l*	*u*
Interaction Tests	*yes*	*yes*	*no*	*No*	*no*	*no*
Branches/ Splits	*2*	*>=2*	*2*	*>=2*	*2*	*>=2*
Pruning	*yes*	*yes*	*yes*	*No*	*yes*	*yes*
User Specified Priors	*Yes*	*yes*	*yes*	*No*	*yes*	*no*

User-Specified Costs	*yes*	*yes*	*yes*	*yes*	*yes*	*no*
Variable Ranking	*yes*	*no*	*no*	*No*	*yes*	*no*
Node Models	*c,k,n*	*c,d*	*c*	*C*	*c*	*c*
Missing Values	*mc*	*mi,s*	*mi*	*mb*	*s*	*pw*
Forests	*yes*	*no*	*no*	*no*	*no*	*No*
Bagging and Ensembles	*yes*	*no*	*no*	*No*	*no*	*No*

mb, missing value branch; c, constant model; d, discriminant model; mi, missing value imputation; k, kernel density model; l, linear splits;mc, missing value category; n, nearest neighbor model; u, univariate splits; s, surrogate splits; pw, probability weights

CHAID functions in another strategy [11]. In the case of an ordered variable P, it is splitting into 10 intervals its data values and to each interval one child node is also assigned. One child node is assigned to each value of P, in the case of P as an unordered variable. To iteratively merge pairs of child nodes, CHAID uses significant tests and Bonferroni corrections. There are two consequences for this approach. The first one is there is chance for some nodes to split into more than two child nodes. The second thing is because of this method is biased towards few distinct values for the selecting variables, it will affect the correctness of the corrections and the sequential nature of the tests.

Splits in linear combinations of all the ordered variables are allows in CART, CRUISE, and QUEST, but combinations of two variables can only split in GUIDE at a time. CRUISE and CART are using alternate splits on other variables if there is a missing values.C4.5 send each observations are for missing values in split for every branching as a probability weighting scheme. Missing values are locally imputes by QUEST. GUIDE treating as separate category for missing values. User specified misclassification costs are accepted by all other than C4.5. User specified class prior probabilities are accepted by CHAID. For predicting Q to be the class with the lowest misclassification cost, all algorithms fit a constant model to each node by default. GUIDE is fit for nearest Neighbor model and bivariate kernel density models. Bivariate linear discriminant model can optionally fit for CRUISE. Ensemble models using bagging[12] and random forest[13] techniques are also can use GUIDE techniques. All the features of algorithms are summarized in table.

CONCLUSION

As the basis of the empirical comparisons published about classification tree algorithms, on an average highest prediction accuracy belongs to GUIDE and the lowest is RPART. The limitation of RPART is, it has fever child nodes comparing QUEST, GUIDE and CRUISE even though C4.5 trees often have the most by far. The accuracy of CRUISE and QUEST are high if it is using the linear combination splits. In the case of C4.5 the computational speed is high by comparing others and is always the fastest one. RPART values is depends on the Z values. If the Z takes more than two value or the unordered variables taking more values RPART can be fast or extremely slow. By comparing piece wise constant models, GUIDE piecewise linear regression tree models have higher prediction accuracy.

1. REFERENCES

1. Fielding A, O'Muircheartaigh CA. Binary segmentation in survey analysis with particular reference to AID. The Statistician 1977, 25:17–28.
2. Messenger R, Mandell L. A modal search technique for predictive nominal scale multivariate analysis. J Am Stat Assoc 1972, 67:768–772.
3. R Development Core Team, R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2009.
4. Therneau TM, Atkinson B. RPART: recursive partitioning. R port by B. Ripley. R package version 3.1-41, 2008.
5. Loh WY, Vanichsetakul N. Tree-structured classification via generalized discriminant analysis (with discussion).
1. J Am Stat Assoc 1988, 83:715–728.
6. Kim H, Loh WY. Classification trees with unbiased multiway splits. J Am Stat Assoc 2001, 96:589–604.

7. Kim H, Loh WY. Classification trees with bivariate linear discriminant node models. J Comput Graphical Stat 2003, 12:512–530.
8. Loh WY, Chen C, Hordle W, Unwin A, eds. Improving the precision of classification trees. Ann Appl Stat 2009, 3:1710–1737.
9. Loh WY, Shih Y. Split selection methods for classification trees. Stat Sin 1997, 7:815–840.
10. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graphical Stat 2006, 15:651–674.
11. Kass GV. An exploratory technique for investigating large quantities of categorical data. Appl Stat 1980, 29:119–127.
12. Breiman L. Bagging predictors. Machine Learning 1996, 24:123–140.
13. Breiman L. Random forests. Machine Learning 2001, 45:5–32.
14. E. Alpaydin. Introduction to Machine Learning. 2nd ed. Boston: MIT Press; 2010.