A Review of Decision Tree Classification Algorithms for Continuous Variables

View the article online for updates and enhancements.

A Review of Decision Tree Classification Algorithms for Continuous Variables

S R Jiao[1], J Song[1a], B Liu[1]

[1]Capital University of Economics and Business, school of statistics, Beijing, 100070, China

[a]songjie@cueb.edu.cn

Abstract. Improving the division accuracy and efficiency of continuous variables has always been an important direction of decision tree research. This article briefly introduces the development of decision tree, focuses on the two types of decision tree algorithms for non-traditional continuous variables — based on CART and based on statistical models. Finally, the future development trend of decision tree algorithms for continuous variables is discussed.

1. Introduction

Decision tree is a classical machine learning algorithm, which is widely used because of its intuitive and clear model. Decision trees were originally applied to categorized data, but in practice, continuous variables are more common. When selecting continuous variables as independent variables, that is, as dividing nodes of the decision tree, the complexity will increase due to the large range of variables. How to improve the performance of decision trees on non-traditional continuous variables has become the focus of discussion among scholars.

In this paper, the continuous variables we discuss are all independent variables, decision trees are used for classification. Decision tree algorithms for continuous variables are mainly divided into two categories — decision tree algorithms based on CART and decision tree algorithms based on statistical models. As shown in Figure 1. When the CART-based decision tree algorithms deal with continuous variables, the division direction and division point are greedily optimized under the goal of maximizing the GINI gain, which mainly considers the degree of difference in data after division. Decision tree algorithms based on statistical model are to apply statistical methods such as statistical test to decision trees. Their division standard is the test of independence and correlation, which mainly considers the correlation between independent variables and target variables after division.
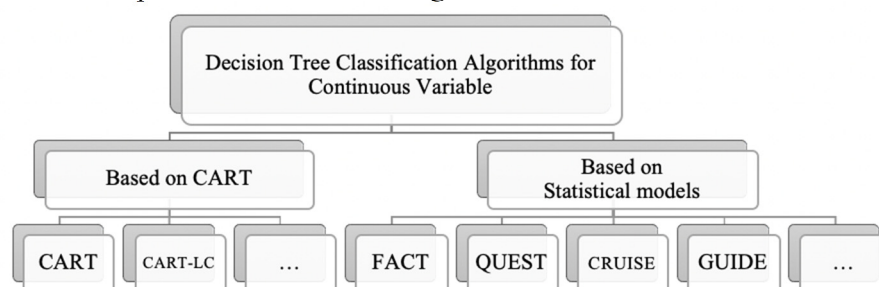

Figure 1. Hierarchical structure diagram of decision tree classification algorithms for continuous variables

2. Development process of decision tree for continuous variables

The decision tree (DT) has been widely used in classification and other fields of data mining, since it was proposed in the 1960s. DT algorithms are playing an increasingly important role with the development of machine learning, while the traditional DT is facing unprecedented challenges with the innovation of branch rules and the selection of variables. Conventional DT algorithms select a single variable for division each time, builds a tree with an unusually large structure. In addition, variable selection with greedy traversal greatly increases the complexity of the model and the difficulty of implementation. In order to improve the predictive ability and reduce the complexity of the model, in the past few decades, many scholars have continuously improved the decision tree algorithm, and introduced many variants of the decision tree algorithm.

Since Hunt (1964) proposed the framework of DT, a series of DT algorithms based on information gain have been proposed, such as ID3 and C4.5[1]. However, this kind of DT neither deal with continuous variables nor solve regression problems. CART (Classification and Regression Trees) was proposed by Breiman et al. (1984), which can build a regression tree with continuous variables [2]. The gain of sum of the squared errors is used as the node division criterion.

When the traditional DT constructs a univariate tree, the division uses only one input dimension, which can only produce a decision surface parallel to the coordinate axis.

Loh and Vanichsetakul (1988) proposed FACT algorithm (Fast Algorithm for Classification Trees), which improved the DT by applying traditional statistical methods in the selection of dividing variables and splitting points. FACT split the node with LDA and select variables with ANOVA [3].

Brodley (1992) proposed a DT algorithm based on linear perceptron, which cannot guarantee convergence, since each observation is called repeatedly to obtain optimization goal of the greatest decrease in impurity [4].

Ciampi (1991) proposed the SUPPORT algorithm (Smoothed and Unsmoothed Piecewise-Polynomial Regression Trees), which extended the dividing criteria of each node to a generalized linear model and the pruning adopted AIC. Chaudhuri et al. (1995) promoted SUPPORT to Poisson and Logistic regression [5-6].

Loh and Shih(1997) proposed QUEST algorithm (Quick, Unbiased, Efficient, Statistical Tree) based on FACT, which removes the bias by using F-test on continuous variables and the contingency table Chi-square test on categorical variables [7].

Kim and Loh (2001, 2003) proposed a classification trees with unbiased multiway splits – CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation). CRUISE algorithm implemented the selection of dividing variables and construction of dividing criteria into two steps based on FACT and QUEST. It corrected the deviation caused by the traditional decision tree using a greedy search and directly used the two-variable linear discriminant model in the leaf nodes to simplify the tree structure [8-9].

Amasyali and Ersoy (2005) proposed a new multivariable tree — Cline, which solved the construction problem of multivariable tree in the case of continuous variables with LDA and KNN [10].

Landwehr (2005) constructed logistic regression tree using Friedman's gradient descent technique (2001) [11-12].

Kim and Loh (2001) proposed GUIDE algorithm (Generalized, Unbiased, Interaction Detection and Estimation) which fine-tunes the test p-value standard based on CRUISE and applies a binary linear kernel and nearest-neighbor method to the segmentation criteria [13].

The article by Wei-Yin Loh (2014) reviews classification and regression trees in the past 50 years [14].

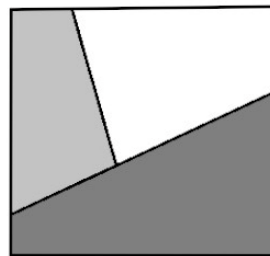## 3. Review of decision tree classification algorithms for continuous variables

3.1. Decision tree algorithm based on CART

CART (Classification and Regression Trees) is proposed by Breiman et al. (1984), it is the first algorithm to build a decision tree using continuous variables. Instead of using stopping rules, it grows a large tree
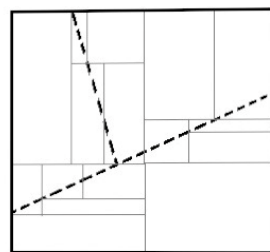
and then prunes the tree to a size that has the lowest cross-validation estimate of error. CART adopts a top-down greedy algorithm — recursive binary splitting method, where "top-down" here means that it splits the predictor variable space sequentially from the top of the tree, and each split generates two new branches. "Greedy" here means that in each step of growing a tree, the determination of the optimal classification is limited to a certain step process, rather than selecting the split points that can build a better tree in the future process.

CART algorithm is constructed using the simple "IF...THEN..." idea, and important variables can be filtered out without a particularly tedious process. But such algorithms also have disadvantages. One is that the CART algorithm has low accuracy when the boundary of data division is complex.

CART-LC is the first decision tree that uses all input dimensions to divide in each decision node, instead of using only one input dimension. It reduces the dimensionality and complexity of nodes through subset selection, and fine-tunes the weights one by one to reduce impurity. Compared with the univariate decision tree which approximately uses segmented hyperplanes parallel to the axis, CART-LC is not looking for an optimal partition for each non-leaf node, but trying to build a suitable linear classifier, so it produces an oblique boundary, which greatly simplifies the decision tree model and get better classification results, as shown in Figure 2.

(a) Distribution of data with three categories in two-dimensional space

(b) Decision boundary: CART (solid line) and CART-LC (dotted line)
Figure 2. Comparison of decision boundary between CART and CART-LC in two-dimensional space

3.2. Decision tree algorithms based on statistical model
A series of algorithms are proposed to solve the problem of continuous variables in decision tree by introducing statistical method into decision tree under the leadership of FACT. Traditional decision tree algorithms induce biases in variable selection, such as CART and C4.5, and tend to select a categorical variable with multiple values. The FACT algorithm selects variables using F-test, which can not only solve the problem of continuous variables, but also fixes the bias. FACT ranks the variables with F-test and then applies LDA to the most significant variable to split the node. However, FACT convert categorical variables into continuous variables since F-test cannot test the difference between categorical variables. First, the categorical variables are converted into dummy variables, and then they are projected into continuous variables on the largest discriminant coordinates. Finally, the projected continuous variables are Box-Cox transformed. FACT is unbiased if all the variables are continuous, but it is biased when there are categorical variables. This is because it uses LDA to convert categorical variables into continuous variables, so it is more effective to split categorical variables with LDA. In order to correct this bias, QUEST uses F-test for continuous variables and uses the contingency table

Chi-square test to select categorical variables. Different from FACT obtains trees with multiway splits, QUEST yields binary trees by merging the classes into two superclasses in each node, and obtains split point by either exhaustive search or quadratic discriminant analysis.

CRUISE is a descendent of QUEST, with multiway splits. CRUISE not only tests the main effects between outcome variable and predictor variables, but also considers the pairwise interaction between predictor variables, when the algorithm selects variables with the contingency table Chi-square test. Before testing the continuous variables, the data is divided into four groups at the sample quartiles of continuous variables, and then a contingency table with classes as rows and groups as columns is constructed. Specifically, CRUISE directly uses the bivariate linear discriminant model in the leaf nodes to simplify the tree structure [8-9].

GUIDE algorithm fine-tunes the test p-value standard based on CRUISE. Specifically, GUIDE tests the pairwise interaction only if no main effect test is significant at a Bonferroni-corrected level. The algorithm applies a binary linear kernel and nearest-neighbor method to the segmentation criteria [13].

The comparison of several decision trees for continuous variables based on statistical methods is shown in Table 1.

Table 1. The comparison of several decision trees based on statistical methods

Algorithm	Variable selection	Split point	Characteristics of the generated tree
FACT	F-test	LDA	A binary tree
QUEST	F-test, Chi-squared test	Exhaustive search, QDA	A binary tree
CRUISE	F-test, Chi-squared test, marginal test and interaction test	LDA	A multi-way tree
GUIDE	F-test, Chi-squared test	Binary linear kernel and nearest-neighbor method	A multi-way tree

4. The conclusion

At this stage, decision tree algorithms have been applied in many fields. With the in-depth research on decision trees for continuous variables, decision trees are also being favored by more fields. However, the expansion of variable range will make the complexity problem more obvious, the simplicity and accuracy of decision trees for continuous variables still need to be improved. At present, the research of decision tree for continuous variables is developing in several directions: as for the accuracy, multivariate decision trees are becoming more and more popular (such as oblique decision tree). As for the efficiency, the development of decision tree is more integrated in order to process big data (such as Adaboost). Otherwise, decision tree is combined with other algorithms to improve itself (such as methods based on statistical model). How to deal with and solve the problem of precision and complexity requires further development and research.

References:
[1] Hunt E B. Concept Learning: An Information Processing Problem[J]. The American Journal of Psychology, 1964, 77(2).
[2] Breiman L, Friedman J, Olshen R, et al. Classification and Regression Trees[M]. New York: Chapman & Hall, 1984.
[3] Loh W Y, Vanichsetakul N. Tree-Structured Classification via Generalized Discriminant Analysis[J]. Journal of the American Statistical Association, 1988, 83(403).
[4] Brodley C E, Utgoff P E. Multivariate versus Univariate Decision Trees[M]. 1992.
[5] Ciampi A. Generalized regression trees[J]. Computational Statistics and Data Analysis, 1991,

12(12).

[6] Chaudhuri, P，Huang, M. C，Loh, W. Y, et al. Piecewise-Polynomial Regression Trees[J]. Statistica Sinica, 1994, 4(1).

[7] Loh W Y, Shih Y S. Split selection methods for classification trees.[J]. Statistica Sinica, 1997, 7(4).

[8] Kim H, Loh W Y. Classification Trees With Unbiased Multiway Splits[J]. Journal of the American Statistical Association, 2001, 96(454).

[9] Kim H, Loh W Y. Classification Trees With Bivariate Linear Discriminant Node Models[J]. Journal of Computational and Graphical Statistics, 2003, 12(3).

[10] Amasyali M F, Ersoy O. Cline: new multivariate decision tree construction heuristics[C]. Istanbul, Turkey: Icsc Congress on Computational Intelligence Methods & Applications, 2005.

[11] Landwehr N, Hall M, Frank E. Logistic Model Trees[J]. Machine Learning, 2005, 59(1-2).

[12] Friedman J H. Greedy Function Approximation: A Gradient Boosting Machine[J]. Annals of Statistics, 2001, 29(5):1189-1232.

[13] Loh W Y. Improving the precision of classification trees[J]. Ann Appl Stat, 2009, 3(4):1710–1737

[14] Loh W Y. Fifty years of classification and regression trees (with discussion)[J]. International Statistical Review, 2014, 34.