Available online at www.elixirpublishers.com (Elixir International Journal)

Bio Technology

Elixir Bio Tech. 40 (2011) 5430-5432

ISSN: 2229-712X

Classification trees using new criteria for two or more categories

Qamruz Zaman[1], Muhammad Azam[2], Muhammad Iqbal[1], Shah Khusro[3] and Karl Peter Pfeiffer[4]

[1]Department of Statistics, University of Peshawar, Pakistan

[2]Department of Statistics, Forman Christian College, Lahore, Pakistan

[3]Department of Computer Science, University of Peshawar, Pakistan

[4]Department of Medical Statistics, Informatics and Health Economics, Innsbruck, Austria.

ARTICLE INFO

Article history:

Received: 20 August 2011;

Received in revised form:

25 October 2011;

Accepted: 5 November 2011;

Keywords

Classification trees,

Node splitting methods,

Misclassification rates,

Deviance.

ABSTRACT

Different node splitting criteria are available for the construction of classification and regression trees. Two of these criteria i.e. Gini and Twoing criterion have been implemented in CART (Salford systems, 1995), Entropy function in C4.5 (Quinlan, 1993) etc. These criteria provide similar results especially for the small number of categories but not for large. To overcome this problem, we proposed a new node splitting method for the construction of classification trees. The performance of the new method is compared with conventional methods using two real life datasets and it is observed that the new method gives better results in terms of misclassification and deviance.

Introduction

During the last two decades, the use of classification and regression tree (CART) analysis has been increased. Classification and regression trees techniques are better choices for the analysis of complex and high dimensionality datasets. CART procedures are used in different fields such as medical, engineering, marketing, sociology and economics research. CART use to predict the categorical and continuous variables. Its algorithm generates simple binary nodes from root node and finally one gets a binary tree structure and uses it to classify the instances with respect to their categories. It was introduced by group of well known researchers Breiman, Friedman, Olshen and Stone in 1984.

Finding a best node splitting point is one of the most important tasks while constructing classification as well as regression trees. There are different nodes splitting methods available in the literature for the construction of binary classification trees. Some of these are the Gini index and Twoing rule used by Breiman et al. (1984), the Entropy method used by Ciampi et al. (1987), Clark and Pregibon (1992) and Quinlan (1993). The approach used for split selection in classification trees is to search through all possible splits generated by predictor variables. A splitting criterion is defined to evaluate those splits and select the one which maximizes the criterion value and finalize the samples into corresponding descendent nodes.

The predictor variables in a binary classification tree may or may not be categorical variables. There are $L-1$ (L is the number of distinct values) possible number of split points for a continuous predictor and $\left(2^{k-1}-1\right)$ (where k is the possible number of categories) for a categorical predictor. The chosen points split the instances, where the predictor variable satisfies the condition $X \le x$, if X is numerical or $X \varepsilon S$, if X is categorical. The value of x or subset S (a subset from all possible combination) is chosen which maximizes the criterion value.

In this paper, we proposed new criteria for the selection of best split point. The new proposed criteria can be used for the construction of classification as well as regression trees.

Tree growing procedure has been discussed in section 2. We described some of the commonly used node split selection criteria with their properties in section 3. New proposed criteria for the selection of best split point have been discussed in section 4. The performance of proposed criteria as compared to traditional approaches has been shown in section 5 using two real life datasets and the results have been discussed in section 6.

Tree Growing Procedure

There are four elements which are required to construct a tree.

A set b_Q of binary questions? Questions may be about sex (male/ female), marital status (married/ unmarried), a disease (high risk/ low risk) etc. when X is categorical. The values which are less than or equal to any specified value x goes to L.H.S node otherwise R.H.S node, when X is continuous.

The second most important step in the tree construction is the application of goodness of split criteria and its numerical measurement. A goodness of split criterion $\delta(s, t)$ can be computed for any split s of any node t.

A stop-splitting rule. In this step one has to decide when and where node splitting should be stopped.

A particular category is assigned to each terminal node on the basis of majority vote.

Node Splitting Approaches

The concept of node impurity has been explained by number of authors like Breiman et al. (1984), Berzal et al. (2003) etc. If the instances of a dataset falls into a particular node belongs to the same category and none of the instance

Tele:

E-mail addresses: ayanqamar@gmail.com

belongs to the other category, we say that there is no impurity or the node is pure. The impurity of a node increases and reaches its maximum, as the proportion of instances belong to different categories becomes closer and closer to each other in any node. Almost all the node splitting approaches yields the similar results, although the selected approach for the construction of classification as well as regression trees may affect the misclassification rates, deviance etc. (Berzal et al., 2003). Few of the most commonly used impurity based node split selection approaches are:

Gini Index

It was introduced by Leo Breiman in 1984 and is implemented in CART (Salford Systems, 1995). Gini index for a node t is given by

$$i(t)_{Gini} = 1 - \sum_{j=1}^{J} P_j^2,$$

where j is total number of categories, P_j is the proportion of jth category in a node t, such that $\sum_{j=1}^{J} P_j = 1.$

For the left and right descendent nodes we have Gini index function as

$$i(t_L)_{Gini} = 1 - \sum_{j=1}^{J} P_{j|L}^2 \quad \text{and} \quad i(t_R)_{Gini} = 1 - \sum_{j=1}^{J} P_{j|R}^2,$$

where t_L and t_R are two left and right descendent nodes, $P_{j|L}$ and $P_{j|R}$ are the proportions of jth category on left and right descendent node respectively, such that $\sum_{j=1}^{J} P_{j|L} = \sum_{j=1}^{J} P_{j|R} = 1.$

After computing the values of node impurity function (1) and (2), the measure of "Goodness of split" which is usually known as criterion value for predictor variable x, the chosen split s at node t is

$$\delta(x,s,t)_{Gini} = i(t)_{Gini} - P_L\, i(t_L)_{Gini} - P_R\, i(t_R)_{Gini}, \tag{3}$$

where $P_L = \frac{N_L}{N}$ and $P_R = \frac{N_R}{N}$, such that $P_L + P_R = 1.$ P_L and P_R are the proportions of instances on the left and right descendent node.

Entropy Function

Quinlan (1983) proposed an evaluation function for the measurement of node impurity and is implemented in C4.5 (Quinlan, 1993), which is based on classical formula

$$i(t)_{Quin} = -\sum_{j=1}^{J} P_j \ln P_j,$$

For the left and right descendent nodes, Entropy function is

$$i(t_L)_{Quin} = -\sum_{j=1}^{J} P_{j|L} \ln\left(P_{j|L}\right) \quad \text{and} \quad i(t_R)_{Quin} = -\sum_{j=1}^{J} P_{j|R} \ln\left(P_{j|R}\right).$$

The measure of goodness of split using Entropy function is

$$\delta(x,s,t)_{Quin} = i(t)_{Quin} - P_L\, i(t_L)_{Quin} - P_R\, i(t_R)_{Quin}.$$

The interpretation of all the symbols is same as discussed in section 2.1. An impurity based non negative function $i(t)$ computes the impurity of a given node of a tree. Suppose a classification problem with j different categories whose probabilities are P_1, P_2, K, P_J respectively. Any function $i(t)$ is said to be an impurity function if it possesses the following properties.

(i) $i(t)$ reaches its maximum only at the point $\left(\frac{1}{J}, \frac{1}{J}, \text{L}, \frac{1}{J}\right)$ It means that the node impurity function achieves its maximum value, when same number of instances fall into a node belongs to J number of categories.

(ii) $i(t)$ reaches its minimum only at the point $(1, 0, \text{L}, 0), (0, 1, \text{L}, 0), \text{L}, (0, 0, \text{L}, 1)$ i.e. node impurity is minimum, when all the instances of a node belong to a single category.

(iii) $i(t)$ is a symmetric function of $(P_1, P_2, \text{K}, P_J)$

Twoing Criteia

Another splitting criterion which directly measures the goodness of split value is introduced by Leo Breiman in 1984. The Twoing rule has also been implemented in CART (Salford systems, 1995). This is a measure of the difference in probability that a category appears in the left descendant rather than the right descendant node. The criteria select a best splitting value, which maximizes the function

$$\delta(x,s,t)_{Twoing} = \frac{P_L P_R}{4}\left[\sum_{j=1}^{J}\left|P_{j|L} - P_{j|R}\right|\right]^2. \tag{2}$$

The main objective is to get a probability that a category j instance goes to the left as different as possible from the probability that it goes to the right. The factor $P_L P_R$ is designed to favor relatively even splits. This factor takes a maximum value of 0.25 when $P_L = P_R = 0.5$, it declines if any of the proportion is close to 0 or 1.

Proposed Node Splitting Approach

In section 2, we discussed two impurity based node splitting criteria and a criteria which is directly used for the measurement of goodness of split. We introduced the idea of using the proportion of existing categories at the top node, which yields similar or improved results as compared to traditional node splitting criteria. The proposed node splitting criteria for the construction of classification trees is

$$\delta(t)_{New} = \frac{1}{4}\frac{P_L}{P_R}\sum_{j=1}^{J}\left|\left(P_j - P_{j|L}\right) + \left(P_j - P_{j|R}\right)\right| \quad \text{if } P_L \le P_R$$

$$\delta(t)_{New} = \frac{1}{4}\frac{P_R}{P_L}\sum_{j=1}^{J}\left|\left(P_j - P_{j|L}\right) + \left(P_j - P_{j|R}\right)\right| \quad \text{else} \tag{4}$$

where j is total number of categories, P_j is the proportion of jth category in a node t, such that $\sum_{j=1}^{J} P_j = 1.$ (5)

$\delta(t)_{New}$ satisfies all the desired properties of node splitting approaches. (6)

Therefore, to select the best split s^* at each node is provided by splitting variable x^* that maximizes the goodness of split measure.

$$\delta(x^*, s^*, t)_{New} = \text{Maximum of } \delta(x,s,t)_{New}$$

Empirical Study

We compared the performance of proposed node splitting approach with two traditional and most commonly used approaches i.e. Gini index and Twoing criteria. For this purpose, we used two well known real life datasets available in R. We did programming in R for the evaluation of performance of all criteria.

Iris Plants Database

This dataset consist of 150 plants. Four of the predictors have been used for the classification of species of the plants (response variable). Three of the categories of species with equal number of plants are

1 setosa
2 versicolor
3 virginica

We constructed three different tables based on the two datasets by using the misclassification criteria, number of terminal nodes and deviance. First row of each table shows the results of first dataset. Here we get the same misclassification rate while the results of terminal nodes favor the new method. This fact is also shown in Fig. 1 and 2.

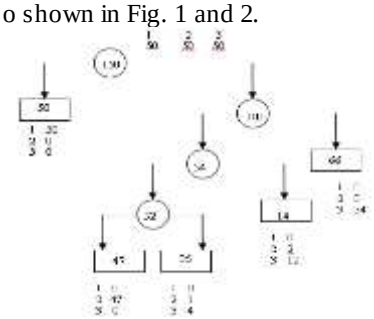

Fig 1 Iris tree produced by using the new criteria. The values under each terminal node give the number of each category

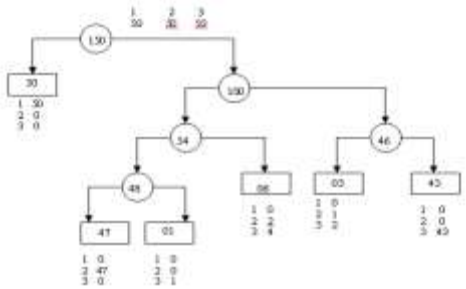

Fig. 2. Iris tree produced by using the Gini and Twoing criteria. The values under each terminal node give the number of each category

Pima.tr Database

Dataset consists of 200 female patients at least 21 year old of Pima Indian heritage. Seven of the predictors have been used for the classification of status of the patient (response variable). Two of the categories with ratio 132:68 of status are

0 tested negatively for diabetes
1 tested positively for diabetes

The number of misclassified instances of new method is 16 which is considerably smaller than the Gini and Twoing. The result of number of terminal nodes does not favor the new approach, which may be due to nonparametric behavior of the approach (Fig. 3 to 4). The deviance row of Pima.tr dataset reveals that the value of our approach is smaller than conventional methods.

The classification tree has been constructed by using three node splitting approaches for two datasets. Number of misclassified instances, number of terminal nodes and deviance has been calculated for each classification tree.

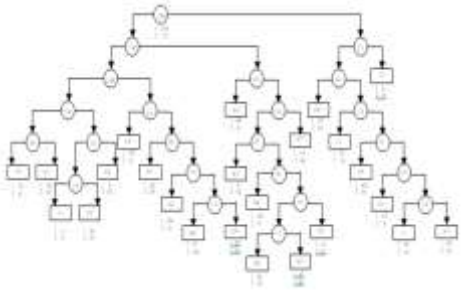

Fig. 3. Pima.tr tree produced by using the New criteria. The values under each terminal node give the number of each category

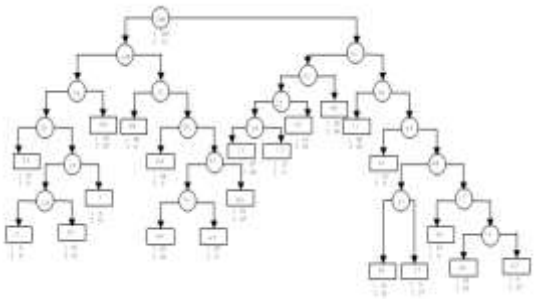

Fig. 4. Pima.tr tree produced by using the Gini and Twoing criteria. The values under each terminal node give the number of each category

Discussion

We proposed a new node splitting criteria for the construction of classification trees. The proposed criteria gives the less number of misclassified instances and deviance for Pima.tr dataset as compared to traditional approaches and equal number of misclassified instances but more deviance for the iris dataset. Therefore we can conclude that the proposed criterion is better as compared to traditional approaches in producing classification trees.

References

Berzal, F., Cubero, J. C., Cuenca, F and Martin-Bautista, M. J. (2003) On the quest of easy to understand splitting rules. Data & Knowledge Engineering, 44, 31-44.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) Categoryification and regression trees. Wadsworth International Group, Belmont, CA.

CART. Salford Systems, San Diego, California, USA, 1995.

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, 23-50, D. Reidel, New York.

Clark, L. A. and Pregibon, D. (1992) Tree based models. In: J. M. Chambers and T. J. Hastie (eds) Statistical models in S, Wadsworth and & Brooks/Cole, Pacific Grove, CA.

Quinlan, J.R. (1983). Learning efficient categoryification procedures and their application to chess endgames. In R.S. Michalski, J.G. Carbonell & T.M. Mitchell, (Eds.), Machine learning: An artificial intelligence approach. Palo Alto: Tioga Publishing Company.

Quinlan, J. R. (1993) Programs for machine learning. Morgan Kaufmann Publishers, San Mateo, California.