Expert Systems With Applications 69 (2017) 21–28

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Regression trees for multivalued numerical response variables

Antonio D'Ambrosio[a,*], Massimo Aria[a], Carmela Iorio[b], Roberta Siciliano[b]

[a] Department of Economics and Statistics, University of Naples Federico II, Italy
[b] Department of Industrial Engineering, University of Naples Federico II, Italy

ARTICLE INFO

Article history:
Received 3 May 2016
Revised 10 October 2016
Accepted 11 October 2016
Available online 15 October 2016

Keywords:
Regression trees
Multivalued variables
Modal variables
Earth mover distance
Mallows distance

ABSTRACT

In the framework of regression trees, this paper provides a recursive partitioning methodology to deal with a non-standard response variable. Specifically, either multivalued numerical or modal response of the type histogram will be considered. These data are known as symbolic data, which special cases are classical data, imprecise data, conjunctive data as well as fuzzy data. In spite of pre-processing data in order to deal with standard regression tree methodology, this paper provides, as main contribution, a definition of the impurity measure and of the splitting criterion allowing for building the regression tree for multivalued numerical response variable. We analyze and evaluate the performance of our proposal, using simulated data as well as a real-world case studies.

1. Introduction

Starting from the pioneer work of Morgan and Sonquist (1963), regression trees have been extensively developed in the CART (Classification and Regression Trees) methodology by Breiman, Friedman, Olshen, and Stone (1984) and justified within the Statistical Learning Paradigm outlined by Hastie, Friedman, and Tibshirani (2001) and defined by Vladimir and Vapnik (1995). Main issue of the tree-based approach to a regression problem is to deal with complex data, namely non linear dependence relations between a response numerical variable and a set of predictors of different typology (qualitative and/or quantitative), distribution free assumption, non parametric estimation, huge data or very large sample size. CART approach to regression trees consists into two main procedures: first, the definition of a recursive partitioning method to build up an exploratory binary tree to understand the dependence relationships among the variables; second, the identification of the suitable decision tree based rule for assigning a response value to new cases for that only the predictors measurements are known. So far, proposals in literature have focus the attention on alternative splitting criteria (Galimberti, Pillati, & Soffritti, 2007; Mola & Siciliano, 1992; Siciliano & Mola, 1996; 2000), assessment of decision rules (Breiman, 1996; Freund & Schapire, 1997; Vezzoli, 2011), strategy of analysis based on the complementary use of regression trees with parametric models (Siciliano, Aria, & D'Ambrosio, 2008), the application of regression trees to special fields such data-editing (D'Ambrosio, Aria, & Siciliano, 2012), etc. All these proposals have always considered standard data, either categorical or numerical for the predictors and obviously numerical for the response. Nowadays, a lot of interest has received nonstandard data types, such as fuzzy data, imprecise data, conjunctive data, which can be also related to a more general definition of symbolic data (Bock & Diday, 2000).

Literature on symbolic data has included several contributions involving also tree-based methods (Limam, Diday, & Winsberg, 2003; Mballo & Diday, 2005), other than other approaches such as factor analysis (Cazes, Chouakria, Diday, & Schektman, 1997), regression analysis (de Carvalho, Neto, & Tenorio, 2004; Neto, de Carvalho, & Freire, 2005), clustering methods (de Carvalho, Brito, & Bock, 2006; Cozza, Guarracino, Maddalena, & Baroni, 2011; Irpino & Verde, 2008; Irpino, Verde, & De Carvalho, 2014). As a matter of fact, tree-based methods have been used with interval data as predictors. Main idea has been to convert interval data into a standard data type upon a suitable pre-processing. The pre-processing consists either in taking either the lower or the upper bound of each interval, as well as the mean value of each interval, such to proceed with a standard regression tree growing procedure. With respect to CART the main difference was to consider as impurity measure the Kolmogorov–Smirnov measure.

One of the first works dealing with regression trees for multivariate response variable was introduced by Segal (1992), in which longitudinal data were used as multivariate response variable.

Regression trees with probability density function as a response variable have been introduced in the framework of functional data analysis. In this case splitting criteria are based on the sum of

* Corresponding author.
E-mail addresses: antdambr@unina.it (A. D'Ambrosio), aria@unina.it (M. Aria), carmela.iorio@unina.it (C. Iorio), roberta@unina.it (R. Siciliano).

http://dx.doi.org/10.1016/j.eswa.2016.10.021
0957-4174/

dissimilarities between the densities or deviations of the densities from their mean (Lane & Robinson, 2011; Nerini & Ghattas, 2007).

More recently, other authors have moved into another direction of research, that is to define suitable tree-based methods dealing directly with symbolic data upon suitable definition of splitting criteria to grow classification trees where predictors are symbolic data and/or functional data (Siciliano, Aria, D'Ambrosio, & Cozza, 2016).

This paper provides a recursive partitioning methodology when the response variable can be either multivalued numerical or modal response of the type histogram. As main result, we will define the impurity measure and the splitting criterion allowing for building the so-called regression tree for multivalued numerical response. The remainder of the paper is organized as follows: in Section 2 we briefly recall the regression trees. In Section 3 we introduce the framework of symbolic data. Section 4 is devoted to introduce and explain our proposal. Section 5 is dedicated to both a simulation study and applications on real problems. In Section 6 we close the paper with some concluding remarks.

2. Standard regression trees

Data can be hierarchically organized in a connected and oriented graph, the so-called tree, characterized by a set of linked nodes, in which any two nodes are connected by exactly one simple path, the starting-node is the *root* and the end-nodes are the *leaves* also known as *terminal nodes*, all the others are *internal nodes*. Main idea is to associate to the root node the starting sample of cases such that the tree structure provides a partition of the starting sample into H groups corresponding to the total number of leaves of the tree.

In CART framework, binary regression trees can be built up as a recursive binary partitioning of n cases into two subgroups which are internally homogeneous and externally heterogeneous with respect to the numerical response variable. In other words, at any internal node of the tree, a splitting criterion is based on the search of that split that generates the most different descendant nodes in terms of mean (or median) value of the response variable. Predictors play the role to generate the set of candidate splitting variables to be considered in the splitting criterion. Specifically, categorical predictors with m categories generate $2^{m-1}-1$ splitting variables if not ordered and $m-1$ if ordered; numerical predictors with m distinct values also generate $m-1$ splitting variables. The splitting criterion is defined as the maximum decrease of impurity when passing from a parent node to the two children nodes, upon a suitable definition of the impurity measure (i.e., the variation or the deviance). To each leave of the tree is assigned as label the mean value of the response distribution within the terminal node. In this way, it is possible to understand the dependence relationships of the response variable on the predictors by analyzing the different tree paths starting from the root node until the different leaves. Regression trees can be also considered for prediction of a new object for that only the predictors' measurements are known. This object can slide down the tree until falling into a terminal node with a given label value assigned. The quality of this prediction can be evaluated in terms of the mean squared error estimates considering either the learning or training sample, or the test sample, alternatively cross-validation. CART adopts the following strategy: first, the recursive partitioning provides the maximum expanded tree such that leaves cannot be further split (i.e., threshold value to be fixed on the percentage of cases within the leave, alternatively on the decrease of impurity and so on); second, the pruning procedure provides to identify a set of nested decision trees upon removing at turn the weakest link on the basis of the trade off between the increase in error prediction and the decrease in the tree size complexity; third, a selection of the most accurate decision tree to be considered for prediction on new cases on the basis of an independent test sample of cases as well as on cross-validation sample. More recent methods improve the accuracy of the decision tree based rules by ensemble methods (Breiman, 1996; Freund & Schapire, 1997).

3. Symbolic data

Symbolic data were defined by Diday (1993) and formally definitively formalized by Billard and Diday (2003). The data descriptions of the units are called symbolic when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data need more complex data tables called symbolic data tables because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical values.

A classical value or realization for the random variable Y_j, $j = 1,\dots,p$, on individual $i = 1,\dots,n$, will be denoted by x_{ij} if it is a classical variable, and $Y_j(i) = \xi_{ij}$ if it is a symbolic variable (Billard & Diday, 2003).

Symbolic data are a hypercube in the p-dimensional space or Cartesian product of distributions in which a generic variable is so defined:

Let Y_j the domain of Y_j.

1. **Interval-valued** variable is a variable which assumes two real values for each observation. These values represent the boundaries of an interval.
   $\xi_{ij} = [a, b]$ where a and b are two numerical values with $a < b$.
2. **Multivalued** variable is one whose possible value takes one or more values from the list of the value in its domain Y.
   The complete list of possible values in Y is finite so that $\Upsilon_j = \{a, b, c, d, e\}$ where i.e. $\xi_{ij} = \{a, b\}$. A multivalued variable is categorical when its domain is defined by qualitative attributes on the contrary it is called quantitative multivalued variable when Y is composed by numerical values.
3. **Modal** variable or multi-state variable is a variable with frequency, probability, or weight attached to each specific value in the data.
   $\xi_{ij} = \{U_j(i), \pi_{ij}\}$ for $i \in \Omega$ where π_{ij} are non-negative measures or a distribution on the domain Y_j of possible observation values and $U_j(i) \subseteq \mathrm{Y}_j$ is the support of π_{ij}. More generally, ξij may be a histogram, an empirical distribution function, a probability distribution, a model, or so on.

Following this definition, classical data can be treated as a special case of symbolic data in which $\xi_{ij} = [a, a]$ where a is an qualitative or quantitative attribute.

4. Distance-based impurity criterion

In this paper we consider the case in which the response variable can be either multivalued numerical or modal variable of the type histogram. The impurity criterion must assure that each split returns children nodes purest than their father node. What we need is a numerical evaluation of the comparison of such symbolic variables. We need a distance-based impurity criterion (D'Ambrosio & Heiser, 2016; De'ath, 2002).

4.1. The earth mover distance

Roughly speaking, the Earth Mover Distance (Rubner, Tomasi, & Guibas, 1998) can be assumed as the minimal amount of work needed to transport earth or mass from one position (properly spread in space) to the other