Computational Statistics and Data Analysis 54 (2010) 1516–1524

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Binary trees for dissimilarity data

Raffaella Piccarreta *

Department of Decision Sciences, Bocconi University, Milan, Italy

ARTICLE INFO

Article history:
Received 27 June 2009
Received in revised form 22 December 2009
Accepted 22 December 2009
Available online 14 January 2010

Keywords:
Dissimilarity matrix
Classification and regression trees
Binary segmentation
Multivariate responses
Perception data
Ecological data

ABSTRACT

Binary segmentation procedures (in particular, classification and regression trees) are extended to study the relation between dissimilarity data and a set of explanatory variables. The proposed split criterion is very flexible, and can be applied to a wide range of data (e.g., mixed types of multiple responses, longitudinal data, sequence data). Also, it can be shown to be an extension of well-established criteria introduced in the literature on binary trees.

1. Introduction

There are many practical applications when the analysis of dissimilarity data is of interest. Examples are subjective ratings of the differences between 'items' (e.g., goods, political candidates) as expressed by one or more observers (e.g., customers, voters), or differences between gene-expression profiles or texts. Alternatively, dissimilarities can be computed from a set of variables (binary, nominal, ordinal, numerical, or a combination of these; Kauffman and Rousseeuw, 1990) or on the basis of more complex 'objects', as for example time series (see e.g., Kakizawa et al., 1998; Chouakria and Nagabhushan, 2007; D'Urso, 2000), longitudinal data (Segal, 1992), or life courses represented as sequences (Abbott, 1995; Elzinga, 2003, 2005).

Pairwise dissimilarities between cases are arranged in the so-called dissimilarity matrix, **D**, which is usually analyzed using Cluster Analysis (Everitt et al., 2009) or Multidimensional Scaling (MDS; Borg and Groenen, 2005). Cluster analysis attempts at finding groups of cases as homogeneous as possible. MDS estimates latent factors which can be considered as responsible for the observed dissimilarities.

In this work we are interested to study the relation between dissimilarities in **D** and a set of explanatory variables. The problem is relevant both from an exploratory and from a predictive point of view. The first aspect is particularly important when one considers perceived dissimilarities and wants to evaluate their dependency upon some objective characteristics of cases. For example, Bergmann Tiest and Kappers (2006) study the relation between haptic perception of materials and their roughness and compressibility. An example of the prediction problem can be found in McVicar and Anyadike-Danes (2001), who consider the sequences of the monthly activities (school, training, employment, unemployment) experienced by young people from Northern Ireland after the end of compulsory school. Dissimilarities between sequences (calculated using Optimal Matching Analysis; Abbott, 1995) are related to a set of background family and individual characteristics (gender, religion, performance at school, working conditions of parents). The goal is to assess whether certain background characteristics are more likely to lead to sequences with particular features, for example dominated by unemployment.

* Corresponding address: Department of Decision Sciences, Bocconi University, via G. Röntgen 1, 20136, Milano, Italy. Tel.: +39 0258365659.
E-mail address: raffaella.piccarreta@unibocconi.it.

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.12.011

The problem of analyzing the dependency of dissimilarities on explanatory variables is thus not new in the literature, and some interesting approaches are described below.

McVicar and Anyadike-Danes (2001) suggest to cluster cases on the basis of dissimilarities, and to relate them to the explanatory variables through a multinomial model. On the one hand, this is reasonable if clusters are highly homogeneous, so that it makes sense to assign the same response value to all the cases in the same cluster. On the other hand, multinomial model can perform unsatisfactory when some levels of the response have low frequency. Hence, a trade-off may exist between considering a small number of clusters so as to avoid estimation and predictions problems (as suggested by McVicar and Anyadike-Danes, 2001) and achieving a satisfactory level of homogeneity possibly considering a high number of clusters.

Alternatively, MDS can be applied. To each case a vector of MDS factor scores is thus associated, generally permitting to distinguish between cases which are similar (and hence clustered together) but not identical one to another. The relation between MDS factors and the explanatory variables could be analyzed, for example through linear models. Nevertheless, factorial techniques are usually considered (see e.g., Green et al., 1989; Cox and Cox, 1994; Bergmann Tiest and Kappers, 2006); external variables are projected onto the MDS factorial space, to visualize their relation with MDS factors. Since projections are linear combinations of the involved variables, a linear relation is implicitly assumed.

The previous proposals proceed in a sequential fashion. **D** is first simplified through clusters or MDS factors, which are then related to the explanatory variables. This simplification has the advantage that it ignores some of the noise possibly characterizing dissimilarities. It also provides an 'estimation' of the response/s underlying dissimilarities. However, explanatory variables have no influence on clusters or MDS factors. As Kiers et al. (2005) point out, this can be a drawback 'because the MDS (*or the cluster*) solution may sometimes be based primarily on other relations in the data, and therefore, the MDS (*or the cluster*) solution may (to some extent) miss the relation of the observed dissimilarities with the external variables' (italics are ours). An approach simultaneously extracting MDS factors and projecting external variables onto the MDS space can be found in Heiser and Meulman (1983). Kiers et al. (2005) refine this proposal, and provide an algorithm which simultaneously determines clusters and MDS factors, and relates *clusters centroids* to the external variables through a linear model. These last techniques are intended to improve the interpretation of MDS factors, which can be difficult especially when perception data are considered. Hence, the exploratory analysis of the relation with the external variables is prevailing here. Actually in Kiers et al. (2005) cluster centroids rather than cases are related to external variables, and no rules are provided to assign new cases to clusters. In this sense, the *prediction* problem is not explicitly taken into account.

More importantly, all the considered methods inspect the dependency of the (possibly simplified) response/s upon the explanatory variables under the assumption of linearity, which in some applications (for example when perceived dissimilarities are considered) may be too far fetched.

Our method differs from those illustrated above under different perspectives. To study the relation between dissimilarity data and explanatory variables, in Section 2 we introduce an extension of the Classification and Regression Trees (CART) procedure (Breiman et al., 1984). We thus refer to a tree-predictor, obtained by recursively partitioning the sample into more and more homogeneous (with respect to the 'response') sub-sets, through a sequence of binary splits taking the form of conditions on the values of the explanatory variables. *Predicted* homogeneous groups of cases are consequently obtained *on the basis of the explanatory variables*, thus simultaneously simplifying and explaining dissimilarities. Tree-predictors are not based upon the assumption of linearity (or of other analytical relations). Also, the prediction problem is explicitly taken into account. A prediction is assigned to each node, and new cases can be assigned to the nodes on the basis of the explanatory variables. Also, validation procedures are considered to prevent over-fitting of the obtained tree to the data it is based upon.

Our procedure can thus be considered as an alternative method to relate dissimilarities to external variables. Furthermore, in Section 2 we also show that our method can be regarded as an extension of some popular and well-established criteria, used to grow trees for single or multiple responses. Under this point of view, our proposal can be regarded as a generalized (CART) criterion which can be used to grow trees when dealing with complex responses, provided that pairwise dissimilarities can be suitably measured. This permits to take advantage of the impressive number of contributions introduced in the literature to properly measure dissimilarities on the basis of complex data.

An application of our method to ecological data is discussed in Section 3. Section 4 summarizes and concludes, outlining directions of future research.

2. Trees for dissimilarity matrices

CART, introduced by Breiman et al. (1984), is one of the most popular methods to build trees. Tree-structured predictors are defined by recursively partitioning a sample into sub-sets, called *nodes*, through a sequence of binary splits, taking the form of conditions on the values of the explanatory variables.

The nodes obtained throughout the sequence of splits should be as homogeneous as possible with respect to the considered response. Hence, an *impurity* function has to be defined to measure the heterogeneity within a node g on the basis of the dissimilarities in **D**. Here we consider the average of the (squared) dissimilarities within g:

$$i(g|\mathbf{D}) = \frac{1}{n_g^2} \sum_{i \in g} \sum_{h \in g} d^2(i, h|\mathbf{D}), \tag{1}$$

where n_g is the size of g and $d(i, h|\mathbf{D})$ is the (i, h)th entry of **D**.