Abstract ampliado

RESUMEN AMPLIADO

Title: Decision Trees with Interactive Basis Functions for Non-Orthogonal Decision Boundaries

Authors and e-mail of them: Antonio Páez; Fernando López (Fernando.lopez@upct.es) ; Manuel Ruiz; Máximo Camacho

Department: Métodos Cuantitativos e Informáticos

University: McMaster University, Universidad Politécnica de Cartagena, Universidad de Murcia

Subject area: ***Big Data en Ciencia Regional***

Resumen: *(mínimo 1500 palabras)*

Decision Trees (DTs) are a popular machine learning technique used both for regression and classification purposes (Loh 2011; James et al. 2013). A DT is trained by means of a training dataset that provides a set of independent variables (or features) used to create recursive partitions of the decision space. This is achieved by locally optimizing at each step a loss function that depends on the type of problem (i.e., regression or classification) and/or the specific implementation of the algorithm (i.e., an entropy function for C4.5 and a gini index for CART; see Loh 2011). Decision Trees find applications in a variety of domains, including, inter alia, transportation (e.g., Ghasri, Rashidi, and Waller 2017), physical geography (e.g., Praskievicz 2018), engineering (e.g., Bektas, Carriquiry, and Smadi 2013; Suhail, Denton, and Zwiggelaar 2018), and environmental sciences (e.g., Choubin et al. 2018). There are several characteristics that make DTs an appealing modeling approach. Notably, DTs are more intuitive than linear/logistic regression (James et al. 2013, 315) and have much greater interpretability than, for instance, artificial neural networks and support vector machines (Yang et al. 2017, 354). In addition, in some settings DTs provide a reasonable heuristic for human decision making (James et al. 2013, 315). Finally, although no single technique can be expected to be uniformly superior in every case, the performance of DTs has been

shown to be competitive with, and in some cases superior to, alternatives such as linear regression, logistic regression, support vector machines, and artificial neural networks (e.g., Kurt, Ture, and Kurum 2008; Choubin et al. 2018; Yang et al. 2017). One characteristic of DTs as conventionally implemented, is that partitions of the variable space are usually done orthogonally to the features. In this way, the partitions are a set of rectangular p-dimensional prisms,or hyperboxes. While this is done to reduce the search space of the algorithm, it has the downside that it may fail to find appropriate partitions, and in some extreme cases, to find any partitions at all. In this case, the performance of the algorithm tends to be mediocre. Accordingly, a number of proposals have aimed at ameliorating this situation by inducing oblique partitions (e.g., Murthy, Kasif, and Salzberg 1994; Wickramarachchi et al. 2016; Cantu-Paz and Kamath 2003). The objective of this paper is to introduce a novel strategy for non-orthogonal partition of variable space in the training of DTs. The approach is based on the use of interactive basis functions (IBFs). We will show that oblique partitions result as a particular case of an IBF. Moreover, depending on the basis function selected, non-linear partitions are also possible. The modeling strategy proposed in this paper is attractive because the basis functions can be precalculated and then used as an input to any decision tree algorithm. Since only the inputs to the algorithm change, this implies that 1) the underlying algorithm is not changed and therefore any existing DT software can be used; and 2) basis functions can 2 be used in many existing implementations of DTs, including evolutionary trees, bagging, and boosting. The structure of the paper is as follows. In the background section we first review some technical aspects of decision trees and motivate the problem. This is followed by a discussion of [basis functions][Interactive Basis Functions] and how they can be employed to induce oblique linear and non-linear partitions. Next, we discuss some practical aspects of implementing IBFs before conducting a benchmarking experiment the assess the performance or DTs with IBFs by means of a set of publicly available empirical datasets. The results indicate that inducing oblique and/or non-linear partitions using basis functions can improve the performance of the technique and/or produce more parsimonious models. We then illustrating the application of IBFs by means of three empirical examples. Finally, we conclude the paper by summarizing our findings and suggesting some directions for future research. Given the simplicity and ease of implementation of the modeling strategy, the development presented in this paper

should be of interest for users of DTs who wish to improve the performance of their models at a relatively modest computational cost.

Palabras Clave: churn prediction, insurance, spatial autocorrelation, spatial logit model, Madrid

Clasificación JEL: