Multi-Class Budget Tree
as Weak Learner for Ensemble Procedures (⋆)

Procedure di aggregazione ensemble per Multi-Class Budget Trees

Gianfranco Giordano, Antonio D’Ambrosio
Dipartimento di Matematica e Statistica, Università degli Studi di Napoli Federico II
e-mail: gianfranco.giordano@unina.it

Keywords: tree-based methods, latent budget models, multi-class variables, ensemble procedures

1. Introduction

During the last few years, classification has shown a great diffusion in many different fields such as medical science, customer satisfaction, pattern recognition, text mining, etc. In this context, tree methods have played an important role since the monography of Breiman et al. (1984). Later, new algorithms such as Two-Stage (Mola and Siciliano, 1992), FAST (Fast Algorithm for Statistical Trees, Mola and Siciliano, 1997), LBT (Latent Budget Trees for Multiple Classification, Siciliano, 1999) have been proposed to improve classification trees. In ordinary classification problems, each observation belongs to only one class of target (response) variable. Often, the problem concerns with the analysis of fuzzy data where objects cannot be placed in distinct groups without any strain, and as consequence the interest moves to the evaluation of the strength with which an observation falls into a particular class rather than another one. For example, when we classify a set of documents, the same document (i.e. a scientific paper) may be relevant to more than one topic. This kind of problem is called multi-class problem. There are many tree algorithms designed to handle single-class problems, but less for multi-class data (Zhang, 1998). Until now, common technique, which could be applied to the multi-class problems, consists on the decomposition of data into multiple independent binary classification problems. In 2005 Aria proposed an interesting multi-response classification, based on the combined use of segmentation tree methodologies and on the latent budget models, called Multi-Class Budget Trees (MCB). In this paper, starting from this approach, we study the problem of the robustness of the method, implementing an ensemble procedure which focus its attention on the boosting philosophy (Freund and Schapire, 1997).

2. The methodology

A segmentation procedure consists on a recursive (r-way or binary) partition of N cases into subgroups where the response variable is internally homogeneous and externally heterogeneous. In this work, a procedure is proposed to design an ensemble tree dealing with multi-class response. The idea is to use Multi-Class budget tree as a weak learner, in resampling algorithm, to define a robust supervised classifier dealing with multiple response dependent variable.

(*This work has been supported by MIUR funds)
2.1. Multi-class budget tree

The MCB methodology consists on a two-stage recursive partition criteria which uses conditional latent budget models (LBM) to define the best split of objects in classification field when the response is multi-class. The purpose is to choose, at each node, the most explicative predictor (or a subgroup of predictors) respect to the $Y$ variable. Then a conditional LBM is applied to find the best partition of the observations in $K$ groups, where $K$ represents the number of latent budgets. We can choose different strategies for the analysis from which depends the measure of $K$. A first way is to use $K = 2$ or $K = 3$ in order to obtain respectively a binary or a ternary tree. A second way to proceed should imply not to fix a value for $K$ but to choose, at each node, the most parsimonious model for the data analyzed. In both cases, the methodology grows a classification tree characterized by a sequence of latent budget models assigned recursively to the internal nodes of the tree.

2.2. Ensemble procedure

A well-known problem in literature concerning the tree-based method is the instability. It can be considered even more true when we work in contexts of multiple response variable analysis. In this paper, as a consequence we propose, to overcome this matter, an ensemble procedure based on the boosting philosophy, which using as weak-learner the MCB, allows to obtain a more accurate and robust classification. This kind of approach, allows also to overcome the typical problem of the parameters identification of LBM. This methodology has been tested on a great number of simulated and real datasets, underlining the characteristics of accuracy and robustness rather than the weak procedure.

References


