Partial Predictability on Optimal Scaling Trees

Predizione parziale sugli alberi optimal scaling

Valerio A. Tutore

Dipartimento di Matematica e Statistica, Università di Napoli Federico II

e-mail: v.tutore@unina.it

Keywords: segmentation methods, nonlinear canonical correlation analysis, categorical predictors, instrumental variable

1. Introduction

A standard binary segmentation procedure aims to find at each node the best split of objects into two sub-groups which are internally the most homogeneous and externally the most heterogeneous with respect to the given output. The best split is found among all possible splits that can derive from the given inputs, namely partitioning the categories of the input into two sub-groups so to provide the corresponding binary split of the objects. Classical tree-based methodologies are characterized by using one or just a small subset of original instances to define the final data partition. In certain contexts it is very important to investigate the role that each single variable plays in explaining the response. For example, when in presence of complex data structures, characterized by groups of co-variates internally correlated with each others and hierarchically connected to a synthesis framework, the need of a better interpretative value is particularly felt. We build up a tree-based method whose nodes splitting are given by splits obtained from Nonlinear canonical correlation analysis object scores. This method with $g$ sets of variables is a generalization of linear CCA with $g$ sets. The generalizations used depend on subjective choices. In this short paper the approach of Gifi is followed. The idea is to consider the two-stage splitting criterion based on the predictability $\tau$ index of Goodman and Kruskal for two-way cross-classifications.

2. The idea

The Optimal scaling tree, that consists on the definition of a splitting criteria with optimization criterion, using Nonlinear Canonical Correlation Analysis, allows to reduce the dimensionality of the analysis, shifting the attention towards a set of latent predictors synthesis of the original variables. The new latent variables will be the object scores extracted by Overals method. We consider as output the variable $Y$ which summarizes the optimal split of the objects. Optimal scaling means that for each categorical variable a nonlinear transformation is permitted, such that it maximizes the analysis criterion. We find the NLCCA’s object scores minimizing

$$\sum_{t=1}^{k} tr(XX'X_QA_t)'(XX'X_QA_t)$$

where $X$ are object scores, $Q_t$ are the transformed variables from original variable matrix $H$ and $A_t$ are the collection of multiple and single category quantifications across

(*) This work has been supported by European FP6 Project iWebCare IST-4-028055.
variables and sets.
In the second step the method identifies the best split of latent variables respect to
the response considering $Z_O$ as the stratifying object variable with $K$ categories. This
instrumental variable has to be external to original predictors. We consider the two-stage
splitting criterion based on the predictability $\tau$ index of Goodman and Kruskal: in the
first stage, the best predictor is found maximizing the global prediction with respect to
the response variable; in the second stage, the best split of the best predictor is found
maximizing the local prediction. It can be demonstrated that skipping the first stage
maximizing the simple $\tau$ index is equivalent to maximizing the decrease of impurity
in CART approach. In the following, we extend this criterion in order to consider
the predictability power explained by each predictor/split with respect to the response
variable conditioned by the instrumental variable $Z_O$. To this purpose, we consider the
predictability indexes used for three-way cross-classifications, namely the multiple $\tau_m$
and the partial $\tau_p$ predictability index of Gray and Williams. At each node, in the first
stage, among all available predictors $X_m$ for $m = 1, \ldots, M$, we maximize the partial
index $\tau_p(Y|X_m, Z_O)$ in order to find the best predictor $X^*$ conditioned by the instrumental
variable $Z_O$:
\[
\tau_p(Y|X_m, Z_O) = \frac{\tau_m(Y|X_m, Z_O) - \tau_s(Y|Z_O)}{1 - \tau_s(Y|Z_O)}
\]
(2)
where $\tau_m(Y|X_m, Z_O)$ and $\tau_s(Y|Z_O)$ are the multiple and the simple predictability
measures. In the second stage, we find the best split $s^*$ of the best predictor $X^*$
maximizing the partial index $\tau_s(Y|s, Z_O)$.
The results of our simulation study have been very promising.

References
Trees, Belmont C.A. Wadsworth.
Gifi A. (1981) Nonlinear Multivariate Analysis, University of Leiden, Department of
Data Theory.
Goodman L.A., Kruskal W.H. (1979) Measures of Association for Cross Classifications,
Springer.
segmentation, in: Computational Statistics: COMPSTAT ’92, Y. Dodge, J. Whittaker
(Eds.), Physica Verlag, Heidelberg (D), 179-184.
Analysis 2007, Book of short papers (Macerata, September 12-14, 2007), EUM
Macerata, 231-234.
Tutore V.A., Mooijaart A (2007) Optimal scaling trees, in: Classification and Data
Analysis 2007, Book of short papers (Macerata, September 12-14, 2007), EUM
Macerata, 359-362.
instrumental variables, Advances in Intelligent Data Analysis, 163-173.