Decision Trees for Three-Way Data

Valerio A. Tutore,
Department of Mathematics and Statistics
University of Naples Federico II
v.tutore@unina.it

Abstract: *The framework of this short paper is tree-based models for three-way data. Three-way data are data that are classified in three ways. Three way data are obtained when prior information play a role in the analysis. Thus, they can be derived when a stratifying variable is used to distinguish either groups of variables or groups of objects. Three way data can be analysed by exploratory methods as well as confirmatory methods. Recently, we have introduced a methodology for classification and regression trees in order to deal specifically with three-way data. So far we have developed only partitioning procedures discarding the induction point of view.*

Keywords: Three-Way Data, Classification Trees, Pruning

1. Introduction

Three-way data are data that are classified in three ways. Longitudinal data, i.e., are three way, because of repeated observation of the same variables on the same objects. So far segmentation methods for classification and regression trees have been proposed as supervised approach to analyze data sets where a response variable and a set of predictors are measured on a sample of objects or cases. Classification and regression trees have becoming a fundamental approach to data mining and prediction (Hastie et al., 2001). From the exploratory point of view, in binary segmentation, the aim is to find the best split of a predictor to split the cases into two sub-groups such to reduce the impurity of the response within each sub-group. The recursive splitting of the cases yields a tree structure. From the confirmatory point of view, pruning algorithms (Breiman et al., 1984) or ensemble methods (Breiman, 1996) allow to define a decision tree model to classify/predict new cases of unknown response on the basis of the measured predictors. Following the pioneer work (Tutore et al., 2006, 2007), this paper provides the decision tree induction methodology for the analysis of three-way data sets. Such data sets can be described by a cube, namely a set of variables (including both predictors and responses) is measured on a sample of objects in a number of distinct situations, also called occasions. Each slide of the cube is a two-way data matrix, i.e. units times variables. Typically, the occasions are associated to modalities of a categorical variable. Whereas in previous work we have proposed partitioning procedures for exploratory trees dealing with three-way data, in the following we introduce decision trees.

2. The data and the two-stage splitting criterion

The three ways of the data set are cases, attributes and situations, respectively. Let **D** be the three-way data matrix of dimensions N, V, Q, where N is the number of cases, objects or units, V is the number of variables, Q is the number of situations. Assume that the V variables can be distinguished into two groups, namely there are M predictor variables $X_1,\ldots,X_m,\ldots,X_M$ and C response variables $Y_1,\ldots,Y_c,\ldots,Y_C$ where $M+C=V$. The Q situations refer to modalities of a stratifying variable, which is called *instrumental variable*. Predictors can be of categorical and/or numerical type whereas responses can be either categorical or numerical.
The two-stage splitting criterion for $C = 1$ can be defined as follows:

$$\max_m \sum_q \gamma_Y\left(t\middle|_q X_m\right) p_Y\left(t|q\right) \qquad (1)$$

$$\max_m \sum_q \gamma_Y\left(t|s\right) p_Y\left(t|q\right) \tag{2}$$

for q = 1,...,Q (i.e. subsamples), m = 1,...,M (i.e. predictors), s = 1,...,S (i.e. splitting variables), with $\sum_q p_Y\left(t|q\right)$, where $\gamma_Y\left(t\left|{}_q X_m\right.\right)$ is the global impurity proportional reduction measure of Y due to each predictors X_m and $\gamma_Y\left(t|s\right)$ the local impurity proportional reduction measure of Y due to each splitting variable s. The former is a weighted average of the measures calculated across the Q occasions. A suitable weighting system $p_Y\left(t|q\right)$ can be given by the percentage of the total impurity of the response in each subsample. Analogously, it can be defined the local impurity proportional reduction measure due to each splitting variable.

3. The partial predictability tree partitioning

Let Y be the output, namely the response variable, and let $\mathbf{X} = \{X_1,...,X_M\}$ be the set of M inputs, namely the predictor variables. In addition, let Z_O be the stratifying object variable with K categories. The response variable is a nominal variable with J classes and the M predictors are all categorical variables (or categorized numerical variables). The sample is stratified according to the K categories of the instrumental variable Z_O. We consider the two-stage splitting criterion based on the predictability τ index of Goodman and Kruskal (1979) for two-way cross-classifications: 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 τ 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. For that, we consider the predictability indexes used for three-way cross-classifications, namely the multiple τ_m and the partial τ_p predictability index of Gray and Williams. At each node, in the first stage, among all available predictors X_m for m = 1,...,M, we maximize the partial index $\tau_p\left(Y|X_m,Z_0\right)$ to find the best predictor X^* conditioned by the instrumental variable Z_O:

$$\tau_p\left(Y|X_m,Z_0\right) = \frac{\tau_m\left(Y|X_m,Z_0\right) - \tau_s\left(Y|Z_0\right)}{1-\tau_s\left(Y|Z_0\right)} \tag{3}$$

where $\tau_m\left(Y|X_m,Z_0\right)$ and $\tau_s\left(Y|Z_0\right)$ 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 $\tau_p\left(Y|s,Z_0\right)$ among all possible splits of the best predictor.

4. The decision tree

In the following, we introduce an extended version of CART pruning procedure in order to define the decision tree based on the partial predictability tree partitioning. As well known, pruning trees is necessary to remove the most unreliable branches and improve understand ability. For the definition of the pruning criterion it is necessary to introduce a measure $R^*(.)$ that depends on the size (number of terminal nodes) and the accuracy (error rate) both. In particular, let T_t be the branch departing from the node t having $\left|\tilde{T}_t\right|$ terminal nodes. The criterion is such that prune node t if

$$R^*\left(t\right) \le R^*\left(T_t\right) \tag{4}$$

We define the following error-complexity measure for the node t and for the branch T_t as

$$R_\alpha\left(t\right) = \sum_q r_q\left(t\right) p_q\left(t\right) + \alpha, \tag{5}$$

$$R_\alpha(T_t) = \sum_{h \in |\overline{T}_t|} \sum_q r_q(h)\, p_q(h) + \alpha |\overline{T}_t|, \quad (6)$$

where α is the penalty for complexity due to one extra terminal node in the tree, $r_q(t)$ is the error rate (the proportion of cases in node t which are misclassified into the q-th subgroup of objects), $p_q(t)$ is the proportion of cases belonging to the q-th subgroup in node t and $|\overline{T}_t|$ is the number of terminal nodes of T_t. Basically, the branch T_t should be pruned if

$$R_\alpha(t) \le R_\alpha(T_t) \quad (7)$$

Thus, using a down-top algorithm and a training set the criterion is to prune each time the branch T_t that provides the lowest reduction in error per terminal node (i.e., the weakest link) as measured by

$$\alpha = \frac{R(t) - R(T_t)}{|\overline{T}_t| - 1} \quad (8)$$

On the basis of the error-complexity measure $R_\alpha(.)$ a sequence of nested optimally pruned trees is generated pruning at each step the sub-tree with the minimum value of α_t.

5. The real world application

There are several fields in which this methodology can be applied with good results. In this short paper, we present an application about a Tourist Satisfaction dataset about Province of Naples in May 2007. The data regard a survey with N = 1876. Predictors of Tourist Satisfaction dataset are: Professionalism, Structures, Competitivity, Appearance, Security maintenance, Cultural environmental attractions, Accessibility of destination, Information and welcome, Mobility. The response variable has been recoded into a dummy variable, satisfied and unsatisfied tourists. A proper classification rule should consider the different nationality of tourists. This can be considered as the instrumental Z_O having two different categories, Italian and Foreigners. Figure 1 shows the best pruned tree with 15 terminal nodes.

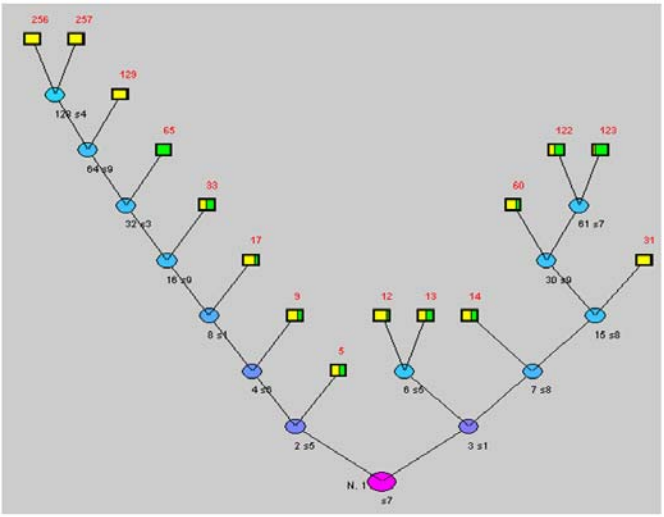

Figure 2: The best pruned tree

Table 1 provides summary information concerning the best pruned tree. In particular, we report the response classes distribution of the objects within the two categories of Z_O, for the predictor selected in each node. With *Node* we indicate the label, with n the number of objects in each node, with *Predictor* the variable splits the node, with z_1 and z_2 the two categories of the instrumental variable (nationality), with S and U respectively the satisfied and the unsatisfied tourists, with *Badclassified* the percentage of bad classified objects in each node discarding the instrumental variable and with *Badclassified z* the percentage of bad classified objects considering the instrumental variable. These two measures agree only when there is no difference in class assignment in relation to the instrumental variable.

Response Classes Distribution			z_1		z_2			
Node	**n**	**Predictos**	**S**	**U**	**S**	**U**	**Badclassified**	**Badclassified z**
1	1876	*s7*	381	343	774	378	38,43	38,43
2	901	*s5*	217	69	494	121	21,09	21,09
3	975	*s1*	164	274	280	257	**45,54**	**43,18**
4	797	*s6*	198	49	457	93	17,82	17,82
5	104	*Terminal node*	23	15	33	33	46,15	46,15
6	422	*s5*	102	86	156	78	38,87	38,87
7	553	*s8*	71	182	115	185	33,63	33,63
8	622	*s1*	172	35	361	54	14,31	14,31
9	175	*Terminal node*	35	13	87	40	30,29	30,29
12	170	*Terminal node*	42	16	83	29	26,47	26,47
13	252	*Terminal node*	70	52	63	67	**47,22**	**45,63**
14	94	*Terminal node*	21	20	37	16	38,30	38,30
15	459	*s5*	50	162	78	169	27,89	27,89
16	478	*s9*	141	20	284	33	11,09	11,09
17	144	*Terminal node*	31	15	77	21	25,00	25,00
30	449	*s9*	47	161	72	169	26,50	26,50
31	10	*Terminal node*	3	1	6	0	10,00	10,00
32	471	*s3*	141	18	281	31	10,40	10,40
33	7	*Terminal node*	0	2	3	2	**42,86**	**28,58**
60	21	*Terminal node*	6	3	8	4	33,33	33,33
61	428	*s7*	41	158	64	165	24,53	24,53
64	443	*s9*	141	6	281	15	10,21	10,21
65	28	*Terminal node*	0	12	0	16	0,00	0,00
122	263	*Terminal node*	36	89	44	94	30,42	30,42
123	165	*Terminal node*	12	59	13	81	15,15	15,15
128	299	*s4*	100	3	187	9	4,01	4,01
129	144	*Terminal node*	36	3	99	6	6,25	6,25
256	169	*Terminal node*	55	1	109	4	2,96	2,96
257	130	*Terminal node*	42	2	81	5	5,38	5,38

Table 1: Nodes in the best pruned tree (Tourist Satisfaction Dataset)

Bibliography

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), Classification and Regression Trees, Wadsworth, Belmont CA.

Breiman L. (1996), Bagging Predictors, Machine Learning, 24, 123-140.

Hastie T., Friedman, J. H., Tibshirani R. (2001), The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer.

Siciliano R., Aria M., Conversano C. (2004), Tree Harvest: Methods, Software and Applications, in Antoch J. (ed.): COMPSTAT 2004 Proceedings, Springer, 1807-1814.

Tutore V.A., Siciliano R. Aria, M. (2007), Conditional Classification Trees using Instrumental Variables, in: Advances in Intelligent Data Analysis, Springer-Verlag, pp 163-173.

