`decisionTree.SDA (symbolicDA)`

Decision tree for symbolic data

The general decision tree algorithm for symbolic data can be described as follows [see Bock, Diday (Eds.) 2000, p. 244-265]:

1. Construction of the contingency table for classical nominal and ordinal data and symbolic multinominal data (sets of categories).

The symbolic multinominal variables are treated as modal variables with uniform probability functions, for example:

a) colour ={red, blue, green, magenta} $\Rightarrow$ colour~{¼ (red), ¼ (blue), ¼ (green), ¼ (magenta)}

b) owned car = {Skoda} $\Rightarrow$ owned car ~ {1 (Skoda)}

The elements of contingency table for these variables are equal to probabilities.

For symbolic multinominal variables with weights (probabilities, frequencies) the elements of contingency table are equal to these weights (probabilities, frequencies).

2) Set the size of the terminal node to n^* and the quality (admissibility) criterion value to W^*

If the size of the node is lower than n^* it is a terminal node. If for any admissible question the quality criterion W is greater than W^* this question can be applied.

3) Construction of the binary questions for each of j variables $(j = 1,\ldots,m)$ and calculation of probability of assigning the object to the left and right node.

For symbolic interval valued data there is a need to compute all possible mid points for such variables. These mid points will be used as cutting threshold (cutoff) values c. If the value c:

a) is within the interval of the *j*-th variable, the probability of assigning the object to the left node is calculates as follows:

$$p_k(l) = \frac{c - \underline{v}_{kj}}{\overline{v}_{kj} - \underline{v}_{kj}},$$

where: $k = 1,\ldots,n$ - number of symbolic object, $\overline{v}_{kj}$ - upper interval bound, $\underline{v}_{kj}$ - lower interval bound for *j*-th variable and *k*-th symbolic object.

b) is below the lower interval bound: $p_k(l) = 0.$

c) is greater than the upper interval bound: $p_k(l) = 1.$

In case of ordinal and nominal data the c values are equal to each if the categories of *j*-th variable, except the last category. For each object there is a need to sum up the frequencies for

each category that are less or equal than c. The sum is treated as a probability of assigning the object to the left node.

Multinominal symbolic variables, that are treated in the same way as modal variables with uniform probability functions, and multinominal symbolic variables with weights, the c values and probabilities are calculated in the same way as in ordinal and nominal case.

Regardless of the variable type the probability of assigning symbolic object to the right node is calculated as follows:

$$p_k(r) = 1 - p_k(l)$$

4) The quality criterion W for each c for each of variables is calculated as follows:

$$W_j(t,c) = \log \prod_{k=1}^{n} \left[p_k(l) \cdot P_l(s) + p_k(r) \cdot P_r(s) \right],$$

where: $j = 1, \ldots, m$-variable number; t - node number; c - cutoff value; $p_k(l)$ - probability that *k*-th object is assigned to the left node; $p_k(r)$ - probability that *k*-th object is assigned to the right node; $P_l(s)$ - conditional probability that within the left node contains the cluster of *k*-th object (it is the product of sum of probabilities of assigning the objects from *s*-th cluster to left node and the sum of assigning all other object to this node), $P_r(s)$ - conditional probability that within the right node contains the cluster of *k*-th object (it is the product of sum of probabilities of assigning the objects from *s*-th cluster to right node and the sum of assigning all other object to this node).

5) The highest W, that:

$$W_j(t,c) > W^*$$

is chosen. The node t is cut into to sub-nodes – left and right under condition that the size of the node is admissible $\left(n_t > n^*\right)$.

6) The steps 4-5 need to be repeated until receiving terminal nodes.

7) Calculation the rate of correct predictions.

References:

2. Billard L., Diday E. (Eds.) (2006), *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley & Sons Ltd, Chichester.
3. Bock H.-H., Diday E. (Eds.) (2000), *Analysis of symbolic data. Explanatory methods for extracting statistical information from complex data*. Springer Verlag, Berlin-Heidelberg.

4. Périnel E. (1996), *Segmentation et Analyse des Données Symboliques. Application à des données probabilistes imprécises*. Thése de doctorat de l'Université Paris IX Dauphine.
5. Périnel E. (1999), *Construire un arbre de discrimination binaire à partir de données imprécises*, Revue de Statistique Appliquée, 47 no. 1, p. 5-30.