CHARLES UNIVERSITY PRAGUE

faculty of mathematics and physics

The Prospector - Beyond Classification

Pavel Vaněček
vanecek@karlin.mff.cuni.cz
Department of Statistics Charles University, Prague

Summary

The Prospector is an algorithm for data description standing on the border between clustering and classification:
We are searching for clusters with respect to a target variable.

This work was partially supported by grant GAČR 201/05/H007.

Classification

The target variable is predicted from the explanatory variables. We use an algorithm based on averaging of classification trees called *Random Forests* proposed by Professor Emeritus Leo Breiman from University of California. This method benefits from averaging a lot of classification trees, each of them being trained on different subsample of the data. Moreover, the best splitting variable in each node of the tree is selected from a random subset of all possible variables that preserves from overfitting.

Categorization

Continuous variables are categorized into a few distinct categories and ordinal categorical variables are possibly recoded into fewer number of categories. This can be efficiently done by genetic algorithm:
- *pool of chromosomes* 0010100100010001…00, each gene stands for a distinct value of variable, 0 = no split, 1 = split
- *fitness of chromosomes* (given the target variable) is equal to either its correlation or Cramer's V

$$V = \sqrt{\frac{c^2}{n \cdot \min(M-1, N-1)}}$$

where *n* is number of cases, *M, N* are number of categories
- offspring is created using *crossover* operator, combination of parent genes step by step, or *mutation*

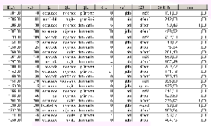
Clustering

Categories of each variable are sorted with respect to the target variable. This creates *n-dimensional grid.* Each grid point is evaluated by a criterion based on its *radius, purity, weight* and then we try to incorporate its nearest neighbours to improve the criterion. This creates so called *nuggets* and they are compared to each other and (a few) best ones are selected.

Case study

A brewery wants to define status of its brand of beer.
Outlets are described according to their type, size, volume of all beer they sell, type of cuisine they provide and so on.
Some of the outlets are scored as either "relevant to the brand" or "not relevant to the brand".
The task is to find the status (characterization, description) of the outlets that are "relevant to the brand".

Input data



Diagnostic tools from classification

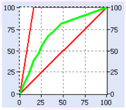


Fig. 1: Gains chart

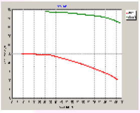


Fig. 2: Coverage vs. accuracy optimizer

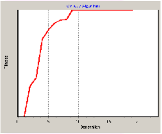


Fig. 3: Fitness of the best chromosome in population

Fig. 4: XML output from the categorization

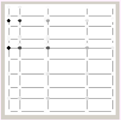


Fig. 5: Projection of the grid into 2-dimensions

References:
[1] Dobývání znalostí z databází
Berka P.
Academia, 2003
[2] The Elements of Statistical Learning
Hastie T., Tibshirani R., Friedman J.
Springer, 2003