



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

Outlet	Type	Size	Volume	Cuisine	Status
1	Pub	Small	Low	Fast Food	Not Relevant
2	Pub	Small	Low	Fast Food	Not Relevant
3	Pub	Small	Low	Fast Food	Not Relevant
4	Pub	Small	Low	Fast Food	Not Relevant
5	Pub	Small	Low	Fast Food	Not Relevant
6	Pub	Small	Low	Fast Food	Not Relevant
7	Pub	Small	Low	Fast Food	Not Relevant
8	Pub	Small	Low	Fast Food	Not Relevant
9	Pub	Small	Low	Fast Food	Not Relevant
10	Pub	Small	Low	Fast Food	Not Relevant
11	Pub	Small	Low	Fast Food	Not Relevant
12	Pub	Small	Low	Fast Food	Not Relevant
13	Pub	Small	Low	Fast Food	Not Relevant
14	Pub	Small	Low	Fast Food	Not Relevant
15	Pub	Small	Low	Fast Food	Not Relevant
16	Pub	Small	Low	Fast Food	Not Relevant
17	Pub	Small	Low	Fast Food	Not Relevant
18	Pub	Small	Low	Fast Food	Not Relevant
19	Pub	Small	Low	Fast Food	Not Relevant
20	Pub	Small	Low	Fast Food	Not Relevant
21	Pub	Small	Low	Fast Food	Not Relevant
22	Pub	Small	Low	Fast Food	Not Relevant
23	Pub	Small	Low	Fast Food	Not Relevant
24	Pub	Small	Low	Fast Food	Not Relevant
25	Pub	Small	Low	Fast Food	Not Relevant
26	Pub	Small	Low	Fast Food	Not Relevant
27	Pub	Small	Low	Fast Food	Not Relevant
28	Pub	Small	Low	Fast Food	Not Relevant
29	Pub	Small	Low	Fast Food	Not Relevant
30	Pub	Small	Low	Fast Food	Not Relevant
31	Pub	Small	Low	Fast Food	Not Relevant
32	Pub	Small	Low	Fast Food	Not Relevant
33	Pub	Small	Low	Fast Food	Not Relevant
34	Pub	Small	Low	Fast Food	Not Relevant
35	Pub	Small	Low	Fast Food	Not Relevant
36	Pub	Small	Low	Fast Food	Not Relevant
37	Pub	Small	Low	Fast Food	Not Relevant
38	Pub	Small	Low	Fast Food	Not Relevant
39	Pub	Small	Low	Fast Food	Not Relevant
40	Pub	Small	Low	Fast Food	Not Relevant
41	Pub	Small	Low	Fast Food	Not Relevant
42	Pub	Small	Low	Fast Food	Not Relevant
43	Pub	Small	Low	Fast Food	Not Relevant
44	Pub	Small	Low	Fast Food	Not Relevant
45	Pub	Small	Low	Fast Food	Not Relevant
46	Pub	Small	Low	Fast Food	Not Relevant
47	Pub	Small	Low	Fast Food	Not Relevant
48	Pub	Small	Low	Fast Food	Not Relevant
49	Pub	Small	Low	Fast Food	Not Relevant
50	Pub	Small	Low	Fast Food	Not Relevant
51	Pub	Small	Low	Fast Food	Not Relevant
52	Pub	Small	Low	Fast Food	Not Relevant
53	Pub	Small	Low	Fast Food	Not Relevant
54	Pub	Small	Low	Fast Food	Not Relevant
55	Pub	Small	Low	Fast Food	Not Relevant
56	Pub	Small	Low	Fast Food	Not Relevant
57	Pub	Small	Low	Fast Food	Not Relevant
58	Pub	Small	Low	Fast Food	Not Relevant
59	Pub	Small	Low	Fast Food	Not Relevant
60	Pub	Small	Low	Fast Food	Not Relevant
61	Pub	Small	Low	Fast Food	Not Relevant
62	Pub	Small	Low	Fast Food	Not Relevant
63	Pub	Small	Low	Fast Food	Not Relevant
64	Pub	Small	Low	Fast Food	Not Relevant
65	Pub	Small	Low	Fast Food	Not Relevant
66	Pub	Small	Low	Fast Food	Not Relevant
67	Pub	Small	Low	Fast Food	Not Relevant
68	Pub	Small	Low	Fast Food	Not Relevant
69	Pub	Small	Low	Fast Food	Not Relevant
70	Pub	Small	Low	Fast Food	Not Relevant
71	Pub	Small	Low	Fast Food	Not Relevant
72	Pub	Small	Low	Fast Food	Not Relevant
73	Pub	Small	Low	Fast Food	Not Relevant
74	Pub	Small	Low	Fast Food	Not Relevant
75	Pub	Small	Low	Fast Food	Not Relevant
76	Pub	Small	Low	Fast Food	Not Relevant
77	Pub	Small	Low	Fast Food	Not Relevant
78	Pub	Small	Low	Fast Food	Not Relevant
79	Pub	Small	Low	Fast Food	Not Relevant
80	Pub	Small	Low	Fast Food	Not Relevant
81	Pub	Small	Low	Fast Food	Not Relevant
82	Pub	Small	Low	Fast Food	Not Relevant
83	Pub	Small	Low	Fast Food	Not Relevant
84	Pub	Small	Low	Fast Food	Not Relevant
85	Pub	Small	Low	Fast Food	Not Relevant
86	Pub	Small	Low	Fast Food	Not Relevant
87	Pub	Small	Low	Fast Food	Not Relevant
88	Pub	Small	Low	Fast Food	Not Relevant
89	Pub	Small	Low	Fast Food	Not Relevant
90	Pub	Small	Low	Fast Food	Not Relevant
91	Pub	Small	Low	Fast Food	Not Relevant
92	Pub	Small	Low	Fast Food	Not Relevant
93	Pub	Small	Low	Fast Food	Not Relevant
94	Pub	Small	Low	Fast Food	Not Relevant
95	Pub	Small	Low	Fast Food	Not Relevant
96	Pub	Small	Low	Fast Food	Not Relevant
97	Pub	Small	Low	Fast Food	Not Relevant
98	Pub	Small	Low	Fast Food	Not Relevant
99	Pub	Small	Low	Fast Food	Not Relevant
100	Pub	Small	Low	Fast Food	Not Relevant

Diagnostic tools from classification

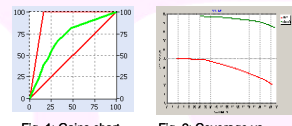


Fig. 1: Gains chart

Fig. 2: Coverage vs. accuracy optimizer

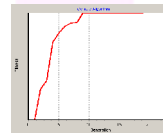


Fig. 3: Fitness of the best chromosome in population



Fig. 4: XML output from the categorization

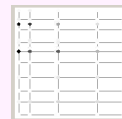


Fig. 5: Projection of the grid into 2-dimensions

References:

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