Documentation: MIP Function — Feature Ranking for Structured Targets

The feature ranking algorithm for structured target variables.

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Dragi Kocev
Homepage	http://source.ijs.si/hbp/clus/wikis/home
Documentation	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-clus-feature-ranking.pdf http://source.ijs.si/hbp/clus/wikis/documentation
Support	http://source.ijs.si/hbp/clus/issues
Source Code	http://source.ijs.si/hbp/clus.git http://source.ijs.si/hbp/mipfunctions/tree/master/r-clus-franking
License	GNU GPL
Current Version	2.12
All Versions	2.12

Description

Methods for feature ranking are used in many domains with many descriptive variables, i.e., high-dimensional problems. The obtained rankings provide an additional insight about the importance of the variables for the target and/or reduce the dimensionality of the problem. Many real-life problems have structured targets that need to be predicted. However, the task of feature ranking in the context of predicting structured target variables is more complex than the same task for simple classification or regression. Typical approaches for this task decompose the output to primitive components, perform feature ranking on these smaller problems, and then aggregate the resulting rankings into a single ranking. They are computationally intractable for large output spaces (e.g., genetic or imaging data) and ignore the dependencies between components of the output. We have developed efficient feature ranking methods in the context of predicting structured targets. The developed methods are based on the ensemble learning paradigm.

This MIP function implements two algorithms for feature ranking for structured targets: (1) RF-RANK exploits the random forests mechanism and (2) GENIE3 exploits the variance reduction at each tree node from the ensemble. For the latter method, the ensemble could be random forest or an ensemble of extra trees. The both methods use predictive clustering trees as base predictive models.

The development of the algorithm and its implementation was partly paid by HBP.

References

[1] D. Kocev, I. Slavkov, S. Džeroski: Feature ranking for multi-label classification using predictive clustering trees, Proceedings of the Workshop - Solving complex machine learning problems with ensemble methods held in conjunction with ECML/PKDD2013, pp. 56-68, 2013

Usage

Input: A set of examples with known values for descriptive and target variables.

Parameters: Descriptive variables, target variables, number of trees in the ensemble, size of the feature subsets considered at each node.

Output: A list of the descriptive variables and their importance to the structured target.

Example

We can illustrate the use of the algorithm on the ADNI data.

Task:

Given a data set consisting of examples (i.e., patients) described in terms of several descriptive variables and labelled with several target variables we want to learn a feature ranking that will order the descriptive variables based on their importance for the target variables.

Input data:

Given a subset of ADNI data with 916 patients described with 34 different variables, collected at baseline evaluation, we split the variables in two groups. In the first group we have descriptive variables: APOE4, Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp, ICw, FDG, Aw45, CDRSB, ADAS13, MMSE, RAwLT_immediate, RAwLT_learning, RAwLT_forgetting, RAwLT_perc_forgetting, FAQ, MOCA. In the second group we have the target variables: EcogPtMem, EcogPtLang, EcogPtwisspat, EcogPtPlan, EcogPtOrgan, EcogPtDiwatt, EcogPtTotal, EcogSPMem, EcogSPLang, EcogSPwisspat, EcogSPPlan, EcogSPOrgan, EcogSPDiwatt, EcogSPTotal. We decided to learn a feature ranking with RF-GENIE parametrized as follows: random forest of 100 trees using SQRT of the descriptive features at each node of the trees.

1	FAQ_bl	80.78
2	CDRSB_bl	73.56
3	ADAS13_bl	28.49
4	MOCA_bl	20.59
5	wentricles_bl	19.9
6	RAwLT_immediate_bl	18.88
7	Entorhinal_bl	18.51
8	Hippocampus_bl	18.48
9	FDG_bl	17.14
10	WholeBrain_bl	16.75

11	Fusiform_bl	16.33
12	Aw45_bl	15.71
13	MidTemp_bl	15.48
14	ICw_bl	13.41
15	MMSE_bl	13.19
16	RAwLT_perc_forgetting_bl	12.8
17	APOE4	11.29
18	RAwLT_learning_bl	8.59
19	RAwLT_forgetting_bl	8.38