Documentation: MIP Function — Multi-Target Regression on Data Streams

The decision tree algorithm for streaming data.

Metadata

Category	MIP function, data analysis algorithm
Maintainers	Aljaž Osojnik
Homepage	http://moa.cms.waikato.ac.nz/
Documentation	http://source.ijs.si/hbp/mipfunctions/raw/master/doc/r-moa-trees.pdf
Support	http://source.ijs.si/hbp/mipfunctions/issues
Source Code	http://source.ijs.si/hbp/moa.git http://source.ijs.si/hbp/mipfunctions/tree/master/r-clus-trees
License	GNU GPL
Current Version	15.10
All Versions	15.10

Description

Methods for data stream mining are often used in Big Data problems, as they offer the ability to quickly process large amounts of data. However, the models obtained using this paradigm are often not interpretable for humans. We use trees learned in an online manner, which allows us to produce accurate and highly interpretable models, at high speeds. Through the use of the Hoeffding bound, the model can infer statistically supported hypothesis and use them to construct a decision tree.

This MIP function implements the FIMT-DD and iSOUP-Tree algorithms for learning decesion trees from data streams, the former for single-target prediction and the latter for multi-target prediction. Both of these algorithms produce models in the form of a model tree, i.e., a decision tree, which uses linear functions in the leaves to achieve better performance.

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

References

Elena Ikonomovska, J. Gama, and S. Džeroski - Incremental multi-target model trees for data streams, Proceedings of the 2011 ACM Symposium on Applied Computing, pages 988-993, ACM, New York, 2011.

Aljaž Osojnik, P. Panov, and S. Džeroski – Comparison of tree based models for multi-target regression on data streams, Proceedings of the 2015 Workshop on New Frontiers in Mining Complex Patterns, in print.

Usage

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

Parameters: Descriptive variables, target variables.

Output: A predictive model in a tree form.

Example

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

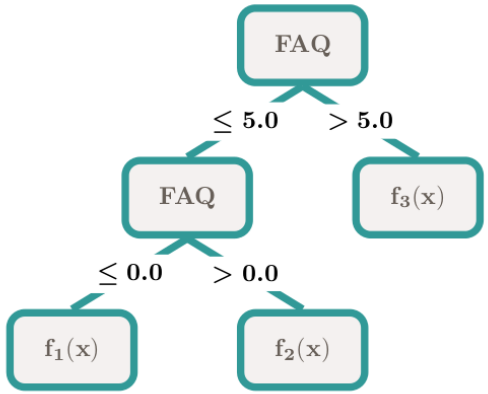
Task:

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 rule ensemble that will predict the values of target variables from the values of descriptive variables. In addition, we want to identify groups of patients that are similar in terms of all the target variables and describe them in terms of descriptive 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.

Example output:



Functions $\mathbf{f_1(x)}$, $\mathbf{f_2(x)}$ and $\mathbf{f_3(x)}$ are linear vector functions of the input values, represented by vector $\mathbf{x}$. For example, $\mathbf{f_1}^1\mathbf{(x)}$, the part of $\mathbf{f_1}$ that calculates the value of EcogPtMem, the first target variable, is defined as:

$$f_1^1(\mathbf{x}) = \text{EcogPtMem}(\mathbf{x}) = 0.0014 * \text{No} + 0.0884 * \text{APOE4} - 0.0402 * \text{wentricles} + 0.0027 * \text{Hippocampus} + 0.1032 * \text{WholeBrain} + 0.0321 * \text{Entorhinal} + 0.0578 * \text{Fusiform} + 0.1114 * \text{MidTemp} + 0.0579 * \text{ICw} + 0.0983 * \text{FDG} + 0.0132 * \text{Aw45} - 0.0177 * \text{CDRSB} + 0.0111 * \text{ADAS13} - 0.0081 * \text{MMSE} - 0.0490$$