Running the URLearning programs

More details explaining the various program options will be added soon. Please follow Issue #3 to see the latest updates about documentation.

Local score calculations

The `score` program calculates the candidate parent sets used as input to the solvers. It accepts a csv file as input and produces a "parent set scores" (pss) file as output. Both URLearning and GOBNILP can use pss files as input.

The `score` program treats each column as a categorical discrete variable, and each unique string is treated as a separate categorical value. In particular, the program *does not* perform any sort of discretization, normalization, etc. Furthermore, it *does not* remove records with missing values. String like "?", "NA", etc., will simply be treated as other categories for the variables for which they appear. Thus, it is likely custom preprocessing scripts will be necessary before using the `score` program.

By default, candidate parent sets which cannot possibly be optimal are pruned at the end of the `score` program; this behavior can be disabled with the `--do-not-prune` flag (see below). For more details about this pruning, please see:

Teyssier, M. & Koller, D. Ordering-Based Search: A Simple and Effective Algorithm for Learning Bayesian Networks Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, 2005 (the penultimate paragraph in Section 3.2).

```
score <input> <output> [-c/--constraints <constraints>] [-d/--delimiter <delimiter>] [-m/--
```

Command line options

- `input`. The input csv file
- `output`. The output pss file
- [`--constraints`]. A file containing simple constraints on the parent sets. Please see the constraints description for more details on this file.
- [`--delimiter`]. The character which separates columns in the `input` file. **N.B.** The actual data type is `char`, so only a single character is allowed. Default: `,`
- [`--has-header`]. Add this flag if the first line of `input` gives the name of the variables. It should use the same `delimiter` as the rest of the file. Default: variables are given names like `Variable_0`.

- [`--r-min`]. Internally, the program uses a sparse AD-tree to collect the necessary counts. This parameter controls the minimum number of records in the leaves of the AD-tree. Default: 5. For more details, please see:

  Moore, A. & Lee, M. S. Cached sufficient statistics for efficient machine learning with large datasets. *Journal of Artificial Intelligence Research*, 1998, 8, 67-91 (Section 5).

- [`--function`]. The scoring function to use. Default: `BIC`. Choices: `BIC`, `fNML`, `BDeu` (case-insensitive). **N.B.** For all scoring functions, the natural logarithm is used for all relevant calculations, and `log(0)` is taken to be `0`.

- [`--ess`]. The equivalent sample size for BDeu. This parameter is ignored for the other scoring functions. Default: 1.0

- [`--max-parents`]. The maximum number of parents to consider for local scores. Additionally, for BIC, the literature shows a hard limit on the number of parents as a function of the number of records. If the `--do-not-prune` flag is given, `max_parents` will not be updates; otherwise, `max_parents` will be set to this limit for `BIC`. A value less than 1 indicates no limit on the sizes of the candidate parent sets. Default: 0. For more details, please see:

  Suzuki, J. Learning Bayesian Belief Networks Based on the MDL Principle: An Efficient Algorithm Using the Branch and Bound Technique. *IEICE Transactions on Information and Systems*, 1999, E82-D, 356-367.

- [`--enable-de-campos-pruning`]. Add this flag to *enable* de Campos and Ji-style score pruning during the search. This feature is experimental (unpublished) for fNML, and it appears to have a bug for large parent limits and BDeu. Please follow Issue #20 for updates about the correctness. Default: all local scores are calculated, up to the `max_parents` limit. For more details, please see:

  de Campos, C. P. & Ji, Q. Efficient Learning of Bayesian Networks using Constraints. *Journal of Machine Learning Research*, 2011, 12, 663-689.

- [`--do-not-prune`]. Add this flag to *disable* standard score pruning after score calculations. Additionally, this flag *disables* the automatic `max_parent` calculation for `BIC` (as described for `--max-parents`). Default: scores are pruned according to the Teyssier and Koller reference given above, and the parent limit for `BIC` is set according to the rule given by Suzuki.

- [`--threads`]. The number of concurrent threads to use for score calculations. The parallelism is very simple; each thread calculates the candidate parent sets for one of the variables and writes them to a temporary pss file. After all candidate parent sets have been calculated, the pss files are concatenated into the final pss file containing all scores. Default: 1

- [`--time`]. The maximum amount of time (in seconds) to use for calculating local scores for each variable. After the specified time has elapsed, local score calculations will end after the calculation of the "current" score, and the candidate parent sets will be found using Teyssier and Koller pruning (unless the `--do-not-prune` flag was given). A value less than 1 indicates no time limit and all scores will be calculated according to the specified parameters. Default: 0.