

Specifying pattern databases in files

In previous work [Fan et al., AAAI 2014; Fan and Yuan, AAAI 2015], we have described strategies for finding good groupings of variables for the pattern databases. These groupings can be given to the various solvers via a file describing all desired pattern databases.

When multiple pattern databases are specified, the bound is calculated according to all of them; the minimum is used for the h estimate for the node.

Format

The (line-based) format of the file is as follows.

- Blank lines are allowed.
- The hash (“#”) begins comments. Everything on the rest of the line is ignored.
- Each non-blank line specifies one pattern database as follows.

```
line := <database_type>;<parameters>
<database_type> := s|d
<parameters> := <static_parameters>|<dynamic_parameters>
<dynamic_parameters> := int
<static_parameters> := [<pattern>;]*[<pattern>]
<pattern> := [int,]*[int]
```

Example

The following example specifies three pattern databases for use with the iris dataset. The first line (d;2) indicates that a dynamic pattern database with k=2 should be created. The second line (s;0,2,3;1,4) indicates that a static pattern database with two grouping (variables {0,2,3} as the first group and {1,4} as the second) should also be created. The third line specifies another static pattern database with groups {0,4} and {1,2,3}.

```
# file pattern database for iris
```

```
d;2 # a dynamic pd
s;0,2,3;1,4 # a static pd
s;0,4;1,2,3 # another static pd
```

Semantics

As the format and example above shows, the file allows specification of both static and dynamic pattern databases (please see [Yuan and Malone, UAI 2012] for more details).

- Dynamic pattern databases

In the case of dynamic pattern databases, the only parameter is k , the size of the patterns. Experimental results suggest that values in the range of 2 to 4 or 5 are reasonable; larger patterns significantly increase the time to construct the pattern database and do not typically show improvement during the search.

- Static pattern databases

Static pattern databases require the groupings of the variables. These are given as comma-separated lists separated by semicolons. Currently, the variable *indices* must be given; the variable *names* will not work. Please see Issue #22 for updates about this behavior. The variable indices are 0-based and exactly match the order of the variables in the input CSV file for calculating scores.

CAUTION

Currently, the validation of the file is quite limited; error checking is very limited and will cause the solver to crash. Similarly, the semantics of the file are not validated. In particular, the code does not currently validate that each variable occurs in exactly one of the groups of a static pattern database.

Thus, extreme care should be used with the file pattern databases. Please follow Issue #23 for updates about semantic validation.