Algorithm Selection Library: Format Specification

August 16, 2016

Bernd Bischl[a], Lars Kotthoff[b], Marius Lindauer[c], Yuri Malitsky[b], Alexandre Frechétte[d], Holger Hoos[d], Frank Hutter[c], Pascal Kerschke[e], Kevin Leyton-Brown[d], Kevin Tierney[f], Joaquin Vanschoren[g]

[a] *University of Dortmund; Germany*
[b] *Cork Constraint Computation Centre, Ireland*
[c] *University of Freiburg, Germany*
[d] *University of British Columbia, Vancouver, Canada*
[e] *University of Münster; Germany*
[f] *University of Paderborn, Germany*
[g] *Eindhoven University of Technology, The Netherlands*

---

1. Directory / File Layout

- Every algorithm selection scenario is stored in an individual folder on the server.
- The folder contains at most one copy of the following, optional files are marked with an (*)
  - *readme.txt*: Human-readable with general information.
  - *description.txt*: Global definitions for the scenario.
  - *feature_values.arff*: Values of features used to predict solvers.
  - *feature_runstatus.arff*: Technical info about features, e.g., aborts.
  - *feature_costs.arff* (*): Costs of feature (step) calculation.
  - *algorithm_runs.arff*: Algorithm runs and performance measurements.
  - *cv.arff* (*): Cross-validation splits for evaluation. Users are allowed to (if they have good reason) to submit this file, but in general it will be generated on the repository server automatically, when the rest of the data set is submitted.

Email addresses: bischl@statistik.tu-dortmund.de (Bernd Bischl), larsko@4c.ucc.ie (Lars Kotthoff), manju@cs.uni-potsdam.de (Marius Lindauer), y.malitsky@4c.ucc.ie (Yuri Malitsky), afrechet@cs.ubc.ca (Alexandre Frechétte), hoos@cs.ubc.ca (Holger Hoos), fh@informatik.uni-freiburg.de (Frank Hutter), kerschke@uni-muenster.de (Pascal Kerschke), kevinlb@cs.ubc.ca (Kevin Leyton-Brown), tierney@dsor.de (Kevin Tierney), j.vanschoren@tue.nl (Joaquin Vanschoren)

– *ground_truth.arff* (*): Information about solution of instances.
– *citation.bib* (*): Citation file for data set.

2. General Remarks

Please keep the following in mind when generating the files.

- Every file base name / column name / factor level name adheres to: ASCII characters; if whitespaces or commas are used, the name has to be quoted
- Instance IDs can also contain "/" to represent an original file path.
- We **highly** recommend using descriptive instance names rather than using numerical IDs. One possible purpose for this is that such descriptive instance names can later be used to reproduce experiments.
- All pairs of instance ID and repetition number have to be unique.
- Use "?" to denote missing values conforming to the rules of the ARFF mechanism. Each such "?" must be explained in the *readme* making it clear why the missing values occurred where they did. Also, use the "runstatus" concept, explained below, for features and algorithms. Never use anything else except "?" for missing values.
- relations of all *arff*-files should begin with the information type and end with scenario_id (see *description.txt*); e.g., if we have instance features of the 2013 SAT Competition, the relation name of the *feature_values* should be `@RELATION FEATURE_VALUES_2013-SAT-Competition`. In case of doubt, this helps to identify files which belong together

3. readme.txt

This is a human readable file meant to allow the creator to explain any and all peculiarities of how the data came into existence. Most importantly, this file is where the creators must explain the reason for any missing data in the set. In addition to this, consider the following as a starting list of things that must be mentioned in this file. But please be aware that explaining things here is no excuse to not have perfectly machine-readable information in the other files or for circumventing format specifications!

- Must describe where and how the data originated.
- Describe the problem that the algorithms are supposed to solve and how their performance is measured.
- Describe the used solvers, their configurations, versions and origins.
- Reference the feature generator.
- Try to include all relevant details of the underlying experiment.

4. description.txt

A global description file containing general information about the scenario in YAML format.

- scenario_id [string]: Name of the scenario
  This id is necessary for human bookkeeping and making sure that all other files in the directory are for the same set of experiments. In the ARFF files, this id will be present under the "@RELATION" line.

- performance_measures [list of strings]: Name of performance metrics measured for the algorithms
  While in many cases only a single entry might be reasonable, there are times when an experiment may yield secondary objectives. As such, all these measures must be listed in decreasing order of importance, if there is such an order, the first one being the one of primary interest.

- maximize [list of Booleans]: true / false
  For each of the objectives listed in the *performance_measures* line, this line specifies if the objective is to minimize or maximize the corresponding value. The number of entries must match number of entries in "performance_measures".

- performance_type [list of factors]: runtime / solution_quality For each of the objectives listed for "performance_measures", this line specifies whether the corresponding metric is"runtime" or "solution_quality". Number of entries must match number of entries in "performance_measures".

- algorithm_cutoff_time [integer]: Cut-off time in seconds
  Algorithms are terminated after this time if they did not produce a successful solution. Might be used both for "runtime" and "solution_quality". Put "?" if it was not set for experiment.

- algorithm_cutoff_memory [integer]: Cut-off memory in megabytes
  Algorithms are terminated if their memory exceeded this value. Might be used both for "runtime" and "solution_quality". Put "?" if it was not set for experiment.

- features_cutoff_time [integer]: Cut-off time in seconds
  Feature set computation is terminated if it exceeds this time. Might be used both for "runtime" and "solution_quality". Put "?" if it was not set for experiment.

- features_cutoff_memory [integer]: Cut-off memory in megabytes
  Feature set computation is terminated if memory exceeded this value. Might be used both for "runtime" and "solution_quality". Put "?" if it was not set for experiment.

- algorithms_deterministic [list of strings]:
  List names of all algorithms which are deterministic.

- algorithms_stochastic [list of strings]:
  List names of all algorithms which are stochastic.

- number_of_feature_steps [integer]: Number of feature steps.

- feature_steps [list of objects describing feature steps]:
  Objects specifying the feature step name, the features it provides, and any dependencies on other feature steps. For instance, probing features need normally a preprocessing step. Therefore probing features depend on the preprocessing step and the probing step. This is helpful for *feature_costs.arff*, as feature generators often calculate features in several steps. Each feature has to be listed in at least one step. To be able to use a feature, all feature steps have to be used which are listed in the feature step's requires section. The definition looks like this:

```
name_1:
        provides:
                - feature1
                - feature2
                - feature4
name_2:
        requires:
                - name_1
        provides:
                - feature3
```

  You are free to use any step names you like, as long as they are unique and do not contain illegal characters.

- default_steps [list of strings]:
  List names of all features which should be used as a default.

- features_deterministic [list of strings]:
  List names of all features processing steps which are deterministic.

- features_stochastic [list of strings]:
  List names of all features which are stochastic.

5. feature_values.arff

This file contains the numerical feature values for all instances. Specifically, it is generally assumed that this is the information that will be used by implemented techniques to differentiate between different instances. Like with all subsequent files, the data stored here should follow the conventions of an ARFF file. It is therefore expected that the rows will correspond to a specific (instance, repetition) pair.

Listing 1: Example description.txt

```
scenario_id: 2013-SAT-Competition
performance_measures:
    - PAR10
    - Number_of_satisfied_clauses
maximize:
    - false
    - true
performance_type:
    - runtime
    - solution_quality
algorithm_cutoff_time: 900
algorithm_cutoff_memory: 6000
features_cutoff_time: 90
features_cutoff_memory: 6000
algorithms_deterministic:
    - lingeling
    - clasp
algorithms_stochastic:
    - sparrow
number_of_feature_steps: 2
feature_steps:
    preprocessing:
        provides:
            - number_of_variables
    local_search_probing:
        requires:
            - preprocessing
        provides:
            - first_local_min_steps
default_steps:
    - preprocessing
features_deterministic:
    - number_of_variables
    - first_local_min_steps
features_stochastic: first_local_min_steps
```

- The first column is called "instance_id" and contains the instance name represented as a string. These names must follow the guidelines of utilizing only ASCII characters.
- The second column specifies the "repetition" and contains integers, starting at 1 and increasing in increments of 1 for each instance with the same name.
- The following columns correspond to instance features. We allow numeric, integer and categorical columns, but only the names defined in *description.txt* can be used. In that file, these names were specified in fields "features_step".
- "?" means the feature value is missing because the feature calculation algorithm crashed or aborted or some other problem occurred. The explanation of such a missing value will need to be registered in *feature_runstatus.arff*.
- In the case of stochastic features, the user might opt to repeat the feature calculation (column "repetition"). The following remarks have to be respected:
  - In case of no repetitions, do not omit the column, simply put a 1 in every entry.
  - If you use repetitions, you can use a different number of repetitions for instances (if you really insist on it), but always the same number of repetitions for features in one row (the latter is enforced by the format definition).
  - If you use repetitions, but some features are deterministic, simply repeat their feature values across the repetitions.
  - Whether features are stochastic or deterministic is defined implicit over the feature steps in "features_steps_deterministic" and "features_steps_stochastic" in in *description.txt*.

Listing 2: Example feature_values.arff with three features

```
1  @RELATION FEATURE_VALUES_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE repetition NUMERIC
5  @ATTRIBUTE number_of_variables NUMERIC
6  @ATTRIBUTE number_of_clauses NUMERIC
7  @ATTRIBUTE first_local_min_steps NUMERIC

9  @DATA
10 inst1.cnf,1,101,24,33
11 inst1.cnf,2,101,24,42
12 inst2.cnf,1,303,105,12
13 inst2.cnf,2,303,105,?
```

```
14 inst3.cnf,1,1002,337,?
15 inst3.cnf,2,1002,337,?
16 ...
```

6. feature_runstatus.arff

This file contains technical information about the feature calculation in general. Specifically, it serves to state whether the execution of feature processing steps was successful or if some crash occurred. The file is designed to allow the user to specify what kind of a crash has transpired, but the reasons for each such situation must be explained in the *readme.txt* file.

We assume that feature processing steps are able to generate features if its execution terminates successfully. If a step terminates because of another reason, e.g., timeout or memout, all features are not available which depend on this processing step as defined in *description.txt*.

- The first column is the "instance_id" and contains the instance name in string format. These names must follow the guidelines of utilizing only ASCII characters.
- The second column specifies the "repetition" and contains integers, starting at 1 and increasing in increments of 1 for each instance with the same name.
- The two columns, "instance_id" and "repetition", must represent exactly the same data as presented in *feature_values.arff*.
- All remaining columns correspond to feature processing steps. Here, the same column names must be adhered to as was defined in *description.txt*. The entries are categoric values specifying the termination condition for each step. The supported range of these values is:
  - "ok" - The step was executed normally.
  - "timeout" - Time ran out before this step terminated.
  - "memout" - Feature computation ran out of memory causing a crash.
  - "presolved" - Instance was solved during feature computation.
  - "crash" - The program failed to execute.
  - "unknown" - Feature computation was not run, probably because an earlier feature computation step presolved the instance.
  - "other" - Some other critical problem occurred while this feature was computed.
- If feature calculation is aborted ("crash" or "other"), you must explain the reasons in the *readme.txt*.

- If at least one used steps has the state "presolved", the instance was solved during feature set calculation.

Listing 3: Example feature_runstatus.arff with two feature processing steps

```
1  @RELATION FEATURE_RUNSTATUS_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE repetition NUMERIC
5  @ATTRIBUTE preprocessing {ok, timeout, memout, presolved, crash, other}
6  @ATTRIBUTE local_search_probing {ok, timeout, memout, presolved, crash, other}

8  @DATA
9  inst1.cnf,1,ok,ok
10 inst1.cnf,2,ok,ok
11 inst2.cnf,1,ok,ok
12 inst2.cnf,2,ok,timeout
13 inst3.cnf,1,ok,memout
14 inst3.cnf,2,ok,presolved
15 ...
```

7. feature_costs.arff

Although usually a minor overhead, feature computation is rarely free. In many cases this cost is in the form of time, but no such requirement is made. Instead, this file specifies the cost of computing each feature processing step.

We strongly recommend to provide information about feature costs. However, you are allowed to omit this file if your scenario does not entail feature costs or you have absolutely no knowledge of them. Please note that certain analysis, e.g., runtime improvement against best single algorithm, are not possible, if feature costs are not provided.

- The first column is the "instance_id" and contains the instance name represented as a string. These names must follow the guidelines of utilizing only ASCII characters
- The second column is an integer specifying the "repetition", starting at 1 and increasing in increments of 1 per each instance repetition.
- The "instance_id" and "repetition" columns must represent exactly the same data as in *feature_values.arff*.
- All columns correspond to the costs of each feature processing steps. The names of these steps must be exactly the same as those specified in *description.txt* in the feature step section. Entries are numerical, they specify the cost of the processing step for that instance / repetition. We recommend to specify a feature processing step for each feature, if the costs for each feature is known. If only the computation cost for all features is known, specify exactly one processing step.

- Put "?" as an entry if the feature computation was not successful due to cut-off or unusual abort.

Listing 4: Example feature_costs.arff

```
1  @RELATION FEATURE_COSTS_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE repetition NUMERIC
5  @ATTRIBUTE preprocessing NUMERIC
6  @ATTRIBUTE local_search_probing NUMERIC

8  @DATA
9  inst1.cnf,1,0.1,10.0
10 inst1.cnf,2,0.1,12.0
11 inst2.cnf,1,0.2,4.0
12 inst2.cnf,2,0.2,?
13 inst3.cnf,1,1.1,?
14 inst3.cnf,2,1.1,?
15 ...
```

8. algorithm_runs.arff

This file contains information on all algorithms evaluated on the instances. This includes, but not limited to, their performance values and runstatus. In practice, certain algorithms may be evaluated multiple times, while others can be performed just once. For example, this is the case for stochastic and deterministic solvers, respectively. To cover both scenarios, each row of this file specifies a single experiment. Please note that different algorithms are allowed to have different numbers of repetitions. Similarly, the number of repetitions used for stochastic algorithms does not have to coincide with the number of repetitions used for stochastic features.

- Each row corresponds to the measurement of 1 algorithm on one instance / repetition.
- The first column is the "instance_id" and contains the instance name as a string.
- The second column is an integer that states the "repetition", starting at 1 and increasing in increments of 1 per each instance repetition. The number of repetitions can vary between the algorithms, e.g., the performance of deterministic algorithms are only measured once and of stochastic ones more than once. However, the combination of algorithm and number of repetition should be the same for all instances.
- "instance_id" must contain exactly the same elements as the corresponding column in *feature_values.arff*.

- The third column is called "algorithm", a string value specifying the name of the evaluated algorithm. Only the names defined in *description.txt* may be used; those listed under the fields "algorithms_deterministic" and "algorithms_stochastic".
- All but the last subsequent columns are numerical, containing the algorithm performance measurements.
- The last column is called "runstatus" and contains whether the algorithm terminated normally or the reason for an abort.
  - "ok" - The algorithm finished properly.
  - "timeout" - Time ran out prior to completion.
  - "memout" - The algorithm required more memory than permitted.
  - "not_applicable" - This algorithm can't be run on this instance.
  - "crash" - The program failed to execute.
  - "other" - Something really unexpected happened.
- If the run was aborted ("crash" or "other"), you have to explain the reasons in the *readme.txt*.
- "?" are not permitted in this file. In case of scenarios with runtime as performance type, the runtime is always known regardless of the runstatus. In case of scenarios with solution quality as performance type, missing values have to be imputed somehow. However, this imputation is domain-specific and hence, it has to be provided by the scenario designer.

Listing 5: Example algorithm_runs.arff with four algorithms

```
1  @RELATION ALGORITHM_RUNS_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE repetition NUMERIC
5  @ATTRIBUTE algorithm STRING
6  @ATTRIBUTE PAR10 NUMERIC
7  @ATTRIBUTE Number_of_satisfied_clauses NUMERIC
8  @ATTRIBUTE runstatus {ok, timeout, memout, not_applicable, crash, other}

10 inst1.cnf, 1, lingeling, 5.0, 24, ok
11 inst1.cnf, 1, clasp, 5.8, 24, ok
12 inst1.cnf, 1, sparrow, 900, 20, timeout
13 inst1.cnf, 2, sparrow, 1.7, 24, ok
14 inst2.cnf, 1, lingeling, 588.1, 105, ok
15 inst2.cnf, 1, clasp, 50, 0, memout
16 inst2.cnf, 1, sparrow, 501, 105, ok
17 inst2.cnf, 2, sparrow, 900, 101, timeout
18 ...
```

9. cv.arff

This file contains information how to split the set of instances according to a cross-validation scheme to assess the performance of an algorithm selector in an unbiased and standard fashion. The file is provided to guarantee that all algorithm selectors are evaluated on the same splits. We support repeated cross-validation, where the process of a single cross-validation is simply repeated. These repetitions are a standard technique for smaller data sets or in cases where other sources of instability might be present in the data.

- The first column is the "instance_id" and contains the instance name as a string. The column must contain exactly the same elements as the corresponding column in *feature_values.arff* at least once.
- The second column is an integer that states the "repetition", starting at 1 and increasing in increments of 1 per each instance repetition. Note that this is for repeated cross-validation and has no connection to repeated algorithm runs or repeated feature evaluations. The column is always present, for a simple cross-validation it is always 1.
- The third column is an integer that states the "fold" of the cross-validation, starting at 1.

Listing 6: Example cv.arff for two times a 2-fold cross validation

```
1  @RELATION CV_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE repetition NUMERIC
5  @ATTRIBUTE fold NUMERIC

7  inst1.cnf, 1, 1
8  inst2.cnf, 1, 1
9  inst3.cnf, 1, 2
10 inst4.cnf, 1, 2
11 inst1.cnf, 2, 1
12 inst2.cnf, 2, 2
13 inst3.cnf, 2, 2
14 inst4.cnf, 2, 1
15 ...
```

10. ground_truth.arff

This is an optional file meant to keep the known information about the instance. This information can be used both as a sanity check to make sure the solvers return correct solutions, but can also be used for cases where a new evaluation metric depends on the optimal solution. Each row specifies exactly one instance, and no duplicates are allowed.

- Rows correspond to instances.
- First column is called "instance_id" and contains a string representing the instance name.
- The "instance_id" column has exactly the same data as in *feature_values.arff*.
- The other columns provide information about ground truths of each instance, e.g.,
  - Satisfiable or Unsatisfiable for SAT instances.
  - Optimal quality of optimization problems.
  - Any other known optimal value of a metric.
- Put "?" if you do not know the ground truth for a particular instance.
- Omit this file if you do not have this information for any instance.

Listing 7: Example ground_truth.arff

```
1  @RELATION GROUND_TRUTH_2013-SAT-Competition

3  @ATTRIBUTE instance_id STRING
4  @ATTRIBUTE SATUNSAT {SAT,UNSAT}
5  @ATTRIBUTE OPTIMAL_VALUE NUMERIC

7  @DATA
8  inst1.cnf, SAT, 24
9  inst2.cnf, UNSAT, 105
10 ...
```

11. citation.bib

Please provide a bibtex file containing the references that should be cited when using this data