Package 'aslib'

November 25, 2016

Title Interface to the Algorithm Selection Benchmark Library

Description Provides an interface to the algorithm selection benchmark library at <http://www.aslib.net> and the 'LLAMA' package (<https://cran.r-project.org/web/packages/llama/index.html>) for building algorithm selection models.

Author Bernd Bischl <bernd_bischl@gmx.net>, Lars Kotthoff <larsko@cs.ubc.ca>, Pascal Kerschke <kerschke@uni-muenster.de> [ctb]

Maintainer Lars Kotthoff <larsko@cs.ubc.ca>

URL https://github.com/coseal/aslib-r/

BugReports https://github.com/coseal/aslib-r/issues

License GPL-3

Imports BatchExperiments, BatchJobs, BBmisc, checkmate, corrplot, ggplot2, llama, mlr, parallelMap, ParamHelpers, plyr, reshape2, RWeka, stringr, yaml

Suggests testthat, rpart

LazyData yes

ByteCompile yes

Version 0.1

RoxygenNote 5.0.1

NeedsCompilation no

Repository CRAN

Date/Publication 2016-11-25 08:42:53

R topics documented:

`ASScenarioDesc` *S3 class for ASScenarioDesc.*

Description

Object members

Details

scenario_id [`character(1)`] Name of scenario.

performance_measures [`character`] Names of measures.

maximize [**named** `character`] Maximize measure?

performance_type [**named** `character`] Either "runtime" or "solution_quality".

algorithm_cutoff_time [`numeric(1)`] Cutoff time for an algorithm run.

algorithm_cutoff_memory [`numeric(1)`] Cutoff memory for an algorithm run.

features_cutoff_time [`numeric(1)`] Cutoff time for a feature run.

features_cutoff_memory [`numeric(1)`] Cutoff memory for a feature run.

feature_steps [**named** `list` **of** `character`] Names of feature processing steps, the other feature steps they require, and the features they provide.

metainfo_algorithms [named `list` **of lists of** `character`] Names of algorithms and meta-information about them.

`checkDuplicatedInstances` *Checks the feature data set for duplicated instances.*

Description

Potentially duplicated instances are detected by grouping all instances with equal feature vectors.

Usage

```
checkDuplicatedInstances(asscenario)
```

Arguments

`asscenario`	[ASScenario] Algorithm selection scenario.

Value

`list` of `character` . List of instance id vectors where corresponding feature vectors are the same. Only groups of at least 2 elements are returned.

`convertAlgoPerfToWideFormat`

Converts `algo.runs` *object of a scenario to wide format.*

Description

The first 2 columns are "instance_id" and "repetition". The remaining ones are the measured performance values. The feature columns are in the same order as "features_deterministic", "features_stochastic" in the description object. codeNA means the performance value is not available, possibly because the algorithm run was aborted. The data.frame is sorted by "instance_id", then "repetition".

Usage

```
convertAlgoPerfToWideFormat(desc, algo.runs, measure)
```

Arguments

desc	[ASScenarioDesc] Description object of scenario.
algo.runs	[data.frame] Algo runs data.frame from scenario.
measure	[character(1)] Selected performance measure. Default is first measure in scenario.

Value

data.frame .

---

convertToLlama *Convert an ASScenario scenario object to a llama data object.*

Description

For features, mean values are computed across repetitions. For algorithms, repetitions are not supported at the moment and will result in an error.

Usage

```
convertToLlama(asscenario, measure, feature.steps)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure to use for modeling. Default is first measure in scenario.
feature.steps	[character] Which feature steps are allowed? Default are the default feature steps or all steps in case no defaults were defined.

Details

Note that feature step dependencies are currently not supported explicitly by LLAMA. The conversion checks that all dependencies are satisfied, but subsequent feature selection on the LLAMA data frame may not work as expected.

Value

Result of calling input.

convertToLlamaCVFolds *Convert an ASScenario scenario object to a llama data object with cross-validation folds.*

Description

For features, mean values are computed across repetitions. For algorithms, repetitions are not supported at the moment and will result in an error.

Usage

```
convertToLlamaCVFolds(asscenario, measure, feature.steps, cv.splits)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure to use for modelling. Default is first measure in scenario.
feature.steps	[character] Which feature steps are allowed? Default are the default feature steps or all steps in case no defaults were defined.
cv.splits	[data.frame] Data frame defining the split of the data into cross-validation folds, as returned by createCVSplits. Default are the splits asscenario$cv.splits

Value

Result of calling input with data partitioned into folds.

createCVSplits *Create cross-validation splits for a scenario.*

Description

Create a data.frame that defines cross-validation splits for a scenario,

and potentially store it in an ARFF file.

The mlr package is used to generate the splits, see makeResampleDesc and makeResampleInstance.

Usage

```
createCVSplits(asscenario, reps = 1L, folds = 10L, file = NULL)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
reps	[integer] CV repetitions. Default is 1.
folds	[integer] CV folds. Default is 10.
file	[character] If not missing, where to save the returned splits as an ARFF file via write.arff. Default is no saving.

Value

data.frame . Splits as defined in the algorithm benchmark repository specification text. Has columns: "instance_id", "fold", "rep". Defines which instances go into the test set for each replication / fold during CV. The training set are the remaining instances, in exactly the order as given by the data.frame for the current repetition.

---

findDominatedAlgos *Creates a table that shows the dominance of one algorithm over another one.*

Description

If NAs occur, they are imputed (before aggregation) by base + 0.3 * range. base is the cutoff value for runtimes scenarios with cutoff or the worst performance for all others.

Stochastic replications are aggregated by the mean value.

Usage

```
findDominatedAlgos(asscenario, measure, reduce = FALSE, type = "logical")
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure for algorithm performance. Default is first measure in scenario.
reduce	[logical(1)] Should the resulting matrix be reduced to algorithms that a are either dominated by or dominate another algorithm? Default is FALSE.
type	[character(1)] Data type of the result object. "logical": Logical matrix, TRUE means row algorithm dominates column algorithm. "character": Same information but more human-readable. States how the row relates to the column.

Value

matrix . See above.

---

fixFeckingPresolve *Bakes presolving stuff into a LLAMA data frame.*

Description

Determines whether any of the feature groups in the LLAMA data frame presolve any of the instances. If so, the performances of all algorithms in the portfolio are set to the runtime of the first used feature group that presolves the respective instance. Furthermore, the success of all algorithms on those instances is set to true.

Usage

```
fixFeckingPresolve(asscenario, ldf)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
ldf	[LLAMA data frame] LLAMA data frame to modify.

Details

These modifications are done on the main LLAMA data and on any test splits. They are *not* done on the training data. This function should only ever be used to evaluate the performance of an actual selector that uses features (i.e. not VBS or single best). Using it in polite company is to be avoided.

Value

The LLAMA data frame with presolving baked into the algorithm performances.

---

getAlgorithmNames *Returns algorithm names of scenario.*

Description

Returns algorithm names of scenario.

Usage

```
getAlgorithmNames(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

character .

`getCosealASScenario` *Retrieves a scenario from the Coseal Github repository and parses into an S3 object.*

Description

Uses subversion export to retrieve a specific scenario from the official Coseal Github repository. The scenario is checked out into a temporary directory and parsed with `parseASScenario`.

Usage

```
getCosealASScenario(name)
```

Arguments

name	[character(1)] Name of benchmark data set.

Value

ASScenario . Description object.

Examples

```
## Not run:
  sc = getCosealASScenario("CSP-2010")

## End(Not run)
```

`getCostsAndPresolvedStatus` *Return wether an instance was presolved and which step did it.*

Description

Return wether an instance was presolved and which step did it.

Usage

```
getCostsAndPresolvedStatus(asscenario, feature.steps)
```

Arguments

`asscenario`	[ASScenario] Algorithm selection scenario.
`feature.steps`	[character] Which feature steps are allowed? Default is all steps.

Value

`list` . In the following, `n` is the number of instances. All following object are ordered by "instance_id".

`is.presolved [logical(n)]`
: Was instance presolved? Named by instance ids.

`solve.steps [character(n)]`
: Which step solved it? NA if no step did it. Named by instance ids.

`costs [numeric(n)]`
: Feature costs for using the steps. Named by instance ids. NULL if no costs are present.

`getDefaultFeatureStepNames` *Returns the default feature step names of scenario.*

Description

Returns the default feature step names of scenario.

Usage

```
getDefaultFeatureStepNames(asscenario)
```

Arguments

asscenario [ASScenario]
Algorithm selection scenario.

Value

character .

---

getFeatureNames *Returns feature names of scenario.*

Description

Returns feature names of scenario.

Usage

```
getFeatureNames(asscenario)
```

Arguments

asscenario [ASScenario]
Algorithm selection scenario.

Value

character .

---

getFeatureStepNames *Returns feature step names of scenario.*

Description

Returns feature step names of scenario.

Usage

```
getFeatureStepNames(asscenario)
```

Arguments

asscenario [ASScenario]
Algorithm selection scenario.

Value

character .

getInstanceNames *Returns instance names of scenario.*

Description

Returns instance names of scenario.

Usage

```
getInstanceNames(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

character .

getNumberOfCVFolds *Returns number of CV folds.*

Description

Returns number of CV folds.

Usage

```
getNumberOfCVFolds(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

integer(1) .

getNumberOfCVReps *Returns number of CV repetitions.*

Description

Returns number of CV repetitions.

Usage

```
getNumberOfCVReps(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

integer(1) .

getProvidedFeatures *Return features that are useable for a given set of feature steps.*

Description

Return features that are useable for a given set of feature steps.

Usage

```
getProvidedFeatures(asscenario, steps)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
steps	[character] Feature steps. Default are all feature steps.

Value

character .

getSummedFeatureCosts *Returns feature costs of scenario, summed over all instances.*

Description

Returns feature costs of scenario, summed over all instances.

Usage

```
getSummedFeatureCosts(asscenario, feature.steps)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
feature.steps	[character] Sum costs only for these selected steps. Default are all feature steps.

Value

character .

imputeAlgoPerf *Imputes algorithm performance for runs which have NA performance values.*

Description

The following formula is used for imputation: base +- range.scalar * range.span + N(0, sd = jitter * range.span
With range.span = max - min.

Returns an object like algo.runs of asscenario, but drops the runstatus and all other measures.

Usage

```
imputeAlgoPerf(asscenario, measure, base = NULL, range.scalar = 0.3,
  jitter = 0, impute.zero.vals = FALSE)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure to impute. Default is first measure in scenario.

base [numeric(1)]
See formula. Default is NULL, which means maximum of performance values if measure should be minimized, or minimum for maximization case.

range.scalar [numeric(1)]
See formula. Default is 0.3.

jitter [numeric(1)]
See formula. Default is 0.

impute.zero.vals
[logical(1)]
Should values which are exactly 0 be imputed to 1e-6? This allows to take the logarithm later on, handy for subsequent visualizations. Note that this really only makes sense for non-negative measures! Default is FALSE.

Value

data.frame .

---

parseASScenario *Parses the data files of an algorithm selection scenario into an S3 object.*

Description

Object members

Let n be the number of (replicated) instances, m the number of unique instances, p the number of features, s the number of feature steps and k the number of algorithms.

desc [ASScenarioDesc] Description object, containing further info.

feature.runstatus [data.frame(n, s + 2)] Runstatus of feature computation steps. The first 2 columns are "instance_id" and "repetition", the remaining are the status factors. The step columns are in the same order as the feature steps in the description object. The factor levels are always: ok, presolved, crash, timeout, memout, other. No entry can be NA. The data.frame is sorted by "instance_id", then "repetition".

feature.costs [data.frame(n, s + 2)] Costs of feature computation steps. The first 2 columns are "instance_id" and "repetition", the remaining are numeric costs of the feature steps. The step columns are in the same order as the feature steps in the description object. codeNA means the cost is not available, possibly because the feature computation was aborted. The data.frame is sorted by "instance_id", then "repetition". If no cost file is available at all, NULL is stored.

feature.values [data.frame(n, p + 2)] Measured feature values of instances. The first 2 columns are "instance_id" and "repetition". The remaining ones are the measured instance features. The feature columns are in the same order as "features_deterministic", "features_stochastic" in the description object. codeNA means the feature is not available, possibly because the feature computation was aborted. The data.frame is sorted by "instance_id", then "repetition".

algo.runs [data.frame] Runstatus and performance information of the algorithms. Simply the parsed ARFF file. See convertAlgoPerfToWideFormat for a more convenient format.

algo.runstatus [`data.frame(n, k + 2)`] Runstatus of algorithm runs. The first 2 columns are "instance_id" and "repetition", the remaining are the status factors. The step columns are in the same order as the feature steps in the description object. The factor levels are always: ok, presolved, crash, timeout, memout, other. No entry can be `NA`. The data.frame is sorted by "instance_id", then "repetition".

cv.splits[`data.frame(m, 3)`] Definition of cross-validation splits for each replication of a repeated CV with folds. Has columns "instance_id", "repetition" and "fold". The instances with fold = i for a replication r constitute the i-th test set for the r-th CV. The training set is the "instance_id" column with repetition = r, in the same order, when the test set is removed. The data.frame is sorted by "repetition", then "fold", then "instance_id". If no CV file is available at all, `NULL` is stored, and a warning is issued, although this should not happen.

Usage

```
parseASScenario(path)
```

Arguments

`path`	[`character(1)`] Path to directory of benchmark data set.

Value

`ASScenario` . Description object.

See Also

`writeASScenario`

Examples

```
## Not run:
  sc = parseASScenario("/path/to/scenario")

## End(Not run)
```

`plotAlgoCorMatrix` *Plots the correlation matrix of the algorithms.*

Description

If NAs occur, they are imputed (before aggregation) by `base + 0.3 * range`. `base` is the cutoff value for runtimes scenarios with cutoff or the worst performance for all others.

Stochastic replications are aggregated by the mean value.

Usage

```
plotAlgoCorMatrix(asscenario, measure, order.method = "hclust",
  hclust.method = "ward.D2", cor.method = "spearman")
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure to plot. Default is first measure in scenario.
order.method	[character(1)] Method for ordering the algorithms within the plot. Possible values are "hclust" (for hierarchical clustering order), "FPC" (first principal component order), "AOE" (angular order of eigenvectors), "original" (original order) and "alphabet" (alphabetical order). See corrMatOrder. Default is "hclust".
hclust.method	[character(1)] Method for hierarchical clustering. Only useful, when order.method is set to "hclust", otherwise ignored. Possible values are: "ward.D2", "single", "complete", "average", "mcquitty", "median" and "centroid". See corrMatOrder. Default is "ward.D2".
cor.method	[character(1)] Method to be used for calculating the correlation between the algorithms. Possible values are "pearson", "kendall" and "spearman". See cor. Default is "spearman".

Value

See corrplot.

---

plotAlgoPerf *EDA plots for performance values of algorithms across all instances.*

Description

If NAs occur, they are imputed (before aggregation) by base + 0.3 range + jitter. base is is the cutoff value for runtimes scenarios with cutoff or the worst performance for all others.

For the CDFs we only show the visible area where successful runs occurred.

Stochastic replications are aggregated by the mean value.

Usage

```
plotAlgoPerfBoxplots(asscenario, measure, impute.zero.vals = FALSE,
  log = FALSE, impute.failed.runs = TRUE, rm.censored.runs = TRUE)

plotAlgoPerfCDFs(asscenario, measure, impute.zero.vals = FALSE, log = FALSE,
  rm.censored.runs = TRUE)

plotAlgoPerfDensities(asscenario, measure, impute.failed.runs = TRUE,
  impute.zero.vals = FALSE, log = FALSE, rm.censored.runs = TRUE)

plotAlgoPerfScatterMatrix(asscenario, measure, impute.zero.vals = FALSE,
  log = FALSE, rm.censored.runs = TRUE)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Measure to plot. Default is first measure in scenario.
impute.zero.vals	[logical(1)] Should values which are exactly 0 be imputed to 1e-6? This allows to take the logarithm later on, handy for subsequent visualizations. Note that this really only makes sense for non-negative measures! Default is FALSE.
log	[logical(1)] Should the performance values be log10-transformed in the plot? Default is FALSE.
impute.failed.runs	[logical(1)] Should runtimes for failed runs be imputed? Default is TRUE.
rm.censored.runs	[logical(1)] Should runtimes for censored runs (i.e. runs that have hitted the walltime) be removed (and eventually be imputed along with the remaining NAs)? Default is TRUE.

Value

ggplot2 plot object.

---

runLlamaModels *Creates a registry which can be used for running several Llama models on a cluster.*

Description

It is likely that you need to install some additional R packages for this from CRAN or extra Weka learner. The latter can be done via e.g. `WPM("install-package", "XMeans")`.

Feature costs are added for real prognostic models but not for baseline models.

Usage

```
runLlamaModels(asscenarios, feature.steps.list = NULL, baselines = NULL,
  learners = list(), par.sets = list(), rs.iters = 100L,
  n.inner.folds = 2L)
```

Arguments

asscenarios	[(list of) ASScenario] Algorithm selection scenarios.
feature.steps.list	[list of character] Named list of feature steps we want to use. Must be named with scenario ids. Default is to take the default feature steps from the scenario.
baselines	[character] Vector of characters, defining the baseline models. Default is c("vbs", "singleBest", "singleBestByPar", "singleBestBySuccesses").
learners	[list of Learner] mlr learners to use for modeling. Default is none.
par.sets	[list of ParamSet] Param sets for learners to tune via random search. Pass an empty param set, if you want no tuning. Must be in of same length as learners and in the same order. Default is none.
rs.iters	[integer(1)] Number of iterations for random search hyperparameter tuning. Default is 100.
n.inner.folds	[integer(1)] Number of cross-validation folds for inner CV in hyperparameter tuning. Default is 2L.

Value

BatchExperiments registry.

---

summarizeAlgoPerf *Creates summary data.frame for algorithm performance values across all instances.*

Description

Creates summary data.frame for algorithm performance values across all instances.

Usage

```
summarizeAlgoPerf(asscenario, measure)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
measure	[character(1)] Selected measure. Default is first measure in scenario.

Value

data.frame .

---

summarizeAlgoRunstatus *Creates summary data.frame for algorithm runstatus across all instances.*

Description

Creates summary data.frame for algorithm runstatus across all instances.

Usage

```
summarizeAlgoRunstatus(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

data.frame .

---

summarizeFeatureSteps *Creates a data.frame that summarizes the feature steps.*

Description

Creates a data.frame that summarizes the feature steps.

Usage

```
summarizeFeatureSteps(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

data.frame .

summarizeFeatureValues *Creates summary data.frame for feature values across all instances.*

Description

Creates summary data.frame for feature values across all instances.

Usage

```
summarizeFeatureValues(asscenario)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.

Value

data.frame .

writeASScenario *Writes an algorithm selection scenario to a directory.*

Description

Splits an algorithm selection scenario into description, feature values / runstatus / costs, algorithm performance and cv splits and saves those data sets as single ARFF files in the given directory.

Usage

```
writeASScenario(asscenario, path = asscenario$desc$scenario_id)
```

Arguments

asscenario	[ASScenario] Algorithm selection scenario.
path	[character(1)] Path to write scenario to. Default is the name of the scenario.

See Also

parseASScenario

Index