

# **The TDMR Framework:** **Tuned Data Mining in R**

**Wolfgang Konen, Patrick Koch**

**Cologne University of Applied Sciences**

**Last update: February 2012**

## **Contents**

The TDMR Framework: Tuned Data Mining in R.....	1
Using TDMR .....	3
Overview.....	3
TDMR Workflow .....	3
Phase 1: DM without SPOT.....	3
Phase 2: Tuned Data Mining in R (TDMR) .....	3
Phase 3: “The Big Loop”: Several TDMs with Unbiased Evaluations.....	4
TDMR Experiment Concept .....	4
TDMR Data Reading and Data Split in Train / Validation / Test Data .....	6
Data Reading .....	6
Training / Validation / Test Set.....	7
Examples.....	8
TDMR Important Variables.....	8
TDMR <code>opts</code> Concept .....	9
TDMR <code>RGain</code> Concept.....	10
Example Usage .....	10
Phase 1: DM on task SONAR .....	10
Phase 2: SPOT tuning on task SONAR.....	12
Phase 3: “The Big Loop” on task SONAR .....	15
Extending TDMR .....	16
TDMR Tuner Concept .....	16
How to use different tuners .....	16
How to integrate new tuners .....	17
TDMR Design Mapping Concept .....	17
How to add a new tuning variable .....	18

Details .....	18
TDMR parallel computing concept.....	18
How to use parallel computing .....	18
Environment envT .....	19
Details .....	20
TDMR seed Concept .....	20
Details .....	21
TDMR Graphic Device Concept .....	22
Utility Functions tdmGraphic* .....	22
GD.RESTART, Case 1: main_TASK solo .....	22
GD.RESTART, Case 2: During SPOT-Run “auto” .....	22
GD.RESTART, Case 3: During unbiased runs.....	23
Summary .....	23
References .....	23

## Using TDMR

### Overview

The TDMR framework is written in R with the aim to facilitate the training, tuning and evaluation of data mining models. It puts special emphasis on tuning these data mining models as well as simultaneously tuning certain preprocessing options. TDMR is especially designed to work with SPOT [Bart10e] as the preferred tuner, but it offers also the possibility to use other tuners, e.g., CMA-ES [Hans06], LHD [McKay79] or direct-search optimizers [BFGS, Powell] for comparison.

This document

- gives a short overview over the TDMR framework,
- explains some of the underlying concepts and
- shows an example usage: how to use TDMR on a new data mining task.

This document concentrates more on the software usage aspects of the TDMR framework. For a more scientific discussion of the underlying concepts and the results obtained, the reader is referred to [Kone10a, Kone11b].

### TDMR Workflow

#### Phase 1: DM without SPOT

Two kinds of DM tasks, classification or regression, can be handled.

For each DM task TASK, create one task-specific function `main_TASK(opts=NULL)`, as short as possible. If called without any parameter, `main_TASK()` should set default parameters for `opts` via `tdmOptsDefaultsSet()`. `main_TASK()` reads in the task data, does the preprocessing if necessary and then calls with the preprocessed data `dset` the task-independent functions `tdmClassifyLoop` or `tdmRegressLoop`, which in turn call the task-independent functions `tdmClassify` or `tdmRegress`.

A template may be copied from `inst/demo02sonar/main_sonar.r`. [Here and in the following `inst/` refers to the directory where the package TDMR is installed. Use `.find.package("TDMR")` to locate this directory.] The template is invoked with

```
result <- main_sonar();
```

[See Sec. "Example Usage" \(Phase 1\) for a complete example.](#)

[See Table 3 for an overview of elements in list `result`.](#)

#### Phase 2: Tuned Data Mining in R (TDMR)

A TDMR task consists of a DM task (Phase 1) plus a tuner configuration (decision which parameters to tune within which ROI, which meta parameters to set for the tuner, ...).

It is recommended to create for each DM task TASK a separate subdir. In this subdir the files shown in **Table 1** have to be created for each tuner configuration (each TDMR task). In the case of SPOT as a tuner, this may look like:

**Table 1: Configuration files for a SPOT run**

.apd	problem design: all <code>opts</code> -settings
.roi	SPOT ROI file, specifies which parameters to tune in which ROI
.conf	SPOT configuration file, usually with <code>alg.func = "tdmStartSpot"</code> . Furthermore, <code>io.apdFileName</code> and <code>io.roiFileName</code> should specify the two files above.

Templates for these three files may be copied from `inst/demo02sonar/sonar_01.*`.

The whole SPOT tuning can be started with a small file `script_phase2.r` in the user directory. A template can be found in `inst/demo02sonar`:

```
source("script_phase2.r");
```

This script will define a `main_TASK` in `tdm$mainCommand`, reads the `.apd` file and calls SPOT. SPOT reads the `.conf` file, calls the generic function `tdmStartSpot(spotConfig)`, which finally executes `tdm$mainCommand`.

The only requirement on `tdm$mainCommand` is that it returns in `result$y`

a suitable quantity to be **minimized** by SPOT.

If `spot.fileMode==T`, SPOT will generate `.des` and `.aroi` files (needed by SPOT internally) and the output files `.bst` and `.res`.

If `spot.fileMode==F`, `tdmStartSpot` will read the design from `spotConfig$alg.currentDesign` and it writes the `.res` data frame onto `spotConfig$alg.currentResult`.

[See Sec. "Example Usage" \(Phase 2\) for a complete example.](#)

### *Phase 3: "The Big Loop": Several TDMs with Unbiased Evaluations*

"The Big Loop" is a script to start several Phase-2-TDMR tasks (usually on the same DM task), optionally with several tuners ([see here for a list of tuners](#)) and compare their best solutions with different modes of unbiased evaluations, e.g. on unseen test data (`tdm$umode = "TST"`) or by starting a new, independent CV (`tdm$umode = "CV"`) or by starting a new, independent resubsampling (`tdm$umode = "RSUB"`).

To start the Big Loop, only one file has to be created in the user directory: `script_all.R`. A template may be copied from [inst/demo02sonar/script\\_all.R](#).

It is invoked with

```
source("script_all.R")
```

This will specify in `runList` the list of TDMR tasks and a list of tuners. For each TDMR task and each tuner

- (a) the tuning process is started (if `spotStep="auto"`) or a previous tuning result is read in from file (if `spotStep="rep"`) and
- (b) one or more unbiased evaluations are started. This is to see whether the result quality is reproducible on independently trained models and / or on independent test data.

The result is a data frame `theFinals` with one row for each TDMR task / each tuner and several columns measuring the success of the best tuning solution in different unbiased evaluations, see **Table 2**. The data frame `theFinals` is written to `tdm$finalFile`.

[See Sec. "Example Usage" \(Phase 3\) for a complete example.](#)

## **TDMR Experiment Concept**

[TDMR Phase 3](#) ("The Big Loop") allows

- (a) to conduct experiments, where different `.conf` files, different tuners, different unbiased evaluations, ... are tried on the same task;
- (b) to repeat certain experiments of kind (a) multiple times with different seeds (`tdm$NExperim>1`).

The concept behind the experiments is as follows:

Model building:

- During model building (training) and tuning the user starts with a data set, which is partitioned into [training and validation set](#).
- The [relative gain](#) achieved on the validation set acts as *performance measure* for the tuning process.
- In the case of “cv” or in the case `opts$NRUN>1` multiple models are build, each with its own training and validation set. In this case multiple relative gains are averaged to get the performance measure for the tuning process.

#### Tuning:

- The above model building process is started several times with different model and preprocessing parameters (design points). The tuning process uses the *performance measure* returned to guide the search for better parameters.
- As a result of the tuning process, a best parameter set emerges. It has a certain performance measure attached to it, but this measure might be too optimistic (e.g. due to averaging over different runs or different folds or due to extensive searching for good solutions in a noisy environment)

#### Unbiased Evaluation (Test):

- Once a best parameter set is established, an unbiased performance test is recommended.
- This evaluation is done by calling `unbiasedRun()` with one or several values for parameter `umode`. The values are in `tdm$umode` (a vector). The possible choices for `umode` are:
  - `umode="RSUB"`: Run the model building anew, using the same data set as in the tuning process, but with another random subsample having `tdm$ststFrac` (usually 20%) test data, the rest for training.<sup>1</sup>
  - `umode="CV"`: Run the model building again with the same data set as in the tuning process, but using cross validation with `tdm$nfold` (usually 10) folds and report the CV-[relative gain](#)
  - `umode="TST"`: This is the recommended way, if you have independent test data set aside, which were not used during the whole tuning process. All the data of the tuning process (usually read from `opts$filename`) are now used for training. The test data are read from `opts$filetest`, the trained model is applied to them and the performance ([relative gain](#)) on these test data with a “fresh” model is an unbiased estimator of the model’s performance on new data.
  - `umode="SP_T"`: Similar to `umode="TST"`, but the split is done in a random fashion by TDMR.
- If `tdm$nrune>1`, each of the above values in `tdm$umode` is repeated in multiple runs and the performance measure returned is the average over all runs.

**Table 2: Elements of data frame `finals`**

<b>finals\$</b>	<b>Description</b>	<b>Condition</b>
<<columns obtained from the tuning process>>		
CONF	the base name of the .conf file	
TUNER	the value of <code>tdm\$tuneMethod</code>	
{PARAMS}	all tuned parameters appearing in .roi file	if <code>tdm\$withParams==T</code>
NEVAL	tuning budget, i.e. # of model evaluations during tuning (rows in data frame res)	
RGain.bst	best solution ( <a href="#">RGain</a> ) obtained from tuning	

<sup>1</sup> This is not the recommended way, but sometimes you have no other data than the training data and you want at least to assess, whether the model build on another random sample reaches the same performance

RGain.avg	average RGain during tuning (mean of res\$Y)	
<<columns obtained from the unbiased runs>>		
NRUN	# of runs with different test & train samples in unbiasedBestRun_*.r or # of unbiased CV-runs. Usually NRUN = tdm\$nrnrun, see fct map.opts in tdmMapDesign.r.	
RGain.OOB	mean OOB training error (averaged over all unbiased runs)	if opts\$method = *.RF
sdR.OOB	std. dev. of RGain.OOB	if opts\$method = *.RF
RGain.TRN	mean training error (averaged over all unbiased runs)	if opts\$method ≠ *.RF
sdR.TRN	std. dev. of RGain.TRN	if opts\$method ≠ *.RF
RGain.RSUB	mean test RGain (test set = random subsample)	if tdm\$umode has "RSUB"
sdR.RSUB	std. dev. of RGain.RSUB (averaged over all unbiased runs)	if tdm\$umode has "RSUB"
RGain.TST	mean test RGain (test set = separate data, user-provided)	if tdm\$umode has "TST"
sdR.TST	std. dev. of RGain.TST (averaged over all unbiased runs)	if tdm\$umode has "TST"
RGain.CV	mean test RGain (test set = CV, cross validation with tdm\$nfold CV-folds)	if tdm\$umode has "CV"
sdR.CV	std. dev. of RGain.CV (averaged over all unbiased runs)	if tdm\$umode has "CV"

More details:

- Each experiment of kind (a) initially deletes file tdm\$finalFile, if it exists, and then writes for each combination {.conf file, tuner} it encounters a line to tdm\$finalFile (usually a file with suffix .fin). This line is a one-row data frame `finals` which is built in `unbiasedBestRun_C.r` (classification) and contains the columns listed in **Table 2**.
- In the case of regression experiments (`unbiasedBestRun_R.r`) each "RGain" has to be replaced by "RMAE" in the table above, see [here](#) for further explanation.
- If `tdm$experFile` is not NULL, then the same one-row data frame `finals` is also appended to the file `tdm$experFile`. Usually, `tdm$experFile` is a file with `.exp` as suffix. This file is never deleted by the TDMR system, only the user may delete it. `tdm$experFile` serves the purpose to accumulate experiments carried out multiple times (with different random seeds). This multiple-experiment execution may be done either directly, within one 'big-loop' experiment, if `tdm$nrnrun>1`, or it may be done subsequently by the user when starting `script_all.R` again at a later point in time with the same `tdm$experFile` defined.
- An `.exp` file can be analyzed with scripts like `exp_summ.r` in `TDM.SPOT.d/appAcid/`.

## TDMR Data Reading and Data Split in Train / Validation / Test Data

### Data Reading

TDMR reads the task data from `opts$filename` and optionally also from `opts$filetest`, if `opts$READ.TST=TRUE`.

It is recommended to read the data prior to tuning with the setting `opts$READ.INIT=TRUE` (the default) and using `opts$READ.CMD`. If `opts$READ.CMD` is not defined, the default is:

```
"read.csv(file=paste(opts$dir.txt, filename, sep="\\"), nrow=opts$READ.NROW)"
```

which includes the defaults `header=TRUE`, `sep=" "`, `dec="."`.

Then TDMR will read the data into `dset` at the beginning of `tdmCompleteEval` with

```
dset <- tdmReadData(opts);
```

and let the function `tdmReadData` do the read work (including the options `opts$READ.CMD`, `opts$READ.TST`, `READ.TXT` and `READ.NROW`). `dset` is passed on with `main_TASK(..., dset=dset)`.

For downward compatibility and for special cases it is also possible to set `opts$READ.INIT=FALSE`, then `dset=NULL` and the data reading is done in `main_TASK`, for *each* tuning step anew.

### Training / Validation / Test Set

In data mining we know three kind of data or data sets:

1. **Training set:** the data for learning or model training.
2. **Validation set:** the data used to obtain a performance measure of the trained model. The performance on the validation data is used to guide the tuning process.
3. **Test set:** When training and tuning is finished, we build a final model. To estimate the quality of the model for new data, we test its performance on test data. Usually, the test data were not seen by the model or the tuner. The user should NOT use the performance on the test data in any way to tune the model further.

Usually, the split into test set on the one side and training/validation set on the other side is done once prior to the tuning process. During tuning, many tuning steps are possible, each containing at least one model training and each step may have a new separation of the training/validation set into a training part and a validation part.

How can we split the data into test set and training/validation set `dsetTrnVa`? – TDMR offers three options here, where options 2 and 3 require `opts$READ.INIT=TRUE` (the default):

1. **No test set splitting.** The whole data is used for training/validation and later also for unbiased testing. This is NOT the recommended way since the test set is already visible during training and tuning, but sometimes you may have only very few data and cannot afford to set test data aside.  
To use this option, set `tdm$umode="RSUB"` or `tdm$umode="CV"` and `opts$TST.testFrac` to the desired random fraction to be set aside (default is 10%).
2. **User-defined test set splitting.** Here we allow two sub-options: (a) The train/validation data are in `opts$filename` and the test data are in `opts$filetest`. TDMR reads both and adds a new column `opts$TST.COL` to the data frame with '0' for train/validation data and '1' for test data. (b) The user provides one data frame containing already a column `opts$TST.COL` with the appropriate '0's and '1's.  
To use option (a), set `tdm$umode="TST"` and `opts$READ.TST=TRUE`. Have `opts$filetest` and `opts$TST.COL` set to meaningful string values.  
To use option (b), set `tdm$umode="TST"` and `opts$READ.TST=FALSE`. The string `opts$TST.COL` has to name a column of the data frame read which contains already the appropriate '0's and '1's.
3. **TDMR sets a random fraction of the data aside for testing.** This is done once before the tuning starts. The test set (the data set aside for testing) is used only in the unbiased evaluation. The whole procedure can be repeated (if `tdm$nExperim>1`) and another random test set is set aside.  
This is the recommended way, it has a completely independent test set and allows to assess the variability due to varying test set selection.  
To use this option, set `tdm$umode="SP_T"` and `opts$TST.testFrac` to the desired random fraction to be set aside (default is 10%). The splitting is coded in the column `dset$tdmSplit` with '0' for training/validation and '1' for test data.

How can we further split `dsetTrnVa` into training set and validation set? – TDMR offers again three options here, which are distinguished by the value of `opts$TST.kind`:

1. **"rand" = Random Subsampling:** Sample a fraction `opts$TST.valiFrac` from `dsetTrnVa` and set it aside for validation, use the rest for training.
2. **"cv" = Cross Validation:** Split `dsetTrnVa` into `opts$TST.nfold` folds and use them for cross validation.
3. **"col" = User-Defined Column:** All records with a '0' in column `opts$TST.COL` are used for training, the rest for validation.

The split into training and validation data is done in `tdmClassifyLoop`, i.e. for each call of `main_TASK`. “col” is normally not recommended (do not use test data for tuning), but is o.k. in the special case of `opts$MOD.method="RF"` or “MC.RF”: Then we need no validation data, since we have OOB.

### Examples

<pre>opts\$READ.INI=TRUE opts\$READ.TST=TRUE opts\$filename="dmc2010_train.txt" opts\$filetest="dmc2010_test.txt" opts\$TST.kind="col" opts\$TST.COL="TST" opts\$MOD.method="RF" tdm\$umode="TST"</pre>	<p>Read the data prior to tuning, with train/validation-set from <code>dmc2010_train.txt</code>, test set from <code>dmc2010_test.txt</code>. This is coded with ‘0’ and ‘1’ in column TST of the data frame <code>dset</code>. The test set is also handed over to the tuning process as ‘validation data’, but it is not used by the tuner, because in the case of RF the performance measure is “OOB on the training set”.</p>
<pre>opts\$READ.INI=TRUE opts\$filename="sonar.txt" opts\$TST.kind="cv" opts\$TST.nfold=5 opts\$TST.testFrac=0.15 tdm\$umode="SP_T"</pre>	<p>Read the data prior to tuning from <code>sonar.txt</code>, split them by random subsampling: 15% into test set and 85% into train+validation set. This is coded with ‘0’ and ‘1’ in column “<code>tdmSplit</code>” of data frame <code>dset</code>. During tuning, the train+validation set is further split by cross validation with 5 folds (new split in each tuning step). The unbiased run uses <u>all</u> 85% train+validation data for training and reports the performance on the 15% test set data.</p>

## TDMR Important Variables

**Table 3: Overview of important variables in TDMR**

<code>opts</code>	<p>list with DM settings (used by <code>main_TASK</code> and its subfunctions). Parameter groups:</p> <ul style="list-style-type: none"> <li>• <code>opts\$READ.*</code> # reading the data</li> <li>• <code>opts\$TST.*</code> # training / validation / test set and resampling</li> <li>• <code>opts\$PRE.*</code> # preprocessing</li> <li>• <code>opts\$SRF.*</code> # sorted random forest (or similar other variable rankings)</li> <li>• <code>opts\$MOD.*</code> # genera model issues</li> <li>• <code>opts\$CLS.*</code> # classification issues</li> <li>• <code>opts\$RF.*</code> # Random Forest</li> <li>• <code>opts\$SVM.*</code> # Support Vector Machine</li> <li>• <code>opts\$GD.*</code> # graphic device issues</li> </ul>
<code>dset</code>	preprocessed data set (used within <code>main_TASK</code> and its subfunctions)
<code>result</code>	<p>list with results from Phase 1: In the case of regression, this list of class <b>TDMregressor</b> contains (see <code>tdmRegress.r</code>):</p> <ul style="list-style-type: none"> <li>• <code>opts</code> # with some settings perhaps adjusted in <code>tdmRegress</code></li> <li>• <code>lastRes</code> # last run, last fold: list of class <b>tdmRegr</b> from <code>tdmRegress</code></li> <li>• <code>R_train</code> # RMAE on training set (vector of length NRUN)</li> <li>• <code>S_train</code> # RMSE on training set (vector of length NRUN)</li> <li>• <code>T_train</code> # Theil's U for RMAE on training set (vector of length NRUN)</li> <li>• <code>*_test</code> # --- similar, with test set instead of training set</li> <li>• <code>y</code> # what to be minized by SPOT, usually <code>mean(R_test)</code></li> </ul>

	<p>In the case of classification, this list of class <b>TDMclassifier</b> contains (see <code>tdmClassify.r</code>):</p> <ul style="list-style-type: none"> <li>• <code>opts</code> # with some settings perhaps adjusted in <code>tdmClassify</code></li> <li>• <code>lastRes</code> # last run, last fold: list of class <b>tdmClass</b> from <code>tdmClassify</code></li> <li>• <code>C_train</code> # classification error on training set (vector of length NRUN)</li> <li>• <code>G_train</code> # gain on training set (vector of length NRUN)</li> <li>• <code>R_train</code> # relativ gain (% of max. gain) on training set (vector of length NRUN)</li> <li>• <code>*_test</code> # --- similar, with test set instead of training set</li> <li>• <code>*_test2</code> # --- similar, with test2 set instead of training set</li> <li>• <code>y</code> # what to be minized by SPOT, usually <code>mean(-R_test)</code></li> </ul>
<code>lastRes</code>	<p>list with results from <code>tdmClassify/tdmRegress</code>:</p> <p>In the case of classification, this list of class <b>tdmClass</b> contains (see <code>tdmClassify.r</code>):</p> <ul style="list-style-type: none"> <li>• <code>opts</code> # with some settings perhaps adjusted in <code>tdmRegress</code></li> <li>• <code>d_train</code> # training set + predicted class column(s)</li> <li>• <code>d_test</code> # test set + predicted class column(s)</li> <li>• <code>d_dis</code> # disregard set + predicted class column(s)</li> <li>• <code>sumEVAL</code> # list of evaluation measures, summed over all response variables</li> <li>• <code>alleVAL</code> # data frame with evaluation measures, one row for each target</li> <li>• <code>lastCm*</code> # confusion matrix for * = train or test</li> <li>• <code>lastModel</code> # the trained model</li> <li>• <code>lastPred</code> # name of prediction column</li> <li>• <code>lastProbs</code> # a list with three probability matrices (row: records, col: classes) # <code>v_train</code>, <code>v_test</code>, <code>v_dis</code> if the model provides probabilities.</li> </ul> <p>In the case of regression, this list of class <b>tdmRegr</b> and contains (see <code>tdmRegress.r</code>):</p> <ul style="list-style-type: none"> <li>• <code>opts</code> # with some settings perhaps adjusted in <code>tdmClassify</code></li> <li>• <code>last_res</code> # last run, last fold: list of class <b>tdmClass</b> from <code>tdmClassify</code></li> <li>• <code>C_train</code> # classification error on training set (vector of length NRUN)</li> <li>• <code>G_train</code> # gain on training set (vector of length NRUN)</li> <li>• <code>R_train</code> # relativ gain (% of max. gain) on training set (vector of length NRUN)</li> <li>• <code>*_test</code> # --- similar, with test set instead of training set</li> <li>• <code>*_test2</code> # --- similar, with test2 set instead of training set</li> <li>• <code>y</code> # what to be minized by SPOT, usually <code>mean(-R_test)</code></li> </ul>
<code>tdm</code>	<p>list with settings for Phase 2 and 3. Elements are</p> <ul style="list-style-type: none"> <li>• <code>mainFile</code> (with path relative to current dir)</li> <li>• <code>mainCommand</code> (string, e.g. <code>"result &lt;- main_sonar(opts)"</code>)</li> <li>• <code>unbiasedFunc</code> (string, e.g. <code>"unbiasedBestRun_C"</code>)</li> <li>• <code>umode</code>: list of unbiased evaluation modes, with elements from <code>{"TST","RSUB","CV"}</code>, see <code>tdmMapDesign.r</code>, <code>tdmCompleteEval.r</code></li> <li>• <code>finalFile</code>, (string, e.g. <code>"sonar.fin"</code>)</li> <li>• <code>withParams</code>: T/F, has theFinals columns with best parameters?</li> </ul>
<code>finals</code>	see <b>Table 2</b>
<code>envT</code>	environment, see <b>Table 5</b>

`lastRes`: The five items `last*` are specific for the `*last*` model (the one built for the last response variable in the last run and the last fold).

## TDMR `opts` Concept

`opts` is a long list with many parameters which control the behaviour of `main_TASK`, i.e. the behaviour of Phase 1. To give this long list a better structure, the parameters are grouped with key words after `"opts$"` and before `"."` (see **Table 3** above).

There are some other parameters in `opts` which do not fall in any of the above groups, e.g.

- `opts$NRUN`
- `opts$VERBOSE`

and others.

You might either specify all `opts`-parameters in your application (i.e. `main_TASK.r` or `*.apd`) or you might use `tdmOptsDefaultsSet()` and specify only those of the `opts`-parameters which differ from this defaults or you enter `main_TASK.r` with a partially filled `opts` and leave the rest to function `tdmFillOptsDefaults` (in `tdmOptsDefaults.r`), which is called from `main_TASK` after the user's `opts`-settings (because some settings might depend on these settings of the user).

Details:

- If the list `opts` is extended by element `X` in the future, you need only to add a default specification of `opts$X` in function `tdmOptsDefaultsFill`, and all functions called from `main_TASK` will inherit this default behaviour.
- TODO: why two function `set..` and `fil...` : set only if `opts==NULL`
- `opts$TST.kind="rand"` triggers random resampling for the division of `dset` into training and test set. In the case of classification this resampling is done by **stratified sampling**: each level of the response variable appears in the training set in proportion to its relative frequency in `dset`, but at least with one record. This last condition is important to ensure proper functioning also in the case of 'rare' levels (most DM models will crash if a certain level does never appear in the training set). In the case of regression the sample is drawn randomly (without stratification).

## TDMR RGain Concept

For **classification**: The `R_`-elements (i.e. `result$R_train` and `result$R_test`) can contain different things, depending on the value of `opts$rgain.type`:

- "rgain" or NULL [def.]: the relative gain in percent, i.e. the gain actually achieved divided by the maximal achievable gain on the given data set,
- "meanCA": mean class accuracy: For each class the accuracy on the data set is calculated and the mean over all classes is returned,
- "minCA": same as "meanCA", but with min instead of mean. For a two-class problem this is equivalent to maximizing the `min(Specificity,Sensitivity)` (see [here](#)).

In each classification case, TDMR seeks to minimize "`-result$R_train`", i.e. to maximize `result$R_train`.

For **regression**: The `R_`-elements (i.e. `result$R_train` and `result$R_test`) can contain different things, depending on the value of `opts$rgain.type`:

- "rmae" or NULL [def.]: the relative mean absolute error RMAE, i.e. the mean  $<|y - y_{(pred)}|>$  divided by the mean  $<|y|>$ ,
- "rmse": root mean square error.

In each regression case, TDMR seeks to minimize `result$R_train`.

## Example Usage

The usage of the TDMR workflow is fairly easy. We show it for the three workflow phases and for the example of the SONAR classification task.

### Phase 1: DM on task SONAR

If you want to build a DM classification model for the SONAR data (see UCI repository or package `mlbench` for further info on SONAR), you write a file `main_sonar.r` in directory `ClassifyTemplate`:

```
main_sonar <- function(opts=NULL, dset=NULL) {
  tdmPath <- "../tdm";
  source(paste(tdmPath,"source.tdm.r",sep="/")); source.tdm(tdmPath);

  if (is.null(opts)) {
    opts = tdmOptsDefaultsSet();      # set initial defaults for many elements of opts. See tdmOptsDefaults.r
  }
}
```

```

# for the list of those elements and many explanatory comments
  opts$filename = "sonar.txt"
  opts$data.title <- "Sonar Data"
}
opts <- tdmOptsDefaultsFill(opts); # fill in all opts params which are not yet set (see tdmOptsDefaults.r)

gdObj<-tdmGraAndLogInitialize(opts); # init graphics and log file

#=====
# PART 1: READ DATA
#=====
if (is.null(dset)) {
  cat1(opts,opts$filename,": Read data ...\n")
  dset <- tdmReadData(opts);
}
names(dset)[61] <- "Class"          # 60 columns V1,...,V60 with input data,
                                   # one response column "Class" with levels ["M" (metal) | "R" (rock)]

response.variable <- "Class"        # which variable is response variable

# which variables are input variables (in this case all others):
input.variables <- setdiff(names(dset), c(response.variable))

#=====
# PART 2: Model building and evaluation
#=====
result <- tdmClassifyLoop(dset,response.variable,input.variables,opts);

# print summary output and attach certain columns (here: y, sd.y, dset) to list result:
result <- tdmClassifySummary(result,opts,dset);

tdmGraAndLogFinalize(opts,gdObj); # close graphics and log file

result;
}

```

This function is invoked with

```
result <- main_sonar();
```

The control flow will pass through the branch `if (is.null(opts))`, where all defaults for `opts` are set with function `tdmOptsDefaultsSet()`. This specifies for example, that an RF model will be built. The dataset will be divided in a training part (90%) and test part (10%), based on `opts$TST.kind="rand"`, `opts$TST.frac=0.1`. You need to specify only those things which differ from `tdmOptsDefaultsSet()`: in this case the filename of the SONAR dataset. Since you do not specify anything from the `opts$SRF`-block, you use the default SRF variable ranking (`opts$SRF.kind="xperc"`, `opts$SRF.Xperc=0.95`). This means that the most important columns containing about 95% of the overall importance will be selected.

You need to specify what column in `dset` is response variable (classification target) and what columns are used for input (in this case all the others, because the SONAR dataset does not have ID columns or otherwise irrelevant columns).

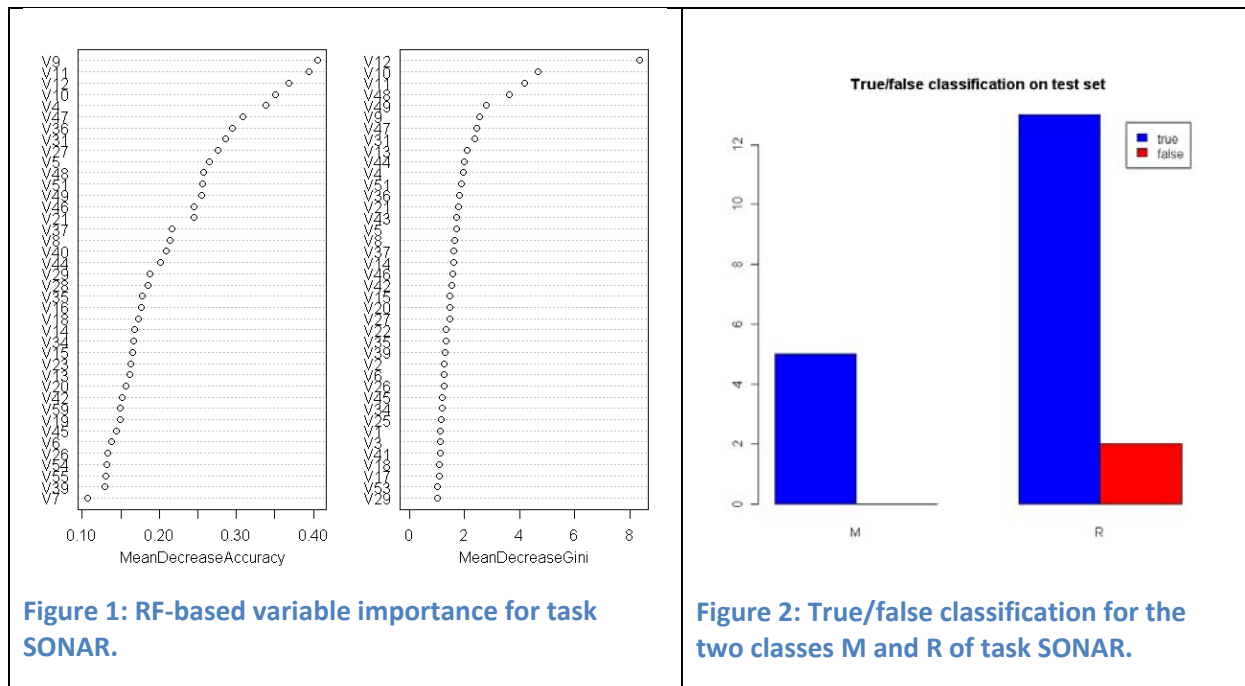
Function `tdmClassifyLoop()` is started, it builds an RF model using the training data and evaluates it on training and test data.

Some output:

```
sonar.txt : Train RF (importance) ...
```

Target levels: M R  
 sonar.txt : Saving sorted RF importance to file ./Output/sonar.txt.SRF.Class.Rdata ...  
 Variables sorted by importance (60):  
 [1] "V9" "V11" "V12" "V10" ...  
 Dropped columns (8 [= 4.8% of total importance]):  
 [1] "V1" "V30" "V60" "V56" "V41" "V24" "V25" "V57"  
 Training cases ( 188):  
     predicted  
 actual M R  
       M 98 8  
       R 22 60  
 total gain: 158.0 (is 0.840% of max. gain = 188.0)

Some graphics output:



The two plots in Figure 1 show the RF-based importance, where MeanDecreaseAccuracy, which has V9, V11 and V12 as the most important variables, is the more reliable measure. The right plot in Figure 2 shows the true/false classifications on the test set (which is here however rather small, so the results are not very reliable, a more reliable test set classification would be obtained with CV).

### Phase 2: SPOT tuning on task SONAR

If you want to do a SPOT tuning on task SONAR, you should follow the steps described in [TDMR Workflow, Phase 2](#) and create the three small files sonar.conf, sonar.apd and sonar.roi. The files' content may look for example like this:

sonar\_01.conf:

```
alg.language = "sourceR"
alg.path = "."
alg.func = "tdmStartSpot"
alg.resultColumn = "Y"
alg.seed = 1235
```

```
io.apdFileName = "sonar_01.apd"
io.roiFileName = "sonar_01.roi"
```

```

io.verbosity = 3;
auto.loop.steps = 50; # number of SPOT's sequential generations
auto.loop.nevals = 100; # concurrently, max number of algo evaluations may be specified

init.design.func = "spotCreateDesignLhd";
init.design.size = 10; # number of initial design points
init.design.repeats = 1; # number of initial repeats

seq.merge.func <- mean;
seq.design.size = 100;
seq.design.retries = 15;
seq.design.maxRepeats = 2;
seq.design.oldBest.size <- 1;
seq.design.new.size <- 3;
seq.predictionModel.func = "spotPredictRandomForest";

report.func = "spotReportSens"

```

#### sonar\_01.apd:

```

if (is.na(match("tdm",ls())))) tdm <- list();
tdm$mainFile <- "../ClassifyTemplate/main_sonar.r";
tdm$mainCommand <- "result <- main_sonar(opts)";

opts = tdmOptsDefaultsSet(); # set initial defaults for many elements of opts.
opts$filename = "sonar.txt"
opts$data.title <- "Sonar Data"
opts$RF.mtry = 4
opts$NRUN = 1 # how many runs with different train & test samples - or -
              # how many CV-runs, if TST.kind="cv"
opts$GRAPHDEV="non";
opts$GD.RESTART=F;
opts$VERBOSE= opts$SRF.verbose = 0;

```

#### sonar\_01.roi:

```

name low high type
CUTOFF1 0.1 0.80 FLOAT
CLASSWT2 5 15 FLOAT
XPERC 0.90 1.00 FLOAT

```

The three parameter CUTOFF1, CLASSWT2 and XPERC are tuned within the borders specified by sonar\_01.roi. Usually you should set `opts$GRAPHDEV="non"` and `opts$GD.RESTART=F` to avoid any graphic output and any graphics device closing from `main_sonar.r`, so that you get only the graphics made by SPOT.

After this preparation, the SPOT tuning is started with:

#### script\_phase2.r:<sup>2</sup>

```

tdm=list(mainCommand="result<-main_sonar(opts)",mainFile="main_sonar.r");
opts <- NULL;
source("sonar_01.apd"); # read in opts-settings
source(tdm$mainFile);
spotUserConfig = list(tdm=tdm,spot.fileMode=F, opts=opts);
spotConfig = spot("sonar_01.conf","auto",spotConfig=spotUserConfig);

```

---

<sup>2</sup> Template available from `inst/demo02sonar`.

It will generate the usual SPOT result files (see SPOT manual [Bart10e])

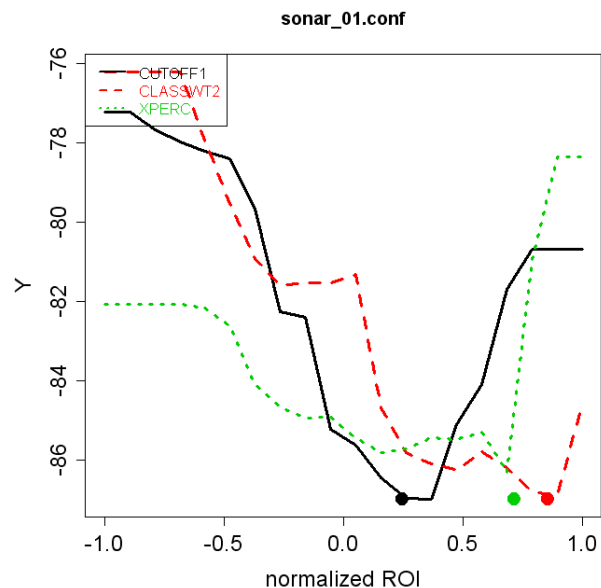
- sonar\_01.res
- sonar\_01.bst

The tuning will stop after 16 sequential steps with the configuration CONFIG=58, because the budget of auto.loop.nevals=100 evaluations is exhausted. The best solution can be seen from the last line of sonar\_01.bst (or alternatively from the printout of spotConfig\$alg.currentBest).

With

```
spot("sonar_01.conf", "rep");
```

the results from a prior tuning run producing sonar\_01.res are read in again and a report including a sensitivity plot (see Figure 3) is made.



**Figure 3: Sensitivity plot. Each ROI [low,high] is mapped to the normalized ROI [-1,1] on the abscissa.**

#### Details:

When spot("sonar\_01.conf","auto") is invoked, the following things happen:

- SPOT is started, reads from sonar\_01.conf that it has to call the inner function  
alg.func = "tdmStartSpot".
- tdmStartSpot(spotConfig) gets with the elements tdm and opts of spotConfig the required information  
tdm\$mainFile <- "main\_sonar.r";  
tdm\$mainCommand <- "result <- main\_sonar(opts)";  
changes to the directory of tdm\$mainFile (current dir in the case of this example) and invokes the DM template main\_sonar.r to with result <- main\_sonar(opts).
- Now tdm\$mainCommand is started and runs the data mining process. The DM template main\_sonar is provided by the user. The only requirement of SPOT or other tuners for the function main\_sonar is that it returns in  
result\$y  
a suitable quantity to be **minimized** by SPOT.

#### More details:

- For a new task TASK, the [opts](#)-part of .apd can usually be copied from the [opts](#)-part of main\_TASK.
- Usually, TASK\_02.apd, TASK\_03.apd, ... will start with source(TASK\_01.apd, local=T) and will only specify those elements of [opts](#) which need to be different.
- For reproducibility of experiments each TDMR task should get its own task name TASK\_01, TASK\_02, ... and the associated set of files (.apd, .conf, .roi ...) should be kept unchanged for further reference. DO NOT alter later the settings in a TDMR task file (unless you want to delete and overwrite the old experiment), but create a new TASK\_xx with its own set of files.
- If a new parameter appears in a .roi file which never appeared in any other .roi file before, a line has to be added to tdmMapDesign.csv, specifying the mapping of this parameter to the corresponding element of [opts](#). ([more details here](#))
- For the current TDMR package version, SPOT is loaded from the CRAN package version, which has been installed as library in the usual way. This is the case when tdm\$theSpotPath is either NA or NULL.  
If you want to load SPOT from source files in pre-defined locations (see [source.tdm.R](#)), set  
tdm\$theSpotPath <- "USE.SOURCE";  
If you want to load SPOT from your own source directory, set tdm\$theSpotPath to this directory.

- How does SPOT handle it, if confFile and spotConfig are both present, e.g. in a call `spot(confFile, "auto", spotConfig)`:
  - Initial defaults for all elements in spotConfig are set inside SPOT (see `spotGetOptions.R`).
  - If confFile exists (only then!), it is read and settings found in confFile overwrite the defaults from `spotGetOptions.R`.
  - If spot is called with parameter spotConfig present, then the elements found in this command line parameter overwrite the settings of step 2.

### Phase 3: “The Big Loop” on task SONAR

To start “The Big Loop”, you configure a file `script_all.R` (a template is in [inst/demo02sonar/](#)), which may look like this:

```
tdm <- list(tdmPath=NULL # from where to load TDMR: if NULL, load package TDMR,
            # else: source R-files from this dir
            , unbiasedFunc="unbiasedBestRun_C"
            , umode=c("CV") # "RSUB"
            , mainFile="main_sonar.r"
            , mainCommand="result <- main_sonar(opts)"
            , tuneMethod=c("spot","lhd") # ["spot" "cmaes" "bfgs" "lhd" ]
            , finalFile="sonar.fin"
            , experFile=NULL # "sonar.exp"
            , nrun=3, nfold=2 # repeats and CV-folds for the unbiased runs
            , optsVerbosity=0 # the verbosity for the unbiased runs
            , withParams=TRUE
            , nExperim=2
            , parallelCPUs = 1 # [1] 1: sequential, >1: parallel with snowFall and this many cpus
            );
runList = c("sonar_02.conf","sonar_03.conf");
spotList = NULL # list(); # =NULL: all in runList; =list(): none
spotStep = "auto";

envT <- tdmCompleteEval(runList,spotList,spotStep,tdm);
```

`script_all.R` specifies a list of TDMR tasks in `runList` (a list of .conf files), specifies a list of tuners in `tdm$tuneMethod`, e.g. `c("spot","cmaes")`, sets other values of `tdm` and calls `tdmCompleteEval`.

Here, this script will trigger the following sequence of experiments:

- sonar\_02.conf is started with tuner (a) lhd and (b) spot
- sonar\_03.conf is started with tuner (a) lhd and (b) spot

This sequence of 4 tuning experiments is repeated `nExperim=2` times. The corresponding 8 result lines are written to `tdm$finalFile`. If (`tdm$experFile != NULL`), these lines are also appended to file `tdm$experFile`. The switch `withParams=TRUE` is only sensible if both .conf files have the same set of parameters in their .roi file.

The result `theFinals` from the last experiment (4 result lines) is in file `sonar.fin`:

CONF	TUNER	CLASSWT2	XPERC	NRUN	NEVAL	RGain.bst	RGain.avg	RGain.OOB	sdR.OOB	RGain.RSUB	sdR.RSUB
sonar_02	lhd	12.026543	0.930197	3	36	86.70213	84.3676	84.4311	1.03715	83.73984	5.63268
sonar_02	spot	14.713475	0.981312	3	36	86.96809	84.6926	85.6287	1.03715	86.99187	7.04085
sonar_03	lhd	8.037636	0.954494	3	36	81.91489	78.6643	80.4391	1.82937	79.67480	7.45134
sonar_03	spot	7.375221	0.914740	3	35	81.91489	78.7082	78.8423	0.34571	74.79675	2.81634

Here CLASSWT2 and XPERC are the tuning parameters, the other columns are defined in **Table 2**.

In the case of the example above, the tuning process had a budget of NEVAL=36 model trainings, resulting in a best solution with class accuracy RGain.bst (in %). The average class accuracy (mean w.r.t. all design points) during tuning is RGain.avg. When the tuning is finished, the best solution is taken and NRUN=3 unbiased evaluation runs are done with the parameters of the best solution. Since the classification model in this example is RF (Random Forest), an OOB-error from the 3 trainings is returned, with average RGain.OOB and standard deviation sdR.OOB. Additionally, NRUN=3 trainings are done with random subsampling (RSUB) of the data set in training and test set, resulting in an average class accuracy on the test set RGain.RSUB and the corresponding standard deviation in sdR.RSUB.

In this case the interpretation of the results is quite clear: The best configuration is sonar\_02.conf with TUNER spot, since this line contains the maximum for all columns RGain.bst, RGain.avg, RGain.OOB and RGain.RSUB. Note that the standard deviation sdR.RSUB is in this case quite large (because the test set is very small). A more reliable result might be obtained with "CV" instead of "RSUB".

Details:

- script\_all.R should be created in and called from the TASK subdir (e.g. TDM.SPOT.d/sonar/). The .conf files in runList should reside in the same directory and should be given w/o path (since TDMR will infer other files, e.g. sonar\_01.apd, from it).
- spotList is a list of .conf files for which the tuners will be started (NULL for all from runList). If a tuner is not started for a certain .conf file it is assumed that its .bst file already exists from a prior run.
- spotStep is a list of strings (may be shorter than runList, then it is cyclically reused) which specifies the SPOT step to be invoked. If e.g. the step is "rep" ("report"), then it is assumed that the .bst file already exists.

Some more details for the developer:

- The unbiased evaluations are done for each element of `tdm$umode` by calling the function `unbiasedBestRun_*` (`...umode,...`) [`*=C` for classification and `*=R` for regression]. The function `unbiasedBestRun_*` reads in the best solution of a tuning run from .bst file, performs a re-run (training + test) with these best parameters.
- script\_all.R in the form given above assumes that library TDMR is loaded. If you want instead to 'source' all necessary R-files from `tdm$tdmPath` (e.g. during development) OR if you want to prepare for parallel execution ([more details ...](#)), then add the following lines after the definition of `tdm` to script\_all.R:

```
tdm$theSpotPath <- NA;
start.tdm.path <- ifelse(is.null(tdm$tdmPath),
  .find.package("TDMR"), paste(tdm$tdmPath, "inst", sep="/"));
source(paste(start.tdm.path, "start.tdm.r", sep="/"), local=T);
```

This will locate the script start.tdm.r (either in the directory of library TDMR or in the sources) and execute it. In this way, script\_all.R can be started directly at the beginning of an R-session, w/o a prior `require(TDMR)`. The line `tdm$theSpotPath <- NA;` means that SPOT is loaded as a package. If it specifies instead the path to the SPOT source, then SPOT's R-files will be 'sourced' from there.

- script\_all.R starts the definition of list `tdm`. If some elements are not def'd, suitable defaults will be added later to `tdm` at the beginning of `tdmCompleteEval` with function `tdmDefaultsFill(tdm)`.

## Extending TDMR

### TDMR Tuner Concept

#### *How to use different tuners*

If you want to tune a TDMR-task with two tuners SPOT and CMA-ES: Simply specify

```
tdm$tuneMethod = c("spot", "cmaes")
```

in `script_all.R` and set the variable `spotStep` to "auto". The tuning results (.bst and .res files) will be copied into subdirs "spot" and "cmaes" of the directory from which you start `script_all.R`.

**Table 4: Tuners available in TDMR**

<code>tdm\$tuneMethod</code>	Description
spot	Sequential Parameter Optimization Toolbox
lhd	Latin Hypercube Design (truncated SPOT, all budget for the initial step)
cmaes	Covariance Matrix Adaption ES
powell	Powell's Method (direct & local search)
bfgs	Broyden, Fletcher, Goldfarb and Shannon method (direct & local search)
bobyqa	direct & local search, with constraint handling

### How to integrate new tuners

Originally TDMR was only written for SPOT as tuning method.

In November 2010, we started to add other tuners to aid the comparison with SPOT on the same footing. As the first other tuner, we introduced CMA-ES (Niko Hansen, R-package by Olaf Mersmann and others). Since comparison with SPOT is the main issue, CMA-ES was wrapped in such a way in `tdmDispatchTuner.r` that the behaviour and output is very similar to SPOT.

This has the following implications which should also be obeyed when adding other tuners to TDMR:

- Each tuning method has a unique name (e.g. "spot", "cmaes"): this name is an allowed entry for `tdm$tuneMethod` and `finals$TUNER` and it is the name of a subdir in `TDM.SPOT.d/TASK/`.
- Each tuner writes result files (.bst, .res) in a fashion similar to SPOT. These result files are copied to the above mentioned subdir at the end of tuning. This facilitates later comparison of results from different tuners.
- Each tuner supports at least two values for `spotStep`: "auto" and "rep" ("report"). In the latter case it is assumed that .bst and .res already exist (in their subdir) and they are usually analysed with `spot(confFile,"rep",...)`.
- Each tuner reads in the .conf file and infers from `spotConfig` the tuner settings (e.g. budget for function calls, max repeats, ...) and tries to make its tuning behaviour as similar to these settings as possible.

For the current CMA-ES tuner the relevant source code for integration in TDMR is in functions `tdmDispatchTuner` and `cmaesTuner` (both in `tdmDispatchTuner.r`) and in `tdmStartCMA.r`.

These functions may be used as templates for the integration of other tuners in the future.

### TDMR Design Mapping Concept

Each variable appearing in .roi file (and thus in .des file) has to be mapped on its corresponding value in list `opts`. This is done in the file `tdmMapDesign.csv`:

roiValue	optsValue	isInt
MTRY	<code>opts\$RF.mtry</code>	1
XPERC	<code>opts\$SRF.Xperc</code>	0
...	...	...

If a variable is defined with `isInt=1`, it is rounded in `opts$...` to the next integer, even if it is non-integer in the design file.

The base file **tdmMapDesign.csv** is read from `<packageDir> = .find.package("TDMR")`.<sup>3</sup> If in the `<dir_of_main_task> = dirname(tdm$mainFile)` an additional file **userMapDesign.csv** exists, it is additionally read and added to the relevant data frame. The file **userMapDesign.csv** makes the mapping modifiable and extendable by the user without the necessity to modify the corresponding source file **tdmMapDesign.r**.

These files are read in when starting **tdmCompleteEval** via function **tdmMapDesLoad** and the corresponding data frames are added to `envT$map` and `envT$mapUser`, resp. This is for later use by function **tdmMapDesApply**; this function can be called from the parallel slaves, which might have no access to a file system.

### How to add a new tuning variable

- add a new line to **userMapDesign.csv** [user] or to **tdmMapDesign.csv** [developer]
- (optional, for developer) add a line to **tdmOptsDefaultsSet()**, if it is a new variable `opts$...` and if all existing and further tasks should have a default setting for this variable

### Details

- We have in [tdmMapDesign.r](#) beneath `{tdmMapDesLoad, tdmMapDesApply}` a second pair of functions `{tdmMapDesSpot$load, tdmMapDesSpot$apply}` with exactly the same functionality. Why? – The second pair of functions is for use in **tdmStartSpot(spotConfig)** where we have no access to `envT` due to the calling syntax of `spot()`. Instead the object **tdmMapDesSpot** stores the maps in local, permanent storage of this object's environment. – The first pair of functions `{tdmMapDesLoad, tdmMapDesApply}` is for use in **tdmStartOther**, especially when called by a *separate* R process when using the tuner `cma.j`. In this case the local, permanent storage mechanism does not work across different R sessions. Here we need the `envT`-based solution of the first pair, since the environment `envT` can be restored across R sessions easily via `save & load`.

## TDMR parallel computing concept

### How to use parallel computing

TDMR supports parallel computing through the packages `snow` and `snowfall` [Knaus08, Knaus09]. Parallelization of TDMR's phase-3-tasks is very easy, you simply have to set `tdm$parallelCPUs` to a suitable value `>1`.

- We parallelize the **tdmDispatchTuner**-calls which are currently inside the 3-fold loop `{tdm$NExperim, runList, tdm$tuneMethod}`. Therefore, these loops are written as `sapply` commands, which can be transformed to `sfSapply`.
- Four operating modes:

<code>tdm\$parallelCPUs</code>	<code>tdm\$fileMode</code>	mode
<code>=1</code>	<code>FALSE</code>	sequential, everything is returned via environment <code>envT</code> , no files are written
<code>=1</code>	<code>TRUE</code>	sequential, everything is returned via environment <code>envT</code> , and logged on several files
<code>&gt;1</code>	<code>FALSE</code>	parallel, everything is returned via environment <code>envT</code> , no files are written or read
<code>&gt;1</code>	<code>TRUE</code>	parallel, everything is returned via environment <code>envT</code> , and logged on several files

`{=1,TRUE}` is the current state of the source code (May'2011).

`{>1,TRUE}` is the parallel mode viable on `maanvs`-clusters at GM.

`{>1,FALSE}` is the parallel mode needed for LIDO (TU DO). It requires more software redesign, since the code should make no file access (no sourcing, no data set reading!) below the call to **tdmDispatchTuner**.

<sup>3</sup> resp. from `tdm$tdmPath/inst/` for the developer version.

The switch `tdm$fileMode==FALSE` is not yet ready (as of June'2011), but should be available in the near future.

### *Environment envT*

The environment `envT` is used to pass necessary information to and back from the parallel slaves. It replaces in nearly all cases the need for file reading or file writing. (File writing is however still possible for the sequential case or for parallel slaves supporting file access. File writing might be beneficial to trace the progress of parallel or sequential tuning processes while they are running and to log the resulting informations.)

Environment **envT** is constructed in `tdmCompleteEval`. **Table 5** shows its elements and it shows in the 3<sup>rd</sup> column which function adds these elements to `envT`:

**Table 5: Elements of environment envT**

variable	remark	function
<code>bst</code>	data frame with contents of <b>last</b> .bst file	<code>tdmStartOther</code> or <code>spotTuner</code> , <code>lhdTuner</code>
<code>bstGrid</code>	list with all bst data frames, <code>bstGrid[[k]]</code> retrieves the kth data frame	<code>tdmCompleteEval</code> or <code>populateEnvT</code>
<code>getBst(conf,tuner,n)</code>	function returning from <code>bstGrid</code> the bst data frame for configuration file <code>conf</code> , tuning method <code>tuner</code> and experiment <code>n</code>	<code>tdmCompleteEval</code>
<code>res</code>	data frame with contents of <b>last</b> .res file	<code>tdmStart*</code> or <code>tdmCompleteEval</code>
<code>resGrid</code>	list with all res data frames, <code>resGrid[[k]]</code> retrieves the kth data frame	<code>tdmCompleteEval</code> or <code>populateEnvT</code>
<code>getRes(conf,tuner,n)</code>	function returning from <code>resGrid</code> the res data frame for configuration file <code>conf</code> , tuning method <code>tuner</code> and experiment <code>n</code>	<code>tdmCompleteEval</code>
<code>result</code>	list with results of <code>tdm\$mainCommand</code> as called in the <b>last</b> unbiased evaluation, see <b>Table 3</b>	<code>unbiasedBestRun_C</code> or <code>unbiasedBestRun_R</code>
<code>theFinals</code>	data frame with one row for each res file, see <b>Table 2</b>	<code>tdmCompleteEval</code> or <code>populateEnvT</code>
<code>tdm</code>		
<code>opts</code>		
<code>tunerVal</code>	the value of <code>tdmDispatchTuner</code> (which can be a long list in case of SPOT)	<code>tdmDispatchTuner</code>
<code>spotConfig</code>		<code>tdmCompleteEval</code>
<code>nGrid</code>	number of elements in <code>bstGrid</code> , <code>resGrid</code>	<code>tdmCompleteEval</code>
<code>nTuner</code>	<code>length(tdm\$tuneMethod)</code>	<code>tdmCompleteEval</code>
<code>nRunList</code>	<code>length(runList)</code>	<code>tdmCompleteEval</code>
<code>runList</code>	<code>runList</code>	<code>tdmCompleteEval</code>

`envT` is used to pass information back and forth between different fcts of TDMR, where `envT$opts` and `envT$tdm` pass info into `tdmStart*`, while `envT$res` and `envT$bst` are used to pass info back from `tdmStart*` to the main level.

## Details

- We have in `tdmCompleteEval` only one parallelization mode (parallel over experiments, tuners and `.conf` files). We decided that it is sufficient to have one strategy for parallelization, for all values of `tdm$parallelCPUs`. We decided that it is dangerous to have nested `sfSapply`-calls.
- When does `sfSapply` return? – The snowfall manual says that `sfSapply` first hands out `nCPU` jobs to the CPUs, then waits for all (!) jobs to return and then hands out another `nCPU` jobs until all jobs are finished. `sfSapply` returns when the last job is finished. Therefore it is not clear what happens with nested `sfSapply`-calls and we make our design in such a way that no such nested calls appear.
- We added column `NEXP (=envT$nExp)` to `tdm$finalFile` and `tdm$experFile`. So it might be that older `.fin` and `.exp` files can no longer be merged with the new ones.
- File writing is no longer necessary for the process, because all data needed are logged in [environment envT](#). But it may be beneficial for tracing the progress of a long-running process. If `tdm$fileMode=TRUE`, each parallelizable branch makes its writing in a separate directory (e.g. `spot/`, `cmaes/`, ... for different tuner branches and/or `01/`, `02/`, `03/`, ... for different experiments with seeds 1,2,3,...). A master file might collect the information from the different files in the end.
- In case `tdm$nExperim>1` we write now on different `.fin` files, e.g.  
`sonar-e01.fin`, `sonar-e02.fin`, ...  
 This is to avoid that parallel executing tasks will remove or write on the same `.fin` file concurrently.
- How and when is the res data frame passed back from SPOT? (we get an error with `spot.fileMode=F`). The bst data frame is in `spotConfig$alg.currentBest`. – Answer: With the new SPOT package version (`>0.1.1372`) and with `spot.fileMode==F`, the res data frame is passed back in **`spotConfig$alg.currentResult`**. The user function `spotConfig$alg.func` is responsible for writing this data frame. We do this for both values of `spot.fileMode`: we start in fcts `spotTuner` and `lhdTuner` a new data frame **`spotConfig$alg.currentResult`** (initially `NULL`) and fill it consecutively in `tdmStartSpot`.

## TDMR seed Concept

In a TDMR task there are usually several places where non-deterministic decisions are made and therefore certain questions of reproducibility / random variability arise:

- 1) Design point selection of the tuner,
- 2) Test/training-set division and
- 3) Model training (depending on the model, RF and neural nets are usually non-deterministic, but SVM is deterministic).

Part 1) is in the case of SPOT tuning controlled by the variable `spot.seed` in the `.conf` file. You may set `spot.seed={any fixed number}` for selecting exactly the same design points in each run. (The design point selection is however dependent on the DM process: If this process is non-deterministic (i.e. returns different `y`-values on the same initial design points, you will usually get different design points from sequential step 2 on.) Or you set `spot.seed=tdmRandomSeed()` and get in each tuning run different design points (even if you repeat the same tuning experiment and even for a deterministic DM process).

In the case of CMA-ES or other tuning algorithms, we use `set.seed(spotConfig$spot.seed)` right before we randomly select the initial design point and ensure in this way reproducibility.

Part 2) and 3) belong to the DM process and the TDMR software supports here three different cases of reproducibility:

- a) Sometimes you want two TDMR runs to behave exactly the same (e.g. to see if a certain software change leaves the outcome unchanged). Then you may set `opts$TST.SEED={any fixed number}` and `opts$MOD.SEED={any fixed number}`.
- b) Sometimes you want the test set selection (RSUB or CV) to be deterministic, but the model training process non-deterministic. This is the case if you want to formulate problem tasks of exactly the same difficulty and to see how different models – or the same model in different runs – perform on these tasks. Then you may set `opts$TST.SEED={any fixed number}`, `opts$MOD.SEED=NULL`.
- c) Sometimes you want both parts, test set selection and model training, to be non-deterministic. This is if you want to see the full variability of a certain solution approach, i.e.

if you want to measure the degree of reproducibility in a whole experiment. Then you may set `opts$TST.SEED= NULL; opts$MOD.SEED=NULL`.

(The case `{TST.SEED= NULL; MOD.SEED=any value}` is a fourth possibility, but it has – as far as I can see – no practical application).

When `opts$*.SEED` is `NULL`, then TDMR will call `opts$*.SEED=tdmRandomSeed()` in `tdmClassify` before each usage of `opts$*.SEED`. (\* = MOD, TST)

Here **tdmRandomSeed** is a function which returns a different integer seed each time it is called. This is even true, if it is called multiple times within the same second (where a function like `Sys.time()` would return the same number). This can easily happen in parallel execution mode, where processes on different slaves usually will be started in the same second.

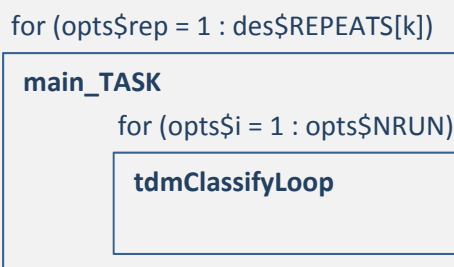
A second aspect of random variability: We usually want each run through `main_TASK` (loop over `i` in `1:opts$NRUN` in `tdmClassifyLoop`) and each repeat during tuning (loop over `r` in `1:des$REPEATS[k]` in `tdmStart*`) to explore different random regions, even in the case where all seed settings (`spot.seed`, `opts$TST.SEED` and `opts$MOD.SEED`) are fixed. We achieve this by storing the loop variables `i` and `r` in `opts$i` and `opts$rep`, resp., and use in `tdmClassify.r` the specific seeds

```
newseed=opts$MOD.SEED+(opts$i-1)+opts$NRUN*(opts$rep-1);
```

and

```
newseed=opts$TST.SEED+(opts$i-1)+opts$NRUN*(opts$rep-1);
```

In this way, each run through `main_TASK` gets a different seed. If `opts$*.SEED` is {any fixed number}, the whole process is however exactly reproducible.



Why is `opts$MOD.SEED=tdmRandomSeed()` and `opts$MOD.SEED=NULL` different? – The first statement selects a random seed at the time of definition time of `opts$MOD.SEED`, but uses it then throughout the whole tuning process, i.e. each design point evaluation within this tuning has the same `opts$MOD.SEED`. The second statement, `opts$MOD.SEED=NULL`, is different: Each time we pass through `tdmClassify` (start of response.variable-loop) we execute the statement

```
set.seed(tdmRandomSeed())
```

which selects a new random seed for each design point and each run.

**New Jan'2012:** When `opts$*.SEED` (\* = MOD, TST) is the string `"algSeed"`, then TDMR will set the relevant seed to `opts$ALG.SEED`, which is the seed `spotConfig$alg.seed+r` from SPOT, where `spotConfig$alg.seed` is set by the user (reproducibility) and `r` is the repeat-number for the design point in question (ensure that each repeat gets another seed to explore the random variability).

## Details

(RNG = random number generator)

- If `TST.SEED=NULL`, the RNG seed will be set to (a different) number via [tdmRandomSeed\(\)](#) in each pass through the `nrun-loop` of `tdmClassifyLoop` / `tdmRegressLoop` (at start of loop).
- If `MOD.SEED= NULL`, the RNG seed will be set to (a different) number via [tdmRandomSeed\(\)](#) in each pass through the response.variable-loop of `tdmClassify` / `tdmRegress` (at start of step 4.3 "model training").

- Before Nov'2010 the TDMR software would not modify RNG seed in any way if TST.SEED=NULL. But we noticed that with a call from SPOT two runs would exactly produce the same results in this case. The reason is that SPOT fixes the RNG seed for each configuration in the same way and so we got the same model training and test set results. To change this, we moved to the new behaviour, where each \*.SEED=NULL leads to a "random" RNG-seed at appropriate places.

## TDMR Graphic Device Concept

### Utility Functions *tdmGraphic\**

These functions are defined in `tdmGraphicUtils.r` and should provide a consistent interface to different graphics device choices.

The different choices for `opts$GRAPHDEV` are

- "pdf": plot everything in one multipage pdf file `opts$GRAPHFILE`
- "png": each plot goes into a new png file in `opts$GD.PNGDIR`
- "win": each plot goes into a new window (`X11()`)
- "non": all plots are suppressed (former `opts$DO.GRAPHICS=F`)

	<b>opts\$GRAPHDEV</b>			
<b>utility function</b>	<b>"pdf"</b>	<b>"png"</b>	<b>"win"</b>	<b>"non"</b>
<code>tdmGraphicInit</code>	open multipage pdf	(create and) clear PNGDIR	-	-
<code>tdmGraphicNewWin</code>	-	open new png file in PNGDIR	open new window	-
<code>tdmGraphicCloseWin</code>	-	close png file	-	-
<code>tdmGraphicCloseDev</code>	close all open pdf devices	close all open png devices	close all devices ( <code>graphics.off()</code> )	-

`tdmGraphicCloseWin` does not close any `X11()`-window (because we want to look at it), but it closes the last open .png file with `dev.off()`, so that you can look at this .png file with an image viewer.

### *GD.RESTART, Case 1: main\_TASK solo*

if `GD.RESTART==F`: No window is closed, no graphic device restarted.

If `GD.RESTART==T` we want the following behaviour:

- close initially any windows from previous runs
- not too many windows open (e.g. if `NRUN=5`, `nfold=10`, the repeated generation of windows can easily lead to s.th. like 250 open windows)
- the important windows should be open long enough to view them (at least shortly)
- in the end, the last round of windows should remain open.

We achieve this behaviour with the following actions in the code for the case `GD.RESTART==T`:

- close all open windows when starting `main_TASK`
- close all open windows before starting the last loop (`i==NRUN`, `k=the.nfold`) of `tdmClassify`
- close all open windows when starting the graphics part (Part 4.7) of `tdmClassify` UNLESS we are in the last loop (`i==NRUN`, `k=the.nfold`); this assures that the windows remain open before the graphics part, that is during the time consuming training part.
- if `GD.CLOSE==T` and `GD.GRAPHDEV!="win"`: close in the end any open .png or .pdf

### *GD.RESTART, Case 2: During SPOT-Run "auto"*

This will normally have `GD.RESTART=F`: No window is closed, no graphic device restarted; but also `GD.GRAPHDEV="non"`, so that no graphic is issued from `main_TASK`, only the graphics from SPOT.

### *GD.RESTART, Case 3: During unbiased runs*

This will normally have also `GD.RESTART=F` and `GD.GRAPHDEV="non"`: No graphics. But you might as well set `GD.RESTART=T` and choose any of the active `GD.GRAPHDEV`'s before calling `unbiaseBestRun_*`, if you want the plots from the last round of `unbiasedBestRun_*`.

## Summary

This report has shown how to use TDMR, the **T**uned **D**ata **M**ining framework in **R**. The examples shown should make the reader familiar with the concepts and the workflow phases of TDMR. They are deliberately made with fairly small datasets in order to facilitate quick reproducibility. For results on larger datasets the reader is referred to [Kone10a, Kone11b].

## References

- [Bart10e] T. Bartz-Beielstein. SPOT: An R package for automatic and interactive tuning of optimization algorithms by sequential parameter optimization. Technical Report <http://arxiv.org/abs/1006.4645>. CIOP Technical Report 05-10, FH Köln, June 2010.
- [Hans06] N. Hansen. The CMA evolution strategy: a comparing review. In: J. Lozano, P. Larranaga, I. Inza, and E. Bengoetxea, editors, Towards a new evolutionary computation. Advances on estimation of distribution algorithms, pages 75-102. Springer, 2006.
- [Kone10a] W. Konen, P. Koch, O. Flasch, T. Bartz-Beielstein. [Parameter-tuned data mining: A general framework](#). In F. Hoffmann and E. Hüllermeier, editors, Proceedings 20. Workshop Computational Intelligence. Universitätsverlag Karlsruhe, 2010.
- [Kone11b] W. Konen, P. Koch, O. Flasch, T. Bartz-Beielstein. [Tuned Data Mining: A Benchmark Study on Different Tuners](#). CIOP Technical Report 02-11, FH Köln, January 2011.
- [Knaus08] Jochen Knaus, Parallel computing in R with sfCluster/snowfall, TR IMBI Uni Freiburg, <http://www.imbi.uni-freiburg.de/parallel/>.
- [Knaus09] Knaus, J. and Porzelius, C. and Binder, H. and Schwarzer, G., Easier parallel computing in R with snowfall and sfCluster. The R Journal, 1:5459, 2009.
- [McKay79] McKay, M.D.; Beckman, R.J.; Conover, W.J. (May 1979). ["A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code"](#) (JSTOR Abstract). *Technometrics* (American Statistical Association) **21** (2): 239–245. doi:[10.2307/1268522](https://doi.org/10.2307/1268522). OSTI [5236110](https://www.osti.gov/purl/5236110). ISSN [0040-1706](https://www.osti.gov/purl/5236110). <http://www.jstor.org/pss/1268522>.