

Readme

for the R script used in article "Some remarks on a set of information theory features used for on-line signature verification" (for article details see: <http://www.ms.sapientia.ro/~manyi/mcyt.html>)

Script setup

Modify the first `setwd()` command in the file according to the path of the unzipped package.

Each test case has to be set manually:

- sample size = 5, 10 or 15 (`smsize` variable);
- feature sets:
 - `rf369`: 6 state of the art features,
 - `entrp`: 6 information theory based features (`feat_name` variable);
- skilled or random forgery test is selected by the `FG` variable ('skilled' or 'random').

Other options are controlled by setting the following variables according to the comments in the R file:

`WScore` : scores are stored in csv files, default directory is `./SCORES` . Some directories are created here, for example file: `SCORES/InftFeat/skilled/5/positive_libsvm.txt` will contain the scores for the positive samples for test case information theory features, skilled forgeries, 5 samples, and the `negative_libsvm.txt` file in the same directory for the negative samples accordingly. The files are generated in order to be used for EER calculation with other programs, like `perfcurve` function from MATLAB. Directory `SCORES_OK` in the zip file contains the scores for all runs, and is not used by the R script. EER values calculated with `perfcurve` are identical as those calculated with the help of the `ROCR` package.

`TUNE_PAR` permits tuning LibSVM parameters, results obtained by this option were not used in the article.

`PLOT_FARFRR`, `PLOT_ROC`, `PLOT_SUM` permit displaying some plots with FAR/FRR and ROC curves for each user (EER is calculated and displayed), and the global EER curve (`PLOT_SUM`) calculated on all users. If variable `SP` is set to `TRUE`, all plots are saved as JPG files in the `IMG` directory. To control plots according to your system, modify the `dev.new()` device opening calls and the surrounding R instructions. Directory `IMG_OK` contains FAR/FRR and ROC curves for the global EERg of all test cases. Directory `IMG_rf369` in the zip file contains curves for all users for the test case `rf369`, skilled, 15 samples and directory `IMG_entrp` for the test case `entrp`, skilled, 15 samples.

Variable `ALL` permits to run the script for some selected users. `TRUE` value will set the script to loop over all users.

When the script is executed, a line is displayed for user after the evaluation:

USER 81 nu: 0.1 g: 0.05 EER: 0.18 AUC: 0.92 , Hit/Miss Neg: 25 / 0 Poz: 5 / 5
with the information:

USER 81: index of the user, starting with 1

nu and g (gamma) values of LibSVM nu parameters

EER: equal error rate calculated from the users data, 0.18 as 18%

AUC: area under the curve value from the users data

and Hit/Miss counts for the negative and positive test samples.

At the end of the script a summary line:

ALL: entrp s 15 g: 0.05 EERa: 0.180 std: 0.121 EERg: 0.209 AUC_a: 0.880 std AUC: 0.111

displays the EER_a : average EER calculated from users individual EER values and EER_g, the global EER calculated from the ROC curve for all scores.

Output of the script goes to a log file in the log directory in case of uncommenting line beginning with: `#sink`.

DATA

DATA directory contains the input files: `mcyt_genuine_41.csv` and `mcyt_forgery_41.csv` with genuine and forgery samples for 100 users, 25 samples / user each. The R script uses two transformed files:

`mcyt41_R_skilled_traintest.csv` : data for skilled forgery tests, 50 samples for each user (25 genuine marked A and 25 forgery marked B in the last column).

`mcyt41_R_random_traintest.csv` : data for random forgery tests, 124 samples for each user (25 genuine marked A and 99 forgeries marked B).

These files do not contain a user id column, users data are in sequence (user 1,2, 3, ...).

ERRATA

The EERg values published in the article were calculated erroneously for the LibSVM classifier. This was the consequence of an error calculating the size of the score vector for the negative samples.

The EERa values were calculated and published correctly.

The error is now eliminated, consequently the EERg values for the LibSVM classifier are the following (in %):

| Random forgeries | | Skilled forgeries | |
|-------------------|------------------------|-------------------|------------------------|
| Inf. theory feat. | State-of-the-art feat. | Inf. theory feat. | State-of-the-art feat. |
| 5 samples | | | |
| 19.85 | 11.46 | 23.31 | 15.09 |
| 10 samples | | | |
| 17.74 | 4.47 | 21.75 | 8.93 |
| 15 samples | | | |
| 17.3 | 4.42 | 20.89 | 7.8 |

All of the values are similar to the results produced by the distance based classifiers. These results do not change the overall results and conclusions of the article.

Contact email: lszabo@ms.sapientia.ro