

README: implementation for the EWUSC (error-weighted shrunken centroid) algorithm

The byte code files (.class files from java) for the EWUSC algorithm are available.

Theoretically, these byte code files should be platform independent, but we make absolutely no guarantee that they will run properly on your system.

To uncompress the archives:

- On linux:
`tar -xvzf shcentroid_code.tar.gz`
- On windows: WinZip can extract these files. There are probably other windows applications that can do it as well.

Format of input files:

Tab-delimited text files:

- The first row is a header row.
- Column 1: gene identifiers (eg. Probe sets or gene names)
- Column 3 onwards: expression values and variability estimates are in consecutive columns
- Example input file: training set and test set for multiple tumor data (These files are available under “Gene expression data sets used”)

Format of output files:

- Relevant genes selected: `genelist_< Δ bin #>_< ρ_0 bin #>.txt`
Eg. `genelist_14_8.txt`
- Number of classification errors for each Δ and each ρ_0 :
 - on the entire training set: `perclass_training_1fold.txt`
 - on the cross validation data: `perclass_randomCV.txt`
 - on the test set: `perclass_testing.txt`

The number of classification errors for each class and each Δ and each ρ_0 is shown. The numbers of errors for all the classes are separated by “+”, and the total number of errors is shown after “+”. For example, “0+2+0+0+0+2+0+3+0+0+0=7” means that there are 2 errors in class #2, 2 errors in class #6, and 3 errors in class #8.
- Files to be read into Matlab:
 - `matlab_numG_<blah>.txt`: number of relevant genes for each Δ and each ρ_0 .
 - `matlab_randomCV_numErr_<blah>.txt`: number of classification errors on cross validation data each Δ and each ρ_0
 - `matlab_test_numErr_<blah>.txt`: number of classification errors on test set each Δ and each ρ_0
 - `matlab_trainCV_numErr_<blah>.txt`: number of classification errors on full training set each Δ and each ρ_0

How to run the EWUSC executable?

- General format:

```
java ewusc -r <# genes> -c <# experiments> -err <error flag> -rep <# repeated
measurements> -errOp <sd or cv or in> -fold <m-fold CV> -numClass <# classes> -
label <file of class labels> -class <label for each experiment> -perm <permutation file>
-delta <upper range of delta> -bin <# delta bins> -loocv <LOOCV flag> -prior <equal
prior flag> -#test <# test experiments> -test <test data> -classTest <labels for each test
experiment> -corr <range of correlation threshold> inputfilename
```

- To see all these options:

```
java ewusc -
```

- Interpretation of options:

- **-err** <error flag>:

- **-err 0** means that the USC (unweighted) algorithm is used
- **-err 1** means that the EWUSC algorithm is applied.

- **-errOp** <sd or cv or in>: ignored if **-err 0** is used.

- **-errOp sd** means that standard deviation will be used to compute variability in variability-weighted approach.
- **-errOp cv** means that coefficient of variation will be used to compute variability in variability-weighted approach.
- **-errOp in** means that error estimates are given for each array measurements in the input file. In this case, # of repeated measurements is set to 2, where 2 values are given for each experiment in consecutive columns such that

```
<measured value for expt 1><tab><error for expt1><tab><measured value for expt 2><tab><error for
expt2><tab> etc...
```

- **-fold** <m-fold CV>:

- For the multiple tumor data, we use “**-fold 4**”: 4-fold CV
- For the 2-class breast cancer data, we used “**-fold 10**”
- This option is ignored if LOOCV is on

- **-label** <file of class labels>

- labels for each class
- example, see “multiple_11class.txt”
- Each label name is on its own line.
- Number of lines = number of classes

- **-class** <label for each experiment>

- label for each experiment
- example, see “label96.txt” or “label27.txt”
- Number of lines = number of experiments
- The class label must exactly match that in the “-label” option.

- **-perm** <permutation file>

- The goal is to make sure that we used the exact same random CV data in our comparison of EWUSC and USC.
- Example “mypermfile0”.
- Each permutation file contains a random permutation of each class size.

- Permutation files used in our experiments are available from the directory “RandomPermutation”.
 - *-delta* <upper range of delta> *-bin* <# delta bins>
 - Initially, the upper range of delta is trial-and-error, until we reach 0 genes in the number of genes chosen.
 - We usually use 50 delta bins in our experiments.
 - In the NCI 60 data, the upper range of Δ is 5.
 - In the multiple tumor data, the upper range of Δ is 20.
 - In the breast cancer data, the upper range of Δ is 2.5.
 - In practice, the upper range of Δ is determined by trial and error until the number of selected genes is reduced to 0.
 - *-loocv* <LOOCV flag>
 - “**-loocv 1**” indicates that LOOCV is on, ie., we do leave-one-out cross validation, which can be quite computationally intensive for large data sets.
 - We used “**-loocv 0**” for most of our experiments.
 - *-=prior* <equal prior flag>
 - This flag indicates if we want to choose equal prior for the classes. The default is “**-=prior 0**”, which is used in our experiments.
 - *-corr* <range of correlation threshold>
 - The default is from 0 to 1, at increment of 0.1. We used the default in all our experiments.
- Examples:
 - The training set, test set, class file, and label files can be downloaded under “pre-processed data”. In order to run these examples, please copy all these files to the current directory containing the bytecode files. You might also want to read “README” under the directory “RandomPermutation” and copy the necessary permutation files to the current directory as well.
 - To apply the EWUSC algorithm to the multiple tumor data:


```
java ewusc -r 7129 -c 96 -c_start 2 -err 1 -rep 2 -errOp in -fold 4 -numClass 11
-label multiple_11class.txt -class multiple_label96.txt -perm mypermfile0 -delta 20
-bin 50 -loocv 0 -#test 27 -test test_multiple_tumor_7129_27.txt -classTest
multiple_label27.txt combined_multiple_tumor_7129_96.txt
```
 - To apply the USC algorithm to the multiple tumor data:


```
java ewusc -r 7129 -c 96 -c_start 2 -err 0 -rep 2 -errOp in -fold 4 -numClass 11
-label multiple_11class.txt -class multiple_label96.txt -perm mypermfile0 -delta 20
-bin 50 -loocv 0 -#test 27 -test test_multiple_tumor_7129_27.txt -classTest
multiple_label27.txt combined_multiple_tumor_7129_96.txt
```
 - To apply the USC algorithm to the NCI 60 data:


```
java ewusc -r 5244 -c 61 -err 0 -rep 1 -fold 3 -numClass 8 -label NCI60_8class.txt
-class NCI60_label61.txt -perm mypermfile0 -delta 5 -bin 50 -loocv 0 -#test 0
transposed_NCI60_5244_61.txt
```