# How To Use These Data

EE6222

September 2015

We provide 121 UCI dataset which are used in [Fernández-Delgado et al.2014]. All these data can be downloaded from UCI Machine Learning Repository [Lichman2013]. The details of each data is presented in Table 1. Each data has 3 flies in the folder. Randomized stratified sampling is employed to make sure one training and one test set are generated (each with 50% of the available patterns), where each class has the same number of training and test patterns. Parameter tuning is performed on this couple of sets to identify parameters with the best performance on the test set. There are two parameters in commonly used RVFL configuration. One is the number of hidden neurons $N$ (you can tuned it with [3,203] with a step-size of 20). The other one is $\lambda$ in ridge regression (you can set it to be $2^C$ and C is $-5:1:14$). Then, with the selected values for the tunable parameters, a 4-fold cross validation is developed using the whole data. However, for some datasets where the training-testing partition is already available (such as annealing and audiology-std, among others), the classifier is trained on the predefined training set and evaluated on the test set. In this case, the test result is calculated on the test set [Fernández-Delgado et al.2014].

For somedata where training-testing partition is not available, there are 3 files in the data folder. Take "abalone" data for an example:

- *abalone_conxuntos.m*: the index for training-testing partitioning. "index1" in the file are the indexes for the training data and "index2" are the indexes for testing data. This partitioning is used to perform parameter-tuning.
- *abalone_conxuntos_kfold.m*: the partition for the 4-fold cross validation. 4-flod cross validation means divide the data into 4 parts without overlapping where each part has roughly equal number of data samples. Then you do the training-testing 4 times. Each time you take one part as the testing data and the rests as the training data. This 4-fold cross validation is used to evaluate your model.
- *abalone_R.m*: the data file. The first column is the data index the last column is the label. The rest columns are features.

For some other data where the training-testing partition is already available, there are also 3 files in the data folder. Take "adult" data for an example:

- *adult_train_R.m*: the training data file. The first column is the data index the last column is the label. The rest columns are features.
- *adult_conxuntos.m* : further partition for the training set for parameter tuning. "index1" defines a subset of training data which can be used to tune your parameter based on the results on another subset of training data defined by "index2".

- *adult_test_R.m* : the testing data file. The first column is the data index the last column is the label. The rest columns are features.

You can normalize each feature by removing the mean value and dividing by its $l2$ norm. For example:
$mean\_X = mean(dataX, 1);$
$dataX = dataX - repmat(mean\_X, size(dataX, 1), 1);$
$norm\_X = sum(dataX.^2, 1);$
$norm\_X = sqrt(norm\_X);$
$norm\_X = repmat(norm\_X, size(dataX, 1), 1);$
$dataX = dataX./norm\_X;$

We also provide a demo for each case, you can refer to *demo_abalone.m* or *demo_adult.m* for details.

# References

[Fernández-Delgado et al.2014] Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(1):3133–3181.

[Lichman2013] Lichman, M. (2013). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Table 1: Datasets used in this work

| Datasets | Patterns | Features | Classes |
|---|---|---|---|
| abalone | 4177 | 8 | 3 |
| ac-inflam | 120 | 6 | 2 |
| acute-nephritis | 120 | 6 | 2 |
| adult | 48842 | 14 | 2 |
| annealing | 798 | 38 | 6 |
| arrhythmia | 452 | 262 | 13 |
| audiology-std | 226 | 59 | 18 |
| balance-scale | 625 | 4 | 3 |
| balloons | 16 | 4 | 2 |
| bank | 45211 | 17 | 2 |
| blood | 748 | 4 | 2 |
| breast-cancer | 286 | 9 | 2 |
| bc-wisc | 699 | 9 | 2 |
| bc-wisc-diag | 569 | 30 | 2 |
| bc-wisc-prog | 198 | 33 | 2 |
| breast-tissue | 106 | 9 | 6 |
| car | 1728 | 6 | 4 |
| ctg-3classes | 2126 | 21 | 3 |
| ctg-10classes | 2126 | 21 | 10 |
| chess-krvk | 28056 | 6 | 18 |
| chess-krvkp | 3196 | 36 | 2 |
| congress-voting | 435 | 16 | 2 |
| conn-bench-sonar | 208 | 60 | 2 |
| conn-bench-vowel | 528 | 11 | 11 |
| connect-4 | 67557 | 42 | 2 |
| contrac | 1473 | 9 | 3 |
| credit-approval | 690 | 15 | 2 |
| cylinder-bands | 512 | 35 | 2 |
| dermatology | 366 | 34 | 6 |
| echocardiogram | 131 | 10 | 2 |
| ecoli | 336 | 7 | 8 |
| energy-y1 | 768 | 8 | 3 |
| energy-y2 | 768 | 8 | 3 |
| fertility | 100 | 9 | 2 |
| flags | 194 | 28 | 8 |
| glass | 214 | 9 | 6 |
| haberman-survival | 306 | 3 | 2 |
| hayes-roth | 132 | 3 | 3 |
| heart-cleveland | 303 | 13 | 5 |
| heart-hungarian | 294 | 12 | 2 |
| heart-switzerland | 123 | 12 | 2 |
| heart-va | 200 | 12 | 5 |
| hepatitis | 155 | 19 | 2 |
| hill-valley | 606 | 100 | 2 |
| horse-colic | 300 | 25 | 2 |
| ilpd-indian-liver | 583 | 9 | 2 |
| image-segmentation | 210 | 19 | 7 |
| ionosphere | 351 | 33 | 2 |
| iris | 150 | 4 | 3 |
| led-display | 1000 | 7 | 10 |
| lenses | 24 | 4 | 3 |
| letter | 20000 | 16 | 26 |
| libras | 360 | 90 | 15 |
| low-res-spect | 531 | 100 | 9 |
| lung-cancer | 32 | 56 | 3 |
| lymphography | 148 | 18 | 4 |
| magic | 19020 | 10 | 2 |
| mammographic | 961 | 5 | 2 |
| miniboone | 130064 | 50 | 2 |
| molec-biol-promoter | 106 | 57 | 2 |
| molec-biol-splice | 3190 | 60 | 3 |

| Datasets | Patterns | Features | Classes |
|---|---|---|---|
| monks-1 | 124 | 6 | 2 |
| monks-2 | 169 | 6 | 2 |
| monks-3 | 3190 | 6 | 2 |
| mushroom | 8124 | 21 | 2 |
| musk-1 | 476 | 166 | 2 |
| musk-2 | 6598 | 166 | 2 |
| nursery | 12960 | 8 | 5 |
| oocMerl2F | 1022 | 25 | 3 |
| oocMerl4D | 1022 | 41 | 2 |
| oocTris2F | 912 | 25 | 2 |
| oocTris5B | 912 | 32 | 3 |
| optical | 3823 | 62 | 10 |
| ozone | 2536 | 72 | 2 |
| page-blocks | 5473 | 10 | 5 |
| parkinsons | 195 | 22 | 2 |
| pendigits | 7494 | 16 | 10 |
| pima | 768 | 8 | 2 |
| pb-MATERIAL | 106 | 4 | 3 |
| pb-REL-L | 103 | 4 | 3 |
| pb-SPAN | 92 | 4 | 3 |
| pb-T-OR-D | 102 | 4 | 2 |
| pb-TYPE | 105 | 4 | 6 |
| planning | 182 | 12 | 2 |
| plant-margin | 1600 | 64 | 100 |
| plant-shape | 1600 | 64 | 100 |
| plant-texture | 1600 | 64 | 100 |
| post-operative | 90 | 8 | 3 |
| primary-tumor | 330 | 17 | 15 |
| ringnorm | 7400 | 20 | 2 |
| seeds | 210 | 7 | 3 |
| semeion | 1593 | 256 | 10 |
| soybean | 307 | 35 | 18 |
| spambase | 4601 | 57 | 2 |
| spect | 80 | 22 | 2 |
| spectf | 80 | 44 | 2 |
| st-aus-credit | 690 | 14 | 2 |
| st-german-credit | 1000 | 24 | 2 |
| st-heart | 270 | 13 | 2 |
| st-image | 2310 | 18 | 7 |
| st-landsat | 4435 | 36 | 6 |
| st-shuttle | 43500 | 9 | 7 |
| st-vehicle | 846 | 18 | 4 |
| steel-plates | 1941 | 27 | 7 |
| synthetic-control | 600 | 60 | 6 |
| teaching | 151 | 5 | 3 |
| thyroid | 3772 | 21 | 3 |
| tic-tac-toe | 958 | 9 | 2 |
| titanic | 2201 | 3 | 2 |
| rains | 10 | 28 | 2 |
| twonorm | 7400 | 20 | 2 |
| vc-2classes | 310 | 6 | 2 |
| vc-3classes | 310 | 6 | 3 |
| wall-following | 5456 | 24 | 4 |
| waveform | 5000 | 21 | 3 |
| waveform-noise | 5000 | 40 | 3 |
| wine | 179 | 13 | 3 |
| w-qua-red | 1599 | 11 | 6 |
| w-qua-white | 4898 | 11 | 7 |
| yeast | 1484 | 8 | 10 |
| zoo | 101 | 16 | 7 |

Details of the datasets. Some keys are: ac-inam=acute-inammation, bc=breastcancer, congress-vot= congressional-voting, ctg=cardiotocography, conn-benchsonar/ vowel= connectionist-benchmark-sonar-mines-rocks/vowel-deterding, pb=pittsburg-bridges, st=statlog, aus=australian, vc=vertebral-column, w-qua=wine-quality.