Performance of COR model in UCI data sets - Updated

Paula Parpart
Department of Experimental Psychology, University College London

November 10, 2015

1 Simulation

The model was fit to data sets from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) and the KEEL Data set repository (http://sci2s.ugr.es/keel/datasets.php). The chosen data sets are classification problems with a small number of continuous attributes and binary dependent variable. When there was initially an uneven class distribution, i.e., different number of instances with either class label, these were evened out by shortening the larger class to the size of the smaller class. By doing this, we could take care of the previous problem that sometimes smaller training set sizes outperformed larger training set sizes (The model does not include an intercept, and when classes (1,-1) are very uneven this can cause problems).

The COR model is cross-validated on each data set by splitting the total number of instances randomly into training and test set. The size of the training set is varied between 10, 20, 30, 50, 100 and 500 of all instances, and the test set represents the complementary set of instances always. For each training set size, the cross-validation split into training and test set is repeated k=1000 times and performance is averaged across all of them. Plots below demonstrate the generalization performance of the COR model for a range of penalization parameters $\theta = [0, 700]$, and as a function of the training set size.

This updated version of the document takes care of the mcmc convergence issue due to a bug in the original mcmc code, which results in faster mcmc convergence. Additionally, the number of train-test set separations was increased from k=10 to k=1000 which decreased the error bars.

The decision rule applied here does not reflect the heuristic decision rules (TTB or Tallying) , but a standard linear classification rule which sums the outputs and thresholds at zero (i.e., like Linear Regression classification). When the COR decision rule predicts the outcome to be exactly 0, the model currently does not guess, i.e., choose a random class, but it is kept at 0 which lowers overall performance rates.

D: Number of dimensions (i.e. cues)

$$\hat{y} = 1\left[0 < \textstyle\sum_{i=1}^{D} x\hat{\beta}^{(i)}\right]$$

1.1 Banknotes: corrected

Number of instances (after evening classes): 1220
Number of instances (before evening classes): 1372
Number of cv splits: 100
Number of attributes: 4
Attribute Information:
1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer) : genuine versus forged
Description: Data were extracted from images that were taken for the evaluation of an authentication procedure for banknotes.

Average (absolute) correlation between attributes: 0.43
Minimum (absolute) correlation between attributes: 0.26
Maximum (absolute) correlation between attributes: 0.79

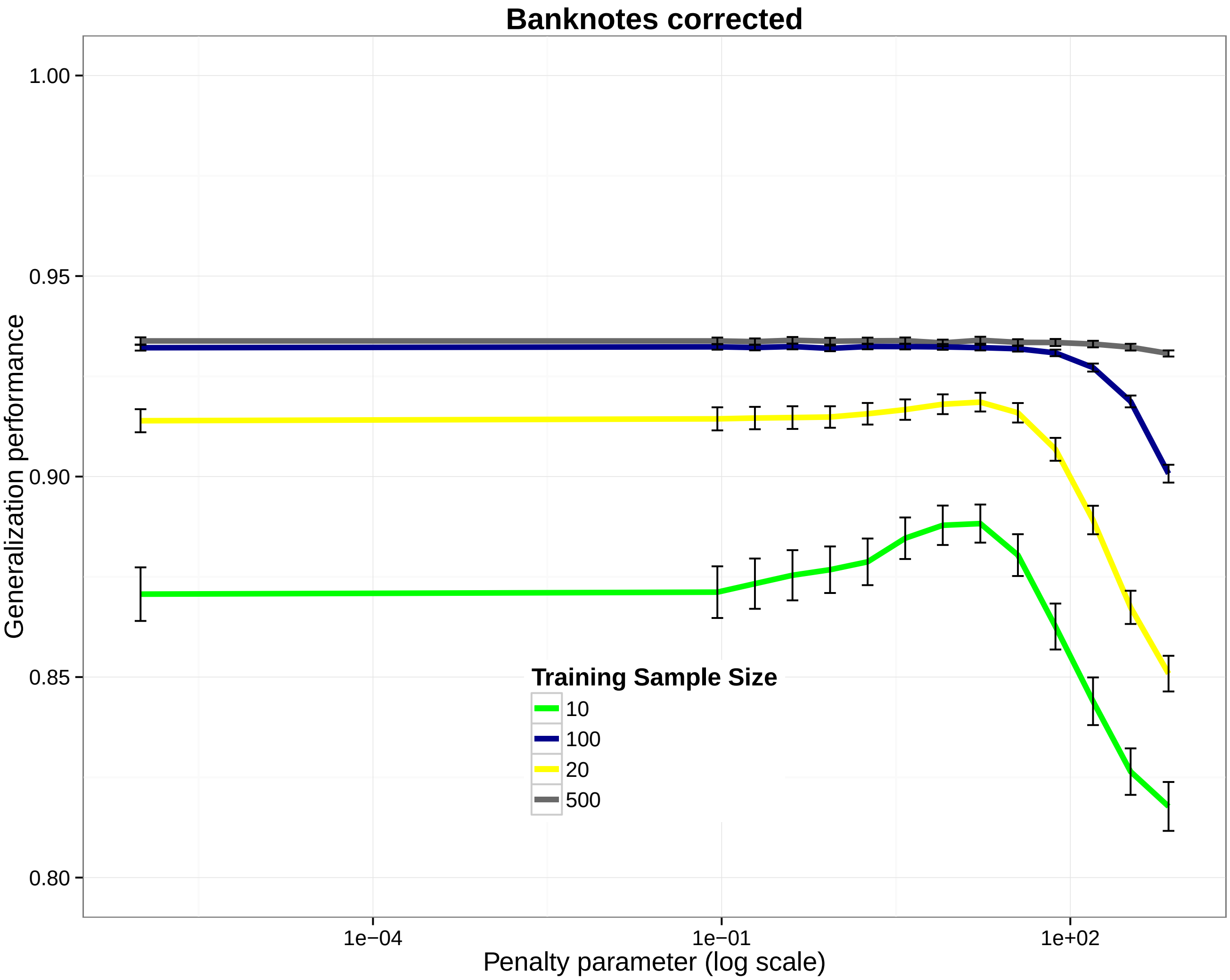

Figure 1: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 and 500 of all instances.

1.2 Haberman's Survival Data: corrected

Number of instances (after evening classes): 162
Number of instances (before): 306
Number of cv splits: 100
Number of attributes: 3
Attribute Information:
1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
– 1 = the patient survived 5 years or longer
– 2 = the patient died within 5 year

Description: The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. Weights are not very large at penalty = 0, i.e. Linear Regression.

Average (absolute) correlation between attributes: 0.05
Minimum (absolute) correlation between attributes: 0.04
Maximum (absolute) correlation between attributes: 0.054

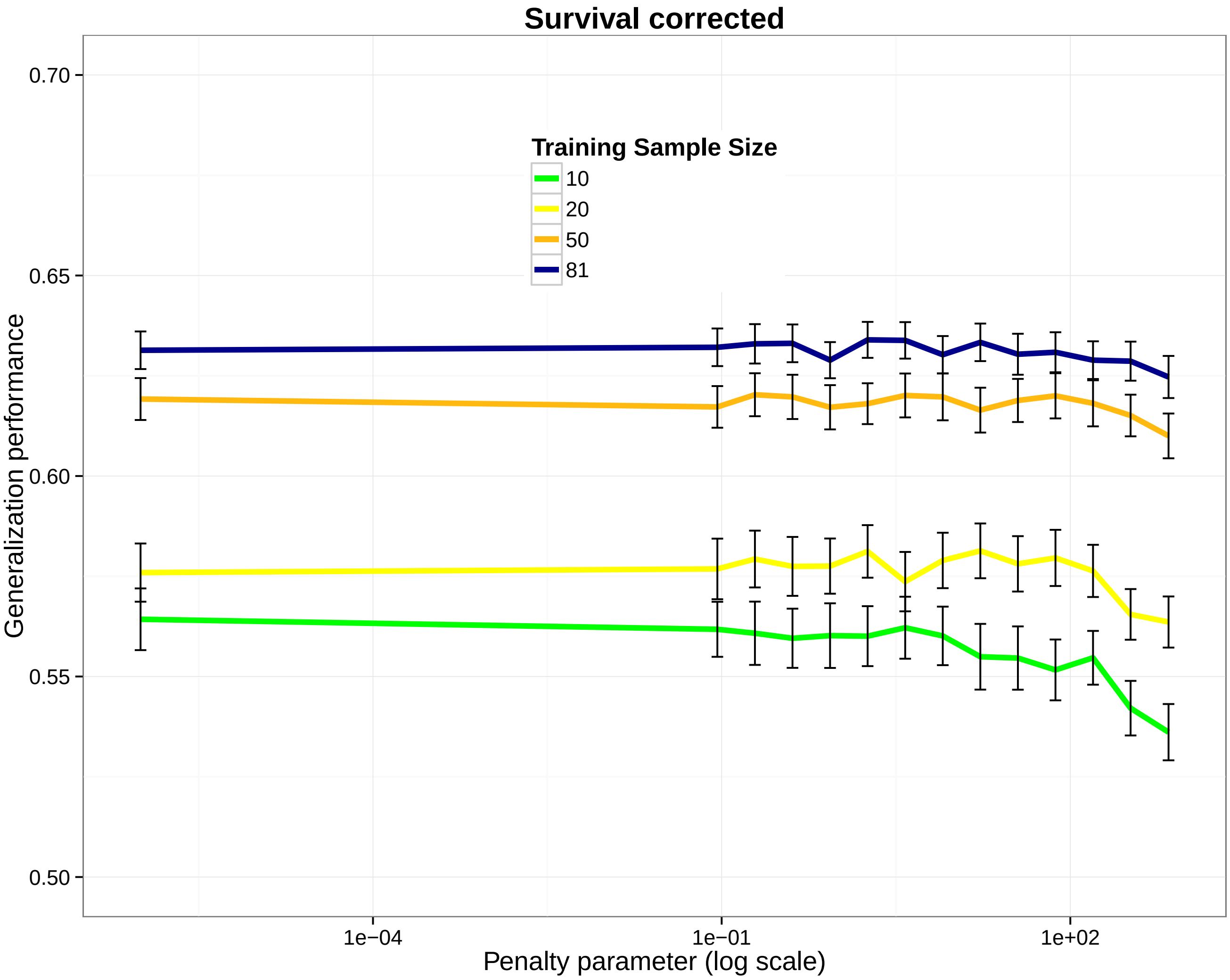

Figure 2: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 50, and 81 of all instances.

1.3 Phoneme: corrected

Number of instances (after evening classes): 3172
Number of instances (before): 5404
Number of cv splits: 100
Number of attributes: 5
Attribute Information:
1. Aa 2. Ao 3. Dcl 4. ly 5. Sh Class: 0,1
The aim of this dataset is to distinguish between nasal (class 0) and oral sounds (class 1).
Average (abs) corr: 0.12. Min (abs) corr: 0.008. Max (abs) corr: 0.28

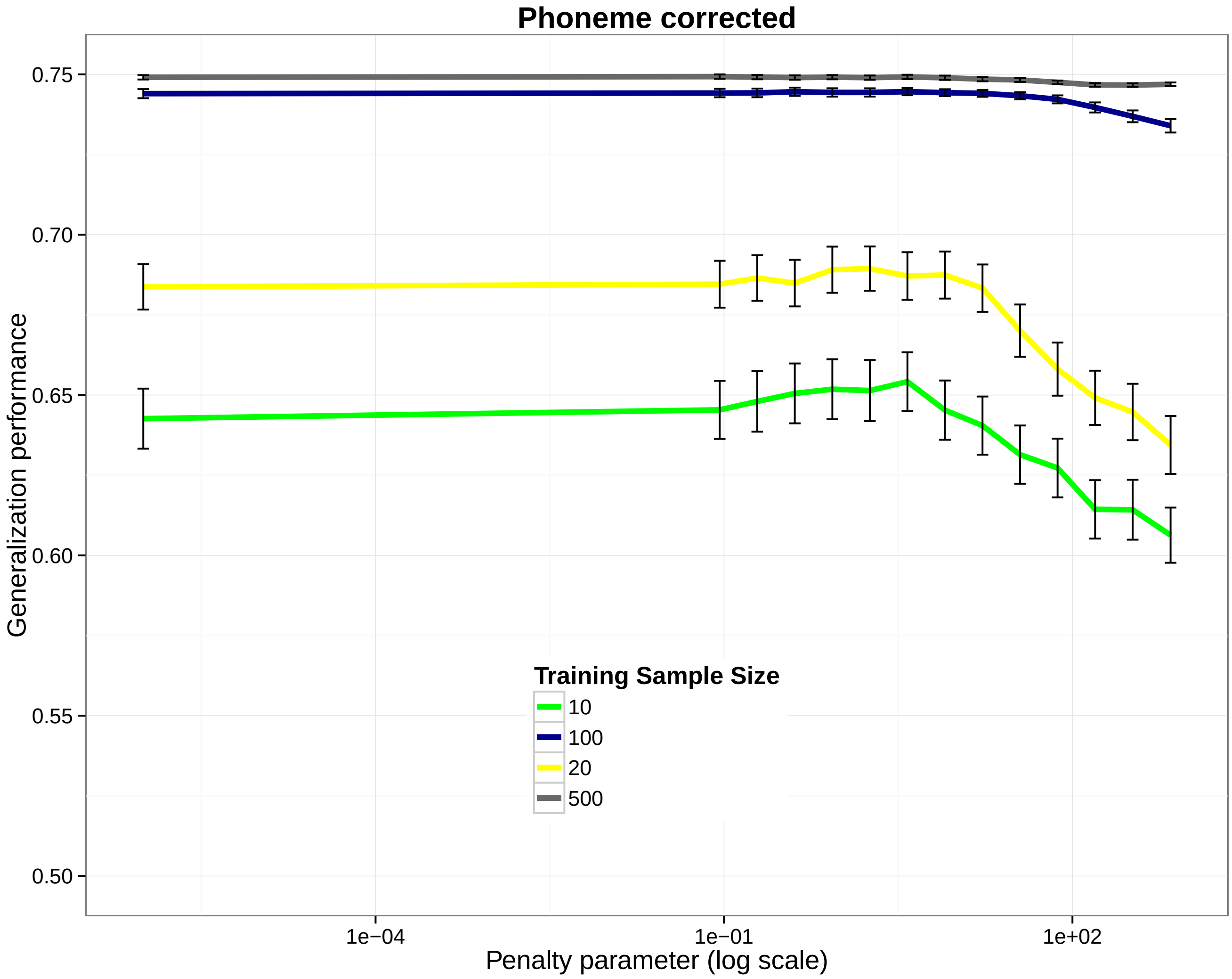

Figure 3: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 100, and 500 of all instances.