Stability-based multivariate mapping using SCoRS

Jane Maryam Rondina*, John Shawe-Taylor *, Janaina Mourão-Miranda * †

*Department of Computer Science, University College London, London - United Kingdom
†Centre for Neuroimaging Sciences, Institute of Psychiatry, King's College London, UK

***Abstract*—Recently we proposed a feature selection method based on stability theory (SCoRS - Survival Count on Random Subspaces) and showed that the proposed approach was able to improve classification accuracy using different datasets. In the present work we propose: (i) an extension of SCoRS using reproducibility instead of model accuracy as the parameter optimization criterion and (ii) a procedure to estimate the rate of false positive selection associated with the set of features obtained. Our results using the proposed framework showed that, as expected, the optimal parameter was more stable across the cross-validation folds, the spatial map displaying the features selected was less noisy and there was no decrease in classification accuracy. In addition, our results suggest that the estimated false positive rate for the features selected by SCoRS is under 0.05 for both optimization approaches, nevertheless lower when optimizing reproducibility in comparison with the standard optimization approach.**

***Keywords*-Optimization, brain mapping, interpretability, feature selection, stability, reproducibility.**

I. Introduction

Feature selection has been used in neuroimaging based pattern recognition approaches with the primary aim of potentially increasing the model performance by eliminating irrelevant features from the model. However, another important role of feature selection methods is to facilitate interpretation by identifying sets of meaningful features with high predictive power. Therefore it is important to select relevant features which are also stable.

Most methods used for selecting features require the optimization of one or more parameter(s). As generalization ability is an important property of learning algorithms, classification and regression methods usually maximize predictive performance in their optimization processes for tuning parameters. Given that the number of available examples in neuroimaging applications is usually small, cross-validation (CV) framework is commonly used for evaluating the model performance and nested cross-validation is usually employed for parameter optimization. Consequently another issue commonly associated with feature selection methods embedded in a nested cross-validation framework is how to summarize the different models resulting from each cross-validation fold (CV-fold).

There are a couple of issues associated with parameters optimization based only on predictive performance for feature selection methods in neuroimaging. In some applications there is a high variability among training examples. In these cases, sets of features selected in each CV-fold tend to have low rates of overlap, even when one single pair of examples is left out in each fold (leave-one-pair-out CV). This issue becomes more evident in case of sparse models.

In the current work we propose an optimization framework based on maximizing the reproducibility among CV-folds. We illustrate this approach for optimizing the threshold level in SCoRS, a feature selection method previously proposed by our group [1]. We implemented the same procedure with two optimization strategies: the standard approach (maximizing classification accuracy) and the proposed approach (maximizing reproducibility across CV-folds). Our results showed that when optimizing the threshold using reproducibility, the optimal threshold became very stable among CV-folds (varying from 0.5 to 0.6). In addition, the model accuracy did not decrease (actually it slightly increased regardless of the fact that the optimization was not based on accuracy). We also addressed the issue of having a different model for each CV-fold by proposing a summarization approach based on the stability of the features selected across folds.

II. Material and Methods

A. Selecting features with SCoRS

SCoRS (Surviving Count on Random Subspaces) [1] is a feature selection method based on Stability Selection theory [2]. It consists of successive iterations where a sparse regression method is applied to sub-samplings of both examples and features obtained randomly from the data. The regression method used was the Lasso [3], although other multivariate regression methods that produce sparse solutions could be applied.

We explain the fundamentals of SCoRS in the algorithm 1, where p is the total number of features, $ntrain$ is the number of training examples, β_i is the coefficient of the feature i and R is the total number of repetitions. The vector c is a counter for the number of times each feature was randomly chosen in a subset from the entire set of features and the vector s is a counter for the number of times each feature was selected by regression in all subsets it took part.

Figure 1. SCoRS algorithm

$X \leftarrow DataMatrix(n_{train}, p)$;
$Y \leftarrow LabelsVector(n_{train})$;
$r \leftarrow 0$;
$s_i = 0$ and $c_i = 0$, $\forall i, i = 1 : p$;
repeat
  Randomly select a subset of features rp out of p;
  $c_i(rp) \leftarrow c_i(rp) + 1$;
  Randomly select a subset of examples rn out of $ntrain$;
  $RX \leftarrow X(rn, rp)$;
  Apply regression to RX;
  $s_i \leftarrow s_i + 1 \forall i | \beta_i \neq 0$
  $r \leftarrow r + 1$;
until $r = R$
Select feature i if $(s_i./c_i) > th$, where $0 < th < 1$ is a threshold value;

After several iterations a subset of the features will "survive" (i.e., they will have a count different from zero in s). However, some of those features will have low counts, meaning that they were not relevant in the most combinations of features in which they took part. For this reason, after all iterations a threshold is applied to the selection frequency, i.e., the ratio between the number of times each feature "survived" (kept in vector s) divided by the number of times it was randomly chosen in the sub-samplings (kept in vector c). The optimal threshold value is selected through a nested cross-validation, as described in the next section.

B. Threshold optimization

The threshold tuning is an important step for selecting relevant features in SCoRS. A nested leave-one-pair-out cross-validation (LOPO-CV) was used for parameter optimization. In each iteration (CV-fold), a pair of subjects was left out in the outer loop for test while the inner loop was used for parameter optimization according to a specific criterion. The parameter range consisted of 9 threshold levels varying from 0.1 to 0.9 in steps of 0.1.

In the present work we compare two different optimization criteria. In the first approach we used the commonly used criterion based on the model performance (i.e. the threshold level that led to the highest classification accuracy across the inner loops was used in the outer loop). We used the SVM (Support Vector Machine, [4] to evaluated the predictive performance of the selected features, although other classification or regression methods could be used in order to get a performance measure.

In the second optimization approach, we used a criterion based on the overlap between the features sets selected across the inner loops. The threshold level that led to the highest overlap was used in the outer loop. Details on how the overlap was measured are given in the next section.

C. Reproducibility measurement

We computed the reproducibility across cross-validation folds based on an adaptation of the overlap measure proposed by [6]. As we implemented a LOPO-CV with F folds, we averaged all pairwise overlaps O_{ij} between each pair of folds i and j, according to equation 1. S_i (S_j) is the subset of features selected in the fold i (j), F is the number of folds, $\bar{N}_i$ is the number of non-zero features in the subset S_i and E is a correction factor for the fact that for a given model the expected overlap of non-zero features will decrease with the increase of the sparsity. The heuristics given by E was used to calculate this correction, where P is the total number of features.

$$O_{ij} = \frac{S_i \cap S_j - E}{\bar{N}_i} \forall (i,j) < F \text{ and } E = \frac{\bar{N}_i{}^2}{P} \quad (1)$$

D. Estimate of False Positive Rate

An important issue related to the interpretation of the selected features is how to control the number of features falsely selected. In [2] a theoretical approach to provide a bound on the expected number of false selections was proposed. However, SCoRS involves random sub-sampling of both features and examples (instead of only sub-sampling examples, as proposed in Stability Selection theory). Therefore we are proposing an empirical approach to estimate the rate of false positive selection (FP) according to the following procedure:

I) Obtain the set of features S composed of the union of the features selected in at least half of the CV-folds; II) Obtain P, the complementar set of S; III) Permute the examples for all features $p \in P$; IV) Using the data matrix updated with features permuted across the examples, run SCoRS again (using the same nested-CV framework); V) Compute how many features in P are selected (this number corresponds to the estimation of how many features were falsely selected).

The reasoning behind this evaluation is to assess what proportion of the features whose correlation with the label has been destroyed through permutation are still selected by chance. Ideally, none of the permuted features should be selected, as the permutation should destroy the correlation between data and labels. However, if the dataset is small, some correlation might still be kept as the number of possible permutations is limited. It is important to emphasize that all examples belonging to the same subject (four examples in this dataset, as explained in section II-E) are kept together, i.e. not permuted among themselves.

E. Neuroimaging Data

We applied the proposed approach to a real functional MRI dataset acquired during visualization of happy faces,

which is part of a depression study. Thirty patients (diagnosed with Recurrent depressive disorder, Depressive episodes, or Bipolar affective disorder) were matched to 30 comparison subjects according to gender, age, smoking, and handedness. The experimental design consisted of viewing emotional faces in a blocked design. Every block was repeated 4 times in a random order. Face blocks were alternated with blocks showing a white fixation cross. More details regarding the context of the original study where the data were acquired can be found in [7].

Data were pre-processed using SPM5 (Wellcome Department of Cognitive Neurology, UK). Slice-timing correction was applied, images were realigned, spatially normalized and smoothed using an 8 mm FWHM Gaussian isotropic kernel. Additional pre-processing was performed using custom-built Matlab routines: a mask was applied to each image in order to extract only voxels that contain brain tissue in all subjects; then, for each subject, all the voxels inside the mask were linearly detrended. Before selecting the examples (i.e. the BOLD signal images corresponding to the times in which the stimuli were presented), the scans were shifted to accommodate the delay due to hemodynamic response.

Within each block, individual scans were averaged to increase the signal-to-noise ratio, i.e. a temporal compression as proposed by [8] was applied. After pre-processing the resulting data-matrix was composed of 219727 features (voxels) and 240 examples (2 groups, 30 subjects in each group, 4 blocks or examples per subject).

III. Results

Figure 2 (a) presents the average model accuracy obtained for each threshold level in each inner fold according to the nested cross-validation framework described in section II-B. Black stars mark the threshold level for which the maximum accuracy was obtained in each CV-fold. When more than one threshold level resulted in the same maximum accuracy, the median value was used. Figure 2 (b) presents the average reproducibility for each threshold in each inner fold. Please note that figures 2 (a) and (b) have different colour scales, as they present different measures. .

Table I presents a comparison of results obtained from both optimization frameworks. The first three rows contain results related to the model performance. Sensitivity corresponds to the proportion of patients correctly classified, specificity is the proportion of healthy subjects correctly classified and Accuracy is the arithmetic mean of sensitivity and specificity. Table I also includes the *average* number of features (*NF*) selected in each fold, the *union* (i.e. features selected in at least one fold) and the *intersection* (i.e. features selected in all folds). The last row of the table shows the absolute (and the percentage) number of falsely selected features according to the procedure described in section D.

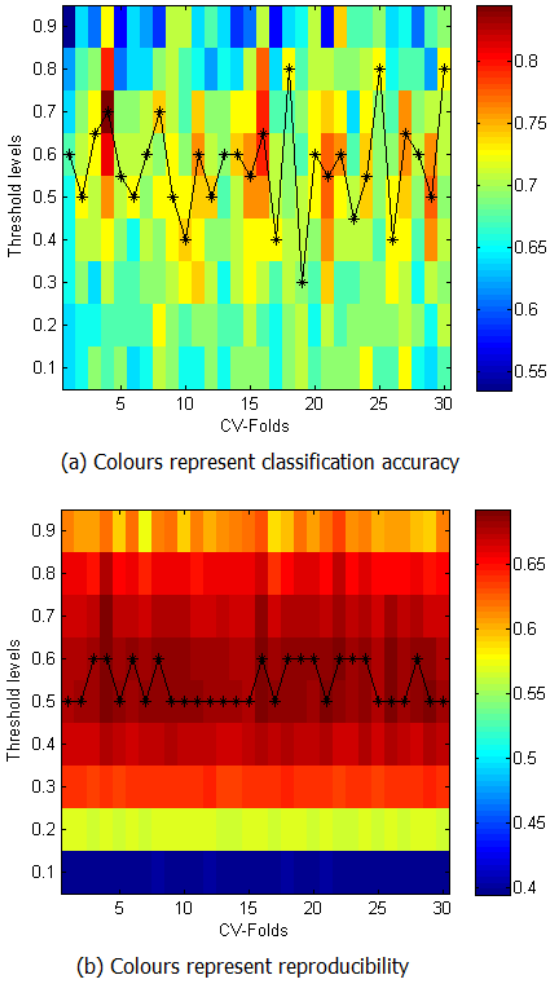

Figure 2. Threshold optimization in a nested cross-validation framework (a) Optimizing accuracy (b) Optimizing reproducibility

Figure 3 displays the set of features S, composed of the union of all features selected in at least half of the CV-folds (as described in section II-D) for both optimization approaches: based on accuracy (a) and based on reproducibility (b). The colours scale represents the relevance calculated by SCoRS, given by the survival rate (the number of times each feature was selected divided by the number of times the feature was chosen in random subsets of features). Survival rates for all features in S were averaged across CV-folds. The selected features with corresponding relevance colours were overlaid on an anatomical template.

IV. Discussion and Conclusion

In the present work we proposed a new criterion for threshold optimization in SCoRS based on reproducibility rather than on classification accuracy. The comparison between the different criteria for parameter optimization showed that optimizing reproducibility did not decrease the

Table I
COMPARING RESULTS OF SCORS OPTIMIZING ACCURACY (APPROACH 1) AND OPTIMIZING REPRODUCIBILITY (APPROACH 2)

Measures	**Approach 1**	**Approach 2**
Sensitivity	77%	77%
Specificity	60%	67%
Accuracy	68%	72%
NF (average)	8513.5	8792.1
NF (union)	33542	23167
NF (intersection)	763	2129
FP	9955 (4.67%)	3995 (1.89%)

model performance in comparison with optimizing accuracy; instead, it slightly increased the accuracy.

As it was expected, optimizing reproducibility increased the overlap between the selected features across CV-folds (*intersection* set) and reduced the spreading of selected features across CV-folds (*union* set). Nevertheless, the average number of features selected across the CV-folds was very similar for both approaches.

It is remarkable to observe that the threshold leading to the maximum reproducibility is very stable among the CV-folds (varying from 0.5 to 0.6). This finding is consistent with the theory of Stability Selection [2].

Regarding the maps showing the localization and relevance of the selected features, it is important to observe that even though no spatial constraints were applied (features were randomly chosen from the whole brain in each iteration), the selected features consist of clusters spatially connected. As neighbor voxels in the brain are correlated due to physiological properties and pre-processing procedures we expect them to share predictive information. The spatial maps obtained using the different parameter optimization approaches were very similar. However the optimization based on reproducibility resulted in slightly less noisy maps then the optimization based on accuracy.

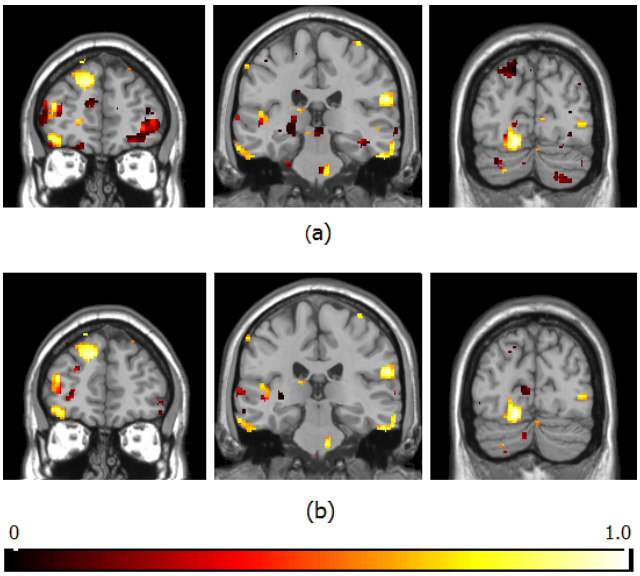

Figure 3. Relevance maps resulting from SCoRS optimized using accuracy (a) and using reproducibility (b).

In addition we proposed a procedure to estimate the rate of false positive selection. Our results showed that the proportion of permuted features included in the model (FP) was smaller when optimizing reproducibility (1.89%) than when optimizing accuracy (4.67%). This is an important indication that optimizing reproducibility leads to higher stability.

Another contribution of this work was to propose an approach to summarize the models from different CV-folds. It is possible to notice from figure 3 that the application of SCoRS with this summarization framework produced maps containing clusters with highly relevant features. The proposed approach can produce maps displaying relevant features with an associated rate of false positive selection, a property very desirable in neuroimaging based pattern recognition applications.

ACKNOWLEDGMENT

JMR was supported by Capes (Coord. for Improvement of Higher Level Personnel), Brazil (3883/11-6). JMM was supported by Wellcome Trust Career Development Fellowship (WT086565/Z/08/Z). The authors would like to thank Tim Hahn and Andreas J Fallgatter from Univ. of Wuerzburg and Frankfurt for kindly providing images for this study.

REFERENCES

[1] J. Rondina, J. Shawe-Taylor, and J. Mourao-Miranda, "A new feature selection method based on stability theory - exploring parameters space to evaluate classification accuracy in neuroimaging data," *LNAI Survey of the state of the art MLINI*, vol. 7263, pp. 58–66, 2012.

[2] N. Meinshausen and P. Buhlmann, "Stability selection," *J. R. Stat. Soc.*, vol. 72, pp. 417–473, 2010.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso." *J. R. Stat. Soc.*, vol. 58, pp. 267–288, 1996.

[4] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers." *5th Annual ACM Workshop*, p. 144152, 1992.

[5] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[6] P. Rasmussen, L. Hansen, K. Madsen, N. Churchill, and S. Strother, "Model sparsity and brain pattern interpretation of classification models in neuroimaging," *Pattern Recognition*, vol. 45, pp. 2085–2100, 2012.

[7] T. Hahn, A. Marquand, A. Ehlis, T. Dresler, S. Kittel-Schneider, T. Jarczok, K. Lesch, P. Jakob, J. Mourao-Miranda, M. Brammer, and A. Fallgatter, "Integrating neurobiological markers of depression," *Arch Gen Psychiatry*, vol. 68, no. 4, pp. 361–368, 2011.

[8] J. Mourão-Miranda, E. Reynaud, F. McGlone, G. Calverte, and M. Brammer, "The impact of temporal compression and space selection on svm analysis of single-subject and multi-subject fmri data," *Neuroimage*, vol. 33, no. 4, pp. 1055–1065, 2006.