An ensemble of classifiers approach with multiple sources of information

Roberto Santana, Concha Bielza, Pedro Larrañaga
Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid
roberto.santana@upm.es, mcbielza@fi.upm.es,
pedro.larranaga@fi.upm.es

Abstract

This paper describes the main characteristics of our approach to the ICANN-2011 Mind reading from MEG - PASCAL Challenge. The distinguished features of our method are: 1) The use of different sources of information as input to the classifiers. We simultaneously use information coming from raw data, channels correlations, mutual information between channels, and channel interactions graphs as features for the classifiers. 2) The use of ensemble of classifiers based on regularized multi-logistic regression, regression trees, and an affinity propagation based classifier.

1 Type of information used for classification

The first building block of our approach is the combination of different sources of information extracted from the MEG signals. We hypothesize that different transformations to the brain signals could reveal diverse types of brain signatures useful for the classification purpose. Therefore, we have tried different information processing variants to unveil this information. In all cases, the starting point was the time series output from the $N = 727$ training cases, for the $k = 204$ channels. For the training set, there are a total 727 cases and 204 time series for each case. The MEG output data corresponds to 200-component numerical vector.

The first type of brain signal representation is constructed by splitting the time series in segments of 5 contiguous time points, and adding the raw signals in each segment. We obtain, for each channel, a vector of 50 features. Therefore, for a fixed frequency, each of the 727 cases will be represented by $204 \times 50 = 10200$ features. We call to this relatively simple transformation of the initial information *raw data*.

For each of the cases, we use its corresponding raw data to compute the correlations between each pair of channels for this case. For example, to compute the correlations between channels i and j, their corresponding vectors of 50 raw values are used. As a result, a symmetric matrix $\mathbf{W}_{204 \times 204}$ is obtained from each case. The final set of features of each case will comprise a vector of $n = \frac{204 \cdot 203}{2} = 20706$ values corresponding to the upper triangular part of the correlation matrix (without the main diagonal). This type of information is called *channels correlations*. This approach intends to compute the interaction between different brain regions during the solution of the recognition task.

In a similar way we compute, for each case, the matrix of mutual information between the channels. First, the continuous data corresponding to two variables, are discretized and from the discretized values the mutual information is obtained. The bin size for discretizing all the data was fixed to equal value of 11. Similarly to the computation of the correlation, the final set of features will comprise vector of $n = \frac{204 \cdot 203}{2} = 20706$ values which are called the *mutual information between channels*. This approach also tries to unveil interaction between different brain regions that could be specific to each mental task.

In the fourth signal processing procedure, the correlation matrix is used to construct interaction graphs between the different channels. The idea is that a further analysis of the graph using topological measures from network theory can serve to reveal local and global information that is not directly recognizable from the correlation values.

The interaction graph $G = (V, A)$ is such that $V = \{v_1, \ldots, v_{204}\}$ is the set of vertices and arc $a_{i,j}$ between vertices v_i and v_j is defined as follows:

$$a_{i,j} = \begin{cases} 1 & if \quad i < j \text{ and } cr_{i,j} > 0.5 \\ -1 & if \quad i < j \text{ and } cr_{i,j} < -0.5 \\ 0 & \text{otherwise} \end{cases}$$

where $cr_{i,j}$ is the correlation coefficient between channels i and j, and

$Information - Freq$	$Full$	$2H$	$5H$	$10H$	$20H$	$35H$
Raw	236	0	0	0	0	0
$Correlation$	547	64	122	501	806	3566
$MutualInf.$	31	5	14	49	98	356
$Interactiongraph$	16	0	0	39	61	349

Table 1. Number of selected features of each type of information and frequencies.

values 1, -1 and 0 for $a_{i,j}$ respectively mean that there is an arc from v_i to v_j, there is an arc from v_j to v_i, or there is no arc between v_i and v_j.

The interaction graph is an arbitrary way to represent strong correlations (below -0.5 or above 0.5) between pairs of channels. We expect that if there are higher order interaction patterns between the channels, at least some of them could be unveiled by a topological analysis of these graphs.

Once correlation graphs have been constructed, a number of (local) topological measures are computed for each node (e.g. clustering coefficient, path length, betweenness centrality, etc.). In addition, a number of global topological measures are computed for the complete graph (e.g. graph density, graph diameter, etc.). The number of local features was $n_{local} = 204 \cdot 13 = 2652$ and the number of global features was $n_{global} = 7$. The total number of topological features extracted for each graph was $n = 2659$. We call to this type of information *channel interactions graphs*.

1.1 Feature selection

In order to identify a reduced set of significant features, we applied, for each feature, a statistical test to determine whether there exists significant different between the 5 different classes for the given feature. The statistical test was applied to each pair of classes. The idea was to identify whether a given feature is effective at identifying differences between any of the 10 possibles pairs of classes. A more stringent requirement would be the identification of features that are significantly different between the 5 classes altogether. However, in our approach we keep features that detect "local" differences between classes.

The statistical test of choice was the Wilconxon rank sum test of equal medians and the parameter $\alpha = 10^{-5}$ was fixed for all the statistical tests. Table 1 shows the number of significant features found for each frequency

	Raw data					Correlation				
Class	1	2	3	4	5	1	2	3	4	5
1	–	**63**	96	10	8	–	61	67	2326	2922
2		–	221	46	0		–	281	2374	3437
3			–	12	55			–	1852	2456
4				–	**0**				–	3102

	Mutual information					Interaction graph				
Class	1	2	3	4	5	1	2	3	4	5
1	–	10	3	249	221	–	**0**	5	92	151
2		–	25	270	254		–	4	281	147
3			–	211	174			–	244	133
4				–	330				–	**278**

Table 2. Number of significant features for all pairs of classes and types of information.

and each type of information. Table 2 shows the number of significant features found for each pair of classes and using all sources of information. Notice that a feature may be significant in the comparison of two or more pairs of variables. Emphasized in bold are the marked differences between the raw data and the interaction graph types of information in terms of the number of relevant features they respectively find for class pairs $(1, 2)$ and $(4, 5)$. These differences confirm our hypothesis that different types of information may reveal different types of brain signatures.

For the classification purpose we use the combined set of all the 6860 relevant features included in Table 1.

2 Classification approaches

Three different classification approaches were used: Elastic net regularized multi-logistic regression [3], regression trees [1] and affinity propagation [2]. The first two methods are supervised classification methods and were initially evaluated in the training set using a 5-fold cross-validation scheme. The second method is an unsupervised classification method that we directly used as a way to classify the test cases similarly as described in [4].

Using 5-fold cross-validation on the training set with the complete set of 6860 variables we observed that elastic net multi-logistic regression was able to reach a 0.83 classification rate for different values of $\beta \in \{0.01, \ldots, 0.9\}$. We then trained the model using the complete set of 727 solutions and used it to classify the test set. 21 different classifications corresponding to different pairs of (α, β), those that achieved and accuracy over 0.98 in the complete training set, were obtained. We called this set of solutions MLRSet.

To evaluate the regression trees, the set of 6860 variables was split into 26 different sets of (overlapping) variables. Each set excluded a subset of features relevant in the identification of 2, 3 or 4 classes, i.e. we used the grouping of variables shown in Table 2 to partition the set of variables. For each subset of features, we used cross-validation on the training set, to learn a regression tree for each subset of features. Of the initial set of 26, three regression trees were removed due to achieve a classification accuracy under 0.48. The remaining 23 were used to create an ensemble of regression trees with the majority vote strategy. Its application, using 5-fold cross-validation on the training set gave an accuracy of 0.6066. The application of each individual tree to the test set produced a set of 23 solutions. We called this set of solutions TreeSet.

Affinity propagation was applied to the combined set of training and test cases. However, by penalizing the preference values of the test cases we enforce that only train cases are allowed to be an exemplar. A test case is classified in the same class its corresponding exemplar belongs to. To evaluate the quality of the classification, we computed the number of non-exemplar training cases that were correctly classified. We have previously observed [4] that this may be an indirect measure of the classification quality for the test cases. 9 different similarity measures were applied to the 26 sets of variables in which the initial set of features was partitioned. As a result, we obtained a set of 234 clusterings. From these clusterings, we selected those for which the number of correctly classified non-exemplar training cases was above 0.60. There were 11 such clusterings. Each cluster determines a assignment to the test cases. We called this set of solutions APSet.

To obtain the final solution, we compute, for each of the three sets produced by the classifiers, the class probability for each test case. The class probability is simply the frequency of each class in the corresponding set for the given test case. The final probability of a case is found as a weighted sum of the probabilities for each of the three sets, i.e.

$p_F = 0.4p_{MLRSet} + 0.3p_{TreeSet} + 0.3p_{APSet}$. The weights were determined according to the accuracies obtained by the two supervised classification algorithms in the training set and we assumed that affinity propagation achieved a classification rate similar to regression trees. The final assignment of a given test case will correspond to the class with the highest class probability in p_F.

Bibliography

[1] L. Breiman. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

[2] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[3] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[4] R. Santana, C. Bielza, and P. Larrañaga. Affinity propagation enhanced by estimation of distribution algorithms. In *Proceedings of the 2011 Genetic and Evolutionary Computation Conference GECCO-2011*, Dublin, Ireland, 2011. Accepted for publication.