Classification of inter-subject fMRI data based on graph kernels

Sandro Vega-Pons*†, Paolo Avesani*†, Michael Andric† and Uri Hasson†
* NeuroInformatics Laboratory (NILab), Fondazione Bruno Kessler, Trento, Italy, Email: vega@fbk.eu
† Centro Interdipartimentale Mente e Cervello (CIMeC), Università di Trento, Italy

***Abstract*—The analysis of human brain connectivity networks has become an increasingly prevalent task in neuroimaging. A few recent studies have shown the possibility of decoding brain states based on brain graph classification. Graph kernels have emerged as a powerful tool for graph comparison that allows the direct use of machine learning classifiers on brain graph collections. They allow classifying graphs with different number of nodes and therefore the inter-subject analysis without any kind of previous alignment of individual subject's data. Using whole-brain fMRI data, in this paper we present a method based on graph kernels that provides above-chance accuracy results for the inter-subject discrimination of two different types of auditory stimuli. We focus our research on determining whether this method is sensitive to the relational information in the data. Indeed, we show that the discriminative information is not only coming from topological features of the graphs like node degree distribution, but also from more complex relational patterns in the neighborhood of each node. Moreover, we investigate the suitability of two different graph representation methods, both based on data-driven parcellation techniques. Finally, we study the influence of noisy connections in our graphs and provide a way to alleviate this problem.**

***Keywords*-brain decoding; connectivity graphs; graph kernels; brain parcellation; inter-subject classification**

I. INTRODUCTION

During the last decade, brain decoding [1] has become a common approach to fMRI data analysis. Generally, this approach aims at predicting whether a perceptual, cognitive, or behavioral stimulus associates with a collected sample of fMRI data. Usually, fMRI data is represented as volumes of four dimensional samples, where each sample is associated to a category. The common vectorial encoding of brain data introduces a bias towards functional segregation studies [2] in contrast to functional integration studies. In segregation studies the inference is driven by regions of voxels with high statistical dependency. This approach is effective when the experiment's design tests a hypothesis investigating brain activity that might be localized to a particular area. In contrast, functional integration studies aim to decode the information captured by relationships between distributed brain regions. Here, an activation pattern is assessed by its network structure, rather than by specificity in a particular region.

Network analyses of brain activation [3] are quite common in brain connectivity studies. Their main purpose is to perform either hypothesis testing on the functional connectivity or inference from topological graph properties such as modularity, node degree distribution or clustering coefficient. All of these methods have in common a graph based encoding of fMRI recording. A recent survey [4] reviews the different ways to model the brain activation as a graph.

Early attempts to approach the brain decoding task with graph encoding were based on vectorial embedding [5]. The graph representation is conceived as an adjacency matrix that is subsequently unfolded into a real vector. There are some restrictive conditions if the method is applied to brain decoding across subjects: the fMRI data has to be registered into a common space, thus leading to graphs with the same number of nodes and a correspondence between nodes across subjects. This constraint is called *fixed-cardinality vertex sequence* property [6].

An alternative approach is based on the notion of graph kernels [7]. In this case, the challenge is to design a graph kernel that is sensitive to the relational information and also computationally efficient. The practical application of graph kernels to the problem of brain decoding overcomes the efficiency issue by reducing the size of the graphs, either by focusing on a region of interest [8] or by computing a parcellation of the brain data [8], [9]. More recently, an efficient graph kernel has been proposed: the Weisfeiler-Lehman kernel [10]. Its reduced computational complexity enables whole brain graph analysis [11]. It has also been successfully applied to a Mild Cognitive Impairment study on resting state [12].

In general, different kinds of graph kernels have been applied in neuroimaging problems: a custom-designed kernel [8], the shortest-path kernel [9] and the previously mentioned Weisfeiler-Lehman kernel [11]. However, there is still no clear evidence on whether they are effective to detect the relational information encoded in the fMRI signal. The graph kernel designed in [8] is only able to detect pairwise relationship information. The computation is done at the level of edge comparison and more complex relations are not considered. In [9], the computation of the graph kernel is affected by the *fixed-cardinality vertex sequence* property. Moreover, it can only be applied on small graphs due to its high computational complexity. In [11], it is shown that the graph kernel may produce above-chance accuracy results. However, it is not clear what kind of information is decisive in the discrimination task.

In this paper, we address the question whether a graph kernel is really exploiting the relational information encoded in a graph representation of fMRI data. More in detail, we investigate whether the information is captured by the node degree distribution of graphs or even by higher order relationships that are difficult to quantify by following other approaches. We focus our work on Weisfeiler-Lehman graph kernel because it supports the whole brain analysis and it does not require the fixed-cardinality vertex sequence property.

The additional contribution of this work is concerned with the relationship between the method to encode the graphs and the graph kernel. We are interested in assessing how the choice of graph encoding can influence the subsequent task of learning from graphs. Moreover, we analyze the impact of node labeling techniques in the graph encoding and how we can alleviate the effect of noisy connections.

In order to address our questions we perform an empirical analysis on a neurocognitive experiment. The experiment aims at assessing how the brain is processing auditory stimuli of different complexities. The protocol was designed as time unlock stimulation and the working assumption is that no specific brain region is devoted to this task.

II. Methods

We study the fMRI brain decoding problem across multiple subjects. Let $\mathbb{T} = \{T_1, T_2, \ldots, T_n\}$ be the set of n trials in a task-related fMRI experiment with m subjects and $Y = \{y_1, y_2, \ldots, y_n\}$ be the corresponding class labels (stimuli). Each trial T_i is composed of a set of voxels $T_i = \{w_1, w_2, \ldots, w_t\}$, where each voxel $w_j = (f_j, c_j)$ is determined by its time series f_j and geometrical coordinates c_j ($< x, y, z >$ spatial coordinates). Our approach for brain decoding is based on three main elements:

- A *graph encoding* method, which builds a graph from every fMRI trial. In other words, each trial T_i is mapped into a graph G_i and the brain decoding problem is transformed into a graph classification problem with the following class-labeled graph dataset $\mathbb{D}_G = \{(G_1, y_1), \ldots, (G_n, y_n)\}$. The graph encoding techniques that we will consider are presented in Section II-A.
- A *graph kernel*, which is used to compute the similarity between each pair of graphs, and therefore a kernel matrix for the classification problem. In Section II-B we discuss our graph kernel choice.
- A *kernel-based classifier* able to work with the kernel matrix. In this case we apply the standard Support Vector Machines (SVM).

A. Graph encoding of fMRI data

We use *simple*, *undirected* and *node-labeled* graphs $G = (V, E, \ell)$ to encode the information in each fMRI trial T_i, where V is a set of nodes, $E \subset V \times V$ a set of undirected edges and $\ell : V \to \Sigma$ is a function that assigns a label from an alphabet Σ to each node in the graph.

We define our graphs in 3 steps. The first step is the *node definition*. We follow the idea of computing a parcellation of the brain data [4] and then assigning a node to each parcel. The goal of parcellation is to reduce the impact of noise in the graph computation and also to reduce the size of the graphs. We adopted two different parcellation methods based on the application of clustering algorithms:

- *Geometrical parcellation*: We apply the Ward's hierarchical clustering algorithm to the set of voxels but only using the geometrical coordinates (c_j) of each voxel as the features. Time series does not come into play when quantifying distance between voxels. It is similar to upscaling the data at certain resolution.
- *Functional parcellation with geometrical constraints*: We apply Ward's algorithm by using the time series (f_j) of each voxel as features. Geometrical features (c_j) of each voxel are used as a constraint to avoid clusters with more than one connected component in the geometrical space.

Once the nodes are defined, the average time series $\widehat{f}$ of all voxels in each parcel is computed and associated to the corresponding node. Afterwards, the edges are determined by computing the *Pearson Correlation Coefficient* ρ between the time series of each pair of nodes and thresholding it with a fixed value τ, i.e. $e_{jk} \in E \Leftrightarrow \rho(\widehat{f}_j, \widehat{f}_k) \geq \tau$.

The node labeling function ℓ is defined such that each node is assigned with its node degree. Moreover, in Section IV-C we explore the use of a more robust labeling mechanism.

B. Graph kernel

Graph kernels have become a popular choice for graph classification problems [7]. They allow the direct use of kernel based classifiers (e.g. SVM) on graph data. A common limitation of graph kernels is their high computational complexity, which makes them mostly useful for the comparison of small graphs. However, recent proposals like the Weisfeiler-Lehman (WL) subtree kernel [10] can be efficiently computed in time $\mathcal{O}(|E|)$. Furthermore, it is a meaningful way of comparing graphs since it is based on the 1-dimensional variant of Weisfeiler-Lehman test of graph isomorphism [10].

The computation of this kernel for two graphs is performed by an iterative process, which starts by comparing the node labels of both graphs. A new artificial label is then computed for each node by compressing the node labels of its neighboring nodes. Afterwards, the graphs with the compressed nodes are compared. This process is repeated until the desired number of iterations is reached. More formally, given two graphs G and G', the WL kernel with h iterations is defined as:

$$WL[h](G, G') = \left\langle \phi_{(h)}(G), \phi_{(h)}(G') \right\rangle \tag{1}$$

where $\phi[h](G)$ is a vector containing the number of occurrences of all existing labels until iteration h is reached for graph G. In our experiments we set $h = 2$ and in Section IV-B we analyze the type of information that the kernel is using in each iteration. Moreover, we use a normalized version of the kernel: $\widehat{WL}(G, G') = \frac{WL(G,G')}{\sqrt{WL(G,G) \cdot WL(G',G')}}$. This way, the kernel takes values in the interval $[0, 1]$ and we avoid the possible adverse effect of the different sizes of the graphs in the comparison. Notice that there is no correspondence between nodes across different subjects and in general all graphs may have different number of nodes and edges.

III. Materials

We use in our analysis the data from 19 healthy participants (with normal hearing) engaged in a passive listening task lacking any executive component. Subjects were presented with two types of auditory stimuli: *Ordered* and *Disordered*, as well as two other conditions not discussed here. These stimuli were designed by using sequences of pure tones at 262, 294, 330 and 349 Hz, corresponding to middle "C", "D", "E" and "F" notes on the Western major scale. The tone sequence order was determined by using a first-order Markov process applied to two transition matrices with different levels of Markov entropy (0.81 and 1.56). The two entropy levels marked the two experimental conditions in this study. Each transition matrix was used to generate 90 sec of auditory stimuli where tones were presented at a rate of 3.3 Hz. The time series collected within these 90 sec were the core data of the study, representing a single trial (sample) in our dataset. Each stimulus was presented once to each participant, therefore our dataset is composed of 38 trials or samples.

All images were acquired using a 4T Bruker/Siemens system. For functional images, we used a single shot echo planar imaging sequence to collect 25 interleaved slices parallel to the AC/PC, with TR = 1500 ms, TE = 33 ms, flip angle = 75 degrees, voxel size = $4 \times 4 \times 4.8$ mm, matrix = 64×64 mm, and slice skip factor = 0.2. We collected 471 of these blood oxygen level dependent scans over a single 706.5 sec run. For anatomical images, we used a 3D T1weighted MPRAGE sequence to collect 176 sagittal slices, with TR = 2700 ms, TE = 4 ms, flip angle = 7 degrees, matrix = 256×224, and isotropic voxel size of 1 mm.

IV. Results and Discussion

In our experiments, we perform a leave-one-subject-out (LOSO) cross-validation, i.e. in each fold, we train with the data of 18 subjects and test on the data of the remaining one. According to a Binomial distribution, any accuracy above 0.65 is significant for this problem, with a p-value below 0.05.

A. Different graph encoding techniques

In this experiment we are comparing the classification accuracies we obtain when using two different node definition techniques for graph encoding: *geometrical parcellation* and *functional with geometrical constraints* (see Section II-A). Both parcellation techniques depend on two parameters:

- cr: Cluster ratio. The hierarchical clustering algorithm requires the number of clusters k to be computed. However, we will not define the value k directly because each subject may have different number of voxels as we are working on individual subject spaces. Thus, a parameter cr is defined such that $k = n_i/cr$, where n_i represents the number of voxels for the i-th subject.
- τ: Threshold for correlation. The threshold to be used for the definition of the edges in the graph.

In order to optimize these parameters and provide a fair estimation of the accuracy, we follow the idea of grid-search. In each fold of the LOSO we do an internal cross validation (again leaving one subject out) to estimate the parameters. We tested the following values in our grid-search, $cr = \{100, 120, 140, 160, 180, 200\}$ and $\tau = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

In Figure 1 we show the accuracy for the best parameters in the grid-search associated to each fold in the external cross-validations procedure.

The average accuracy of the whole LOSO procedure is:

- Geometrical encoding: 73.68%
- Functional encoding: 60.52%

These results were obtained with the best combination of values for the parameters from the grid-search inside each fold of the LOSO cross-validation.

In general, these results could be counter-intuitive at first sight. One could expect a higher accuracy from a parcellation process that uses the functional information of the voxels. However, the use of functional information causes the joining of voxels that belong to different spatial regions in the brain. This avoids the creation of strong links between these regions,

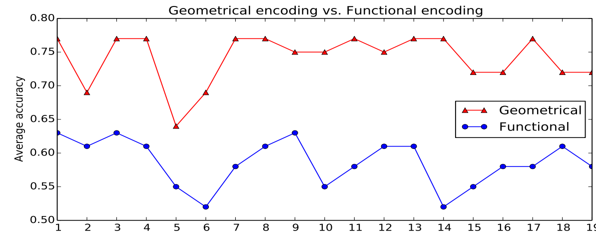

Fig. 1. Comparison of the best accuracy values in the internal grid-search for each one of the 19 folds in the LOSO procedure.

therefore leading to graphs with less discriminative information for the problem. Consequently, we decided to discard the functional parcellation-based encoding and will focus on the geometrical-based encoding.

In order to show the complexity of this problem, in Figure 2 we provide a general description of the graph population we are working with. Analyzing the best parameter configuration for each fold in the cross-validation, we obtained that the most stable parameters for the geometrical-based encoding were $cr = 140$ and $\tau = 0.4$. Thus, the graph dataset shown in this figure, corresponds to the geometrical encoding procedure for this parameter's combination.

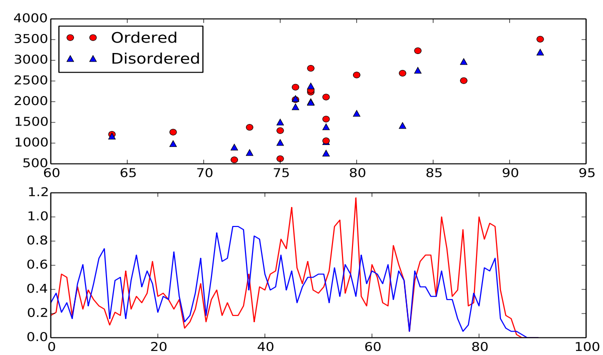

Fig. 2. Top: Number of edges vs. number of nodes for all graphs. Bottom: Average node degree distribution for class.

In the top part of Figure 2, we show the relation between the number of nodes and edges for all graphs in both classes. Notice that any two graphs corresponding to the same subject have the same number of nodes. It can be seen from this figure that there is no clear pattern that allows the discrimination between classes by only looking at the number of nodes and edges. In the bottom panel of this figure, we show the average node degree distribution for all graphs in both classes. Even though the average plot is not very reliable, in this case we can appreciate that one class tends to have nodes with higher node degree than the other.

B. Graph kernel interpretation

In this section, we analyze the type of information that the graph kernel is taking into account. As we mentioned in Section II-A, we are using the node degree value of each node as its node label in our graph encoding. Therefore, we can expect that the node degree distribution of the graphs is a discriminative feature for this problem (this idea is to some extent supported by Figure 2). Analyzing the definition of WL kernel (see Section II-B) we conclude that in fact, the first iteration just compares the node degree distributions of the graphs. Besides the theoretical analysis of the WL kernel, we corroborated this claim with the use of synthetic data.

Nevertheless, this is not the only information taken into account by the WL kernel. In its second iteration, this kernel is sensitive to similar patterns on the neighborhood of each node among the different graphs. All the labels in the neighborhood of a node are used to compute a second order label. Therefore, the matching of two second order labels means the matching of the whole neighborhoods of two nodes in the graphs.

In the computation of the kernel, the final result is obtained by adding the results of the two iterations. Therefore, we can compare the accuracy results of our method when using:

- WL[1,2]: The WL graph kernel with the two iterations (in the way it has been previously used in this section).
- WL[1]: The WL graph kernel with only the first iteration (only node degree distribution).
- WL[2]: The WL graph kernel with only the second iteration (only neighborhood pattern).

In the three cases we use the geometrical encoding with the best parameters we found in our grid search, i.e. $cr = 140$ and $\tau = 0.4$. The average accuracy of the LOSO experiment for each case is:

- WL[1,2]: 73.68%
- WL[1]: 68.42%
- WL[2]: 57.89%

From these results we can conclude that a significant part of the discriminative information comes from the node degree distribution. However, there is also some valuable information in the neighborhood patterns of each node. Hence, the combination of both sources provides the most accurate results.

C. Influence of node labeling

The node degree has been proven to be a meaningful label for the graph encoding. However, it can be sensitive to noise. The node degree of a given node can be affected by just adding or removing one edge from it. This simple variation would affect the result of the WL graph kernel. In this section, we explore the idea of using a more robust label definition for the graph encoding step. The idea is to compute intervals of consecutive node degree values and assign the same label to all possible node degree values in the interval. Therefore, given the node degree distribution of a graph, we are assigning the same label to an interval of p consecutive node degrees values.

In Table I, we explore the influence of the parameter p in the classification accuracy. We compare the average classification accuracy for WL[1,2], WL[1] and WL[2] with $cr = 140$, $\tau = 0.4$ and the following p values: $p = \{1, 3, 5, 7, 9, 11\}$.

TABLE I
AVERAGE CLASSIFICATION ACCURACY OF WL[1,2], WL[1] AND WL[2] FOR DIFFERENT VALUES OF p.

GK \ $p =$	1	3	5	7	9	11
WL[1,2]	73.68	71.05	76.31	**78.94**	73.68	68.42
WL[1]	68.42	71.05	71.05	76.31	71.05	68.42
WL[2]	57.89	52.63	52.63	57.89	57.89	52.63

From the results of Table I we observe an improvement of the accuracy for some values of p. This supports the idea that the performance of the classifier can be improved by using a more stable labeling function. It can also be appreciated that, in most cases, there is a similar relation between WL[1,2], WL[1] and WL[2]. In other words, most of the information comes from the label distribution in the graphs, but some valuable information is also coming from the label patterns in the neighborhood of each node. Despite the results, defining uniform intervals of length p for the label assignment may not always be the best choice. A more complex intervals definition, e.g. taking into account the global node degree distribution of the graphs, could improve the results even more.

V. CONCLUSIONS

We have shown that brain parcellation based on geometrical features is a convenient approach for graph encoding. Moreover, we have found the Weisfeiler-Lehman kernel is a suitable option for brain graph classification. It is able to extract information from the node degree distribution and from more complex relational patterns in the graphs. The node degree distribution can be sensitive to noise, but this problem can be addressed by using a node degree agglomeration technique. In our experiments, we obtained above-chance accuracy results in the inter-subject classification of auditory stimuli. These results were obtained by using the full-brain fMRI data and without any spatial alignment of individual subject's data.

ACKNOWLEDGMENT

This research has been supported by the RESTATE Programme under the FP7 COFUND Marie Curie Action [grant number 267224]; and ERC-STG [grant number 263318].

REFERENCES

[1] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: a tutorial overview." *NeuroImage*, vol. 45, no. 1, pp. 199–209, 2009.
[2] O. Sporns, "Network attributes for segregation and integration in the human brain." *Current opinion in neurobiology*, vol. 23, no. 2, pp. 162–171, 2013.
[3] E. Bullmore and O. Sporns, "Complex brain networks: graph theoretical analysis of structural and functional systems." *Nature reviews. Neuroscience*, vol. 10, no. 3, pp. 186–198, 2009.
[4] G. Varoquaux and R. C. Craddock, "Learning and comparing functional connectomes across subjects," *NeuroImage*, vol. 80, pp. 405–415, 2013.
[5] J. Richiardi, H. Eryilmaz, S. Schwartz, P. Vuilleumier, and D. Van De Ville, "Decoding brain states from fMRI connectivity graphs." *NeuroImage*, vol. 56, no. 2, pp. 616–626, 2011.
[6] J. Richiardi and B. Ng, "Recent advances in supervised learning for brain graph classification," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 907–910.
[7] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph Kernels," *Journal of Machine Learning Research*, vol. 11, pp. 1201–1242, 2010.
[8] S. Takerkart, G. Auzias, B. Thirion, D. Schön, and L. Ralaivola, "Graph-Based Inter-subject Classification of Local fMRI Patterns," in *Machine Learning in Medical Imaging*, ser. LNCS, F. Wang, D. Shen, P. Yan, and K. Suzuki, Eds. Springer Berlin, 2012, vol. 7588, pp. 184–192.
[9] F. Mokhtari and G.-A. A. Hossein-Zadeh, "Decoding brain states using backward edge elimination and graph kernels in fMRI connectivity networks." *Journal of neuroscience methods*, vol. 212, no. 2, pp. 259–268, 2013.
[10] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-Lehman Graph Kernels," *J. Mach. Learn. Res.*, vol. 12, pp. 2539–2561, 2011.
[11] S. Vega-Pons and P. Avesani, "Brain Decoding via Graph Kernels," in *Pattern Recognition in Neuroimaging (PRNI), 2013 International Workshop on*. IEEE, 2013, pp. 136–139.
[12] B. Jie, D. Zhang, C.-Y. Y. Wee, and D. Shen, "Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification." *Human brain mapping*, pp. 1–22, 2013.