Mapping brains on grids of features for Schizophrenia analysis

A.Perina[1], D.Peruzzo[6], M.Kesa[2], N.Jojic[3], V.Murino[1,4], M.Bellani[5], P.Brambilla[7], and U.Castellani[4,⋆]

[1] Istituto Italiano di Tecnologia (IIT), Genova, Italy
[2] Tallinn University of Technology, Tallinn, Estonia
[3] Microsoft Research, Redmond, USA
[4] University of Verona, Department of Computer Science, Verona, Italy
[5] University of Verona, Department of Psychiatry, Verona, Italy
[6] IRCCS "E. Medea" Scientific Institute, Udine, Italy
[7] University of Udine, Department of Experimental and Clinical Medical Sciences, Udine, Italy

Abstract. This paper exploits the embedding provided by the counting grid model and proposes a framework for the classification and the analysis of brain MRI images. Each brain, encoded by a *count* of local features, is mapped into a window on a grid of feature distributions. Similar sample are mapped in close proximity on the grid and their commonalities in their feature distributions are reflected in the overlap of windows on the grid. Here we exploited these properties to design a novel kernel and a visualization strategy which we applied to the analysis of schizophrenic patients. Experiments report a clear improvement in classification accuracy as compared with similar methods. Moreover, our visualizations are able to highlight brain clusters and to obtain a visual interpretation of the features related to the disease.

1 Introduction

Neuroanatomical methods using Magnetic Resonance Imaging (MRI) are largely used to understand the structural brain changes due to a certain disease [6]. A common approach is to discover morphological abnormalities between neuropsychiatric patients and healthy controls in some areas of the brain [5, 6]. Pattern recognition techniques are playing an important role to perform statistics at an individual level on a multivariate feature space [9, 3].
In this paper, we exploit counting grids (CG) [8] for the effective analysis of MRI brain images. We described each brain, as a "bag" of local features c_z represented by cortical thickness values collected from the left temporal lobe[8] [6]. Then we employ the counting grid model, illustrated in Fig.1. It consists of a 2 dimensional grid, where each cell is represented by a probability distribution

⋆ Corresponding author.
[8] Each c_z represent the amount of thickness level z. See Sec. 4

over the features z, and where each brain is mapped in a *window* on this grid, based on how much each feature in the bag c_z agrees with the amount of features z present in the window. The learning algorithm for the grid ensures that similar samples are mapped close by, with their windows likely to intersect; the commonalities of feature distributions can be found in the intersection of the windows.

This paper builds upon this geometric reasoning and it proposes a framework for the classification and analysis of brain MRI images based on the properties of the counting grid outlined in the previous paragraph. As first contribution, we propose a robust strategy to learn the model, aimed at dealing with the reduced number of training samples that typically occurs in the medical domain. Then, we introduce a generative kernel based on the diffusion distance in the counting grid space which reached 83% accuracy in the classification of schizophrenic patients. Furthermore, as a third contribution, we propose a visualization framework to help an expert to discover implicit information in high-dimensional medical data. More specifically in Sec.4, we will show that the embedding provided by CGs represent a natural framework to perform *i)* subject-based and *ii)* feature-based analysis. For *i)* the learning algorithm maps subjects with similar characteristic in close proximity on the CG, making it possible to visualize clusters with specific characteristics. For *ii)* it is possible to analyze which features are more important for the involved disease using the gradients of the learned distributions on the grid and label embedding. Our findings confirm that the main cause of the disease is related with cortical thickness reduction [6].

2 The counting grid model and multiresolution learning

Data samples are often represented as bags of features without particular order [2, 3]. Each t-th observation is characterized by a vector – often called count vector $\{c_z^t\}$ – containing the number of occurrences or the amount of each feature z. The counting grid model, recently introduced in [8], is a generative model for such representations. In this model, individual distributions over words are arranged on a grid (see Figure 1) and each one is relatively sparse, with only a few features having significant probability of occuring. More precisely, a counting grid $\pi_{\mathbf{i},z}$ is a set of normalized counts of features indexed by z, on a 2-dimensional[9] discrete grid. The CG is indexed by $\mathbf{i} = (x, y)$, where $x \in [1 \dots E_x],\ y \in [1 \dots E_y]$. $\mathbf{E} = E_x \times E_y$ describes the extent of the counting grid. Since π is a grid of distributions, $\sum_z \pi_{\mathbf{i},z} = 1$ for every location $\mathbf{i}$ on the grid. Figure 1 is an illustration of the CG geometry. A given bag-of-features, represented by its counts $\{c_z\}$ is assumed to follow a distribution found in a window (and not a point) of the grid. In particular, using a window of dimensions $\mathbf{W} = W_x \times W_y$, each bag can be generated by first selecting a position $\mathbf{i}$ on the grid and then by placing the window in the grid such that $\mathbf{i}$ is its *upper left corner*. Then, all counts in this window are

[9] N-dimensional in general, here we focus on 2 dimensions.

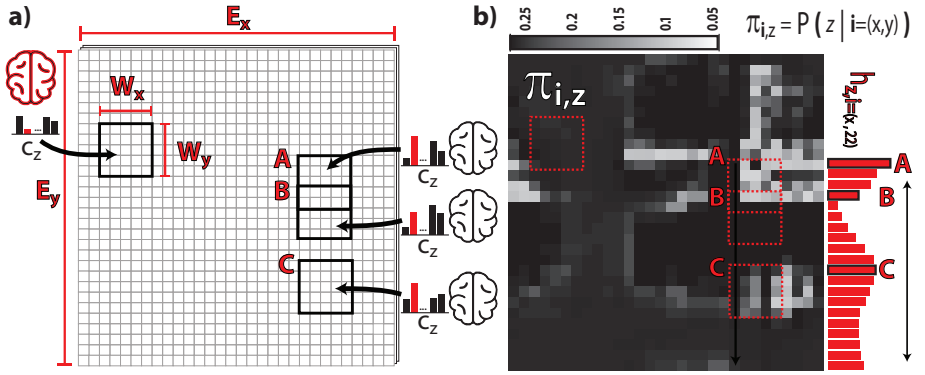

Fig. 1. a) The counting grid geometry. **b)** $\pi_{\mathbf{i},\hat{z}}$, the amount for a feature $\hat{z}$ related to high cortical thickness across the grid. The bar plot on the right represents the variation of the amount of $\hat{z}$ while shifting of 1 grid location, from window 'A' down to window 'C'. Bars correspond to the sum of the expression of $\hat{z}$ in a window of size $W = 5 \times 5$ (e.g., $h_{\mathbf{k},\hat{z}} = \sum_{\mathbf{i} \in W} \pi_{\mathbf{i},\hat{z}}$, see Section 2)

averaged to form the histogram $h_{\mathbf{i},z} = \frac{1}{W_x \cdot W_y} \sum_{\mathbf{k} \in W_\mathbf{i}} \pi_{\mathbf{k},z}$, and finally a set of features in the bag is generated. The window floating over the grid captures well a *slow* and *smooth* evolution of the features, which for example is often found among samples with the same phenotype.
This process is described by Fig. 1b where we show the counting grid $\pi_{\hat{z}}$ for a particular feature $\hat{z}$ (a "slice"). Here, we show with the bars on the left the expected amount of $\hat{z}$ in several windows $\mathbf{W}$ sliding down from position A to the bottom part of the grid (i.e., along column 22 - see Figure 1b). Some other windows are also evidenced and they refer to the three windows A,B and C as shown in the left panel; as visible the amount of feature $\hat{z}$ in the window, correspond with the amount of the feature present in the bags c_z, which is highlighted with the red color in the left panel.

To learn the model, we notice that the position of the window $\mathbf{i}$ in the grid is a latent variable given which the probability of the bag of features $\{c_z\}$ is

$$p(\{c_z\}|\mathbf{i}) = \prod_z h_{\mathbf{i},z}^{c_z} = \prod_z \left(\frac{1}{W_x \cdot W_y} \cdot \sum_{\mathbf{k} \in W_\mathbf{i}} \pi_{\mathbf{k},z}\right)^{c_z} \quad (1)$$

where $W_\mathbf{i}$ indicates the particular window placed at location $\mathbf{i}$ (see Fig.1a, windows marked with A, B and C).

Computing and maximizing the log likelihood of the data turns to be an intractable problem; therefore it is necessary to employ an iterative EM algorithm. Starting from a random initialization of the counting grid π, the E-step aligns all bags of features $\{c_z^t\}$ to grid windows, to match the bags' histograms, inferring $q_\mathbf{i}^t \propto \exp \sum_z c_z^t \cdot \log h_{\mathbf{i},z}$, i.e., where each bag maps on the grid. In practice, $q_\mathbf{i}^t$

is a probability distribution over the locations of CG and, after learning, it is equal to 1 if sample t maps to location $\mathbf{i}$. In the M-step the model parameter, i.e., the counting grid π, is re-estimated. For details on the learning algorithm and on its efficiency, the reader can refer to the original paper [8].

Multiresolution Learning: In the context of MRI brain image analysis, datasets are usually small, and the algorithms to learn generative models are prone to overfitting.
To alleviate this issue we propose multiresolution learning. Starting with a window size $\mathbf{W}^{init} = \mathbf{E} - \mathbf{1}$, at each step, we learn the model using the algorithm described in [8] for 10-20 iterations. Then we decrease the window size $\mathbf{W}^{(s+1)} = \mathbf{W}^{(s)} - \mathbf{1}$, and repeat the procedure using $\pi^{(s)}$ the counting grid at resolution s as initialization for resolution $s+1$. This process is repeated until we reach the desired size of $\mathbf{W}$, when we let the learning algorithm run until convergence.
This procedure helped avoid local minima. Discussing the theoretical motivations goes beyond the scope of this paper, however empirically we found that the classification results (see Sec. 4) obtained by learning the grid using multiresolution learning outperformed the standard procedure [8], with a p-value of $1.2e^{-5}$ (ANOVA Test, considering all complexities).

3 Diffusion kernel on the counting grid geometry

In the last years, hybrid generative discriminative paradigms have been proposed for classification [11, 1]. The idea is to firstly learn a generative model, and then use some of its by-products a features for a discriminative classifier. This usually yields better accuracy than the standard Maximum Likelihood approach. Here, we propose a generative kernel which exploits the geometric reasoning of the underlying generative model. We observe in fact that by construction, each point in the grid depends on its neighborhood, defined by $\mathbf{W}$. Indeed, we propose to consider this aspect when comparing two samples by their difference. In more detail, given two samples t and u and their mapping on the counting grid q^t and q^u, we propose to propagate their difference $\Delta^0 = q^t - q^u$ as:

$$\Delta^s = \Delta^0 * \phi(\mathbf{i}, s), \tag{2}$$

where $\phi(\mathbf{i}, s)$ is a box function defined by s. The idea of this propagation process is to capture the fact that samples close to the grid share feature content, thus are in some way similar. In fact, the size s can be naturally defined as the size of the Counting Grid window $\mathbf{W}$.

The proposed generative kernel is defined as:

$$k(q^t, q^u) = e^{-\rho||\Delta^s||_1} \tag{3}$$

where ρ is the standard bandwidth parameter of the Gaussian Radial Basis (RBF) kernel. We call our kernel *Diffusion CG* kernel[10] since there is a clear

[10] See [10] for a formal demonstration on the validity of the proposed kernel.

analogy with the diffusion distance introduced in [10] for histogram comparison. Here, we differ from [10] since, in order to be coherent with the counting grid estimation process, $\phi(.)$ is a box function rather than a Gaussian. Moreover, the size s of the box takes a fixed and known value, and therefore we can avoid the integration over all the scales like in [10]. Finally, in order to implement the hybrid generative-discriminative scheme the proposed generative kernel is employed with a Support Vector Machine (SVM) [4] as a discriminative classifier.

4 Classification and Analysis of Schizophrenic patients

Clinical Data The study population used in this work includes 42 patients with schizophrenia (21 male, 21 female) and 40 age-matched controls (19 male, 21 female). Diagnoses for schizophrenia were corroborated by the clinical consensus of two psychiatrists. MRI scans were acquired using a 1.5 T Siemens Magneton Symphony Maestro Class, Syngo MR 2002B. A coronal 3D MPR sequence was acquired[11] covering the entire brain. The FreeSurfer software[12] has been used to analyzed MRI images. Following the standard processing pipeline two meshes are built: one of the boundaries between the grey matter and the white matter and one between the grey matter and the cerebrospinal fluid. The cortical thickness is then computed as the shortest distance from each vertex to the complementary surface. The 3D meshes were also automatically divided into gyral based regions of interest (ROIs) following [5] and grouped in macro-ROIs referred to the principal lobes. In this study, we focus the analysis on the whole left temporal lobe since Schizophrenia affects a wide area of the brain usually not concentrated in a single region [6].

Finally, the thickness values observed on the left temporal lobe are accumulated into a histogram by defining 25 bins spanning the range 0-6 mm. Indeed, according to CG paradigm our words are the thickness values of each bin and the count is the histogram itself.

MRI image classification. We compared our method with: *i)* two baselines using SVMs with linear and histogram intersection kernels on $\{c_z^t\}$ [4], *ii*) Nearest Neighbor in the Counting Grid space using the mapping positions unequivocally identified by $q_{\mathbf{i}}^t$ as in [8] , *iii)* the generative-discriminative approach based on topic models of [3], and *iv)* a SVM with RBF kernel on the euclidean embedding (in 2D) produced by Locally Linear Embedding (LLE) [12]. Results are shown in Figure 2 where, as in [8], we identify the CG complexity by its capacity $\kappa = \frac{\mathbf{E}}{\mathbf{W}}$. We evaluate the accuracy using Leave-One-Out protocol by computing the average of 10 repeated tests to be robust against the CG training procedure. We considered CGs of 10 different complexities defined by grid size $\mathbf{E} = [3 \times 3, 6 \times 6, \ldots 30 \times 30]$ and window size $\mathbf{W} = \mathbf{2} \times \mathbf{2}$[13].

[11] TR = 2,140 ms, TE = 3.9 ms, flip angle = 15°, FOV = 176 x 235 mm2, matrix size = 384 x 512 x 144, voxel size 0.45 x 0.45 x 1.25 mm^3, TI = 1,100 ms

[12] version 4.3.1 http://surfer.nmr.mgh.harvard.edu/

[13] Larger windows yielded to slightly lower results up to $\mathbf{W} = \mathbf{5} \times \mathbf{5}$

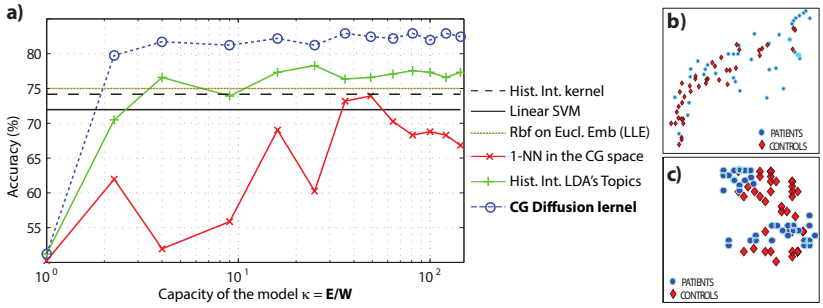

Fig. 2. a) Classification accuracies (logarithmic scale is used for the x-axis). The proposed kernels strongly outperforms all the competitors. For [3], the capacity on the x-axis represents the number of LDA topics. **b)** Euclidean embedding based on LLE (best result) **c)** Embedding (on a grid) provided by Counting Grid (best result)

In all the experiments we estimate the SVM and RBF parameters C and ρ by adopting the standard protocol described on the libSVM web page[14]. We implemented the method in Matlab starting from the code of CG public available[15].

Figure 2 shows clearly how the proposed approach outperforms the competitors [8, 3, 12], reaching an accuracy over 83%. We used permutation testing to evaluate the probability of getting accuracies higher obtained during the crossvalidation procedure by chance [7]. We permuted the labels 500 times without replacement (for each of 10 complexities tried), each time randomly assigning patient and control labels to each subject and repeated the crossvalidation procedure. Then, we counted the number of times the accuracy for the permuted labels were higher than the ones obtained for the real labels. Dividing this number by 500×10 we derived a p-value of $5.3e^{-4}$.
Finally, the panels b) and c) of Figure 2 depict the embedding produced by CG and LLE (in correspondence of the best classification result). The former highlights two clusters of patients and it looks visually better. Moreover, being the samples arranged on a dense space of features (i.e.., each position of the grid is a distribution over the features z), we can exploit the CG embedding to investigate how features vary when moving from areas more visited by patients to regions characterized by a higher presence of controls. This is the goal of the next section.

Exploiting the dense feature embedding. Once the learning phase is carried out, one can embed onto the counting grid any other phenotype y^t of the samples, discrete or continuous, like age, sex, label, etc. The resulting embedding may serve for diagnosis support and are obtained as follows:

$$\gamma_{\mathbf{i}} = \frac{\sum_t \sum_{\mathbf{k}|\mathbf{i}\in\mathbf{W}_\mathbf{k}} q^t_{\mathbf{k}} \cdot y^t}{\sum_t \sum_{\mathbf{k}|\mathbf{i}\in\mathbf{W}_\mathbf{k}} q^t_{\mathbf{k}}} \tag{4}$$

[14] http://www.csie.ntu.edu.tw/cjlin/papers/guide
[15] http://www.alessandroperina.com/

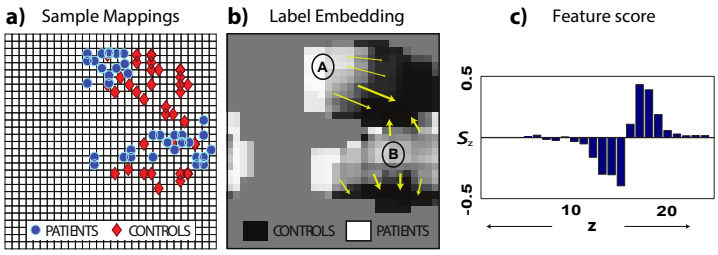

Fig. 3. a) Mapping position of the samples and the relative label embedding **b)**. **c)** Variation in the features moving from patients to controls.

In practice, for each sample, the value of y^t is copied in the window identified by $q^t_{\mathbf{i}}$, and the average across the samples is taken.
Figure 3.b the label embedding is shown, i.e., $y^t \in \{1 = \text{control}\ , -1 = \text{patient}\ \}$. Noticeably, even if the labels *are not* used during the learning phase of the generative model, the embedding clearly exhibits some structures, separating patients and controls. This suggests that CG are actually suitable to describe the latent structures generating the data. Figure 3.b clearly shows how patients and controls are clustered. In particular, the patients cluster in two groups by revealing that patient group is characterized by a larger variation probably due to different patterns. As further analysis we show how the cortical thickness (e.g., our feature) varies between patients and controls. Firstly, we compute the gradient of the label embedding, $\nabla\gamma_{\mathbf{i}}$, which returns information about *where* and *how* the classes are separated on the embedding. In Figure 3.b we draw some arrows to highlight the regions with highest gradient in moving from patients to controls, i.e., the borders between the classes. Secondly, we compute how much the relevance of each thickness value z varies along such borders. To capture this idea mathematically we compute the directional derivatives of $\pi_{z,\mathbf{i}}$ in the direction of the gradient of the label embedding $\nabla\gamma_{\mathbf{i}}$ and sum over all the locations $\mathbf{i}$ in the grid we can compute a feature score for every z:

$$S_z = \sum_{\mathbf{i}} S_{z,\mathbf{i}} = \sum_{\mathbf{i}} \langle \nabla\gamma_{\mathbf{i}}, \nabla\pi_{z,\mathbf{i}} \rangle \tag{5}$$

In practice, we expect to observe a high value S_z if the locations of strong variations of the feature relevance correspond to the areas with a transition between patients and controls. In other words there is a simultaneous variation of both features relevance and classes that makes the feature strictly related to the disease. The value of S_z is shown in Figure3c and it can be interpreted as the variation of the thickness histogram between patients and controls. It is clearly highlighted that features associated to low thickness values decrease moving from patients to controls suggesting that the pathology is characterized by thickness reduction on patients.

5 Conclusions

In this paper we propose a novel approach to analyze brains by exploiting the Counting Grid model. We highlight how CG can naturally reveal relations between subjects by showing that patiens and controls form well separated clusters. In particular, patients are lying on larger areas of the CG by evidencing a strong intra-class variability in Schizophrenic subjects. Moreover, CG can be used to evaluate the importance of involved features in order to better understand the disease. For instance, as expected, in our study it is clearly evidenced a reduction of cortical thickness in patients. In future work we plan to study the heterogeneous aspects of Schizophrenia by investigating the relations between the involved features and the detected clusters of patients on the CG.

References

1. Batmanghelich, N., Taskar, B., Davatzikos, C.: Generative-discriminative basis learning for medical imaging. IEEE Trans. Med. Imaging 31(1), 51–69 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (Mar 2003)
3. Castellani, U., Perina, A., Murino, V., Bellani, M., Rambaldelli, G., Tansella, M., Brambilla, P.: Brain morphometry by probabilistic latent semantic analysis. In: Jiang, T., Navab, N., Pluim, J., Viergever, M. (eds.) Medical Image Computing and Computer-Assisted Intervention (MICCAI). Lecture Notes in Computer Science, vol. 6362, pp. 177–184. Springer Berlin Heidelberg (2010)
4. Chapelle, O., Haner, P., Vapnik, V.: Support vector machines for histogram-based image classifcation. IEEE Transaction on Neural Network 10(5), 1055–1064 (1999)
5. Desikan, R.S., Sgonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L.e.a.: An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31(3), 968–80 (2006)
6. Goldman, A.: Widespread reductions of cortical thickness in schizophrenia and spectrum disorders and evidence of heritability. Arch Gen Psychiatry 66(5), 467–477 (2009)
7. Hirschhorn, J., Daly, M.J.: Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genetic 6(2), 95–108 (2005)
8. Jojic, N., Perina, A.: Multidimensional counting grids: Inferring word order from disordered bags of words. In: Conference on Uncertainty in Artificial Intelligence (2011)
9. Lemma, S., Blankertza, B., Dickhausa, T., Mller, K.R.: Introduction to machine learning for brain imaging. NeuroImage 56(2), 387399 (2011)
10. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: IEEE Computer Vision and Pattern Recognition (CVPR). vol. 1, pp. 246–253 (2006)
11. Perina, A., Cristani, M., Castellani, U., Murino, V., Jojic., N.: Free energy score spaces: using generative information in discriminative classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(7), 1249–1262 (2012)
12. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)