Disentangling Disease Heterogeneity with Max-Margin Multiple Hyperplane Classifier

Erdem Varol, Aristeidis Sotiras, and Christos Davatzikos

Center for Biomedical Image Computing and
Analytics University of Pennsylvania Philadelphia, PA 19104, USA
{erdem.varol,aristeidis.sotiras,christos.davatzikos}@uphs.upenn.edu

Abstract. There is ample evidence for the heterogeneous nature of diseases. For example, Alzheimer's Disease, Schizophrenia and Autism Spectrum Disorder are typical disease examples that are characterized by high clinical heterogeneity, and likely by heterogeneity in the underlying brain phenotypes. Parsing this heterogeneity as captured by neuroimaging studies is important both for better understanding of disease mechanisms, and for building subtype-specific classifiers. However, few existing methodologies tackle this problem in a principled machine learning framework. In this work, we developed a novel non-linear learning algorithm for integrated binary classification and subpopulation clustering. Non-linearity is introduced through the use of multiple linear hyperplanes that form a convex polytope that separates healthy controls from pathologic samples. Disease heterogeneity is disentangled by implicitly clustering pathologic samples through their association to single linear sub-classifiers. We show results of the proposed approach from an imaging study of Alzheimer's Disease, which highlight the potential of the proposed approach to map disease heterogeneity in neuroimaging studies.

1 Introduction

Brain disorders often assume a heterogeneous clinical presentation: Autism Spectrum Disorder (ASD) encompasses neurodevelopmental disorders characterized by deficits in social communication and repetitive behaviors [5]; Schizophrenia can be subdivided into distinct groups by separating its symptomatology to discrete symptom domains [2]; Alzheimer's Disease (AD) can be separated into three subtypes on the basis of the distribution of neurofibrillary tangles [8]; and Mild Cognitive Impairment (MCI) may be further classified based on the type of specific cognitive impairment [11].

Disentangling disease heterogeneity may greatly contribute to our understanding and lead to more accurate diagnosis, prognosis and targeted treatment. However, most commonly used neuroimaging analysis approaches assume a single unifying pathophysiological process and perform a monistic analysis to identify it. These approaches aim to either identify voxels that characterize group differences through mass-univariate statistical techniques [1], or reveal patterns of

N. Navab et al. (Eds.): MICCAI 2015, Part I, LNCS 9349, pp. 702–709, 2015.
DOI: 10.1007/978-3-319-24553-9_86

variability through high-dimensional pattern classification analysis, towards categorizing population with respect to the underlying pathology [10]. Thus, the heterogeneity of the disease is completely ignored.

Contrarily, few research efforts have been focused on revealing the inherent disease heterogeneity. These methods can mainly be classified into two groups. The first class assumes an a priori subdivision of the diseased samples into coherent groups, based on independent criteria, and opts to identify group-level anatomical differences using univariate statistical methods [7, 12]. Thus, multivariate effects are ignored, while the a priori definition of disease subtypes is either difficult to obtain (*e.g.*, from autopsy near the date of imaging), or noisy and non-specific (*e.g.*, cognitive or clinical evaluations). The second class focuses on the diseased population and maps it to distinct anatomical subtypes by applying multivariate clustering driven by considering all image elements [11, 9]. Thus, disease heterogeneity may be confounded due to considering the whole brain anatomy instead of the disease-specific information, and disentangling it may not be possible.

In order to tackle the aforementioned limitations, it is necessary to develop a principled machine learning approach that will allow for the simultaneous identification of a class of images with pathological changes and its separation to coherent subgroups. To the best of our knowledge, only one approach has been proposed in this direction [3], which makes strong assumptions regarding the number of the existing disease subgroups (that there are exactly 2 subgroups). Here, we propose a novel non-linear machine learning algorithm for integrated binary classification and subpopulation clustering. The proposed approach is motivated by recent machine learning approaches that derive non-linear classifiers through the use of multiple-hyperplanes [4, 6]. Multiple max-margin classifiers are combined to form a convex polytope that separates healthy controls from pathological samples, while heterogeneity is disentangled by implicitly clustering pathologic samples through their association to single linear sub-classifiers. By varying the number of estimated hyperplanes (faces of polytope), it is possible to capture multiple modes of heterogeneity.

2 Method

In high dimensional spaces, linear Support Vector Machines (SVMs) are able to separate by a large margin two classes. However, in the case that the one class is drawn from a multimodal distribution (as in the presence of heterogeneity), the classes may be still linearly separated, albeit with a smaller margin. This may be remedied by the use of a non-linear classifier, allowing for larger margins and thus, better generalization. However, while kernel methods, such as Gaussian kernel SVM, provide non-linearity, they lack interpretability when aiming to characterize heterogeneity. Instead, we introduce non-linearity by means of using multiple linear classifiers that form a locally linear hyperplane whose linear segments separate the clusters of negative samples from the positive class (Fig 1). In this way, subjects are explicitly clustered, giving rise to interpretable directions of variability that may be useful in discovering heterogeneity.

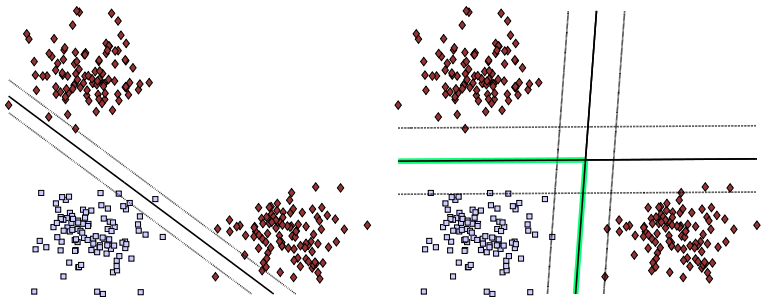

Fig. 1. Heterogeneity due to the presence of two clusters. **Left:** Result obtained by linear SVM (small margin). **Right:** Result obtained by separately classifying each cluster (large margin). Solid lines correspond to the classifier, dashed lines indicate margin, while highlighted linear segments define the separating convex polytope.

Suppose that our dataset consists of n binary labeled d-dimensional data points ($\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n, \mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$). Without loss of generality, we assign the negative class to be pathologic whose heterogeneity we seek to reveal.[1] Our aim is twofold. First, we aim to estimate k hyperplanes that form a convex polytope that separates the two classes with a large margin. Second, we aim to assign each pathologic sample to the hyperplane that best separates it from the normal controls. Towards fulfilling these aims, we introduce the proposed approach by extending standard linear maximum margin classifiers.

2.1 Margin for Multiple Hyperplanes — Polytope

The hypothesis class of standard linear maximum margin classifiers comprises the set of all linear classifiers $\mathbf{w}$ that separate the two classes by a halfspace. Here, we extend the hypothesis class by considering the set of sets of K hyperplanes, generalizing the geometry of the classifier to that of a convex polytope. Due to the interior/exterior asymmetry of the polytope, it is necessary to confine one class to its interior, while restricting the other class to its exterior. Without loss of generality, we confine the positive class to the interior of the polytope. Thus, the search space $\mathcal{F}_K$ is defined as:

$$\mathcal{F}_K \triangleq \{\{\mathbf{w}_j, b_j\}_{j=1}^K \mid \text{if } y_i = +1\ \forall j, \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1,\ \text{if } y_i = -1,\ \exists j : \mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1\}$$

In other words, $\mathcal{F}_K$ comprises all sets of k classifiers such that all classifiers correctly classify all members of the positive class, while for every member of the negative class, there is at least one classifier that correctly classifies it. The latter gives rise to an assignment problem, which can also be seen as a clustering task. Thus, if $\mathbf{S}^- = [s_{i,j}] \in \{0, 1\}^{n^- \times K}$ denotes the binary matrix that describes the assignment of the negative class samples to the jth face of the polytope, then the search space becomes:

$$\mathcal{F}_K(\mathbf{S}^-) \triangleq$$
$$\{\{\mathbf{w}_j, b_j\}_{j=1}^K \mid \text{if } y_i = +1\ \forall j, \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1,\ \text{if } y_i = -1, s_{i,j} = 1 : \mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1\}$$

[1] Label reversal would enable us to seek heterogeneity in the control samples.

Given the assignment $\mathbf{S}^-$, there are K margins; each one corresponding to one face of the polytope. Analogous to the SVM formulation, the margin for the jth face of the polytope is $\frac{2}{\|\mathbf{w}_j\|_2}$. However, due to the piecewise nature of the convex polytope, there are multiple notions of margin for the surface of the polytope. In this work, we aim to maximize the average margin across all the faces of the polytope: $\bar{m} = \frac{1}{K}\sum_{j=1}^{K} \frac{2}{\|\mathbf{w}_j\|_2}$ in order to keep the problem tractable. Thus, for a given dataset $\mathcal{D}$ and assignment $\mathbf{S}^-$ for the negative class, the objective of maximizing polytope margin becomes:

$$\begin{aligned} \underset{\{\mathbf{w}_j, b_j\}_{j=1}^K}{\text{maximize}} \; & \frac{1}{K}\sum_{j=1}^{K} \frac{2}{\|\mathbf{w}_j\|_2} \\ \text{subject to } & \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 \qquad \text{if } \; y_i = +1 \\ & \mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 \qquad \text{if } \; y_i = -1, s_{i,j} = 1 \end{aligned} \tag{1}$$

Note that given the assignments, the objective and the constraints are separable into K independent subproblems. Each subproblem is analogous to the SVM formulation after adding the slack terms $\xi_{i,j}$, or:

$$\begin{aligned} \underset{\mathbf{w}_j, b_j, \xi}{\text{minimize}} \; & \frac{\|\mathbf{w}_j\|_2^2}{2} + C\sum_{i=1}^{n} \xi_{i,j} \\ \text{subject to } & \mathbf{w}_j^T \mathbf{x}_i + b_j \geq 1 - \xi_{i,j} \qquad \text{if } y_i = +1 \\ & \mathbf{w}_j^T \mathbf{x}_i + b_j \leq -1 + \xi_{i,j} \qquad \text{if } y_i = -1, s_{i,j} = 1 \\ & \xi_{i,j} \geq 0 \end{aligned}$$

where C is a penalty parameter on the training error. If we now use the definition of the slack terms as $\xi_{i,j} = \max\{0, 1 - y_i(\mathbf{w}_j^T \mathbf{x}_i + b_j)\}$, and consider all hyperplanes $(\{\mathbf{w}_j, b_j\}_{j=1}^K)$ at the same time, we get the objective function:

$$\begin{aligned} \underset{\{\mathbf{w}_j, b_j\}_{j=1}^K}{\text{minimize}} \; & \sum_{j=1}^{K} \frac{\|\mathbf{w}_j\|_2^2}{2K} \\ + C \sum_{\substack{i | y_i = +1 \\ j}} & \frac{1}{K} \max\{0, 1 - \mathbf{w}_j^T \mathbf{x}_i - b_j\} + C \sum_{\substack{i | y_i = -1 \\ j}} s_{i,j} \max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} \end{aligned} \tag{2}$$

So far, we have assumed that the assignment matrix $\mathbf{S}^-$ is known. However, this is not the case in practice and $\mathbf{S}^-$ has to be estimated too. We relax the 0-1 assignment to a soft assignment; $s_{i,j}$ is allowed to be in the interval $[0, 1]$, satisfying the constraint that $\sum_{j=1}^{K} s_{i,j} = 1$ for all i. Given this relaxation the problem becomes convex with respect to the blocks $\{\mathbf{W}, \mathbf{b}\}$ and $\{\mathbf{S}^-\}$.

For $\mathbf{S}^-$ fixed, the solution to $\mathbf{W}$ can be obtained using K calls to a modified version of LIBSVM[2] that allows for adaptive sample weightings, where the weights are given by

$$c_{i,j} = \begin{cases} C s_{i,j} & \text{if } y_i = -1 \\ \frac{C}{K} & \text{if } y_i = +1 \end{cases} \tag{3}$$

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/weights/

Algorithm 1. — Max-Margin Multiple Hyperplane

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{y} \in \{-1, +1\}^n$ (training signals), C, K (parameters)
Output: $\mathbf{W} \in \mathbb{R}^{d \times K}$ (Classifier); $\mathbf{S}^- \in \mathbb{R}^{n^- \times K}$ (Soft Clustering Assignment)
Initialization: Set rows of $\mathbf{S}^-$ with probability $\mathrm{Dir}(\mathbf{1}_K)$ (Dirichlet Assignment)
Loop: Repeat until convergence (or a fixed number of iterations)
• Fix $\mathbf{S}^-$ — Solve for $\mathbf{W}$ by LIBSVM[1] (sample weights set by equation (Eq. 3)
• Fix $\mathbf{W}$ — Solve for $\mathbf{S}^-$ by equation (Eq. 4)

For $\mathbf{W}$ fixed, the problem of estimating $\mathbf{S}^-$ is a linear program (LP) of assignment which has infinite solutions when the loss function $\max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\}$ is equal to 0 for multiple classifiers j and for the same sample i. In this case, we choose the solution that is proportional to the margin:

$$s_{i,j} = \begin{cases} 0 & \text{if} \quad \max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} > 0 \\ \frac{1 + \mathbf{w}_j^T \mathbf{x}_i + b_j}{\sum_j (1 + \mathbf{w}_j^T \mathbf{x}_i + b_j) \mathbf{1}(\max\{0, 1 + \mathbf{w}_j^T \mathbf{x}_i + b_j\} \leq 0)} & \text{otherwise} \end{cases} \tag{4}$$

where $\mathbf{1}(\cdot)$ is the indicator function. The previous steps are summarized in Algorithm (1). Note that we don't explicitly give solution for the bias terms b_j. This is because all data points $\mathbf{x}_i$ can include a constant unitary component that corresponds to the bias term. In this case, last element of $\mathbf{w}_j$ contains the solution for the bias term.

Once the polytope classifier $[\mathbf{W}, \mathbf{b}]$ is trained, predicting the class y^* of a new instance $\mathbf{x}^*$ is straightforward:

$$y^* = \mathrm{sign}(\min_j \mathbf{w}_j^T \mathbf{x}^* + b_j) \tag{5}$$

In other words, if $\mathbf{x}^*$ is in the interior of the polytope defined by $\mathbf{W}, \mathbf{b}$, then $\mathbf{w}_j^T \mathbf{x}^* + b_j > 0$ for all faces of the polytope resulting in the prediction $y^* = +1$. Otherwise, if $\mathbf{x}^*$ is in the exterior of the polytope defined by $\mathbf{W}, \mathbf{b}$, then $\mathbf{w}_j^T \mathbf{x}^* + b_j < 0$ for at least one face of the polytope, resulting in the prediction $y^* = -1$. Analogously, the prediction score is simply $\min_j \mathbf{w}_j^T \mathbf{x}^* + b_j$. Also, the clustering assignment $s_{*,j}$ is done in the same manner as Eq. (4).

3 Experimental Validation

We validated our approach on both low dimensional synthetic data and clinical data. For all of our experiments, the features were z-normalized and the default parameter setting ($C = 1$) was used for the LIBSVM subroutine of the proposed method. Thus, the only free parameter to be tuned was the number of polytope faces K (note that $K = 1$ corresponds to linear SVM). Increasing K has two effects on the performance of the algorithm: 1) the model complexity increases; and 2) the number of subject clusters increases. To assess the performance of the method, it is important to check for overfitting and clustering stability. These two effects were examined by examining the out-of-sample classification accuracy and the adjusted Rand clustering overlap index.

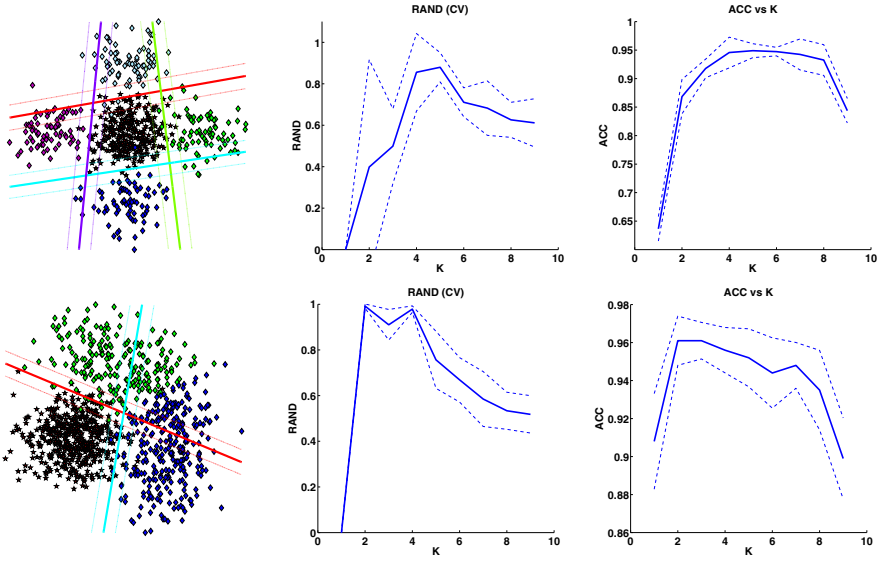

Fig. 2. Sythetic data experiments: **Left:** Data, optimal polytope classifier and the cluster assignments, **Middle:** Cross-validated adjusted Rand index across folds. **Right:** Cross-validated classification accuracy. **Note:** $K = 1$ corresponds to linear SVM.

The first set of experiments consisted of classification of 2 dimensional synthetic data with known ground truth about the underlying clusters. We simulated two cases: 1) a single (+) group with 4 disjoint (−) groups, (Fig. 2 (Top)); and 2) a single (+) group with heterogeneous (−) group distributed along a semicircle (Fig. 2 (Bottom)). The out-of-sample accuracy was computed using 10-fold cross-validation. The cross-validated Rand index was computed using the cluster assignments of the common subjects between folds and taking the average Rand index across all folds. Since only the (−) group was clustered, the (+) samples were ignored in the Rand index calculation.

The synthetic data experiments revealed two key insights. First, it demonstrated that our method is able to separate the two classes, while meaningfully clustering the negative group when K equals the number of underlying subgroups (Fig. 2 (Left)). Second, its performance - as quantified by the Rand index (Fig. 2 (Middle)) and the classification accuracy (Fig. 2 (Right)) - varied smoothly for increasing K, reaching a maximum for the ideal number and thus, allowing us to perform model selection. We note that our method is able to capture heterogeneity in the presence of distinct patterns of variability (case #1), while it provides reasonable estimates in more complex cases (case #2).

Having established a model selection strategy, we evaluate our method using data from the Alzheimer's disease neuroimaging initiative (ADNI[3]). The ADNI dataset comprises the baseline scans of 190 controls (CN), and 133 AD patients. The images were 1.5 Tesla T1-weighted MRI volumetric scans that were processed using an in-house pipeline of 1) bias correction, 2) skull stripping, 3)

[3] `http://adni.loni.usc.edu/`

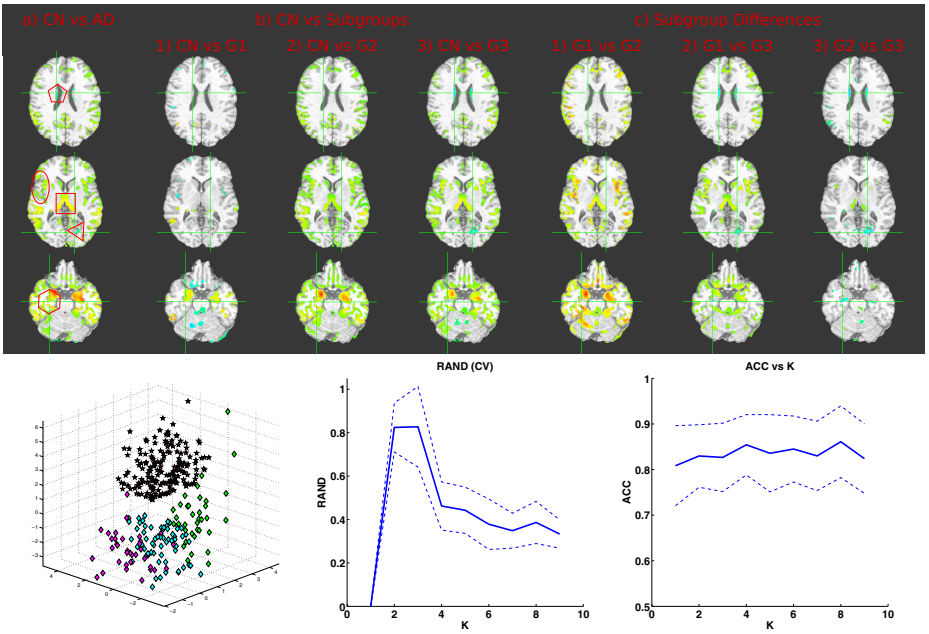

Fig. 3. Top: a) Gray Matter Group Differences (p<0.05) between CN and AD. Shape glossary: pentagon=caudate, ellipse=insula, square=thalamus, triangle=left cuneus, hexagon=right hippocampus **b)** Group differences (p<0.05) between CN vs. 3 subtypes of AD **c)**. Group differences (p<0.05) between 3 different AD subtypes. **Color-map:** Right group compared to left group [Red: loses volume] / [Cyan: gains volume] — **Bottom: Left:** Imaging features projected along the 3 faces of the polytope classifier, CN, AD group 1, AD group 2, AD group 3. **Middle:** Cross-validated adjusted Rand index across folds. **Right:** Cross-validated classification acccuracy.

tissue segmentation and 4) deformable registration that resulted in 151 cortical and sub-cortical anatomical volumes for each subject.

For the ADNI dataset, setting $2 \leq K \leq 9$ resulted in comparable out-of-sample accuracies with statistically insignificant differences (Fig 3 (Bottom right)). The fact that the cross-validation accuracy at $K = 1$ is > 0.80 suggests that the data is already separable and introducing non-linearity will only marginally improve separability. However, the clustering reproducibility analysis revealed that setting $K = 2, 3$ results in stable clusterings (Fig 3 (Bottom middle)) despite the shuffling of samples across folds. This suggests that there may be distinct patterns of variability between controls and these K AD subgroups. The drop of the Rand index for $K > 4$ further strengthens this observation.

In order to investigate the previous observation, we fixed $K = 3$ and found AD subgroups that differed in age composition. Subgroup 1 (G1) comprised younger patients, while subgroups 2 and 3 (G2 and G3) comprised older patients. Then, we performed Voxel-Based Morphometry (VBM) analysis between the CN group and the whole AD population (Fig. 3a); between the CN group and each AD subgroup (Fig. 3b) and between pairs of AD subgroups (Fig. 3c). The VBM analysis allows us to study the structural differences between the respective groups.

Typical CN vs AD VBM analysis reveals a common AD pattern with reduced gray matter in cortical and subcortical regions. However, when examining the AD subgroups separately, we observe a heterogeneous behavior: G1 does not exhibit thalamus or insula atrophy as it is observed for the other two groups; G2 differs from the typical AD pattern in caudate, cuneus and hippocampal regions; and G3 shows a typical AD profile. The differences between the AD subgroups are highlighted by the VBM results shown in Fig. 3c. To further illustrate the heterogeneity of the disease patterns, we projected the imaging features for both the CN and AD subgroups along the $K = 3$ polytope faces. The result is shown in Fig 3 (Bottom left) and emphasizes the segregation of the three subgroups.

4 Conclusion

In this paper, we proposed a novel machine learning method for simultaneous binary classification and subgroup clustering. The proposed method mapped disease heterogeneity in a data-driven way, revealing distinct imaging subtypes in a robust and generalizable fashion.

References

[1] Ashburner, J., Friston, K.J.: Voxel-Based Morphometry–The Methods. NeuroImage 11(6), 805–821 (2000)
[2] Buchanan, R.W., Carpenter, W.T.: Domains of psychopathology: an approach to the reduction of heterogeneity in schizophrenia. The Journal of Nervous and Mental Disease 182(4), 193–204 (1994)
[3] Filipovych, R., Resnick, S.M., Davatzikos, C.: Jointmmcc: Joint maximum-margin classification and clustering of imaging data. IEEE Transactions on Medical Imaging 31(5), 1124–1140 (2012)
[4] Fu, Z., Robles-Kelly, A., Zhou, J.: Mixing linear svms for nonlinear classification. IEEE Transactions on Neural Networks 21(12), 1963–1975 (2010)
[5] Geschwind, D.H., Levitt, P.: Autism spectrum disorders: developmental disconnection syndromes. Current Opinion in Neurobiology 17(1), 103–111 (2007)
[6] Gu, Q., Han, J.: Clustered support vector machines. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, pp. 307–315 (2013)
[7] Koutsouleris, N., et al.: Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study. NeuroImage 39(4), 1600–1612 (2008)
[8] Murray, M.E., et al.: Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. The Lancet. Neurology 10(9), 785–796 (2011)
[9] Noh, Y., et al.: Anatomical heterogeneity of alzheimer disease based on cortical thickness on mris. Neurology 83(21), 1936–1944 (2014)
[10] Vemuri, P., et al.: Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. NeuroImage 39(3), 1186–1197 (2008)
[11] Whitwell, J.L., et al.: Patterns of atrophy differ among specific subtypes of mild cognitive impairment. Archives of Neurology 64(8), 1130–1138 (2007)
[12] Whitwell, J.L., et al.: Neuroimaging correlates of pathologically defined subtypes of Alzheimer's disease: a case-control study. The Lancet. Neurology 11(10), 868–877 (2012)