Published in final edited form as:

IEEE Trans Med Imaging. 2012 May ; 31(5): 1124–1140. doi:10.1109/TMI.2012.2186977.

JointMMCC: Joint Maximum-Margin Classification and Clustering of Imaging Data

Roman Filipovych

Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. Roman.Filipovych@uphs.upenn.edu

Susan M. Resnick

Laboratory of Behavioral Neuroscience, National Institute on Aging, Biomedical Research Center/ 04B317, 251 Bayview Blvd., Baltimore, MD 21224

Christos Davatzikos

Section of Biomedical Image Analysis, Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. Christos.Davatzikos@uphs.upenn.edu

Abstract

A number of conditions are characterized by pathologies that form continuous or nearly-continuous spectra spanning from the absence of pathology to very pronounced pathological changes (*e.g.*, normal aging, Mild Cognitive Impairment, Alzheimer's). Moreover, diseases are often highly heterogeneous with a number of diagnostic subcategories or subconditions lying within the spectra (*e.g.*, Autism Spectrum Disorder, schizophrenia). Discovering coherent subpopulations of subjects within the spectrum of pathological changes may further our understanding of diseases, and potentially identify subconditions that require alternative or modified treatment options. In this paper, we propose an approach that aims at identifying coherent subpopulations with respect to the underlying MRI in the scenario where the condition is heterogeneous and pathological changes form a continuous spectrum. We describe a Joint Maximum-Margin Classification and Clustering (JointMMCC) approach that jointly detects the pathologic population via semi-supervised classification, as well as disentangles heterogeneity of the pathological cohort by solving a clustering subproblem. We propose an efficient solution to the non-convex optimization problem associated with JointMMCC. We apply our proposed approach to an MRI study of aging, and identify coherent subpopulations (*i.e.*, clusters) of cognitively less stable adults.

Index Terms

Semi-supervised classification; clustering; MRI; aging

I. Introduction

Categorizing populations with respect to underlying pathologies manifested by structural or functional characteristics and captured by various imaging methods is of importance for diagnostic and prognostic purposes. In this respect, high-dimensional pattern classification has gained significant attention, and has been found to be a promising technique for capturing complex spatial patterns of pathological changes [1]–[7]. The major advantage of these methods is in their ability to provide highly specific and sensitive markers of diseases on an individual basis.

Typical classification methods assume that the nosological boundaries between the conditions are distinct and well-defined, which may not always be the case. In various studies, populations of patients are highly heterogeneous, and categorization of subconditions within many diseases is yet to be established. Additionally, a number of diseases are characterized by pathologies that form continuous or nearly-continuous spectra spanning from complete absence of pathology to very pronounced pathological changes. Autism Spectrum Disorder (ASD), schizophrenia, Mild Cognitive Impairment (MCI) are but a few examples of conditions where pathologies are highly heterogeneous, as well as form continuous spectra:

- *Autism Spectrum Disorder (ASD)*: ASD is a Pervasive Developmental Disorder (PDD) characterized by deficits in language, presence of stereotypic or repetitive behaviors, and social impairments [8], [9]. While the boundaries between ASD, its comorbidities, and healthy state are blurred, several diagnostic subcategories within ASD were defined: Autism, Aspergers disorder, and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS). At the same time, it has been often argued that the Asperger disorder criteria do not work in the clinic [10], [11]. As the result, there has been an intense debate as to the distinctiveness of Asperger disorder from other subgroups within the autism spectrum [12]. Appropriate computational tools would allow to assess whether Aspergers disorder forms a distinct subgroup within the autism spectrum with respect to, for example, imaging phenotypes. At the same time, it would be preferable to analyze heterogeneity of ASD in a subpopulation where the difference between the underlying subcategories are maximally highlighted. It is therefore important to have the ability to automatically identify subpopulations where the distinct subgroups are likely to be discovered.
- *Schizophrenia*: There is an intense debate in the schizophrenia research community between the proponents of a categorical disease classification system versus the supporters of a dimensional/syndromal perspective because of (1) the lack of evidence for decisive biological boundaries and (2) the considerable similarities in the clinical presentations and courses of schizophrenic and affective psychoses [13]–[17]. The comparative genetic studies suggest that a considerable degree of neurobiological heterogeneity within the psychiatric diseases blurs the established diagnostic divisions. This hypothesis has been supported by findings of (1) distinct neuroanatomical (endo)phenotypes underlying different psychopathological symptom dimensions in schizophrenia [18], (2) a significant sexual dimorphism in the structural abnormalities associated with the disorder [19], and (3) a progression of structural brain abnormalities during the course of the disease [20], [21]. At the same time, the early prodromal state of the disease largely overlaps with depressive syndromes [22] and nonspecific psychopathological phenomena found in the general population [23]. As the result, in order to analyze heterogeneity of schizophrenia, one needs to identify a subpopulation that is not confounded by irrelevant conditions that have similar symptomatic profiles.
- *Aging*: It was shown in the studies of aging that cognitive performance of some older adults may decline gradually at different rates, while some may remain stable for a long period of time [24]. Moreover, some may develop MCI, or even proceed to developing Alzheimer's disease (AD). Additionally, as normal aging increases cognitive heterogeneity [25], [26], there is a need to identify aging populations where the heterogeneity is evident and can be discovered using computational approaches.

In summary, the above diseases and processes, as well as many others, have the following common properties: (1) There is a degree of severity of associated pathology that may be

reflected along several different dimensions (*e.g.*, memory, language, etc.); (2) Regardless of the fact that the boundaries between the healthy state and the diagnosis are blurred, at the extreme end of the disease spectrum there is significant heterogeneity that may manifest itself in the form of different diagnostic or prognostic subcategories.

In the case imaging profiles are driven by the heterogeneity of the condition, the task of categorizing abnormal imaging profiles becomes two-fold: (1) separate individuals that have relatively normal imaging profiles from those that exhibit pathologies (*i.e.*, individuals with pronounced pathological changes); and (2) disentangle heterogeneity in the pathological (*i.e.*, abnormal) subset. With the identified distinct imaging profiles at hand, one can proceed to analyze them with respect to the diagnostic differences. Figure 1 shows a schematic diagram of a heterogeneous aging population that does not classify well under any specific category, but rather evolves continuously from cognitively normal to cognitively declined along two different pathways (*e.g.*, from normal to Mild Cognitive Impairment (MCI) that remains stable, and from normal to MCI that then evolves to Alzheimer's disease). In the context of aging, cognitively declined subpopulations are characterized by significantly worsened performance over time as estimated during cognitive evaluations. In contrast, the cognitive performance of cognitively stable individuals does not worsen significantly over time.

Following the state of the art in pattern recognition, a standard pipeline would be to learn a classifier that separates a subset of images with pronounced pathological changes, and then to apply a clustering algorithm to discover novel categories within it. However, there are several difficulties in applying standard techniques to categorize abnormal subpopulations. If the ultimate task is to identify coherent subpopulations in the population of diseased subjects, then the pathological population has to be selected such, that it is possible to distinguish underlying sub-pathologies. Classifying individual into normal or abnormal groups with respect to their MRI using existing classification algorithms does not guarantee that the heterogeneity in the abnormal population can be disentangled. Additionally, commonly used brain image classification methods usually assume that categorical labels for all training samples are available and reliable [1], [27]–[31]. Unfortunately, this is not often the case for those individuals whose imaging profiles are not severely affected, and at the same time show some level of pathology that prevents from unanimously categorizing them as normal. As the result, one can often be confident in only those labels that correspond to the extreme cases (*i.e.*, extreme normal or extreme abnormal).

In this paper, we propose the *Joint Maximum-Margin Classification and Clustering (JointMMCC)* method that simultaneously separates a class of images with pronounced pathological changes and finds clusters in the obtained abnormal class. Guided by labels of only the most extreme samples, JointMMCC learns a maximum-margin classifier in such a way, that the obtained class of abnormal images can be clustered by a hyperplane with a large margin (*i.e.*, Figure 1). In order to solve the JointMMCC problem, we describe a cutting plane algorithm that transforms the original optimization problem into a nested set of relaxed subproblems. Each of these relaxed subproblems are in turn solved using the *constrained concave-convex procedure* [32]. We use our proposed method to cluster magnetic resonance (MR) brain images of cognitively declined aging populations.

The remainder of this paper is organized as follows. In Section II, we provide the background for the proposed approach, and describe our method in Section III. In Section IV, we provide the details of the data used in our study. We report our results in Section V, and conclude the paper in Section VI.

II. Background

Categorization of populations with respect to the types of underlying pathologies has been extensively studied earlier, and usually aims at assigning individuals to known diagnostic categories (*i.e.*, normal, Mild Cognitive Impairment, Alzheimer's disease, etc.). In this respect, pattern classification methods have begun to provide tests of high sensitivity and specificity on an individual patient basis, in addition to characterizing group differences. These methods were shown to work particularly well in the task of classifying patient populations from normal cohorts in various clinical studies (*e.g.*, Alzheimer's [7], [33]–[35], autism [36], schizophrenia [37], etc.). Here, Support Vector Machines (SVM) [38] offer a solid theoretical framework for designing robust classifiers, and have been widely used in neuroimaging studies. SVM-like methods own their popularity in neuroimaging applications partly due to the remarkable robustness they show in the problems where the number of dimensions is significantly higher than the number of training instances. In fact, SVM have been shown to work well even for the problems where features from all voxels in the brain are used to train a classifier using images of less than a hundred subjects [39].

At the same time, the potential of more exploratory extensions of SVM, like semi-supervised SVM and maximum-margin clustering, in the context of categorizing brain image sets was largely neglected.

Semi-supervised SVM, originally introduced as *Transductive SVM (TSVM)* [40], build upon the theory of SVM and consider partially labeled point sets. Given a set of points $\chi = \{\mathbf{x}_1, \dots, \mathbf{x}_{l+1}, \dots, \mathbf{x}_n\}$ where $\mathbf{x} \in \mathbb{R}^m$, the first l points in χ are labeled as $y_i \in \{-1, +1\}$, and the labels $y_j \in \{-1, +1\}$ of the remaining $u = n - l$ points are unknown, task of finding a separating function $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$, $\mathbf{w} \in \mathbb{R}^m$, $b \in \mathbb{R}$, within the framework of linear semi-supervised SVM could be formulated as the following optimization problem:

$$\begin{aligned} \min_{y_{l+1}, \dots, y_n} \quad & \min_{\mathbf{w}, b, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \beta_l \sum_{i=1}^l \xi_i + \beta_u \sum_{j=l+1}^n \xi_j \\ \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \forall i=1, \dots, l \\ & y_j (\mathbf{w}^T \mathbf{x}_j + b) \geq 1 - \xi_j, \forall j=l+1, \dots, n \\ & \xi_i \geq 0, \xi_j \geq 0, \forall i=1, \dots, l, \forall j=l+1, \dots, n \end{aligned} \quad (1)$$

where the constants β_l and β_u reflect prior confidence in labels ($y_1, \dots, y_l$) and in the cluster assumption, respectively.

Maximum-Margin Clustering (MMC) further extends the theory of SVM to the unsupervised scenario, where instead of finding a large margin classifier given labels of all data points as in SVM, the target is to find the labeling $(z_1, \dots, z_n) \in \{-1, +1\}^n$ that would result in a large margin classifier [41]. In other words, if one were to run SVM with the labeling obtained from MMC, the obtained margin would be maximal among all possible labelings. For $\widehat{\mathbf{w}} \in \mathbb{R}^m$ and $\widehat{b} \in \mathbb{R}$, the problem of MMC can be formulated as follows:

$$\begin{aligned} \min_{z_1, \dots, z_n} \quad & \min_{\widehat{\mathbf{w}}, \widehat{b}, \widehat{\xi}} \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \widehat{\beta} \sum_{k=1}^n \widehat{\xi}_k \\ \text{s.t.} \quad & \widehat{z}_k (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \geq 1 - \widehat{\xi}_k, \forall k=1, \dots, n \\ & \widehat{\xi}_k \geq 0, \forall k=1, \dots, n \end{aligned} \quad (2)$$

Both semi-supervised SVM and MMC are originally formulated as non-convex integer programming problems and are difficult to solve. Building upon the ideas of cutting plane algorithm [42], a fast training method was proposed for linear SVM in [43]. By using the

concepts of cutting plane algorithm and by utilizing *constrained concave-convex procedure (CCCP)*, computationally efficient solutions to TSVM and MMC were proposed by [44] and [45], respectively. In both cases, the algorithms were shown to scale linearly with the number of data samples, and linearly with the dimensionality of the problem.

The method proposed in this paper combines the ideas of both TSVM and MMC. Our method is most closely related to hierarchical SVM classification approaches [46], [47]. Similar to hierarchical SVM, the aim of JointMMCC is to separate nested classes (or clusters) using a successive set of maximum-margin separating hyperplanes. However, the work presented herein is, to our knowledge, the first that casts the problems of semi-supervised classification and maximum-margin clustering within a unified optimization framework. In order to solve the JointMMCC problem, we extend the cutting-plane algorithm from [44] to the case of joint classification and clustering. We apply our approach to the problem of disentangling heterogeneity in the imaging profiles of normal, yet cognitively less stable, older adults.

III. Joint maximum-margin classification and clustering

A. Problem formulation

In this section, we present the joint maximum-margin classification and clustering (JointMMCC) method that performs semi-supervised classification in such a way, that one of the classes contains two clusters that in turn are separable by another hyperplane with a large margin. In what follows we assume without loss of generality that the two target clusters exist in the positive (pathologic) class. More formally, given class labels $(y_1, \dots, y_l) \in \{-1, +1\}^l$ for l out of n data points ($l < n$), we look for a classification hyperplane $\mathbf{w}$ that assigns datapoints $\mathbf{x}$ into one of the two classes. At the same time, the classification hyperplane must be such, that there exists another hyperplane $\widehat{\mathbf{w}}$ separating two clusters within the positive class with a large margin. The JointMMCC problem has the following form:

$$\begin{aligned}
 & \min_{y_{l+1}, \dots, y_n} \quad \min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \beta_l \sum_{i=1}^l \xi_i + \beta_u \sum_{j=l+1}^n \xi_j + \beta \sum_{k=1}^n \widehat{\xi}_k \\
 & \quad z_1, \dots, z_n \quad \widehat{\mathbf{w}}, \widehat{b}, \widehat{\xi} \\
 \text{s.t.} \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\
 & y_j (\mathbf{w}^T \mathbf{x}_j + b) \geq 1 - \xi_j \\
 & z_k (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \geq 1 - \widehat{\xi}_k, \text{ if } (\mathbf{w}^T \mathbf{x}_k + b) \geq 0 \\
 & \xi_i \geq 0, \xi_j \geq 0, \widehat{\xi}_k \geq 0
 \end{aligned} \tag{3}$$

For simplicity, unless noted otherwise, index i runs over $1, \dots, l$, i.e., samples with available class labels, index j runs over $l+1, \dots, n$, i.e., samples with unknown class labels, and index k runs over $1, \dots, n$, i.e., all samples.

The third constraint in Problem 3 is similar to that of classical MMC Problem 2. However, in case of JointMMCC the constraint needs to be satisfied only for the points classified as belonging to the positive class. That is, points from the negative class do not affect the parameters of the clustering hyperplane $\widehat{\mathbf{w}}$. Clearly, due to its nonlinear and non-continuous nature the constraint is very difficult to handle in the current formulation. To make the problem tractable, we use an approximated form of the constraint, which leads to the following JointMMCC problem:

$$\begin{aligned}
& \min_{y_{l+1}, \dots, y_n} \quad \min_{\mathbf{w}, b, \xi} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \beta_l \sum_{i=1}^l \xi_i + \beta_u \sum_{j=l+1}^n \xi_j + \beta \sum_{k=1}^n \widehat{\xi}_k \\
& \quad \quad \quad \min_{z_1, \dots, z_n} \quad \min_{\widehat{\mathbf{w}}, \widehat{b}, \widehat{\xi}} \\
s.t. \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\
& y_j (\mathbf{w}^T \mathbf{x}_j + b) \geq 1 - \xi_j \\
& z_k (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) + \gamma R(\mathbf{x}_k) \geq 1 - \widehat{\xi}_k \\
& \xi_i \geq 0, \xi_j \geq 0, \widehat{\xi}_k \geq 0
\end{aligned} \tag{4}$$

where $R(\mathbf{x})$ in the third constraint is the ramp function defined as follows:

$$R(\mathbf{x}) = \begin{cases} -(\mathbf{w}^T \mathbf{x} + b), & \text{if } \mathbf{w}^T \mathbf{x} + b < 0 \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

and $\gamma \gg 1$ is a large constant that renders the constraint always satisfiable for points not belonging to the positive class. Notice, that $R(\mathbf{x})$ is convex function of $\mathbf{w}$ and b .

B. Cutting plane algorithm for JointMMCC

The difficulty in solving Problem 4 lies in the fact that we have to minimize the objective function w.r.t. the labeling vectors $(y_{l+1}, \dots, y_n)$ and $(z_1, \dots, z_n)$ in addition to the hyperplanes parameters and slack variables. To reduce the number of variables involved, similarly to [45], we reformulate the JointMMCC problem as follows:

THEOREM 1. *Problem 4 is equivalent to*

$$\begin{aligned}
& \min_{\mathbf{w}, b, \xi} \quad \min_{\widehat{\mathbf{w}}, \widehat{b}, \widehat{\xi}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \beta_l \sum_{i=1}^l \xi_i + \beta_u \sum_{j=l+1}^n \xi_j + \beta \sum_{k=1}^n \widehat{\xi}_k \\
s.t. \quad & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\
& \left| \mathbf{w}^T \mathbf{x}_j + b \right| \geq 1 - \xi_j \\
& \left| \widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b} \right| + \gamma R(\mathbf{x}_k) \geq 1 - \widehat{\xi}_k \\
& \xi_i \geq 0, \xi_j \geq 0, \widehat{\xi}_k \geq 0
\end{aligned} \tag{6}$$

where class labels y_j and cluster labels z_k are calculated, respectively, as

$$y_j = \text{sign}(\mathbf{w}^T \mathbf{x}_j + b), \tag{7}$$

and

$$z_k = \text{sign}(\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \tag{8}$$

Similarly to a related theorem in [45], the simple proof of Theorem 1 is achieved by

showing that for every $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ the smallest feasible $\sum_{j=l+1}^n \xi_j$ and $\sum_{k=1}^n \widehat{\xi}_k$ are identical for Problems 4 and 6, and their corresponding labeling vectors are the same.

By reformulating Problem 4 as Problem 6, the number of variables involved in JointMMCC problem is reduced by $2n - l$. To further reduce the number of variables involved in the optimization problem, we propose the following theorem:

THEOREM 2. *Problem 6 can be equivalently formulated as*

$$\min_{\mathbf{w}, \mathbf{b}, \widehat{\mathbf{w}}, \widehat{\mathbf{b}}, \zeta \geq 0} \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \zeta$$

$$s.t. \quad \beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \beta_u \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| + \beta \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| + \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) + \zeta \geq \sum_{i=1}^l c_i + \sum_{j=l+1}^n c_j + \sum_{k=1}^n \widehat{c}_k, \forall \mathbf{c}, \widehat{\mathbf{c}} \in \{0, 1\}^n \quad (9)$$

and any solution $(\mathbf{w}^*, \widehat{\mathbf{w}}^*)$ to Problem 6 is also a solution Problem 9, with

$$\zeta^* = \beta_l \sum_{i=1}^l \xi_i^* + \beta_u \sum_{j=l+1}^n \xi_j^* + \beta \sum_{k=1}^n \widehat{\xi}_k^*. \text{ Here, } \mathbf{c} = (c_1, \dots, c_n) \text{ and } \widehat{\mathbf{c}} = (\widehat{c}_1, \dots, \widehat{c}_n).$$

PROOF. We will show that Problem 6 and Problem 9 have the same objective value and an equivalent set of constraints. Specifically, we will show that for every $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ the

smallest feasible $\zeta^* = \beta_l \sum_{i=1}^l \xi_i^* + \beta_u \sum_{j=l+1}^n \xi_j^* + \beta \sum_{k=1}^n \widehat{\xi}_k^*$ in Problem 6 and ζ in Problem 9 are equal. This means, with $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ fixed, $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \xi_i, \xi_j, \widehat{\xi}_k)$ and $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \zeta)$ are optimal solutions to problems 6 and 9, respectively, and they result in the same objective function value.

For any given $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ the ξ_i, ξ_j and $\widehat{\xi}_k$ in Problem 6 can be optimized individually, and the optimum is achieved as

$$\xi_i^{(1)} = \max \{0, 1 - y_i (\mathbf{w}^T \mathbf{x}_i + b)\} \quad (10)$$

$$\xi_j^{(1)} = \max \{0, 1 - |\mathbf{w}^T \mathbf{x}_j + b|\} \quad (11)$$

$$\widehat{\xi}_k^{(1)} = \max \left\{ 0, 1 - |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \gamma \sum_{k=1}^n R(\mathbf{x}_k) \right\} \quad (12)$$

Similarly, for Problem 9, the optimal ζ is achieved as

$$\begin{aligned} \zeta^{(1)} = \min_{\mathbf{c}, \widehat{\mathbf{c}} \in \{0, 1\}^n} & \left\{ \beta_l \left[\sum_{i=1}^l c_i - \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) \right] \right. \\ & + \beta_u \left[\sum_{j=l+1}^n c_j - \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| \right] \\ & \left. + \beta \left[\sum_{k=1}^n \widehat{c}_k - \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \right] \right\} \quad (13) \end{aligned}$$

Since each c_i, c_j and $\widehat{c}_k$ are independent in Equation 13, they can be optimized individually. Therefore,

$$\begin{aligned} \zeta^{(1)} &= \beta_l \sum_{i=1}^l \max_{c_i \in \{0, 1\}} \{c_i - c_i y_i (\mathbf{w}^T \mathbf{x}_i + b)\} + \beta_u \sum_{j=l+1}^n \max_{c_j \in \{0, 1\}} \{c_j - c_j |\mathbf{w}^T \mathbf{x}_j + b|\} + \beta \sum_{k=1}^n \max_{\widehat{c}_k \in \{0, 1\}} \{\widehat{c}_k - \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \widehat{c}_k \gamma R(\mathbf{x}_k)\} \\ &= \beta_l \sum_{i=1}^l \max \{0, 1 - y_i (\mathbf{w}^T \mathbf{x}_i + b)\} + \beta_u \sum_{j=l+1}^n \max \{0, 1 - |\mathbf{w}^T \mathbf{x}_j + b|\} + \beta \sum_{k=1}^n \max \{0, 1 - |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \gamma R(\mathbf{x}_k)\} \quad (14) \end{aligned}$$

Hence, for any $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ the objective functions for Problems 6 and 9 have the same value given the optimal $\xi_j, \xi_j, \widehat{\xi}_k$ and ζ . Therefore the optima for the two optimization problems are the same.

While Problem 9 has 2^{2n} constraints, it has only one slack variable ζ that is shared across all constraints. Each constraint in this formulation corresponds to a sum of a subset of constraints from 6, and a pair of vectors $\mathbf{C} = (\mathbf{c}, \widehat{\mathbf{c}})$ select the subset [43]. Using Theorem 1 and Theorem 2, we can solve problem 9 instead to find the same classifying hyperplane $\mathbf{w}$ and clustering hyperplane $\widehat{\mathbf{w}}$.

Similarly to [44], [45], we employ an adaptation of the cutting plane algorithm [42] to solve the JointMMCC training problem. The JointMMCC algorithm keeps a subset Ω of working constraints and computes the optimal solution to Problem 9 subject to the constraints in Ω . The algorithm then adds into Ω the most violated constraint in Problem 9 (*i.e.*, constraint resulting in largest ζ). In this way, a better approximation of the original JointMMCC problem is constructed by a cutting plane that cuts off the current optimal solution from the feasible set [42]. The algorithm stops when no constraint in 9 is violated by more than ϵ . Additionally, we have:

THEOREM 3. The most violated constraint in Problem 9 can be computed as follows:

$$c_i = \begin{cases} 1, & \text{if } y_i (\mathbf{w}^T \mathbf{x}_i + b) < 1 \\ 0, & \text{otherwise} \end{cases} \quad (15)$$

$$C_i = \begin{cases} 1, & \text{if } |\mathbf{w}^T \mathbf{x}_j + b| < 1 \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

$$\widehat{C}_k = \begin{cases} 1, & \text{if } |\widehat{\mathbf{w}}^T \mathbf{x}_k + b| + \gamma R(\mathbf{x}_k) < 1 \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

PROOF. In order to fulfill all constraints in Problem 9, the slack variable ζ^* can be calculated as follows

$$\begin{aligned}
\zeta^* = \max_{\mathbf{c}, \widehat{\mathbf{c}} \in \{0,1\}^n} & \left\{ \beta_l \left[\sum_{i=1}^l c_i - \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) \right] \right. \\
& + \beta_u \left[\sum_{j=l+1}^n c_j - \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| \right] \\
& \left. + \widehat{\beta} \left[\sum_{k=1}^n \widehat{c}_k - \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \right] \right\}
\end{aligned} \tag{18}$$

$$\begin{aligned}
= & \beta_l \sum_{i=1}^l \max_{\mathbf{c}_i \in \{0,1\}^n} \left\{ c_i \left[1 - y_i (\mathbf{w}^T \mathbf{x}_i + b) \right] \right\} \\
& + \beta_u \sum_{j=l+1}^n \max_{\mathbf{c}_j \in \{0,1\}^n} \left\{ c_j \left[1 - |\mathbf{w}^T \mathbf{x}_j + b| \right] \right\} \\
& + \widehat{\beta} \sum_{k=1}^n \max_{\mathbf{c}_k \in \{0,1\}^n} \left\{ \widehat{c}_k \left[1 - |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| - \gamma R(\mathbf{x}_k) \right] \right\}
\end{aligned}$$

Hence, the most violated constraint $\mathbf{C} = (\mathbf{c}, \widehat{\mathbf{c}})$ resulting in the largest ζ^* could be calculated as in Equations 15–17.

JointMMCC algorithm iteratively selects the most violated constraint $\mathbf{C}$ under current hyperplanes parameters and adds it into the working constraint set Ω until no violation of constraint is detected. If a point $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \zeta)$ fulfills all constraints up to precision ϵ , i.e.,

$$\begin{aligned}
& \left[\beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \beta_u \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| + \widehat{\beta} \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| \right. \\
& \left. + \widehat{\beta} \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \right] \geq \left[\beta_l \sum_{i=1}^l c_i + \beta_u \sum_{j=l+1}^n c_j + \widehat{\beta} \sum_{k=1}^n \widehat{c}_k \right] - (\zeta + \epsilon)
\end{aligned} \tag{19}$$

then the point $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \zeta + \epsilon)$ is feasible [44]. Additionally, in Problem 9, there is a single slack variable ζ that measures the joint classification and clustering loss. As a result, the approximation accuracy ϵ of this appropriate solution is directly related to the training loss, and can be used to design a stopping criterion.

Assuming that the working constraint set is Ω , JointMMCC could be formulated as the following optimization problem

$$\begin{aligned}
& \min_{\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \zeta \geq 0} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \zeta \\
& s.t. \forall \mathbf{C} = (\mathbf{c}, \widehat{\mathbf{c}}) \in \Omega: \beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) + \beta_u \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| + \widehat{\beta} \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| + \widehat{\beta} \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \geq \sum_{i=1}^l c_i + \sum_{j=l+1}^n c_j + \sum_{k=1}^n \widehat{c}_k - \zeta
\end{aligned} \tag{20}$$

Notice, that Problem 20 has only a subset of constraints from Problem 9. Before providing the details of solving Problem 20, we present the outline of the cutting plane procedure for JointMMCC in Algorithm 1.

C. Class and cluster balance constraints

In maximum-margin clustering (and semi-supervised classification), a trivially “optimal” solution is to assign all samples to the same class, thus making the resulting margin infinite [41]. To eliminate trivial solutions where one class (or cluster) is either empty or consists of a significantly smaller fraction of data points, following [41], one may consider the following class balance constraint:

$$-r \leq \sum_{j=l+1}^n \text{sign}(\mathbf{w}^T \mathbf{x}_j + b) \leq r \quad (21)$$

Similarly, we can consider the following cluster balance constraint for JointMMCC

$$-\widehat{r} \leq \sum_{k=1}^n \text{sign}(\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) I_{(\mathbf{w}_l^T \mathbf{x}_k + b_l \geq 0)} \leq \widehat{r} \quad (22)$$

where constants r and $\widehat{r}$ control the class and cluster imbalance, respectively. $I_{(\cdot)}$ in (22) is the function that returns 1 if the boolean expression $(\cdot)$ is true, and zero otherwise.

The sign function in Equations 21 and 22 is non-linear and complicates the optimization problem. Additionally, non-linear dependence between $(\mathbf{w}, b)$ and $(\widehat{\mathbf{w}}, \widehat{b})$ in (22) is difficult to handle. To simplify the optimization procedure, the constraints can be approximated with their relaxed versions that also prevent trivial solutions

$$-r \leq \sum_{j=l+1}^n (\mathbf{w}^T \mathbf{x}_j + b) \leq r \quad (23)$$

$$-\widehat{r} \leq \sum_{k=1}^n (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \leq \widehat{r} \quad (24)$$

The relaxation of the cluster balance constraint in (24) is somewhat severe as it may allow the clustering hyperplane to be affected by the samples from the negative class. Nevertheless, in our experiments we found that the constraint serves well the purpose of controlling cluster imbalance while allowing to obtain meaningful clustering results.

D. Optimization via the CCCP

At each iteration of JointMMCC we need to solve Problem 20 under the current working constraint set Ω . Notice, that the non-convex constraint in (20) is the difference of two convex functions and can be rewritten as

$$\begin{aligned} & F(\cdot) - G(\cdot) \\ &= \left[\sum_{i=1}^l c_i + \sum_{j=l+1}^n c_j + \sum_{k=1}^n \widehat{c}_k - \zeta - \beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) \right] \\ &\quad - \left[\beta_u \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| + \widehat{\beta} \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| \right] \\ &\quad + \widehat{\beta} \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \leq 0, \end{aligned} \quad (25)$$

where

$$F(\cdot) = \sum_{i=1}^l c_i + \sum_{j=l+1}^n c_j + \sum_{k=1}^n \widehat{c}_k - \zeta - \beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) \quad (26)$$

and

$$G(\cdot) = \beta_u \sum_{j=l+1}^n c_j |\mathbf{w}^T \mathbf{x}_j + b| + \beta \sum_{k=1}^n \widehat{c}_k |\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}| \beta \gamma \sum_{k=1}^n \widehat{c}_k R(\mathbf{x}_k) \quad (27)$$

Optimization problems with concave-convex objective function under concave-convex constraints can be efficiently addressed using *constrained concave-convex procedure* (CCCP) [32]. Given an initial point $(\mathbf{w}_0, b_0, \widehat{\mathbf{w}}_0, \widehat{b}_0)$, the CCCP computes $(\mathbf{w}_{t+1}, b_{t+1}, \widehat{\mathbf{w}}_{t+1}, \widehat{b}_{t+1})$ in the constraint by replacing term $G(\cdot)$ with its first order Taylor expansion in the vicinity of point $(\mathbf{w}_t, b_t, \widehat{\mathbf{w}}_t, \widehat{b}_t)$. As the absolute value components and the ramp function in (25) are non-smooth functions, the gradient of $G(\cdot)$ has to be replaced by a subgradient [48]. Specifically, we need to calculate subgradients of the following types: $|x| = \text{sign}(x)$, and

$$\partial_{\mathbf{w}} R(\mathbf{x}_k) |_{\mathbf{w}=\mathbf{w}_t} = \begin{cases} 0, & \text{if } \mathbf{w}_t^T \mathbf{x}_k + b_t \geq 0 \\ -\mathbf{x}_k, & \text{otherwise,} \end{cases} \quad (28)$$

$$\partial_b R(\mathbf{x}_k) |_{b=b_t} = \begin{cases} 0, & \text{if } \mathbf{w}_t^T \mathbf{x}_k + b_t \geq 0 \\ -1, & \text{otherwise} \end{cases} \quad (29)$$

As a result, solving Problem 20 under the current working constraint set Ω can be achieved by iteratively solving the following *quadratic programming* (QP) problem:

$$\begin{aligned} \min_{\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b}, \zeta \geq 0} & \frac{1}{2} \mathbf{w}^T \mathbf{w} + \frac{1}{2} \widehat{\mathbf{w}}^T \widehat{\mathbf{w}} + \zeta \\ \text{s.t.} & \beta_l \sum_{i=1}^l c_i y_i (\mathbf{w}^T \mathbf{x}_i + b) \\ & + \beta_u \sum_{j=l+1}^n c_j \text{sign}(\mathbf{w}_t^T \mathbf{x}_j + b_t) (\mathbf{w}^T \mathbf{x}_j + b) \\ & + \beta \sum_{k=1}^n \widehat{c}_k \text{sign}(\widehat{\mathbf{w}}_t^T \mathbf{x}_k + \widehat{b}_t) (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \\ & - \beta \gamma \sum_{k=1}^n \widehat{c}_k (\mathbf{w}^T \mathbf{x}_k + b) I_{(\mathbf{w}_t^T \mathbf{x}_k + b_t < 0) + \zeta} \\ & \geq \sum_{i=1}^l c_i + \sum_{j=l+1}^n c_j + \sum_{k=1}^n \widehat{c}_k - r \leq \sum_{j=l+1}^n (\mathbf{w}^T \mathbf{x}_j + b) \leq r - \widehat{r} \leq \sum_{k=1}^n (\widehat{\mathbf{w}}^T \mathbf{x}_k + \widehat{b}) \leq \widehat{r} \end{aligned} \quad (30)$$

Details on the derivation of Problem 30 can be found in the supplemental materials.

The QP Problem (30) can be solved in polynomial time. Following the $CCCP$, the obtained solution $(\mathbf{w}, b, \widehat{\mathbf{w}}, \widehat{b})$ from the QP Problem (30) is then used as $(\mathbf{w}_{t+1}, b_{t+1}, \widehat{\mathbf{w}}_{t+1}, \widehat{b}_{t+1})$ and the next iteration continues until convergence [32], [48]. Additionally, at each iteration of Algorithm 1, we can initialize the $CCCP$ with the values obtained at the previous iteration of the cutting plane algorithm. Similarly to [44], [45], this allows significant reduction in runtime, as in successive iterations Algorithm 1 differs only by a single constraint. The $CCCP$ is summarized in Algorithm 2.

IV. Materials

a) Dataset

We used our algorithm to cluster populations of aging individuals from the Baltimore Longitudinal Study of Aging (BLSA) [49] which has been following a set of older adults with annual and semi-annual imaging and clinical evaluations. In order to eliminate possible bias due to intra-subject image similarity, we considered only images obtained during the first neuroimaging visit (*i.e.*, baseline images). At the time of development of this paper, the dataset contained 143 baseline images. Seventeen of the considered subjects were diagnosed with MCI over the course of the study. A diagnosis of MCI was assigned by consensus conference if a participant had deficits in either a single cognitive domain (usually memory) or had more than one cognitive deficit but did not have functional loss in activities of daily living. Participants were evaluated at the consensus conference if their Blessed Information Memory Concentration [50] score was greater than three or if their informant or subject Clinical Dementia Rating (CDR) [51] score was 0.5 or above.

Cognitive evaluations—In conjunction with each imaging evaluation, every individual's cognitive performance was evaluated on tests of mental status and memory. We selected the following four measures for our analysis: the total score from the Mini-Mental State Exam (MMSE) [52], the immediate free recall score (sum of five immediate recall trials) on the California Verbal Learning Test (CVLT) [53], and long-delay free recall score on CVLT, and the total number of errors from the Benton Visual Retention Test (BVRT) [54]. We focused on these measures because changes in new learning and recall are among the earliest cognitive changes detected during the prodromal phase of AD [55]. More specifically, we used follow-up cognitive evaluations to estimate the rate of change in cognitive performance. As the above cognitive scores are very noisy, one can have relatively higher confidence in extreme (*i.e.*, very low or very high) rates of changes in cognitive performance that may more reliably indicate whether an individual will decline cognitively or remain stable.

b) Preprocessing MRI data

All MR images were preprocessed following a mass-preserving shape transformation framework [56]. Gray matter tissue (GM) was segmented out from each skull-stripped MR brain image by a brain tissue segmentation method [57]. Each tissue-segmented brain image was spatially normalized into a template space [58], by using a high-dimensional image warping method [59]. The total tissue mass was preserved in each region during the image warping, and tissue density maps were generated in the template space. An example of a tissue density map for GM is shown in Figure 2. These tissue density maps give a quantitative representation of the spatial distribution of tissue in a brain, with brightness being proportional to the amount of local tissue volume before warping.

c) Dimensionality reduction

Due to the high computational demand associated with the JointMMCC optimization problem we reduced the dimensionality of the tissue density maps using a wavelet-based data compression approach. For every image, we calculated the Cohen-Daubechies-Feauveau (CDF) 9/7 wavelet transform [60], which resulted in wavelet coefficients equal to the original dimensionality of the data. We then aggregated wavelet transforms over all available images by averaging the respective wavelet coefficients. We selected indices of the coefficients that corresponded to large absolute average values. The selected indices represent the top-ranked wavelet coefficients. Finally, for a wavelet transform of any given image, we retained only the top-ranked coefficients. In this way, a tissue density map was represented with a small set of wavelet coefficients, that on average have high absolute values.

Note, that our choice of the dimensionality reduction technique is not necessarily optimal for our method. Ideally, one would need to select features that result in the largest classification and (or) clustering margins. Nonetheless, the simple dimensionality reduction step serves the purpose of assessing the potential of our approach.

V. Experiments

A. Simulated experiments

d) Simulated data—Simulations allow generation of images that come from precisely known distributions, and to provide a quantitative estimate of the method's performance. In our experiments our goal was to assess the ability of the proposed approach to correctly discover coherent subpopulations of individuals that exhibit abnormal morphological patterns in the brain. To achieve this goal we introduced artificial patterns of volumetric pathology into a subset of available baseline BLSA images. More specifically, we randomly selected 50% of the original images to represent the simulated normal population. The remaining 50% of the images were selected to represent the simulated abnormal population. Half of the images from the selected abnormal subpopulation were further randomly selected to represent the group with “Pathology 1”, while the other half of the selected abnormal images were assigned into “Pathology 2” group. We then created two different patterns of brain pathology, each defined by a triplet of spherical volumetric regions. The two patterns of brain pathology were selected such, that the corresponding spherical regions did not overlap. Figure 3 shows the examples of the volumetric regions that were selected by us to represent two different types of pathology in the brain. By multiplying tissue density values inside the selected spherical volumetric regions by some coefficient one can introduce artificial atrophy in the original brain. Multiplying tissue density values by a coefficient that is close to 1 introduces very slight (*i.e.*, indistinct) atrophy, while multiplication by a value that is significantly different from 1 results in a pronounced atrophy. For a given subject and for a specific severity of the pathology, the multiplication coefficient was drawn from a Gaussian distribution with mean μ and standard deviation $\sigma = 0.07$. The values for the means μ considered in our simulated experiments were selected to be (1.0, 1.1, 1.2, ..., 2.2) and allowed to generate 13 levels of pathology. It is therefore expected that a much better clustering accuracy can be achieved in a simulated dataset generated with larger μ (*i.e.*, very distinct pathology).

Finally, 24 images from the simulated normal subgroup were randomly assigned into the negative labeled set, and 24 images from the simulated abnormal subgroup were randomly selected into the positive labeled group (12 images representing Pathology 1, and 12 images representing Pathology 2). Figure 4 summarizes the setup of the simulated experiment.

There are several properties of our simulated data that make it particularly suitable for assessing the performance of the clustering algorithm: (1) We perform the simulated experiments on 3D images that match the dimensions of the images in the real-life experiments; (2) For each image and each level of pathology distinctiveness, the multiplication values that are used to introduce the pathology in the spherical regions are drawn from a relatively wide Gaussian distribution, which leads to overlaps between the levels of pathology distinctiveness, and thus allows to simulate a spectrum in the simulated abnormal set of images. This property of the simulated data specifically models studies of patients and healthy control subjects where, for example, the severity of symptoms in patients ranges from mild to severe; (3) Each type of pathology is represented by multiple regions in the brain; (4) Although the simulated data was formed from the data that included MCI subjects, in the simulated experiments we are interested in clustering subjects with respect to the *simulated* pathologies. As the result, pathology associated with cognitive decline serves as confound in the simulated experiments. In this way, the simulated experiments closely model the real-life clustering problems where the disease-related patterns are confounded by brain changes that are of no particular interest for the problem (*e.g.*, age, gender).

The task of joint classification and clustering is then to discover the two patterns of pathology. In order to achieve this, the method has to simultaneously classify unlabeled abnormal images, as well as to cluster an abnormal population with respect to the two types of simulated pathology. As the result, JointMMCC attempts to discover a subset of images with simulated pathologies such that the subset can be reliably clustered.

e) Validating results—In order to obtain an unbiased estimate of the method's performance, we randomly split the simulated data into the test and training sets, such that half of the simulated labeled and the simulated unlabeled images from the three simulated groups (*i.e.*, “simulated normal”, “Pathology 1”, and “Pathology 2”) was in the training set, and another half was in the test set. The parameters of the classification/clustering algorithm were then optimized on the training set, and the clustering and classification accuracies were assessed on the test set.

We used a two-stage criterion for parameters optimization: 1) Given the data points classified by the algorithm into the positive class, the cluster where the fraction of images with “Pathology 1” was higher was identified as the Pathology 1 cluster, while the remaining cluster was identified as the Pathology 2 cluster. For the purpose of parameters optimization, the clustering accuracy was then calculated as the fraction of images from the positive class that were assigned correctly into the appropriate clusters. The set of parameters that yielded highest clustering accuracy on the training set was then selected; 2) If two different sets of parameters resulted in the maximum clustering accuracy, then we selected the set of parameters that also yielded higher classification accuracy. Additionally, we discarded any set of parameters that resulted in a clustering solution where one of the clusters contained less than five subjects. Notice, that as the fraction of the correctly clustered images was estimated among all images in the classified abnormal population (which may contain simulated normal images misclassified as abnormal), the solutions where a large number of simulated normal images are classified as abnormal will typically result in a lower clustering accuracy. Therefore, our measure of clustering accuracy biases the parameters estimation toward the solutions that are specific at the cost of lost sensitivity. This is consistent with the expected functionality of the JointMMCC, and thus allows to perform a fair comparison with the alternative methods that perform classification and clustering in a sequence (*e.g.*, SVM and MMC, or TSVM and MMC).

The process of randomly assigning subjects into the training and the test sets was repeated ten times, and the average classification and clustering accuracies were estimated. Table I provides the details of the parameter values that were used in the parameter optimization stage of the experiment. As the number of possible parameter combinations was extremely large, we resorted to a greedy optimization of training parameters. The initial values b_0 and $\widehat{b}_0$ were set to zeros, while the initial vectors w_0 and $\widehat{w}_0$ were selected at random such that each of the elements in vectors w_0 and $\widehat{w}_0$ had absolute value less than $1/d$, where d was the dimensionality of the data (*i.e.*, number of top-ranked features). The precision ϵ was selected to be 0.01, however, no more than 20 iterations were allowed in Algorithm 1 if the desired precision was not reached. In addition to the parameters of the JointMMCC algorithm, the application-specific parameters include the number of top-ranked features (d), and the wavelet decomposition level (m) in the feature extraction step of the approach.

f) Joint maximum-margin classification and clustering—We compared the performance of JointMMCC with the performance of SVM classification followed by a maximum-margin clustering (SVM+MMC), as well as to the sequential application of TSVM and MMC (TSVM+MMC). Although the primary focus of our method is on clustering, which can be performed at the loss of classification sensitivity, we compared the classification performance of our approach with the performances of fully-supervised SVM, as well as with TSVM. Recall, that the parameters of the SVM and TSVM algorithms were estimated jointly with MMC, which biased the parameters estimation process toward the solutions with higher clustering accuracy, possibly at the cost of degraded classification performance. As the result, similarly to JointMMCC, the parameters for the semi- and fully-supervised SVM methods were selected such that the simulated abnormal images were more likely to be misclassified as simulated normal than vice versa. Tables II, III and IV present the medians of the parameters estimated for the three approaches for different distinctiveness levels of the simulated pathology as modeled by the mean μ of the Gaussian distribution from which the multiplication coefficients used to introduce pathologies were drawn. The tables also provide the median parameter values over all levels of distinctiveness.

Figure 5 shows the mean classification accuracy obtained using the proposed JointMMCC approach, semi-supervised SVM, and fully supervised SVM in relationship to the distinctiveness of the simulated pathologies. Additionally, Figures 6 and 7 show the specificity and the sensitivity of the three methods. The specificity and the sensitivity were

calculated, respectively, as $SPE = \frac{TN}{TN+FP}$ and $SEN = \frac{TP}{TP+FN}$, where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, and FN is the number of false negatives.

Overall, Figure 5 suggests that there is no noticeable difference in classification performance between the three approaches. As expected, the specificity of the approaches was higher than the sensitivity, which in the case of SVM+MMC and TSVM+MMC is likely due to the choice of the clustering-based parameters optimization criterion.

While the classification accuracy reflects the ability to correctly identify the abnormal subpopulation, the clustering accuracy reflects the potential to discover clusters in the abnormal cohort, and is the main focus of the method. Figure 8 shows clustering accuracies achieved using JointMMCC, as well as by using SVM (or TSVM) to identify the abnormal subpopulation, and then by applying MMC to cluster the abnormal set. The clustering accuracy was calculated as the accuracy of correctly assigning images from the obtained abnormal class into the two clusters. Recall, that given our measure of clustering accuracy,

incorrect classification of simulated normal images into the abnormal class will reduce the clustering accuracy.

The result in Figure 8 suggests that once the clustering accuracy is beyond chance, the clustering accuracy obtained using JointMMCC is superior to that of the SVM+MMC and TSVM+MMC. Importantly, there is no consistent clear preference between the two alternative methods, with TSVM+MMC performing somewhat better than SVM+MMC on the datasets where the simulated pathology is moderately distinct.

In order to summarize the overall classification and clustering performance of the approaches, Figure 9 shows the average between the classification and clustering accuracies as a function of pathology distinctiveness. The results suggest that, for the problem addressed in this paper, the JointMMCC algorithm performs on a par with or better than the approaches performing classification and clustering in two disjoint stages.

Finally, Figure 10 presents the accuracy of assigning pathologies into clusters, which was calculated as the fraction of the *simulated abnormal* images in the *obtained abnormal* class that were correctly assigned into the two clusters. In contrast to the clustering accuracy shown in Figure 8, the accuracy of assigning pathologies into clusters is not affected by the normal images that were incorrectly classified as abnormal.

B. Clustering cognitively declining older adults

g) Setup for the JointMMCC—The main task in our experiment is to discover clusters of images in the cognitively declined subpopulation (as defined via the rate of change in CVLT score), given only a limited amount of labeled information about the cognitively most stable and least stable cases. The task is two-fold and involves both classifications of unlabeled data into “cognitively stable” and “cognitively declined” groups, as well as clustering into two clusters the imaging profiles of the subjects whose profiles are similar to those of the cognitively declining subjects. Figure 11 shows the setup of the joint classification and clustering as applied to the problem of finding coherent subpopulations in cognitively less stable adults. The cognitively most stable labeled subset was formed from images of 25 subjects who were not diagnosed with MCI during the course of the study and had the highest slopes of CVLT Immediate Free Recall scores. Similarly, images of 25 subjects who were not diagnosed with MCI during the course of the study and had the lowest slope scores were assigned into the labeled cognitively least stable subset. The slope of the CVLT score represents the rate of cognitive decline, and lower slopes of the score indicate higher rates of decline. The remaining 93 subjects were unlabeled. Similarly to the clustering approach [61] that was applied to the same study, the 17 subjects who were diagnosed with MCI at some point during the study were considered to be the seeded subjects that served the purpose of interpreting the clustering results. It has to be noted that while we focused solely on CVLT to define a “cognitively declining” population, more stringent diagnostic criteria could be more appropriate.

The brain regions of most significant voxel-wise differences between the most stable and least stable subjects are shown in Figure 12. None of the regions survived correction for multiple comparisons, which may be due to the relative small sample size. Several regions of relatively reduced volumes of GM in least stable subjects compared to most stable population are evident, including the temporal, parietal and medial frontal cortical regions. The inverse contrast in Figure 12 shows increased periventricular gray tissue in least stable subjects.

In general, the patterns of regional differences in Figure 12 are consistent with previous studies pointing toward accelerated change with age for frontal gray matter, superior,

middle, and medial frontal, and superior parietal regions in MCI subjects [62]. In addition, age-related loss of brain tissue has been observed for occipital regions in the past [49], although the frontal and parietal regions have been shown, in general, to exhibit greater rates of decline in cortical thickness than temporal and occipital areas [63].

h) Joint classification and clustering—As we mentioned earlier, 17 out of 143 BLSA subjects considered in our experiments were diagnosed with MCI at some point during the study. In the absence of ground truth data, as it is often the case in the exploratory clustering analysis, one needs to have certainty in the stability and reproducibility of the clustering configuration before attempting to interpret the results. To ensure that the clustering results are stable with respect to the labeled data, we designed a leave-one-out (LOO) cross-validation procedure shown in Figure 13. In our experiments, at each run of the LOO procedure we left out one of the labeled images, and performed joint classification and clustering. When performing joint classification and clustering using different subsets of the labeled images, we observed that one of the two obtained clusters usually contained a large majority of the subjects that converted to MCI. In particular, on average 74% of the MCI subjects in the cognitively declining class were assigned into one of the clusters. We identified the cluster that contained most of the BLSA-converters as “MCI-like cluster”. The other cluster was identified as the “Non-MCI cluster”. We then applied a voting procedure to aggregate the classification and clustering results over all runs of LOO cross-validation.

Following the results of the simulated experiments that were performed on the images simulated from the real-life dataset, the parameters of the algorithm were selected based on the median overall parameter values in Table II. As the result, we used the possible parameter values in Table I that were closest to the median overall parameter values in Table II, and selected the following: $m = 9$, $d = 1000$, $\beta_1 = 10$, $\beta_u = 0.001$, $\widehat{\beta} = 1$, $r = 10$, $\widehat{r} = 0.001$, and $\gamma = 10$. It is worth mentioning that while the simulated data was generated from the real images, the parameters selection was performed on a task of clustering populations with artificial pathologies, which is different from the current experiment. In our implementation it took approximately 30 seconds to complete the maximum allowed 20 iterations of the CCCP in Algorithm 2.

As the result, unlabeled BLSA subjects were classified into “cognitively stable” or “cognitively declined” classes, and images from the “cognitively declined” class were assigned into “MCI-like” or “Non-MCI” clusters. Table V shows the differences in the distributions of cognitive scores at baseline, as well as slopes of cognitive evaluations, in cognitively stable and cognitively declined classes as estimated using the Wilcoxon signed ranks test. t -test was performed only for the case where the distribution passed the test for normality at 5% significance level as assessed via the Kolmogorov–Smirnov test (*i.e.*, slope of CVLT List A Sum).

Figure 14 shows t -statistics that signify voxel-wise brain difference between the obtained two classes. The regions of reduced GM in cognitively declined individuals relative to the cognitively stable population include the hippocampus, amygdala, and entorhinal cortex, much of the temporal lobe GM and the insular cortex (especially the superior temporal gyrus), posterior cingulate and precuneous, and orbitofrontal cortex. The inverse contrast in Figure 14 shows increased periventricular gray tissue in cognitively declined subpopulation. Notice, that the voxel-wise differences in the groups identified by the classification component of JointMMCC are more pronounced than the differences between labeled groups in Figure 12.

As we mentioned earlier, clustering the abnormal images results in two clusters: MCI-like cluster and non-MCI cluster. Table VI shows the differences in the distributions of cognitive scores in MCI-like and non-MCI clusters.

Figure 15 shows t -statistics that signifies voxel-wise brain difference between the obtained two clusters. The regions of reduced GM in individuals from the MCI-like cluster relative to non-MCI cluster include a number of temporal, parietal, occipital and occipitotemporal and medial cortical regions. The inverse contrast in Figure 15 shows increased periventricular gray tissue in the MCI-like profile. The results indicate that images in the MCI-like cluster potentially reflect much more pronounced tissue loss than was observed in general cognitively declined population in Figures 14 and 15.

VI. Discussion and Conclusion

In this paper, we presented a method for disentangling heterogeneity in imaging profiles that are characterized by spectra of underlying pathologies. Our approach is designed to address several major challenges that are characteristic of a variety of imaging studies: 1) heterogeneity of diseases; 2) absence of clear nosological boundaries between subconditions; and 3) uncertainty in categorical labels. Our method attempts to find coherent subsets of images with respect to the underlying pathologies by incorporating clustering. At the same time, by incorporating semi-supervised classification our approach is able to identify a pathological image set in a way that the heterogeneity in the pathological cohort can be reliably disentangled, and works even if the reliably labeled data is not available. In contrast to the approaches attempting to discern the *known* subcategories of imaging profiles, our focus was on identifying the *unknown* subcategories of images, that in turn may relate to diagnostic subcategories within the condition.

The necessity to identify a clusterable pathological subpopulation is particularly pressing in the problems where the boundaries between healthy and pathological states are blurred. The main assumption of our method is that the more pronounced the pathology is, the easier it is to disentangle the underlying heterogeneity in the imaging profiles. While in this paper we focused on two-cluster problems, the method can be extended to multi-cluster analysis. Our simulated experiments, as well as the consistent results of our approach in the imaging study of aging, show that for the task of clustering image sets characterized by spectra of pathologies it is important to identify a pathological population that is clusterable.

In our approach, the classification hyperplane between the abnormal and normal cases is found in a semi-supervised manner. As the result, the classification decision boundary may be pushed out of the lowest density area in order to obtain an abnormal subpopulation where a large clustering margin can be found (which is a desired property). Intuitively, the method attempts to discard some abnormal images in such a way that the remaining abnormal set is clusterable. As a consequence, the method tends to increase the specificity of the classifier, while decreasing the sensitivity. As the semi-supervised classification hyperplane determines the abnormal image set, the optimal clustering hyperplane changes depending on the position of the classification hyperplane. This balancing between the classification margin and the clustering margin is controlled primarily by the constant γ .

Although our focus was on medical image analysis, our method is general and can be applied to other exploratory clustering problems where there is an expectation to find clusters at an end of a spectrum. Moreover, in medical image analysis, the method is not limited to the clustering of brain images. An obvious biomedical application of our method would be in the analysis of abnormal tissue subtypes in the cases where the differences between the normal and abnormal tissue are not always distinct.

Currently, our method is designed for two-cluster problems. While clustering methods with known number of clusters still fall under the umbrella of unsupervised learning methods, assuming that there are two clusters in the data is a form of prior knowledge. However, even in its current form, our method can be viewed as the means to test the hypothesis that there are two distinct imaging profiles in the data, which in turn can provide a valuable insight into the heterogeneity of the condition. Moreover, the method can be extended to a multi-cluster scenario following, for example, the multi-class clustering formulation and the cutting plane algorithm of [64].

The application of this paper was in clustering cognitively declined aging populations. The problem of analyzing populations of normal older adults is particularly challenging due to the absence of clearly defined categorical labels, unreliability of cognitive evaluations, and heterogeneity of the data. Nonetheless, our method was able to identify imaging profiles characteristic of a subpopulation that is likely to exhibit cognitive decline in the future. It can be expected that, eventually, some of the older adults will be diagnosed with MCI (or even AD), while others will only exhibit cognitive decline associated with normal aging. As the result, it is likely to find two different trajectories of aging in the cognitively declining population. However, as all subjects in our population are healthy (with the exception of the few MCI subjects whom we used as seeding points to accumulate clustering results within the cross-validation), it is difficult to select a proper population of cognitively declining individuals that can be readily analyzed using clustering.

Earlier studies have suggested that verbal memory decline and impairment are evident years before a diagnosis of dementia [55], [65]. Additionally, performance on the CVLT has been specifically implicated in predicting future dementia type and discriminating between normal aging, MCI, and dementia [66], [67]. In this paper, we applied our method within a population of normal individuals who lacked a clear diagnostic or prognostic categorization. We therefore used CVLT as a surrogate for the diagnostic category assignment. At the same time, it has to be noted that while CVLT was shown to have a predictive value, it is not sufficiently specific and sensitive, and other measures could have been used in place of CVLT. Moreover, with the exception of the relatively few extreme cases, it is difficult to tell with certainty if a given subjects is cognitively declining, as the longitudinal CVLT scores used to determine the rate of cognitive decline are extremely noisy. Hence, one ends up having very few labeled and a much larger set of unlabeled samples. This scenario is common in other studies where there are many subjects with mild forms of the condition of interest and only few extreme cases. It has been also shown earlier that including unlabeled data within TSVM can lead to better classification performance of medical images than when training a supervised classifier using only the labeled information [68]. Together, this points toward the potential benefits of semi-supervised classification and clustering in the analysis of older normal populations.

In our analysis, we were able to address the heterogeneity of the cognitively declined subpopulation by identifying two distinct clusters that have significantly different phenotypic, as well as cognitive, profiles. The discovered two clusters were significantly different with respect to the cognitive performance. One of the clusters had relatively more pronounced MCI-like phenotypic profile accompanied by significant gray matter tissue loss in a number of temporal, parietal, occipital and occipitotemporal and medial cortical regions. This MCI-like profile of one of the obtained clusters was further confirmed by the fact that a majority of the subjects who were actually diagnosed with MCI later during the study in fact belonged to that cluster.

The imaging profile of the MCI-like cluster was consistent with the longitudinal pattern of regional brain volume change that have been observed for MCI and normal aging. In

particular, it has been shown in the past that MCI subjects exhibit accelerated volumetric changes in ventricular CSF, temporal gray matter, and orbitofrontal and temporal association cortices [62].

In a previous attempt to analyze the heterogeneity of normal aging it was shown that clustering images of the entire population of healthy older adults reveals a hierarchical structure of the image set, where a healthier subpopulation forms a dense cluster, and the remaining, highly heterogeneous subpopulation forms a rather disperse cluster, which in turn is comprised of a smaller more homogeneous subcluster of relatively healthier subjects and a more disperse subcluster of relatively worse performing subjects, and so on [61]. This result can be viewed as the evidence of the spectrum-like structure of the space formed by the MRI scans of healthy older adults. At the same time, it has been shown that normal aging increases cognitive heterogeneity [25], [26], and, therefore, there is a need to identify aging subpopulations where the heterogeneity is evident and can be disentangled. To achieve this, our method identified an image set that represented subjects with the declining cognitive performance (as assessed via CVLT), while ensuring that the image-based heterogeneity within this cognitively declining cohort can be disentangled. The patterns of tissue loss in the images from the cognitively declining subset as compared to the cognitively stable subset were consistent with the patterns of tissue loss in the pathological subpopulation identified via clustering in [61], and included temporal, parietal and medial frontal cortical regions. At the same time, our analysis goes beyond pointing toward the spectrum-like structure of the MRI in the aging population, and suggests that the cognitively declining population can be characterized by two distinct imaging profiles, one of which is similar to the imaging profile of MCI.

In addition to identifying a clusterable subset of more pathological MRI, our results signify the challenge associated with the uncertainty in the labeled information. In particular, it would be reasonable to expect the voxel-wise differences in baseline images of the subjects in the cognitively most stable and least stable subpopulations to be more evident than the difference between any other cognitively stable and cognitively declining subpopulations (since the most stable and most declined subpopulations exhibit extreme rates of cognitive decline as assessed via cognitive evaluations). However, Figure 14 and the respective figures in the supplemental materials show that, guided by the few labeled instances, our method identified subpopulations exhibiting regional group differences that are more pronounced than the differences between the labeled subjects. This particular result suggests that the labeled subpopulations identified with respect to the extreme rates of change in CVLT may not in fact represent populations with the extreme pathological changes in the brain. At the same time, the differences between the most stable and least stable subpopulations, as well as the differences between the stable and declining subpopulations identified by the method, were consistent with the existing picture of cognitive decline [49], [62], [63].

It has to be mentioned that while the presence of subpathologies may imply differences in etiologies and possibly distinct pathological processes, they may not be discernible depending on the biological specificity of the imaging technique. In some cases, it may be more appropriate, to focus on the differences in pathological paths or timecourses, and possible co-morbid effects that may be discernible using the available imaging data. Moreover, when using an exploratory technique (like clustering), it is difficult to predict whether the heterogeneity in the population can be disentangled using the specific computational tool. At the same time, in the case of analyzing the normal older population we discovered that there may be two imaging profiles in the less cognitively stable population, and one of these profiles is associated with mild cognitive impairment. Had we not discovered any clusters in the population, the results would have suggested that there are

either no distinct subgroups in the data, or that the MRI modality is not suitable for capturing the underlying group differences.

While in our experiments we analyzed baseline neuroimaging evaluations of older adults, we used follow-up cognitive evaluations to identify labeled subjects that are likely to decline cognitively in the future. Alternative strategies can be employed to select the extreme normal and extreme abnormal subpopulations if follow-up evaluations are not available. For example, when studying heterogeneity of MCI populations in relation to progression to Alzheimer's disease, the two labeled extreme populations can be formed from healthy subjects and subjects diagnosed with AD, while subjects with MCI may be unlabeled. Similarly, in the studies of ASD, the extreme labeled subpopulations can be identified based on the performance along one of the relevant dimensions (*e.g.*, learning, social skills).

Studies of ASD and schizophrenia are two additional applications that can be studied using our approach. Both ASD and schizophrenia are characterized by blurred boundaries between the normal and diseased states, and by significant heterogeneity of the pathological populations, which fit well into the formulation of JointMMCC. At the same time, while our framework is based on the soft-margin versions of maximum margin classification and clustering, and thus allows to handle the outliers methodologically, a post-analysis of the mislabeled images with respect to external non-imaging information would be enlightening. However, in the exploratory application considered in this paper we are not able to identify outliers as our population of healthy adults lacks predefined diagnostic subcategories. In this respect, identification and analysis of outliers in the studies of ASD and schizophrenia would be more straightforward. We are currently working on applying the proposed approach to studies of these two diseases. Additionally, extending our method to longitudinal studies and to analyzing timecourses of pathological changes is a very promising direction. However, there are methodological challenges that have to be addressed before the current method can be applied to longitudinal data.

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), and R01-AG14971, N01-AG-3-2124, N01-AG-3-2124.

References

- [1]. Davatzikos C, Xu F, An Y, Fan Y, Resnick SM. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. *Brain*. 2009; 132(8):2026–2035. [PubMed: 19416949]
- [2]. Fan Y, Resnick SM, Wu X, Davatzikos C. Structural and functional biomarkers of prodromal alzheimer's disease: A high-dimensional pattern classification study. *NeuroImage*. 2008; 41(2): 277–285. [PubMed: 18400519]
- [3]. Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, Knopman DS, Petersen RC, Jack J, R. C, O. Alzheimer's Disease Neuroimaging Initiative. MRI and CSF biomarkers in normal, MCI, and AD subjects: Diagnostic discrimination and cognitive correlations. *Neurology*. 2009; 73(4):287–293. [PubMed: 19636048]
- [4]. McEvoy LK, Fennema-Notestine C, Roddey JC, Hagler DJ, Holland D, Karow DS, Pung CJ, Brewer JB, Dale AM. Alzheimer Disease: Quantitative Structural Neuroimaging for Detection and Prediction of Clinical and Structural Changes in Mild Cognitive Impairment1. *Radiology*. 2009; 251(1):195–205. [PubMed: 19201945]
- [5]. Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. Spatially augmented lppboosting for ad classification with evaluations on the adni dataset. *NeuroImage*. 2009; 48(1): 138–149. [PubMed: 19481161]

- [6]. Duchesne S, Caroli A, Geroldi C, Barillot C, Frisoni G, Collins D. Mri-based automated computer classification of probable ad versus normal controls. *IEEE Trans. Med. Imaging.* Apr; 2008 27(4):509–520. [PubMed: 18390347]
- [7]. Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RSJ. Automatic classification of mr scans in alzheimer's disease. *Brain.* Mar; 2008 131(3):681–689. [PubMed: 18202106]
- [8]. Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision). 4th ed.. American Psychiatric Publishing, Inc.; Jul. 1994
- [9]. The ICD-10 classification of mental and behavioural disorders : clinical descriptions and diagnostic guidelines. World Health Organization; Geneva: 1992.
- [10]. Mayes SD, Calhoun S, Crites D. Does dsm-iv asperger's disorder exist? *Journal of Abnormal Child Psychology.* 2001; 29:263–271. [PubMed: 11411788]
- [11]. Miller J, Ozonoff S. The external validity of asperger disorder: lack of evidence from the domain of neuropsychology. *J Abnorm Psychol.* 2000; 109(2):227–38. [PubMed: 10895561]
- [12]. Matson J, Wilkins J. Nosology and diagnosis of asperger's syndrome. *Research in Autism Spectrum Disorders.* 2008; 2(2):288–300.
- [13]. Green EK, Grozeva D, Jones I, Jones L, Kirov G, Caesar S, Gordon-Smith K, Fraser C, Forty L, Russell E, Hamshire ML, Moskvin V, Nikolov I, Farmer A, McGuffin P, Holmans PA, Owen MJ, O'Donovan MC, Craddock N. The bipolar disorder risk allele at cacna1c also confers risk of recurrent major depression and of schizophrenia. *Mol Psychiatry.* 2010; 15(10):1016–22. [PubMed: 19621016]
- [14]. Moskvin V, Craddock N, Holmans P, Nikolov I, Pahwa JS, Green E, Consortium WTCC, Owen MJ, O'Donovan MC. Gene-wide analyses of genome-wide association data sets: evidence for multiple common risk alleles for schizophrenia and bipolar disorder and for overlap in genetic risk. *Mol Psychiatry.* Mar; 2009 14(3):252–260. [PubMed: 19065143]
- [15]. Owen MJ, Craddock N, Jablensky A. The genetic deconstruction of psychosis. *Schizophr Bull.* Jul; 2007 33(4):905–911. [PubMed: 17551090]
- [16]. Craddock N, O'Donovan MC, Owen MJ. Genes for schizophrenia and bipolar disorder? implications for psychiatric nosology. *Schizophr Bull.* Jan; 2006 32(1):9–16. [PubMed: 16319375]
- [17]. Craddock N, O'Donovan MC, Owen MJ. The genetics of schizophrenia and bipolar disorder: dissecting psychosis. *J Med Genet.* Mar; 2005 42(3):193–204. [PubMed: 15744031]
- [18]. Koutsouleris N, Gaser C, Jaumlger M, Bottlender R, Frodl T, Holzinger S, Schmitt GJE, Zetzsche T, Burgermeister B, Scheuerecker J, Born C, Reiser M, Moulmler H-J, Meisenzahl EM. Structural correlates of psychopathological symptom dimensions in schizophrenia: a voxel-based morphometric study. *Neuroimage.* 2008; 39(4):1600–12. [PubMed: 18054834]
- [19]. Davatzikos C. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *Neuroimage.* Sep; 2004 23(1):17–20. [PubMed: 15325347]
- [20]. Koutsouleris N, Schmitt GJE, Gaser C, Bottlender R, Scheuerecker J, McGuire P, Burgermeister B, Born C, Reiser M, Moller H-J, Meisenzahl EM. Neuroanatomical correlates of different vulnerability states for psychosis and their clinical outcomes. *The British Journal of Psychiatry.* Sep; 2009 195(3):218–226. [PubMed: 19721111]
- [21]. Meisenzahl EM, Koutsouleris N, Bottlender R, Scheuerecker J, Jger M, Teipel SJ, Holzinger S, Frodl T, Preuss U, Schmitt G, et al. Structural brain alterations at different stages of schizophrenia: a voxel-based morphometric study. *Schizophrenia Research.* 2008; 104(1–3):44–60. [PubMed: 18703313]
- [22]. Klosterkttter J, Schultze-Lutter F, Gross G, Huber G, Steinmeyer EM. Early self-experienced neuropsychological deficits and subsequent schizophrenic diseases: an 8-year average follow-up prospective study. *Acta Psychiatrica Scandinavica.* 1997; 95(5):396–404. [PubMed: 9197904]
- [23]. Johns LC, Cannon M, Singleton N, Murray RM, Farrell M, Brugha T, Bebbington P, Jenkins R, Meltzer H. Prevalence and correlates of self-reported psychotic symptoms in the british population. *Br J Psychiatry.* 2004; 185:298–305. [PubMed: 15458989]

- [24]. Resnick SM, Sojkova J, Zhou Y, An Y, Ye W, Holt DP, Dannals RF, Mathis CA, Klunk WE, Ferrucci L, Kraut MA, Wong DF. Longitudinal cognitive decline is associated with fibrillar amyloid-beta measured by [11c]pib. *Neurology*. 2010; 74(10):807–15. [PubMed: 20147655]
- [25]. Morse, CK.
- [26]. Christensen H, Mackinnon A, Jorm AF, Henderson AS, Scott LR, Korten AE. Age differences and interindividual variation in cognition in community-dwelling elderly. *Psychology and Aging*. 1994; 9(3):381–390. [PubMed: 7999323]
- [27]. Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, Brammer MJ, Murphy C, Murphy DG. Investigating the predictive value of whole-brain structural mr scans in autism: a pattern classification approach. *Neuroimage*. Jan; 2010 1(49):44–56. [PubMed: 19683584]
- [28]. Ecker C, Marquand A, Mourão Miranda J, Johnston P, Daly EM, Brammer MJ, Maltezos S, Murphy CM, Robertson D, Williams SC, Murphy DG. Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. Aug; 2010 30(32):10 612–10 623.
- [29]. Fan Y, Shen D, Gur RC, Gur RE, Davatzikos C. Compare: Classification of morphological patterns using adaptive regional elements. *IEEE Trans. Med. Imaging*. 2007; 26(1):93–105. [PubMed: 17243588]
- [30]. Liu, Y.; Teverovskiy, L.; Carmichael, O.; Kikinis, R.; Shenton, M.; Carter, CS.; Stenger, VA.; Davis, S.; Aizenstein, H.; Becker, J.; Lopez, O.; Meltzer, C. Discriminative mr image feature analysis for automatic schizophrenia and alzheimer's disease classification. Robotics Institute; Pittsburgh, PA: Mar. 2004 Tech. Rep. CMU-RI-TR-04-15
- [31]. Ramirez J, Gorriz J, Segovia F, Chaves R, Salas-Gonzalez D, Lopez M, Alvarez I, Padilla P. Computer aided diagnosis system for the alzheimer's disease based on partial least squares and random forest spect image classification. *Neuroscience Letters*. 2010; 472(2):99–103. [PubMed: 20117177]
- [32]. Smola, A.; Vishwanathan, S.; Hofmann, T. Kernel methods for missing variables. *Proc. International Workshop on Artificial Intelligence and Statistics*; 2005. p. 325-332.
- [33]. Duchesne S, Bocti C, Sousa KD, Frisoni GB, Chertkow H, Collins DL. Amnesic mci future clinical status prediction using baseline mri features. *Neurobiology of Aging*. 2008 In Press, Corrected Proof.
- [34]. Fan Y, Batmanghelich N, Clark CM, Davatzikos C. Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*. 2008; 39(4):1731–1743. [PubMed: 18053747]
- [35]. Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in mci patients, and their use in prediction of short-term conversion to ad: Results from adni. *NeuroImage*. Feb; 2009 44(4):1415–1422. [PubMed: 19027862]
- [36]. Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, Brammer MJ, Murphy C, Murphy DG. Investigating the predictive value of whole-brain structural mr scans in autism: A pattern classification approach. *NeuroImage*. 2010; 49(1):44–56. [PubMed: 19683584]
- [37]. Fan Y, Gur RE, Gur RC, Wu X, Shen D, Calkins ME, Davatzikos C. Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study. *Biol Psychiatry*. 2008; 63(1):118–24. [PubMed: 17555720]
- [38]. Vapnik, VN. *The nature of statistical learning theory*. Springer-Verlag New York, Inc.; New York, NY, USA: 1995.
- [39]. Kloppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR, Ashburner J, Frackowiak RSJ. Automatic classification of mr scans in alzheimer's disease. *Brain*. Mar; 2008 131(3):681–689. [PubMed: 18202106]
- [40]. Vapnik, VN. *Statistical Learning Theory*. Wiley-Interscience; Sep.
- [41]. Xu, L.; Neufeld, J.; Larson, B.; Schuurmans, D. NIPS. 2005. Maximum margin clustering; p. 1537-1544.
- [42]. Kelley JJE. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*. 1960; 8(4):703–712.

- [43]. Joachims, T. Training linear svms in linear time. KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining; New York, NY, USA: ACM; 2006. p. 217-226.
- [44]. Zhao, B.; Wang, F.; Zhang, C. Cuts3vm: a fast semi-supervised svm algorithm. Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008. p. 830-838.
- [45]. Zhao, B.; Wang, F.; Zhang, C. Efficient maximum margin clustering via cutting plane algorithm. Proc. 8th SIAM International Conference on Data Mining; 2008. p. 751-762.
- [46]. Cesa-Bianchi, N.; Gentile, C.; Zaniboni, L. Hierarchical classification: combining bayes with svm. Proc. 23rd International Conference on Machine Learning; 2006. p. 177-184.
- [47]. Marszalek, M.; Schmid, C. Constructing category hierarchies for visual recognition. Proc. European Conference on Computer Vision; 2008. p. 479-491.
- [48]. Cheung, P-M.; Kwok, JT. A regularization framework for multiple-instance learning. Proceedings of the 23rd international conference on Machine learning, ser. ICML '06; New York, NY, USA: ACM; 2006. p. 193-200.
- [49]. Resnick SM, Pham DL, Kraut MA, Zonderman AB, Davatzikos C. Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain. J. Neurosci. 2003; 23(8):3295–3301. [PubMed: 12716936]
- [50]. Blessed G, Tomlinson B, Roth M. The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. Br J Psychiatry. 1968; 114:797–811. [PubMed: 5662937]
- [51]. Morris JC. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the alzheimer type. International psychogeriatrics IPA. 1997; 9(Suppl 1):173–176. Suppl 1. discussion 177–178.
- [52]. Folstein MF, Folstein SE, McHugh PR. "mini-mental state". a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res. Nov; 1975 12(3):189–198. [PubMed: 1202204]
- [53]. Delis, D.; Kramer, J.; Kaplan, E.; Ober, B. California Verbal Learning Test - Research Edition. The Psychological Corporation; New York: 1987.
- [54]. Benton, A. Revised Visual Retention Test. The Psychological Corporation; New York: 1974.
- [55]. Grober E, Hall CB, Lipton RB, Zonderman AB, Resnick SM, Kawas C. Memory impairment, executive dysfunction, and intellectual decline in preclinical alzheimer's disease. Journal of the International Neuropsychological Society. 2008; 14:266–78. [PubMed: 18282324]
- [56]. Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. Neuroimage. 2001; 14(6):1361–1369. [PubMed: 11707092]
- [57]. Pham DL, Prince JL. Adaptive fuzzy segmentation of magnetic resonance images. IEEE Trans. Med. Imaging. 1999; 18(9):737–752. [PubMed: 10571379]
- [58]. Kabani, NJ.; MacDonald, D.; Holmes, CJ.; Evans, AC. 3d anatomical atlas of the human brain. International Conference on Functional Mapping of the Human Brain NeuroImage; Jun. 1998
- [59]. Shen D, Davatzikos C. Hammer: Hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imag. 2002; 21(11):1421–1439.
- [60]. Cohen A, Daubechies I, Feauveau JC. Biorthogonal bases of compactly supported wavelets. Communications on Pure and Applied Mathematics. 1992; 45(5):485–560.
- [61]. Semi-supervised cluster analysis of imaging data. NeuroImage. 2011; 54(3):2185–2197. [PubMed: 20933091]
- [62]. Driscoll I, Davatzikos C, An Y, Wu X, Shen D, Kraut M, Resnick SM. Longitudinal pattern of regional brain volume change differentiates normal aging from MCI. Neurology. Jun; 2009 72(22):1906–1913. [PubMed: 19487648]
- [63]. Thambisetty M, Wan J, Carass A, An Y, Prince JL, Resnick SM. Longitudinal changes in cortical thickness associated with normal aging. NeuroImage. 2010; 52(4):1215–1223. [PubMed: 20441796]
- [64]. Zhao, B.; Wang, F.; Zhang, C. Efficient multiclass maximum margin clustering. ICML '08: Proceedings of the 25th international conference on Machine learning; New York, NY, USA: ACM; 2008. p. 1248-1255.

- [65]. Howieson DB, Carlson NE, Moore MM, Wasserman D, Abendroth CD, Payne-Murphy J, Kaye JA. Trajectory of mild cognitive impairment onset. *J Int Neuropsychol Soc.* 2008; 14(2):192–8. [PubMed: 18282317]
- [66]. Greenaway MC, Lacritz LH, Binegar D, Weiner MF, Lipton A, Munro Cullum C. Patterns of verbal memory performance in mild cognitive impairment, alzheimer disease, and normal aging. *Cognitive and behavioral neurology official journal of the Society for Behavioral and Cognitive Neurology.* 2006; 19(2):79–84. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/16783130>. [PubMed: 16783130]
- [67]. Royall DR, Palmer R, Chiodo LK, Polk MJ. Decline in learning ability best predicts future dementia type: the freedom house study. *Experimental Aging Research.* 2003; 29(4):385–406. [PubMed: 12959874]
- [68]. Filipovych R, Davatzikos C. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (mci). *Neuroimage.* 2011; 55(3):1109–19. [PubMed: 21195776]

Fig. 1.

Schematic example that might relate to aging. A homogeneous normal population gradually evolving into a heterogeneous cognitively declined population with two distinct clusters (*i.e.*, Diagnosis 1 and Diagnosis 2). Usually, one has confidence in labels corresponding to only the extreme cases (*i.e.*, the people that would have a fairly definitive and reliable diagnosis, such as AD, or else the very healthy). The task is then to build a classification hyperplane such, that the two clusters in the obtained class of cognitively declined individuals can be reliably detected.

Fig. 2.
Example of a smoothed tissue density map of GM.

Fig. 3.
Two patterns of pathology that represent two different directions of disease evolution.

Fig. 4.

Setup of the simulated experiment. 72 images assigned into normal group, and 71 images assigned into abnormal group. Additionally, 36 images from the abnormal group have simulated Pathology 1, while 35 images from the abnormal group have simulated Pathology 2. 24 images from the simulated normal group are labeled as such. Additionally, 24 images from the abnormal group (12 images with Pathology 1 and 12 with Pathology 2) are labeled as abnormal.

Fig. 5. Classification accuracy as a function of pathology distinctiveness, where the pathology distinctiveness was modeled by the mean μ of the Gaussian distribution from which the multiplication coefficients used to introduce pathologies were drawn. The results are shown for JointMMCC, as well as for semi-supervised and fully-supervised SVM. The 95% confidence intervals are also shown.

Fig. 6. Classification specificity as a function of pathology distinctiveness. The results are shown for JointMMCC, as well as for semi-supervised and fully-supervised SVM.

Fig. 7. Classification sensitivity as a function of pathology distinctiveness. The results are shown for JointMMCC, as well as for semi-supervised and fully-supervised SVM.

Fig. 8. Clustering accuracy as a function of pathology distinctiveness. The results are shown for JointMMCC, as well as for the application of SVM (or TSVM) to identify the abnormal population, followed by MMC to discover clusters within it. The 95% confidence intervals are also shown.

Fig. 9. Average between the classification and clustering accuracies as a function of pathology distinctiveness. The results are shown for JointMMCC, as well as for the application of SVM (or TSVM) to identify the abnormal population, followed by MMC to discover clusters within it. The 95% confidence intervals are also shown.

Fig. 10.

Accuracy of assigning abnormal images into clusters as a function of pathology distinctiveness. The measure considers only the simulated abnormal images that were also classified as abnormal. The results are shown for JointMMCC, as well as for the application of SVM (or TSVM) to identify the abnormal population, followed by MMC to discover clusters within it. The 95% confidence intervals are also shown.

Fig. 11.
Setup of the experiment for clustering cognitively less stable populations.

Fig. 12. Voxel-wise differences in baseline images of the subjects in the cognitively most stable and least stable subpopulations. The colormap shows t -statistic and was thresholded at $p < 0.05$. (Differences thresholded at $p < 0.001$ are provided in the supplemental materials).

Fig. 13.

LOO evaluation scheme used in the real-life experiments. At each run of the LOO we remove one labeled image from the experiment. We then apply our JointMMCC algorithm to obtain a classification hyperplane that separates cognitively more stable and cognitively less stable subjects, as well as a clustering hyperplane that identifies clusters in the cognitively less stable subpopulation. The cluster containing most of the seeded MCI subjects is identified as the “MCI-like cluster”. The other cluster is identified as the “Non-MCI cluster”. The classification and clustering results are aggregated over all LOO runs.

Fig. 14. Voxel-wise differences in baseline images of the subjects in the cognitively stable and cognitively declined subpopulations. The colormap shows t -statistic and was thresholded at $p < 0.05$. (Differences thresholded at $p < 0.001$ are provided in the supplemental materials).

Fig. 15.

Voxel-wise differences in baseline images of the subjects in the two clusters. The colormap shows t -statistic and was thresholded at $p < 0.05$. (Differences thresholded at $p < 0.001$ are provided in the supplemental materials).

TABLE I

Sets of parameter values considered in the parameters optimization stage of the approach.

Parameter	Values
β_l	$10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2$
β_u	$10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2$
$\hat{\beta}$	$10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2$
γ	$10^0, 10^1, 10^2, 10^3$
r	$10^{-15}, 10^{-5}, 10^{-3}, 10^{-1}, 10^1$
$\hat{r}$	$10^{-15}, 10^{-5}, 10^{-3}, 10^{-1}, 10^1$
Number of features (d)	10, 100, 500, 1000
Decomposition level (m)	3, 5, 7, 9, 11

TABLE II
JointMMCC: List of median parameter values estimated on the simulated training data.

μ	m	d	β_1	β_u	β	r	$\hat{r}$	γ
1.0	8	500	1	0.001	1	0.00001	0.0005	55
1.1	7	500	10	0.001	5.5	0.1	5.05	100
1.2	10	100	55	0.001	0.1	0.1	0.001	100
1.3	6	1000	10	0.001	10	0.1	0.001	10
1.4	11	1000	55	0.0055	0.505	10	0.001	100
1.5	5	1000	1	0.0055	5.5	5.05	0	1
1.6	5	1000	10	0.01	10	0.001	0	1
1.7	11	1000	10	0.01	5.5	10	0.001	10
1.8	11	1000	5.5	0.001	0.0055	10	0.001	10
1.9	11	1000	1	0.001	5.0005	10	0.001	10
2.0	8	1000	1	0.001	0.001	10	0.001	10
2.1	5	1000	1	0.001	0.001	10	0.001	10
2.2	5	1000	1	0.001	0.001	10	0.001	10
All	8	1000	5.5	0.001	1	10	0.001	10

TABLE III

SVM+MMC: List of median parameter values estimated on the simulated training data.

μ	m	d	β_l	β_u	β	r	$\hat{r}$	γ
1.0	6	300	0.55	–	0.001	–	0.001	–
1.1	7	300	1	–	0.001	–	0.001	–
1.2	9	1000	0.1	–	0.001	–	0.001	–
1.3	9	1000	0.1	–	0.001	–	0.001	–
1.4	10	1000	0.055	–	0.001	–	0.001	–
1.5	7	1000	0.01	–	0.001	–	0.001	–
1.6	11	1000	0.055	–	1	–	1	–
1.7	11	1000	0.01	–	1	–	1	–
1.8	11	1000	0.01	–	1	–	1	–
1.9	11	1000	0.0055	–	1	–	1	–
2.0	11	1000	0.0055	–	1	–	1	–
2.1	11	1000	0.001	–	1	–	1	–
2.2	1	1000	0.001	–	1	–	1	–
All	11	1000	0.01	–	1	–	1	–

TABLE IV
TSVM+MMC: List of median parameter values estimated on the simulated training data.

μ	m	d	β_l	β_u	β	r	$\hat{r}$	γ
1.0	5	750	0.01	0.01	0.001	0.0505	0.001	–
1.1	4	500	1	0.0505	0.001	5.05	0.001	–
1.2	11	750	0.055	0.001	0.001	0.001	0.001	–
1.3	11	1000	0.505	0.001	0.001	0.001	0.001	–
1.4	11	500	1	0.01	0.01	0.001	0.001	–
1.5	11	1000	0.1	0.01	0.1	0.001	0.001	–
1.6	11	1000	1	0.001	100	10	0.001	–
1.7	11	1000	1	0.0055	100	0.1	0.001	–
1.8	11	1000	1	1	5.005	0.001	0.001	–
1.9	11	1000	0.01	1	50.005	0.001	0.001	–
2.0	8	1000	1	0.1	100	0.001	0.001	–
2.1	5	500	1	1	0.001	0.001	0.001	–
2.2	11	1000	1	1	0.001	5.05	0.001	–
All	11	1000	1	0.01	0.01	0.001	0.001	–

TABLE V

Differences in cognitive scores between the obtained classes.

Stable vs. declined	
CVLT List A Sum	$p_w=0.0640$
CVLT Long Delay Free	$p_w=0.4684$
BVRT Errors	$p_w=0.7525$
MMSE	$p_w=0.5427$
Slope of CVLT List A Sum	$p=0.0755$
Slope of CVLT Long Delay Free	$p_w=0.0749$
Slope of MMSE	$p_w=0.3876$

TABLE VI

Differences in cognitive scores between the obtained clusters.

Clusters: MCI-like vs. Non-MCI	
CVLT List A Sum	$p_w=0.1298$
CVLT Long Delay Free	$p_w=0.0417$
BVRT Errors	$p_w=0.1578$
MMSE	$p_w=0.4492$
Slope of CVLT List A Sum	$p=0.0749$
Slope of CVLT Long Delay Free	$p_w=0.0040$
Slope of MMSE	$p_w=0.4545$

Algorithm 1**Cutting Plane Algorithm for JointMMCC**

-
- 1** Select $\beta_f, \beta_w, \hat{\beta}, \epsilon$, and set $\Omega = \emptyset$;
 - 2** Solve optimization Problem 20 under the current working constraint set Ω ;
 - 3** Select the most violated constraint $\mathbf{C} = (\mathbf{c}, \hat{\mathbf{c}})$ under the current classification hyperplane $\mathbf{w}$ and clustering hyperplane $\hat{\mathbf{w}}$ following Equations 15–17;
 - 4** If the selected constraint is violated by no more than ϵ , goto step 5; Otherwise set $\Omega = \Omega \cup \{\mathbf{C}\}$, goto step 2;
 - 5** Return the class labels for unlabeled samples and cluster labels for samples in the positive class using Equations 7 and 8, respectively.
-

Algorithm 2*CCCP* for solving Problem 20

-
- 1 Initialize $(\mathbf{w}_0, b_0, \hat{\mathbf{w}}_0, \hat{b}_0)$ with values from the last iteration of Algorithm 1;
 - 2 Find $(\mathbf{w}_t, b_t, \hat{\mathbf{w}}_t, \hat{b}_t)$ as the solution to the *QP* Problem (30);
 - 3 If convergence criterion satisfied, return $(\mathbf{w}_t, b_t, \hat{\mathbf{w}}_t, \hat{b}_t)$ as the optimal hyperplanes' parameters; otherwise $t = t + 1$, goto step 2.
-