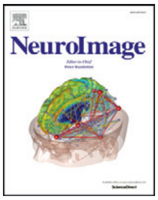
NeuroImage 145 (2017) 346–364

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg



HYDRA: Revealing heterogeneity of imaging and genetic patterns through a multiple max-margin discriminative analysis framework

Erdem Varol *, Aristeidis Sotiras, Christos Davatzikos, for the Alzheimer's Disease Neuroimaging Initiative [1]

Section for Biomedical Image Analysis, Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Accepted 12 February 2016
Available online 23 February 2016

Keywords:
Semi-supervised pattern analysis
Multivariate
Max-margin classification
Convex polytope
SVM
Clustering
MRI
Genetics
Neuroimaging
Heterogeneity
Aging
Alzheimer's disease
ADNI

ABSTRACT

Multivariate pattern analysis techniques have been increasingly used over the past decade to derive highly sensitive and specific biomarkers of diseases on an individual basis. The driving assumption behind the vast majority of the existing methodologies is that a single imaging pattern can distinguish between healthy and diseased populations, or between two subgroups of patients (*e.g.*, progressors vs. non-progressors). This assumption effectively ignores the ample evidence for the heterogeneous nature of brain diseases. Neurodegenerative, neuropsychiatric and neurodevelopmental disorders are largely characterized by high clinical heterogeneity, which likely stems in part from underlying neuroanatomical heterogeneity of various pathologies. Detecting and characterizing heterogeneity may deepen our understanding of disease mechanisms and lead to patient-specific treatments. However, few approaches tackle disease subtype discovery in a principled machine learning framework. To address this challenge, we present a novel non-linear learning algorithm for simultaneous binary classification and subtype identification, termed HYDRA (Heterogeneity through Discriminative Analysis). Neuroanatomical subtypes are effectively captured by multiple linear hyperplanes, which form a convex polytope that separates two groups (*e.g.*, healthy controls from pathologic samples); each face of this polytope effectively defines a disease subtype. We validated HYDRA on simulated and clinical data. In the latter case, we applied the proposed method independently to the imaging and genetic datasets of the Alzheimer's Disease Neuroimaging Initiative (ADNI 1) study. The imaging dataset consisted of T1-weighted volumetric magnetic resonance images of 123 AD patients and 177 controls. The genetic dataset consisted of single nucleotide polymorphism information of 103 AD patients and 139 controls. We identified 3 reproducible subtypes of atrophy in AD relative to controls: (1) diffuse and extensive atrophy, (2) precuneus and extensive temporal lobe atrophy, as well some prefrontal atrophy, (3) atrophy pattern very much confined to the hippocampus and the medial temporal lobe. The genetics dataset yielded two subtypes of AD characterized mainly by the presence/absence of the apolipoprotein E (APOE) $\varepsilon 4$ genotype, but also involving differential presence of risk alleles of CD2AP, SPON1 and LOC39095 SNPs that were associated with differences in the respective patterns of brain atrophy, especially in the precuneus. The results demonstrate the potential of the proposed approach to map disease heterogeneity in neuroimaging and genetic studies.

© 2016 Elsevier Inc. All rights reserved.

Introduction

Automated analysis of spatially aligned medical images has become the main framework for studying the anatomy and function of the human brain. This is typically performed by either employing voxel-based (VBA) or multivariate pattern analysis (MVPA) techniques.

VBA complements region of interest (ROI) volumetry by providing a comprehensive assessment of anatomical differences throughout the brain, while not being limited by *a priori* regional hypotheses. VBA typically performs mass-univariate statistical tests on either tissue composition or deformation fields, aiming to reveal regional anatomical or shape differences (Ashburner et al., 1998; Goldszal et al., 1998; Ashburner and Friston, 2000; Davatzikos et al., 2001; Chung et al., 2001; Fox et al., 2001; Job et al., 2002; Kubicki et al., 2002; Chung et al., 2003; Studholme et al., 2004; Bernasconi et al., 2004; Giuliani et al., 2005; Job et al., 2005; Meda et al., 2008; Ashburner, 2009). However, voxel-wise methods often suffer from low statistical power and more importantly, ignore multivariate relationships in the data.

* Corresponding author at: Section for Biomedical Image Analysis, Center for Biomedical Image Computing and Analytics, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, USA. Fax: +1 215 614 0266.

E-mail address: erdem.varol@uphs.upenn.edu (E. Varol).

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

http://dx.doi.org/10.1016/j.neuroimage.2016.02.041
1053-8119/© 2016 Elsevier Inc. All rights reserved.

On the other hand, MVPA techniques have gained significant attention due to their ability to capture complex relationships of imaging signals among brain regions. This property allows to better characterize group differences and could potentially lead to improved diagnosis and personalized prognosis. As a consequence, machine learning methods have been used with increased success to derive highly sensitive and specific biomarkers of diseases on individual basis (Mourão Miranda et al., 2005; Klöppel et al., 2008; Davatzikos et al., 2008; Vemuri et al., 2008; Duchesne et al., 2008; Sabuncu et al., 2009; McEvoy et al., 2009; Ecker et al., 2010; Hinrichs et al., 2011; Cuingnet et al., 2011).

A common assumption behind both VBA and MVPA methods is that there is a single pattern that distinguishes the two contrasted groups. In other words, most computational neuroimaging analyses assume a single unifying pathophysiological process and perform a monistic analysis to identify it. However, this approach ignores the heterogeneous nature of diseases, which is supported by ample evidence. Typical examples of brain disorders that are characterized by a heterogeneous clinical presentation include both neurodevelopmental and neurodegenerative disorders: autism spectrum disorder (ASD) comprises neurodevelopmental disorders characterized by deficits in social communication and repetitive behaviors (Geschwind and Levitt, 2007; Jeste and Geschwind, 2014); schizophrenia and Parkinson's disease can be subdivided into distinct groups by separating its symptomatology to discrete symptom domains (Buchanan and Carpenter, 1994; Graham and Sagar, 1999; Koutsouleris et al., 2008; Nenadic et al., 2010; Zhang et al., 2015; Lewis et al., 2005); Alzheimer's disease (AD) can be separated into three subtypes on the basis of the distribution of neurofibrillary tangles (Murray et al., 2011); and mild cognitive impairment (MCI) may be further classified based on the type of specific cognitive impairment (Huang et al., 2003; Whitwell et al., 2007).

Disentangling disease heterogeneity may significantly contribute to our understanding and lead to a more accurate diagnosis, prognosis and targeted treatment. However, few research efforts have been focused on revealing the inherent disease heterogeneity. These approaches can be categorized into two distinct classes. The first class assumes an *a priori* subdivision of the diseased samples into coherent groups, based on independent (*e.g.*, clinical) criteria, and opts to identify group-level anatomical or functional differences using univariate statistical methods (Huang et al., 2003; Koutsouleris et al., 2008; Nenadic et al., 2010; Whitwell et al., 2012; Zhang et al., 2015). As a consequence, multivariate relationships in the data are ignored. Moreover, and more importantly, these methods depend on an *a priori* disease subtype definition, which may be either difficult to obtain (*e.g.*, from autopsy near the date of imaging), or noisy and non-specific (*e.g.*, cognitive or clinical evaluations). Methods belonging to the second class apply multivariate clustering (typically driven by all image elements) directly to the diseased population towards segregating subsets of distinct anatomical subtypes (Graham and Sagar, 1999; Whitwell et al., 2007; Lewis et al., 2005; Noh et al., 2014). Such an approach aims to cluster brain anatomies instead of pathological patterns. Thus, it has the potential risk of estimating clusters that reflect normal inter-individual variability, some of which is due to sex, age and other confounds, instead of highlighting disease heterogeneity.

In order to tackle the aforementioned limitations, it is necessary to develop a principled machine learning approach that is able to simultaneously identify a class of pathological samples and separate them into coherent subgroups based on multivariate pathological patterns. To the best of our knowledge, one approach has been previously proposed in this direction (Filipovych et al., 2012). That work tackled disease subtype discovery by simultaneously solving classification and clustering in a semi-supervised maximum margin framework. It jointly estimated two hyperplanes, one that separates the diseased population from the healthy one, and another hyperplane that splits the estimated diseased population into two groups. Thus, only one linear classifier was used to separate patients from controls, thereby limiting its ability to capture heterogeneous pathologic processes. Moreover, it arbitrarily assumed that exactly two disease subgroups exist, rather than attempting to determine the number of subtypes from the data.

Here, we propose a novel non-linear semi-supervised[2] machine learning algorithm for integrated binary classification and subpopulation clustering aiming to reveal heterogeneity through discriminant analysis (HYDRA). To the best of our knowledge, ours is the first algorithm to deal with anatomical/genetic heterogeneity in a supervised-clustering fashion with arbitrary number of clusters. The proposed approach is motivated by recent machine learning methods that derive non-linear classifiers through the use of multiple-hyperplanes (Fu et al., 2010; Gu and Han, 2013; Varol and Davatzikos, 2014; Kantchelian et al., 2014; Takács, 2009; Osadchy et al., 2015). Classification is performed through the separation of healthy controls from pathological samples by a convex polytope that is formed by combining multiple linear max-margin classifiers. Heterogeneity is disentangled by implicitly clustering pathologic samples through their association to single linear sub-classifiers. Multiple dimensions of heterogeneity may be captured by varying the number of estimated hyperplanes (faces of the polytope). This is in contrast to non-linear kernel classification methods which may accurately fit to heterogeneous data in terms of disease prediction, but do not provide any explicit clustering information that can be used to determine subtypes of pathology. HYDRA is a hybrid between unsupervised clustering and supervised classification methods; it can simultaneously fit maximum margin classification boundaries and elucidate disease subtypes, which is not possible with neither unsupervised clustering methods nor non-linear kernel classifiers.

Note that a preliminary version of this work was presented in (Varol et al., 2015). The current paper extends our previous work in multiple ways: (i) A more sophisticated initialization scheme based on determinental point processes is employed (Sec. Appendix A.1); (ii) the sensitivity to initialization due to the non-convexity of the objective function has been improved by using multiple initializations and consensus strategies (Sec. Appendix A.4); (iii) a symmetric version of the algorithm is developed towards accounting for the heterogeneity of the healthy controls and avoiding over-learning (Section 2.4); (iv) a detailed description of the proposed methodology is provided; (v) we extensively evaluate our method, HYDRA, by using additional (imaging and genetic) datasets and comparing it to unsupervised clustering and non-linear classification methods.

The remainder of this paper is organized as follows. In Section 2, we detail the proposed approach. Next, we experimentally validate our method using synthetic (Section 3) and clinical (Section 4) data. We discuss the results in Section 5, while Section 6 concludes the paper with our final remarks.

Method

In high dimensional spaces, the modeling capacity of linear support vector machines (SVMs) is theoretically rich enough to discriminate between two homogeneous classes. However, while two classes are linearly separable with high probability, the resulting margin may be small. This case arises, for example, when one class is generated by a multimodal distribution that models a heterogeneous process (see Fig. 1a). This may be remedied by the use of non-linear classifiers, allowing for larger margins and thus, better generalization. However, while kernel methods, such as Gaussian radial basis function (GRBF) kernel SVM, provide non-linearity, they lack interpretability when aiming to characterize heterogeneity.

[2] The term semi-supervised is in reference to lack of disease subtype labels that must be inferred from data.

Download English Version:

https://daneshyari.com/en/article/5631585

Download Persian Version:

https://daneshyari.com/article/5631585

Daneshyari.com