NeuroImage xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Control-group feature normalization for multivariate pattern analysis of structural MRI data using the support vector machine☆

Kristin A. Linn [a,*], Bilwaj Gaonkar [c], Theodore D. Satterthwaite [b], Jimit Doshi [c], Christos Davatzikos [c], Russell T. Shinohara [a]

[a] Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA
[b] Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
[c] Center for Biomedical Image Computing and Analytics (CBICA), University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 3 October 2015
Accepted 14 February 2016
Available online xxxx

Keywords:
Feature normalization
Multivariate pattern analysis
Structural MRI
Support vector machine

ABSTRACT

Normalization of feature vector values is a common practice in machine learning. Generally, each feature value is standardized to the unit hypercube or by normalizing to zero mean and unit variance. Classification decisions based on support vector machines (SVMs) or by other methods are sensitive to the specific normalization used on the features. In the context of multivariate pattern analysis using neuroimaging data, standardization effectively up- and down-weights features based on their individual variability. Since the standard approach uses the entire data set to guide the normalization, it utilizes the total variability of these features. This total variation is inevitably dependent on the amount of marginal separation between groups. Thus, such a normalization may attenuate the separability of the data in high dimensional space. In this work we propose an alternate approach that uses an estimate of the control-group standard deviation to normalize features before training. We study our proposed approach in the context of group classification using structural MRI data. We show that control-based normalization leads to better reproducibility of estimated multivariate disease patterns and improves the classifier performance in many cases.

1. Introduction

Machine learning classification algorithms such as the support vector machine (SVM) (Cortes and Vapnik, 1995; Vapnik, 2013) are often used to map high-dimensional neuroimaging data to a clinical diagnosis or decision. Structural and functional magnetic resonance imaging (MRI) are promising tools for building biomarkers to diagnose, monitor, and treat neurological and psychological illnesses. Mass-univariate methods such as statistical parametric mapping (Frackowiak et al., 1997; Friston et al., 1991, 1994) and voxel- based morphometry (Ashburner and Friston, 2000; Davatzikos et al., 2001) test for marginal disease effects at each voxel, ignoring complex spatial correlations and multivariate relationships among voxels. As a result, methods have emerged for performing multivariate pattern analysis (MVPA) that leverage the information contained in the covariance structure of the images to discriminate between the groups being studied (Craddock et al., 2009; Cuingnet et al., 2011; Davatzikos et al., 2005, 2008, 2009, 2011; De Martino et al., 2008; Fan et al., 2007; Klöppel et al., 2008; Koutsouleris et al., 2009; Langs et al., 2011; Mingoia et al., 2012; Mourão-Miranda et al., 2005; Pereira, 2007; Richiardi et al., 2011; Sabuncu and Van Leemput, 2011; Vemuri et al., 2008; Venkataraman et al., 2012; Wang et al., 2007; Xu et al., 2009; Reiss and Ogden, 2010; Gaonkar and Davatzikos, 2013). Identifying multivariate structural and functional signatures in the brain that discriminate between groups may lead to a better understanding of disease processes and is therefore of great interest in the field of neuroimaging research.

The SVM is a common choice for estimating multivariate patterns in the brain because it is amenable to high-dimensional, low sample size data. Our focus in this work is on patterns in the brain that reflect structural changes due to disease. However, the methods apply more generally to applications of MVPA using BOLD measurements from fMRI data or measures of connectivity across the brain. The SVM takes as input image-label pairs and returns a decision function that is a weighted sum of the imaging features. The estimated weights reflect the joint contribution of the imaging features to the predicted class label.

Machine learning methods in general, and SVMs in particular, are sensitive to differences in feature scales. For example, a SVM will place more importance on a feature that takes values in the range of [1000,2000] than a feature that takes values in the interval [1,2]. This is because the former tends to have a stronger influence on the Euclidean distance between feature vector realizations and therefore drives the SVM optimization. To give all voxels or regions of interest equal importance during classifier training, it is common practice to implement feature-wise standardization in some way, either by normalizing

☆ For the Alzheimer's disease neuroimaging initiative.
* Corresponding author.
E-mail address: klinn@upenn.edu (K.A. Linn).

http://dx.doi.org/10.1016/j.neuroimage.2016.02.044
1053-8119/

each to have mean zero and unit variance or by scaling to a common domain. For example, (Peng et al., 2016) scale each feature to be in the interval [0,1], and (Hanke et al., 2016; Zacharaki et al., 2009; Etzel et al., 2011; Wang et al., 2012; Sato et al., 2012) normalize to mean zero and unit variance. Such a preprocessing step, while common in practice, tends to be applied without weighing the consequent ramifications in a careful manner. Careful consideration must be given to the choice of feature normalization, as it is directly tied to the relative magnitude of the estimated SVM weights and thus the performance and interpretation of the classifier. While the original idea of feature scaling dates back to the universal approximation theorem from the neural network literature, it has not been explored in detail in the context of neuroimaging and MVPA. This is the object of this manuscript.

The rest of this paper is organized as follows: in Section 2, we provide a brief introduction to MVPA using the SVM, review two popular feature normalization methods, and propose an alternative based on the control-group variability. Using simulations, we compare the performance of different feature normalization techniques in Section 3, followed by an investigation of the effects of feature normalization on an analysis of data from healthy controls and patients with Alzheimer's disease. We include a discussion in Section 4 and concluding remarks in Section 5.

2. Material and methods

2.1. Multivariate pattern analysis using the SVM

Let $(Y_i, X_i^T)^T$, $i = 1, \ldots, n$, denote n independent and identically distributed observations of the random vector $(Y, X^T)^T$, where $Y \in \{-1, 1\}$ denotes the group label, and $X \in \mathbb{R}^p$ denotes a vectorized image with p voxels. A popular MVPA tool used in the neuroimaging community is the SVM (Cortes and Vapnik, 1995; Vapnik, 2013). SVMs are known to work well for high dimension, low sample size data (Schölkopf et al., 2004). Such data are common in the neuroimaging-based diagnostic setting. Henceforth, we focus on MVPA using the SVM.

The hard-margin linear SVM solves the constrained optimization problem

$$\begin{aligned} &\arg\min_{v,b} \frac{1}{2}\|v\|^2 \\ &\text{such that } Y_i(v^T X_i + b) \geq 1 \quad \forall i = 1, \ldots, n, \end{aligned} \tag{1}$$

where $b \in \mathbb{R}$ and $v \in \mathbb{R}^p$ are parameters that describe the classification function. For a given set of training data, let the solution to (1) be denoted by $(\tilde{v}, \tilde{b})$. Then, for a new observation X^{new} with unknown label Y^{new}, the classification function $c(X^{new}) = \text{sign}(\tilde{v}^T X^{new} + \tilde{b})$ returns a predicted group label.

When the data from the two groups are not linearly separable, the soft-margin linear SVM allows some training observations to be either misclassified or fall in the SVM margin through the use of slack variables ξ_i with associated cost parameter C. In this case, the optimization problem becomes

$$\begin{aligned} &\arg\min_{v,b,\xi} \frac{1}{2}\|v\|^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{such that :} \\ &Y_i(v^T X_i + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n, \\ &\xi_i \geq 0 \quad \forall i = 1, \ldots, n, \end{aligned} \tag{2}$$

where $C \in \mathbb{R}$ is a tuning parameter that penalizes misclassification, and $\xi = (\xi_1, \xi_2, \ldots, \xi_n)^T$ is the vector of slack variables. For details about solving optimization problems (1) and (2) we refer the reader to (Hastie et al., 2001).

In high-dimensional problems where the number of features is greater than the number of observations, the data are almost always separable by a linear hyperplane (Orrù et al., 2012). However, when applying MVPA to region of interest (ROI) data such as volumes of subregions in the brain, the data may not be linearly separable. In this case, the choice of C is critical to classifier performance and generalizability. Examples of MVPA using the SVM include classification of multiple sclerosis patients into disease subgroups (Bendfeldt et al., 2012), the study of Alzheimer's disease (Cuingnet et al., 2011; Davatzikos et al., 2011), and various classification tasks involving patients with depression (Costafreda et al., 2009; Gong et al., 2011; Liu et al., 2012). This is only a small subset of the relavant literature, which demonstrates the widespread popularity of the approach.

2.2. SVM Feature normalization for MVPA

The choice of feature normalization affects the estimated weight pattern of a SVM and can lead to vastly different conclusions about the underlying disease process. Two widely implemented approaches are to (*i*) normalize each feature to have mean zero and unit variance, and (*ii*) scale each feature to have a common domain such as [0,1]. Henceforth, we will refer to (*i*) as *standard normalization* and (*ii*) as *domain standardization* (Pedregosa et al., 2011).

Let μ_j and σ_j denote the mean and standard deviation of the j^{th} feature, $j = 1, \ldots, p$. Denote the corresponding empirical estimates by $\bar{X}_j = n^{-1}\sum_{i=1}^{n} X_{i,j}$ and $\hat{\sigma}_j = \{(n-1)^{-1}\sum_{i=1}^{n}(X_{i,j} - \bar{X}_j)^2\}^{1/2}$. Then, subject i's standard-normalized j^{th} feature is calculated as

$$X_{i,j}^Z = \frac{X_{i,j} - \bar{X}_j}{\sigma_j}.$$

Alternatively, subject i's domain-scaled j^{th} feature is calculated as

$$X_{i,j}^U = \frac{X_{i,j} - \min_i X_{i,j}}{\min_i X_{i,j} - \min_i X_{i,j}}.$$

One potential drawback of using domain scaling is the instability of the minimum and maximum order statistics, especially in small sample sizes. This may introduce bias in the estimated weight pattern by up- and down-weighting features in an unstable way. In comparison, the standard normalization may seem relatively stable. However, it implicitly depends on the relative sample size of each group and the separability between groups. To see this, let f_{X_j} denote the marginal distribution of X_j, with mean μ_j and variance σ_j^2. Let $f_{X_j \mid Y=y}$ denote the conditional distribution of X_j given $Y = y$ with mean $\mu_{j,y}$ and variance $\sigma_{j,y}^2$. In addition, let $\gamma = \text{pr}(Y = 1)$. Then, $\mu_j = \gamma\mu_{j,1} + (1-\gamma)\mu_{j,-1}$ and

$$\begin{aligned} \sigma_j^2 &= E\left(X_j - \mu_j\right)^2 \\ &= EX_j^2 - \mu_j^2 \\ &= \int x_j^2 \left\{\gamma f_{X_j \mid Y=1}(x) + (1-\gamma) f_{X_j \mid Y=-1}(x)\right\} dx - \mu_j^2 \\ &= \gamma\left(\sigma_{j,1}^2 + \mu_{j,1}^2\right) + (1-\gamma)\left(\sigma_{j,-1}^2 + \mu_{j,-1}^2\right) - \mu_j^2. \end{aligned}$$

After simplification, the previous expression can be written as

$$\sigma_j^2 = \gamma\sigma_{j,1}^2 + (1-\gamma)\sigma_{j,-1}^2 + \gamma(1-\gamma)\left(\mu_{j,1} - \mu_{j,-1}\right)^2. \tag{3}$$

The right-hand side of expression (3) shows that the variance of feature j depends on a mixture of the conditional variances of both classes and a term that depends on the squared Euclidean distance between their marginal means. Larger marginal separability of feature j will lead to a larger estimate of the pooled standard deviation used for normalization. Thus, normalizing by the pooled standard deviation can in some cases harshly penalize, or down-weight, features that have good

Please cite this article as: Linn, K.A., et al., Control-group feature normalization for multivariate pattern analysis of structural MRI data using the support vector machine, NeuroImage (2016), http://dx.doi.org/10.1016/j.neuroimage.2016.02.044