Published in final edited form as:

Med Image Comput Assist Interv. 2012 ; 15(0 1): 723–730.

Deriving statistical significance maps for SVM based image classification and group comparisons

Bilwaj Gaonkar and Christos Davatzikos

Section for Biomedical Image Analysis, University of Pennsylvania, Philadelphia, PA 19104, USA

Abstract

Population based pattern analysis and classification for quantifying structural and functional differences between diverse groups has been shown to be a powerful tool for the study of a number of diseases, and is quite commonly used especially in neuroimaging. The alternative to these pattern analysis methods, namely mass univariate methods such as voxel based analysis and all related methods, cannot detect multivariate patterns associated with group differences, and are not particularly suitable for developing individual-based diagnostic and prognostic biomarkers. A commonly used pattern analysis tool is the support vector machine (SVM). Unlike univariate statistical frameworks for morphometry, analytical tools for statistical inference are unavailable for the SVM. In this paper, we show that null distributions ordinarily obtained by permutation tests using SVMs can be analytically approximated from the data. The analytical computation takes a small fraction of the time it takes to do an actual permutation test, thereby rendering it possible to quickly create statistical significance maps derived from SVMs. Such maps are critical for understanding imaging patterns of group differences and interpreting which anatomical regions are important in determining the classifier's decision.

1 Significance

Precise quantification of group differences using medical images central in scientific studies of the effects of disease on the human body. The dominant approach addressing this problem involves performing independent statistical testing either pixel/voxel-wise [1] or regions of interest (ROI-wise) in the image. It has been argued that such univariate analysis might miss group difference patterns that span multiple voxels or regions [4]. Hence, replacing univariate methods by multivariate methods such as SVMs [8] [11][7] has been discussed in literature. However, unlike univariate methods [1], SVMs do not naturally provide statistical tests (and corresponding p-values) associated with every voxel/region of an image. Permutation testing has been suggested for interpreting SVM output for such high dimensional data [6]. However, performing these tests is time consuming and computationally costly. Hence, we developed and validated an analytical approximation for SVM permutation tests that allows for tremendous computational speed up. Section 2 of this manuscript presents the theory that allows us to achieve this speed up. Section 3 presents experiments that validate the theory. Section 4 presents a brief discussion and possible avenues for further development.

2 Analysis

2.1 Support Vector Machines: Background

The support vector machine [10] is a powerful pattern classification engine that was first proposed in the nineties. In medical imaging it has been used to distinguish cognitively abnormal people from controls based on their brain MR images. For completeness we briefly explain the concept of the SVM next. To use SVMs we stack preprocessed image data into a large rectangular matrix $X \in R^{m \times p}$ whose rows $\mathbf{x}_i$ index individuals in the

population, and columns index image voxels. Also, with every individual $\mathbf{x}_i$ we associate a binary label $y_i \in +1, -1$ which indicates the presence or absence of disease. Note that the $\mathbf{x}_i$ live in a Euclidean space of dimension p . The SVM finds the largest margin hyperplane parameterized by the direction $\mathbf{w}^* \in \mathbb{R}^p$ that separates the patients from the controls (or two groups, in general). This concept is illustrated for 2-dimensional space in figure 1. The SVM is typically formulated as follows:

$$\begin{aligned} \mathbf{w}^*, \quad b^* = \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \\ \text{subj.to.} \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \quad i=1, \dots, m \end{aligned} \quad (1)$$

For unseen data the SVM simply uses the position of the unseen data relative to the learnt hyperplane, $\mathbf{w}^*, b^*$ to decide upon disease status. Although SVM performance can be assessed by cross validation, it is equally important to be able to interpret which regions/features are important in deriving the classifier's decision. Such interpretations are particularly desirable by clinicians interested in understanding how disease affects anatomy and function and which features they should be attending to in interpreting medical images. Currently the only way to quantify importance of individual voxels to the SVM classification is through the use of permutation tests. The concept of permutation testing is illustrated in figure 1. Briefly, the data labels y_i are permuted randomly. For each random permutation an SVM is trained and a hyperplane $\mathbf{w}$ is found. After many permutations we can generate an approximation to the null distribution of each component of $\mathbf{w}$. Comparing the components of $\mathbf{w}^*$ with these null distributions allows for statistical inference. The inference procedure described above is based on [6]. Permutation testing is computationally expensive and becomes increasingly difficult as dataset size increases. In this paper we develop an analytical approximation of permutation testing using SVMs for medical imaging data.

2.2 The analytical approximation of permutation testing for SVMs

Key assumption—The proposed analytic framework is based on one key observation about permutation testing with SVMs while using medical imaging data, as well as any high-dimensionality data observed via dramatically smaller number of samples. This observation is that for a great majority of the permutations, the number of support vectors (SVs) in the learnt models equal or almost equal the total number of training subjects themselves. This behavior and the corresponding assumption can be partly justified by appealing to a known generalization error bound for SVMs (See eqn 93 of [3])

$$E[P(\text{Error})] \leq \frac{E[\text{Number of support vectors}]}{\text{Number of training samples}} \quad (2)$$

Here $E[P(\text{Error})]$ is a measure of the generalization/test error of the SVM. Since most permutations are completely random, we do not expect the corresponding learnt models to generalize well to new training data. Thus, for most permutations the number of SVs tends to the total number of subjects themselves. This key assumption allows us to simplify the statistical analysis of SVMs while using medical imaging data. Using the theory developed under this assumption we are able to predict the nature of the null distribution for a variety of medical imaging datasets (see figure 2).

Theory—Note that VC-theory dictates that linear classifiers shatter high dimension low sample size data. Hence, for any permutation of $\mathbf{y}$ one can always find a separating

hyperplane that perfectly separates the training data. Thus, we choose use the hard margin support vector machine formulation from [10] instead of (1) for further analysis in this paper. The hard margin support vector machine (see [10]) can then be written as:

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subj.to. } & y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \quad \forall i \in \{1, \dots, m\} \end{aligned}$$

It is required (see [2]) that for the 'support vectors' (indexed by $j \in \{1, 2, \dots, n_{SV}\}$) we have $\mathbf{w}^T \mathbf{x}_j + b = y_j \forall j$. Now, if all our data were support vectors this would allow us to write the constraints in optimization (3) as $\mathbf{X}\mathbf{w} + \mathbf{J}b = \mathbf{y}$ where $\mathbf{J}$ is a column matrix of ones and $\mathbf{X}$ is a super long matrix with each row representing one image. Since this is indeed the case for most of our permutation tests (figure 2), the optimization (3) becomes:

$$\begin{aligned} & \min_{\mathbf{w}, b} \|\mathbf{w}\|^2 \\ \text{subj.to. } & \mathbf{X}\mathbf{w} + \mathbf{J}b = \mathbf{y} \end{aligned} \quad (3)$$

The above formulation is exactly the same as an LS-SVM [9]. This equivalence between the SVM and LS-SVM for high dimensional low sample size data was also previously noted in [12] where it was based on observations about the distribution of such data as elucidated in [5]. Since the LS-SVM, (4) can be solved in the closed form [9], we can now compute $\mathbf{w}$ as:

$$\mathbf{w} = \mathbf{X}^T \left[(\mathbf{X}\mathbf{X}^T)^{-1} + (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{J} \left(-\mathbf{J}^T (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{J} \right)^{-1} \mathbf{J}^T (\mathbf{X}\mathbf{X}^T)^{-1} \right] \mathbf{y} \quad (4)$$

Note that this expresses each component w_j of $\mathbf{w}$ as a linear combination of y_j 's. Thus, we can hypothesize about the probability distribution of the components of $\mathbf{w}$, given the distributions of y_j . If we let y_j attain any of the labels (either +1 or -1) with equal probability, we have a Bernoulli like distribution on y_j with $E(y_j) = 0$ and $Var(y_j) = 1$ (the theory can be readily extended in the case of unequal priors). Note that (5) expresses $\mathbf{w}$ as a linear combination of these y_j we have:

$$E(w_j) = 0 \quad Var(w_j) = \sum_{i=1}^m C_{ij}^2 \quad (5)$$

where C_{ij} are the components of the matrix $\mathbf{C}$, which is defined as:

$$\mathbf{C} \doteq \mathbf{X}^T \left[(\mathbf{X}\mathbf{X}^T)^{-1} + (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{J} \left(-\mathbf{J}^T (\mathbf{X}\mathbf{X}^T)^{-1} \mathbf{J} \right)^{-1} \mathbf{J}^T (\mathbf{X}\mathbf{X}^T)^{-1} \right] \quad (6)$$

At this point we know the expectation and the variance of w_j . We still need to uncover the probability density function (pdf) of w_j . Next, we use the Lyapunov central limit theorem to show that when the number of subjects is large, the p.d.f of w_j can be approximated by a normal distribution. To this end, from (5) and (7), we have:

$$w_j = \sum_{i=1}^m C_{ij} y_i = \sum_{i=1}^m z_i^j \quad (7)$$

where we have defined a new random variable $z_i^j = C_{ij} y_i$ which is linearly dependent on y_i . We can infer the expectation and variance of z_i^j from y_j as:

$$E(z_i^j) = 0 = \mu_i \quad \text{Var}(z_i^j) = C_{ij}^2 \quad (8)$$

Thus, z_i^j are independent but not identically distributed and w_j are linear combinations of z_i^j . Then according to the Lyapunov central limit theorem (CLT) w_j is distributed normally if:

$$\lim_{m \rightarrow \infty} \frac{1}{\left[\sqrt{\sum_{i=1}^m \text{Var}(z_i^j)} \right]^{2+\delta}} \sum_{k=1}^m E[|z_k^j - \mu_k|^{2+\delta}] = 0 \quad \text{for some } \delta > 0 \quad (9)$$

As is standard practice we check for $\delta = 1$.

$$E[|z_k^j - \mu_k|^{2+\delta}] = (1/2) |C_{kj} - 0|^{2+\delta} + (1/2) |-C_{kj} - 0|^{2+\delta} = C_{kj}^3 \quad (10)$$

Thus, we can write the limit in (10) as:

$$\lim_{m \rightarrow \infty} \frac{\sum_{k=1}^m C_{kj}^3}{\left[\sqrt{\sum_{i=1}^m C_{ij}^2} \right]^3} = \sum_{k=1}^m \left(\frac{C_{kj}^2}{\sqrt{\sum_{i=1}^m C_{ij}^2}} \right)^3 = 0 \quad (11)$$

Hence, given an adequate number of subjects, the Lyapunov CLT allows us to approximate the distribution of individual components of $\mathbf{w}$ using the normal distribution as:

$$w_j \xrightarrow{d} \mathcal{N}\left(0, \sum_{i=1}^m C_{ij}^2\right). \quad (12)$$

These predicted distributions fit actual distributions obtained using permutation testing very well. Thus w_j 's computed by an SVM model using true labels can now simply be compared to the distribution given by (13) and statistical inference can be made. Thus, (13) gives us a fast and efficient analytical alternative to actual permutation testing. In the next section we validate our analytical approximation using actual data.

3 Experiments and Results

In order to validate the theory proposed above we performed two experiments using brain imaging data. Both the experiments were done using tissue density maps (TDMs) generated after preprocessing of the raw images (these maps are commonly used in approaches like “modulated VBM” in which the Jacobian determinant multiplies the spatially normalized images). Tissue density maps were used because they inform us about the quantity of tissue present at each brain location in a common template space. These can be generated from registration fields warping a given subject to a template space. Since these maps are usually computed in template space a specific voxel location in a TDM corresponds to the same brain region across multiple subjects. Our two experiments were:

Experiment 1 : Simulated data was generated as follows 1) grey matter tissue density maps were generated from brain images of 152 normal subjects 2) simulated brain shrinkage was introduced in the right frontal lobe of half of these images by a localized reduction in intensity of the corresponding TDMs. The vectorized TDM corresponding to each subject forms the superlong vector $\mathbf{x}_i$ of section 2.1.

Experiment 2 : 278 TDMs were generated. This dataset contained 152 controls and 126 Alzheimer's patients. Grey matter (GM), white matter(WM) and ventricular(CSF) tissue density maps were computed. Maps corresponding to the i^{th} subject are vectorized and the vectors of all 3 tissue types (each living in $R^{1 \times q}$) were concatenated into the super long vector $\mathbf{x}_i \in R^{1 \times 3q}$.

For both experiments we stacked the $\mathbf{x}_i$'s corresponding to controls and patients(either simulated or actual) into the matrix $\mathbf{X}$ as detailed in section 2.1. In both experiments, we performed permutation tests (1000 permutations) and also computed the predicted distribution using (13). Figure 3 shows how the predicted probability distribution function compares with the distribution obtained from the actual permutation tests. We also computed p-values at every voxel location according to the predicted and the actual distribution. These results are shown in figure 4. Notice that the predicted p-values are very close to the actual ones in both simulated and actual data. Furthermore a p-value threshold of 0.05 accurately demarcates the region of simulated brain shrinkage in the simulated data and the hippocampus in the real data (figure 5). The actual permutation tests took approximately 8 hours of computational time for Experiment 1 and 50 hours of computational time for Experiment 2 while the analytical approximation took less than a minute to compute for either experiment.

4 Discussion

An important issue that remains to be addressed is that of dataset size. This method assumes a large dataset size. So the natural question that occurs is "How big is big enough?". This remains to be addressed as part of future work. Another perspective that needs to be explored is that of multiple comparisons. "Do we need to correct for multiple comparisons?". If yes, how do we make such a correction? We intend to address both of these issues in future work.

Nevertheless, the current manuscript provides analytical machinery that potentially replaces computationally intensive permutation testing when using SVMs for multivariate image analysis and classification. This ability to easily associate voxel level statistical significance with the output of an SVM allows us to use it for easily discovering brain regions and networks associated with disease in addition to disease classification.

References

1. Ashburner J, Friston KJ. Voxel-based morphometry—the methods. *Neuroimage*. Jun; 2000 11(6 Pt 1):805–821. <http://dx.doi.org/10.1006/nimg.2000.0582>. [PubMed: 10860804]
2. Bishop, CM. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. 1st ed., corr. 2nd printing edn.. Springer; Oct. 2007 2006
3. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*. 1998; 2:121–167. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.117.3731>.
4. Davatzikos C. Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *Neuroimage*. Sep; 2004 23(1):17–20. <http://dx.doi.org/10.1016/j.neuroimage.2004.05.010>. [PubMed: 15325347]
5. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. Jun; 2005 67(3):427–444. <http://dx.doi.org/10.1111/j.1467-9868.2005.00510.x>.
6. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*. Feb; 2005 6(2):95–108. <http://dx.doi.org/10.1038/nrg1521>. [PubMed: 15716906]

7. Klöppel S, Stonnington CM, Chu C, Draganski B, Scahill RI, Rohrer JD, Fox NC, Jack CR Jr, Ashburner J, Frackowiak RSJ. Automatic classification of mr scans in alzheimer's disease. *Brain*. Mar; 2008 131(Pt 3):681–689. <http://dx.doi.org/10.1093/brain/awm319>. [PubMed: 18202106]
8. Mouro-Miranda J, Bokde ALW, Born C, Hampel H, Stetter M. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. *Neuroimage*. Dec; 2005 28(4):980–995. <http://dx.doi.org/10.1016/j.neuroimage.2005.06.070>. [PubMed: 16275139]
9. Suykens JAK, Vandewalle J. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*. Jun; 1999 9(3):293–300. <http://dx.doi.org/10.1023/A:1018628609742>.
10. Vapnik, VN. The nature of statistical learning theory. Springer-Verlag New York, Inc; New York, NY, USA: 1995. <http://books.google.com/books?id=sna9BaxVbj8C>
11. Wang Z, Childress AR, Wang J, Detre JA. Support vector machine learning-based fmri data group analysis. *Neuroimage*. Jul; 2007 36(4):1139–1151. <http://dx.doi.org/10.1016/j.neuroimage.2007.03.072>. [PubMed: 17524674]
12. Ye J, Xiong T. Svm versus least squares svm. *Journal of Machine Learning Research - Proceedings Track*. 2007:644–651.

Fig. 1.

Left: Concept of support vector machines in 2-D space Right: Permutation testing for support vector machines

Fig. 2.

For most permutations the number of support vectors in the learnt model is almost equal to the total number of samples (a) simulated dataset (b) real dataset with Alzheimer's patients and controls (c) real dataset with schizophrenia patients and controls

Fig. 3.

Experimental(blue histogram) and theoretically predicted(red line) null distributions for two randomly chosen components of w in (a) real dataset with Alzheimer's patients and controls (b) simulated data

Fig. 4. Comparison of p-value maps(top row) with the corresponding image histograms(bottom row), generated using our theoretical framework and actual permutation tests (a) In real data pertaining to Alzheimer's disease (b) In simulated data

Fig. 5.

Regions found by thresholding p-value maps at 0.05 a) in real data pertaining to Alzheimer's disease (left) theoretically predicted (right) actual permutation testing (b) in simulated data (left) ground truth region of introduced brain shrinkage (middle) found using theory (right) found using actual permutation testing