[International Journal of Computer and Communication Technology](https://www.interscience.in/ijcct)

Volume 3 | Issue 2 Article 12

April 2012

Cognitive States from Brain Images : SVM Approach

Ramakrushna Swain
Dept. of C.S.& E., C.V.Raman College of Engineering, Bhubaneswar, India,, rk_swain@rediffmail.com

Lambodar Jena
Dept. of C.S.& E., C.V.Raman College of Engineering, Bhubaneswar, India, lambodarjena@gmail.com

Narendra K. Kamila
Dept. of C.S.& E., C.V.Raman College of Engineering, Bhubaneswar, India,, narendrakkamila@gmail.com

Cognitive States from Brain Images : SVM Approach

*Ramakrushna Swain, Lambodar Jena, Narendra K Kamila
Dept. of C.S.& E., C.V.Raman College of Engineering, Bhubaneswar, India,
* rk_swain@rediffmail.com

Abstract- **As the field of functional human brain mapping has matured, it has become apparent that a comprehensive understanding of the human brain, and its relationship with cognition, will require a quantitative assessment of individual differences in both brain function and structure. To assess brain structure, accurate classification of magnetic resonance images needed. In recent years, functional Magnetic Resonance Imaging (fMRI) has emerged as a powerful new instrument to collect vast quantities of data about activity in the human brain. As in other modern empirical sciences, this new instrumentation has led to a flood of new data and a corresponding need for new data analysis methods. A standard fMRI study gives rise to massive amounts of noisy data with a complicated spatio-temporal correlation structure. In this paper we discuss the analysis of fMRI data, from the angle of support vector machine classification for analysis of complex, multivariate data.**

Keywords: ***fMRI, brain imaging, Support Vector Machine, Classifier, voxels.***

I. INTRODUCTION

The study of human brain function has received a tremendous boost in recent years from the advent of functional Magnetic Resonance Imaging (fMRI), a brain imaging method that dramatically improves our ability to observe correlates of neural brain activity in human subjects at high spatial resolution (several millimeters), across the entire brain. This fMRI technology offers the promise of revolutionary new approaches to studying human cognitive processes, provided we can develop appropriate data analysis methods to make sense of this huge volume of data. A twenty-minute fMRI session with a single human subject produces a series of three dimensional brain images each containing approximately 15,000 voxels, collected once per second, yielding tens of millions of data observations.

In recent years, SVM analysis approach has grown in popularity is the use of machine learning algorithms to train classifiers to decode stimuli, mental states, behaviors and other variables of interest from fMRI data and thereby show the data contain enough information about them. A typical fMRI experiment can produce a three-dimensional image related to the human subject's brain activity every half second, at a spatial resolution of a few millimeters. In machine learning problems with tens of thousands of features and only dozens or hundreds of independent training examples, dimensionality reduction is essential for good learning performance. We describe recent research applying machine learning methods to the problem of classifying the cognitive state of a human subject based on fRMI data.

II. FUNCTIONAL MAGNETIC RESONANCE IMAGING

Functional Magnetic Resonance Imaging (fMRI) is a technique for obtaining three dimensional images related to neural activity in the brain through time. More precisely, fMRI measures the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood with respect to a control baseline, at many individual locations within the brain. It is widely believed that blood oxygen level is influenced by local neural activity, and hence this blood oxygen level dependent (BOLD) response is generally taken as an indicator of neural activity. An fMRI scanner measures the value of the fMRI signal (BOLD response) at all the points in a three dimensional grid, or image. The cells within this three-dimensional image are referred to as voxels (volume elements). The voxels in a typical fMRI study have a volume of a few tens of cubic millimeters, and a typical three dimensional brain image typically contains 10,000 to 15,000 voxels which contain cortical matter and are thus of interest. While the spatial resolution of fMRI is dramatically better than that provided by earlier brain imaging methods, each voxel nevertheless contains on the order of hundreds of thousands of neurons [1] [2].

There are several common objectives in the analysis of fMRI data. These include localizing regions of the brain activated by a task, determining distributed networks that correspond to brain function. Each of these objectives can be approached through the application of suitable statistical methods. This role can range from determining the appropriate statistical method to apply to a data set, to the development of unique statistical methods geared specifically toward the analysis of fMRI data [3].

The fMRI data comprise a sequence of magnetic resonance images (MRI), each consisting of a number of uniformly spaced volume elements, or voxels, that partition the brain into equally sized boxes. The image intensity from each voxel represents the spatial distribution of the nuclear spin density in that area. Changes in brain hemodynamic, in reaction to neuronal activity, impact the local intensity of the MR signal, and therefore changes in voxel intensity across time can be used to infer when and where activity is taking place. During the course of an fMRI experiment, images of this type are acquired between 100–2000 times, with each image consisting of roughly 100,000 voxels.

Further, the experiment may be repeated several times for the same subject, as well as for multiple subjects to facilitate population inference. Though a good number of these voxels consist solely of background noise, and can be excluded from further analysis, the total amount of data that needs to be analyzed is staggering. In addition, the data exhibit a complicated temporal and spatial noise structure with a relatively weak signal. A full spatiotemporal model of the data is generally not considered feasible and dimensionality reduction is taken throughout the course of the analysis. Statistics play an important role in determining in which dimensionality reduction method is appropriate in the various stages of the analysis.

A. *Data Acquisition :*

To construct an image, the subject is placed into the field of a large electromagnet. The magnet has a very strong magnetic field, typically between 1.5–7.0 Tesla, which aligns the magnetization of hydrogen (1H) atoms in the brain. Within a slice of the brain, a radio frequency pulse is used to tip over the aligned nuclei. Upon removal of this pulse, the nuclei strive to return to their original aligned positions and thereby induce a current in a receiver coil. This current provides the basic MR signal. A system of gradient coils is used to sequentially control the spatial in homogeneity of the magnetic field, so that each measurement of the signal can be approximately expressed as the Fourier transformation of the spin density at a single point in the frequency domain, or k-space as it is commonly called in the field. Mathematically, the measurement of the MR signal at the jth time point of a readout period can be written as

$$S(t_j) \approx \int_x \int_y M(x,y) \cdot e^{(-2\pi i(k_x(t_j)x+k_y(t_j)y))}\,dx\,dy,$$

where M(x, y) is the spin density at the point (x, y), and $(k_x(t_j), k_y(t_j))$ is the point in the frequency domain (k-space) at which the Fourier transformation is measured at time t_j . Here $t_j = j\Delta t$ is the time of the jth measurement, where Δt depends on the sampling bandwidth of the scanner; typically it takes values in the range of 250–1000 µs. To reconstruct a single MR image, one needs to sample a large number of individual k-space measurements, the exact number depending on the desired image resolution. For example, to fully reconstruct a 64×64 image, a total of 4096 separate measurements are required, each sampled at a unique coordinate of k-space. There is a time cost involved in sampling each point, and therefore the time it takes to acquire an image is directly related to its spatial resolution.

III. UNDERSTANDING fMRI DATA

The ability to connect the measures of brain physiology obtained in an fMRI experiment with the underlying neuronal activity that caused them will greatly impact the choice of inference procedure and the subsequent conclusions that can be made. Therefore, it is important to gain some rudimentary understanding of basic brain physiology. In addition, since neuronal activity unfolds both in space and time, the spatial and temporal resolution of fMRI studies will limit any conclusions that can be made from analyzing the data and understanding these limitations is paramount. Finally, as relatively small changes in brain activity are buried within noisy measurements, it will be important to understand the behavior of both the signal and noise present in fMRI data and begin discussing how these components can be appropriately modeled.

We 1st performed a preliminary qualitative analysis on the fMRI data produced in the various run of the virtual reality game. As a result we recognized that each sample of data was quite different from the others. The challenge was therefore how to extract the information needed to do good predictions from heterogeneous samples. Our strategy to achieve this goal is the dimensionality reduction and classification and presents an efficient algorithm [4]. In order to obtain good learning performance in these domains, researchers often perform dimensionality reduction before learning a classifier. Two typical approaches to dimensionality reduction are feature selection (e.g., forward stepwise selection) and feature synthesis (e.g., singular value decomposition) [7].

A. *What is a classifier?*

Classification is the analogue of regression when the variable being predicted is discrete, rather than continuous. A classifier is a function that takes the values of various features (independent variables or predictors, in regression) in an example (the set of independent variable values) and predicts the class that example belongs to (the dependent variable) [5]. In a neuroimaging setting, the features could be voxels and the class could be the type of stimulus the subject was looking at when the voxel values were recorded (see Figure 1). We will denote an example by the row vector x = [x1 . . . xv] and its class label as y. A classifier has a number of parameters that have to be

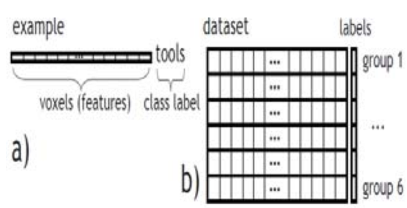

Figure 1a : An example where voxels are features as a row vector (left) and a dataset as matrix of such vectors(right).

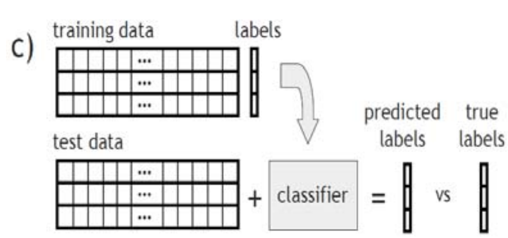

Figure 1b : A classifier is learned from the training set, examples whose labels it can see, and used to predict labels for a test set, examples whose lebels it cannot see. The predicted labels are then compared to the true labels and the classifier's accuracy computed.

learned from training data – a set of examples reserved for this purpose – similarly to how regression parameters are estimated using least squares. The learned classifier is essentially a model of the relationship between the features and the class label in the training set. More formally, given an example x, the classifier is a function f that predicts the label ˆy = f(x). Once trained the classifier can be used to determine whether the features used contain information about the class of the example. This relationship is tested by using the learned classifier on a different set of examples, the test data. Intuitively, the idea is that, if the classifier truly captured the relationship between features and class, it ought to be able to predict the classes of examples it hasn't seen before. The typical assumption for classifier learning algorithms is that the training (and testing) examples are independently drawn from an 'example distribution'; when judging a classifier on a test set we are obtaining an estimate of its performance on any test set from the same distribution.

B. *Which feature selection method works best?*

Given that our classification problem involves very high dimensional, noisy, sparse training data, it is natural to consider feature selection methods to reduce the dimensionality of the data before training the classifier. Feature selection leads to large and statistically significant improvements in classification error.

C. *Approach.*

Within the field of machine learning, the most common approach to feature selection when training classifiers is to select those features that best discriminate the target classes. For example, given the goal of learning a target classification function f: X → Y, one common approach to feature selection is to rank order the features of X by their mutual information with respect to the class variable Y, then to greedily select the n highest scoring features.

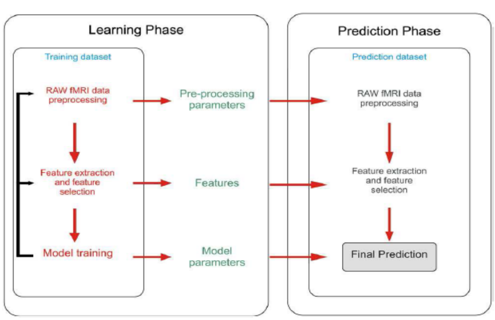

Figure 2 : Learning Phase and Prediction Phase

As in figure 2 main steps of a generic pattern recognition algorithm is used in fMRI data analysis. In the training phase the raw fMRI data are pre-processed and relevant features are selected from the data prior to model training. Prediction is performed using the trained model on a new data set, after this later has been preprocessed in the same way and reduced to same features as in the training.

IV. SUPPORT VECTOR MACHINE (SVM)

SVM is an unsupervised approach based on statistical learning theory. It estimates the optimal boundary in the feature space by combining a maximal margin strategy with a kernel method; this process is called a kernel machine. The machine is trained according to the structural risk minimization (SRM) criterion. The decision boundaries are directly derived from the training data set by learning. The SVM maps input into a high-dimensional feature space through a selected kernel function. Then, it constructs an optimal separating hyper-plane in the feature space. The dimensionality of the feature space is determined by the number of support vectors extracted from the training data (see Figure 3). The SVM can locate all the support vectors, which exclusively determine the decision boundaries. To estimate the misclassification rate (risk), the so called leave-one-out procedure is used. It removes one of N_i training samples, performs training using the remaining training samples, and tests the removed sample with the newly derived hyper plane. It repeats this process for all of the samples, and the total number of errors becomes the estimation of the risk.

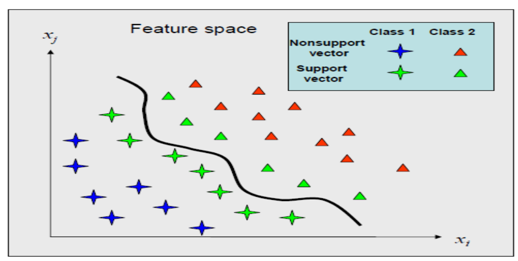

Figure 3. Optimal boundary serached by the SVM.

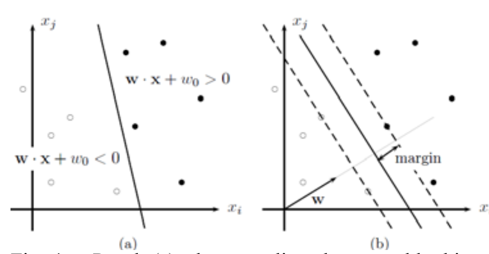

Fig 4 : Panel (a) shows a linearly separable binary classification problem and a separating hyper plane. Panel (b) shows the maximum margin hyper plane.

A. *Support Vector Classification*

For support vector classifiers, the key notion that we need to introduce is that of the maximum margin hyper plane for a linear classifier. Then by using the "kernel trick" this can be lifted into feature space. We consider first the separable case and then the non-separable case [6].

B. The Separable Case

Figure 4(a) illustrates the case where the data is linearly separable. For a linear classifier with weight vector w and offset w0, let the decision boundary be defined by $w \cdot x + w_0 = 0$, and let ˜w = (w, w_0). Clearly, there is a whole version space of weight vectors that give rise to the same classification of the training points. The SVM algorithm chooses a particular weight vector that gives rise to the "maximum margin" of separation. Let the training set be pairs of the form (x_i, y_i) for i = 1, n, where $y_i = \pm 1$. For a given weight vector we can compute the quantity ˜i = y_i(w·x+w0), functional margin which is known as the functional margin. Notice that ˜i > 0 if a training point is correctly classified. If the equation f(x) = w · x + w0 defines a discriminant function (so that the output is sgn(f(x))), then the hyper plane $cw \cdot x + cw_0$ defines the same discriminant function for any c > 0. Thus we have the freedom to choose the scaling of ˜w so that mini ˜i = 1, and in this case ˜w is known as the canonical form of the hyperplane.geometrical margin the geometrical margin is defined as i = ˜i/|w|. For a training point xi that is correctly classified this is simply the distance from x_i to the hyper plane.

To see this, let c = 1/|w| so that ˆw = w/|w| is a unit vector in the direction of w, and ˆ w_0 is the corresponding offset. Then ˆw · x computes the length of the projection of x onto the direction orthogonal to the hyper plane and ˆw·x+ ˆ w_0 computes the distance to the hyper plane. For training points that are misclassified the geometrical margin is the negative distance to the hyper plane.

The geometrical margin for a dataset D is defined as D = mini i. Thus for a canonical separating hyper plane the margin is 1/|w|. We wish to find the maximum margin hyper plane, i.e. the one that maximizes D. By considering canonical hyper planes, we are thus led to the following optimization problem to determine the maximum margin hyper plane:

$$\text{minimize } \frac{1}{2}|\mathbf{w}|^2 \qquad \text{over } \mathbf{w},\ w_0$$
$$\text{subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 \qquad \text{for all } i = 1,\ldots,n. \tag{1}$$

It is clear by considering the geometry that for the maximum margin solution there will be at least one data point in each class for which $y_i(w \cdot x_i + w_0) = 1$, see Figure 4(b). Let the hyper planes that pass through these points be denoted H+ and H− respectively. This constrained optimization problem can be set up using Lagrange multipliers, and solved using numerical methods for quadratic programming (QP) problems. The form of the solution is

$$\mathbf{w} = \sum_i \lambda_i y_i \mathbf{x}_i, \tag{2}$$

where the i's are non-negative Lagrange multipliers. Notice that the solution is a linear combination of the x_i's.

The key feature of the above equation is that i is zero for every x_i except those which lie on the hyper planes H_+ or H_-; these points are called the support vectors. The fact that not all of the training points contribute to the final support vectors solution is referred to as the sparsity of the solution. The support vectors lie closest to the decision boundary. Note that if all of the other training points were removed (or moved around, but not crossing H_+ or H_-) the same maximum margin hyper plane would be found. The quadratic programming problem for finding the i's is convex, i.e. there are no local minima.

To make predictions for a new input x,we compute

$$\mathrm{sgn}(\mathbf{w}\cdot\mathbf{x}_* + w_0) = \mathrm{sgn}\left(\sum_{i=1}^{n} \lambda_i y_i (\mathbf{x}_i \cdot \mathbf{x}_*) + w_0\right). \tag{3}$$

In the QP problem and in the above equation the training points {x_i} and the test point X_* enter the computations only in terms of inner products. Thus by using the kernel trick we can replace occurrences of the inner product by the kernel to obtain the equivalent result in feature space.

C. The Non-Separable Case

For linear classifiers in the original x space there will be some datasets that are not linearly separable. One way to generalize the SVM problem in this case is to allow violations of the constraint $y_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1$ but to impose a penalty when this occurs. This leads to the soft margin support vector machine soft margin problem, the minimization of

$$\frac{1}{2}|\mathbf{w}|^2 + C\sum_{i=1}^{n}(1 - y_i f_i)_+ \tag{4}$$

With respect to w and w_0, where $f_i = f(x_i) = w \cdot x_i + w_0$ and (z) + = z if z > 0 and 0 otherwise. Here C > 0 is a parameter that specifies the relative importance of the two terms. This convex optimization problem can again be solved using QP methods and yields a solution of the form given in eq. (3). In this case the support vectors (those with $\lambda_i \neq 0$) are not only those data points which lie on the separating hyper planes, but also those that incur penalties. This can occur in two ways (i) the data point falls in between H_+ and H_- but on the correct side of the decision surface, or (ii) the data point falls on the wrong side of the decision surface. In a feature space of dimension N, if N > n then there will always be separating hyper plane. However, this hyper plane may not give rise to good generalization performance, especially if some of the labels are incorrect, and thus the soft margin SVM formulation is often used in practice.

In a typical feature synthesis approach we build features based on the distribution of the independent variables in the training data, using algorithms like singular value decomposition (SVD) or independent component analysis (ICA). Unfortunately, feature spaces produced in this manner may not necessarily be good for classification, since they are derived without reference to the quantity we are trying to predict. For example, consider fMR images taken while a subject was thinking about items of different semantic categories, and suppose that our task is to decide which semantic category was present in each example. If we perform an SVD of this data, the components extracted will capture image variability due to awareness, task control, language use, the visual form of the cues given, and many other factors. Most of these dimensions of variability will have little information about the semantic category, and if their variance is too high they may prevent the SVD from noticing the directions of variation which would be useful for classification. On the other hand, the basic idea of the SVD-trying to find a small set of features which accurately describe the test data seems sound. The problem is only that the SVD performs its data reduction without paying attention to the classification problem at hand[5].

D. Support Vector Regression

The SVM was originally introduced for the classification problem, and then extended to deal with the regression case. The key concept is that of the ∊–insensitive error function. This is defined as

$$g_\epsilon(z) = \begin{cases} |z| - \epsilon & \text{if } |z| \geq \epsilon, \\ 0 & \text{otherwise.} \end{cases} \qquad (5)$$

We can interpret exp $(-g_\in(z))$ as a likelihood model for the regression residuals.

For the linear regression case with an ∊–insensitive error function and a Gaussian prior on w, the MAP value of w is obtained by minimizing

$$\frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{n} g_\epsilon(y_i - f_i)$$

w.r.t. w. As for support vector classification, many of the coefficients α_i are zero. The data points which lie inside the ∊-"tube" have $\alpha_i = 0$, while those on the edge or outside have non-zero α_i.

V. EXPERIMENTS

A. Datasets

To evaluate the SVDM, we tested it on data from an fMRI experiment. In this experiment, the subject observes a word displayed on a screen for 3 seconds, followed by 8 seconds of a blank screen. Each word describes either a type of tool or a type of building, and the subject's task is to think about the word and its properties while it is displayed. During an experiment the task repeats 84 times, and a 3D image of the fMR signal is acquired every second. Each image contains 64x64x16 = 65536 voxels, but only approximately 16000 of those contain cortex; hence we only consider this latter number as features (for more details about fMRI please refer to (Mitchell et al., 2004)). The dataset thus contains 84 examples; each example is the average image during a 4 second span while the subject is thinking about a word shown a few seconds earlier [7]. The classification task is to decide which of the two semantic categories, tool or building, the subject was thinking about. We trained three separate SVDMs, one per experimental subject. Figure 1 show, for each of the two categories, one slice of activation in the temporal cortex of a subject, overlaid on the corresponding structural image. The data for this figure comes from another experiment, where the task was done many times in a row and all the images acquired during that period were averaged. With the reduction in noise due to averaging it is easy to see a difference between Tools and Buildings; our interest is to do the same for the current dataset, which is noisier since there was less averaging. That is, we wish to decode the \cognitive state" (Mitchell et al., 2004) of the subject from a brief interval of fMRI data.

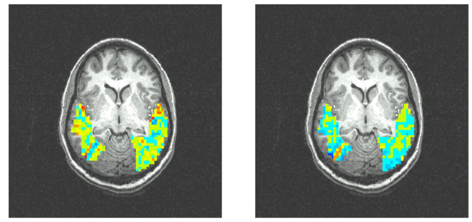

Fig 5 : 2D slices from the average 3D fMRI image acquired while the subject was thinking of either "Tools" (left) or "Buildings" (right) many times in arrow. (more dark means more active)

B. Classification and Prediction

There is a growing interest in using fMRI data as a tool for classification of mental activity leading to different cognitive states. In addition, there has been growing interest in developing methods for predicting stimuli directly from functional data. This would allow for the possibility to infer information from the scans about the subjects thought process and use brain activation patterns to characterize subjective human experience. A particularly controversial application has been the idea of using fMRI for lie detection. The efficient prediction of brain states is a challenging process that requires the application of novel statistical and machine learning techniques. Various multivariate pattern classification approaches have successfully been applied to fMRI data in which a classifier is trained to discriminate between different brain states and then used to predict the brain states in a new set of fMRI data. To date, efficient preprocessing of the data has been shown to be more important than the actual method of prediction. However, this is an area that without a doubt will be the focus of intense research in the future and where statisticians and neurologists are well positioned to make a significant impact.

VI. CONCLUSION

We have presented fMRI studies basing upon Support Vector Machine (SVM) demonstrating the feasibility of training classifiers to distinguish a variety of cognitive states based on fMRI observations. This problem is interesting both because of its relevance to studying human cognition and as a case study of machine learning in high dimensional, noisy, sparse data settings. Combining information from different modalities will be challenging to data analysts, if for no other reason than that the amount of data will significantly increase. In addition, since the different modalities are measuring fundamentally different quantities, it is not immediately clear how to best combine the information. This is an extremely important problem that has already started to become a major area of research. The Support Vector Machine approach has arguably become the dominant way to analyze fMRI data. It models the time series as a linear combination of several different signal components and tests whether activity in a brain region is systematically related to any of these known input functions. As the experimental design and imaging techniques become more sophisticated, the need for novel statistical methodology will only increase, promising an exciting future for statisticians and neurologists in the field.

REFERENCES

[1] Martin A. Lindquist "The Statistical Analysis of fMRI Data", Statistical Science, 2008, Vol. 23, No. 4, 439–464, DOI: 10.1214/09-STS282 c Institute of Mathematical Statistics, 2008

[2] TOMM. MITCHELL, REBECCA HUTCHINSON, RADU S. NICULESCU, FRANCISCO PEREIRA, XUERUI WANG; "Learning to Decode Cognitive States from Brain Images", School of Computer Science, Carnegie Mellon University; tom.mitchell@cmu.edu

[3] Statistical Science 2008, Vol. 23, No. 4, 439–464 DOI: 10.1214/09-STS282 c Institute of Mathematical Statistics, 2008

[4] E. Olivetti, D. Sona, P. Avesani, R. Moretta, F. Zini "Mixture Models for Cognitive Brain State Prediction" IRST - Fondazione Bruno Kessler_ Trento – Italy S. Veeramachaneni Thomson R&Dy Eagan, MN - U.S.A J. Schwarzbach University of Trento Trento – Italy June 29, 2007

[5] Francisco Pereira ,Tom Mitchell "Machine learning classifiers and fMRI: a tutorial overview "

[6] C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. 2006 Massachusetts Institute of Technology. www.GaussianProcess.org/gpml

[7] Francisco Pereira, Geo rey Gordon "The Support Vector Decomposition Machine"

[8] A.K. Jain, R.P.W. Duin and J. Mao, "Statistical pattern recognition: A review", IEEE Trans. on PAMI, vol.22, no.1, pp.4-37, Jan. 2000.

[9] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution", Neural Computation, 7, pp.1129-1159, 1995.

[10] C.J.C. Burges, "A tutorial on support vector machines for pattern recognition", Data Mining and Knowledge Discovery, 2, pp.121-167, 1998.

[11] V. Vapnik, Statistical Learning Theory, John Wiley and Sons, Inc., New York, 1998.

[12] Aguirre, G. K., Zarahn, E. and D'Esposito, M. (1998). The variability of human, BOLD hemodynamic responses. NeuroImage 8 360–369.