The Relevance Voxel Machine (RVoxM): A Bayesian Method for Image-Based Prediction

Mert R. Sabuncu[1,2] and Koen Van Leemput[1,2,3,4]

[1] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, MGH, Harvard Medical School, USA
[2] Computer Science and Artificial Intelligence Laboratory, MIT, USA
[3] Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland
[4] Department of Informatics and Mathematical Modeling, Technical University of Denmark, Denmark

Abstract. This paper presents the Relevance Voxel Machine (RVoxM), a Bayesian multivariate pattern analysis (MVPA) algorithm that is specifically designed for making predictions based on image data. In contrast to generic MVPA algorithms that have often been used for this purpose, the method is designed to utilize a small number of spatially clustered sets of voxels that are particularly suited for clinical interpretation. RVoxM automatically tunes all its free parameters during the training phase, and offers the additional advantage of producing probabilistic prediction outcomes. Experiments on age prediction from structural brain MRI indicate that RVoxM yields biologically meaningful models that provide excellent predictive accuracy.

Keywords: Multivariate Pattern Analysis, MRI.

1 Introduction

Medical imaging commonly entails relating image content to a clinical or experimental condition. Traditional univariate approaches, such as voxel-based morphometry [2], generate anatomical maps of the effects by analyzing each location individually. MVPA methods, in contrast, offer increased specificity and sensitivity for predicting the outcome by considering all voxels *simultaneously* [6,8,10,12,14,16,17,15]. Yet studies on image-based prediction have typically employed generic MVPA methods, such as Support or Relevance Vector Machines (SVMs/RVMs) [5,18], which do not account for the spatial organization of *imaging* data.

As demonstrated in semi-supervised learning, significant performance gains can be obtained by explicitly utilizing the underlying structure of the data [3,4]. One approach to achieve this with images is to impose an a priori model on the covariation of voxel measurements – a strategy that has proven powerful in computer vision [13]. Further motivation for such image-based prediction models is *interpretability*: rather than a "black box" tool, we are also interested in understanding and visualizing the key areas that drive predictions. Although it is

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 99–106, 2011.

possible to display the workings of generic linear MVPA methods as images [12], the results are often scattered and hard to interpret biologically [7].

In this paper, we present the Relevance Voxel Machine (RVoxM), a novel MVPA algorithm that is specifically designed for image-based prediction. It uses a Bayesian approach and builds largely on existing RVM machinery to obtain not only good prediction performance, but also sparse solutions. Unlike RVMs, however, where sparseness is realized by discarding many of the samples, i.e., *training subjects*, our approach removes most *voxels*, retaining only those voxels that are relevant for prediction. Furthermore, our model encourages spatial clustering of these "relevance voxels" and computes predictions as linear combinations of their content, yielding results that are both biologically plausible and intuitive to interpret. Compared to related efforts that incorporate spatial context within the SVM or penalized regression frameworks [15,7], our method inherits all the usual advantages of RVMs over non-Bayesian methods, including providing probabilistic outcomes and the automatic tuning of all free parameters.

We test RVoxM on the problem of estimating the age of healthy subjects from structural brain MRI scans, and show that it achieves high accuracy using a pattern of "relevance voxels" that easily lends itself to biological interpretation.

2 Model

We use a generative model similar to the one of RVM [18]. Let t denote a real-valued *target* variable (e.g., age) that we aim to predict from image data, and x_i a voxel-level measurement (e.g., gray matter density) at the voxel indexed by i. We define a Gaussian conditional distribution for t: $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}), \beta^{-1})$, with variance β^{-1} and a mean that is given by the linear model

$$y(\mathbf{x}) = \sum_{i=1}^{M} x_i w_i + w_0 = \mathbf{w}^{\mathrm{T}}\mathbf{x}, \tag{1}$$

where $\mathbf{w} = (w_0 \cdots w_M)^{\mathrm{T}}$ are adjustable "weights" encoding the strength of each voxel's contribution to the prediction, $\mathbf{x} = (1, x_1, \cdots, x_M)^{\mathrm{T}}$ denotes the vectorized image the prediction is based on, and M is the number of voxels. For notational convenience, we include an extra "voxel" to account for the bias, w_0.

We assume a zero-mean Gaussian prior distribution over $\mathbf{w}$:

$$p(\mathbf{w}|\boldsymbol{\alpha}, \lambda) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{P}^{-1}),$$

where $\mathbf{P}$ is a $(M+1) \times (M+1)$ precision (inverse covariance) matrix defined as

$$\mathbf{P} = \mathrm{diag}(\alpha_0, \cdots, \alpha_{\mathrm{M}}) + \lambda\mathbf{K}.$$

Here, $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_M)^{\mathrm{T}}$ and λ are *hyperparameters*, and $\mathbf{K}$ is a fixed, positive-semidefinite matrix that encourages local spatial smoothness of $\mathbf{w}$. In particular, we use $\mathbf{K} = \boldsymbol{\Upsilon}^{\mathrm{T}}\boldsymbol{\Upsilon}$, where $\boldsymbol{\Upsilon}$ is a sparse matrix in which each row corresponds to a pair of neighboring voxels in the image. For neighboring voxels $\{i, j\}$, the corresponding row has zero entries everywhere expect for the $i^{\mathrm{th}}$and j^{th} column, which have entries -1 and 1, respectively. Re-writing the prior as

$$p(\mathbf{w}|\boldsymbol{\alpha},\lambda) \;\propto\; \exp\Big(-\frac{1}{2}\sum_{i=0}^{M}\alpha_i w_i^2 - \frac{\lambda}{2}\|\boldsymbol{\Upsilon}\mathbf{w}\|^2\Big)$$

shows that it encodes a preference for models that are both *sparse and spatially clustered*: we explicitly seek models that explain t through a small collection of image patches that easily lend themselves to neuroscientific interpretation. The fact that there is a hyperparameter α_i associated with each voxel's weight w_i is responsible for achieving sparsity in those weights – in practice many of the α_i's tend to very large values, forcing the corresponding weights to zero and "switching off" the contribution of many voxels. Importantly, we also explicitly take the spatial structure of image data into account by penalizing large entries in the vector $\boldsymbol{\Upsilon}\mathbf{w}$, which represent large differences between the weights of neighboring voxels. Thus, we encode a preference for spatial clusters of "switched-on" voxels, as these are both biologically more plausible and easier to interpret than speckles of isolated voxels scattered throughout the image area.

3 Hyperparameter Estimation

Given training data, $\{\mathbf{x}_n, t_n\}_{n=1}^N$, where $\mathbf{x}_n$ is the n^{th} training image and t_n its target variable, our first goal is to determine the values of the hyperparameters $\boldsymbol{\alpha}$, λ, and β. Using type-II maximum likelihood, we estimate the hyperparameters by maximizing the marginal likelihood function obtained by integrating out $\mathbf{w}$:

$$\begin{aligned} p(\mathbf{t}|\mathbf{X},\boldsymbol{\alpha},\lambda,\beta) &= \int_{\mathbf{w}}\Big(\prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\beta)\Big)p(\mathbf{w}|\boldsymbol{\alpha},\lambda)\mathrm{d}\mathbf{w} \\ &= \int_{\mathbf{w}}\Big(\frac{\beta}{2\pi}\Big)^{N/2}\exp(-\frac{\beta}{2}\|\mathbf{t}-\mathbf{X}\mathbf{w}\|^2)\frac{|\mathbf{P}|^{1/2}}{(2\pi)^{M/2}}\exp(-\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{P}\mathbf{w})\mathrm{d}\mathbf{w} \\ &= \frac{|\boldsymbol{\Gamma}|^{-1/2}}{(2\pi)^{N/2}}\exp(-\frac{1}{2}\mathbf{t}^{\mathrm{T}}\boldsymbol{\Gamma}^{-1}\mathbf{t}), \qquad (2) \end{aligned}$$

where $\mathbf{t} = (t_1, \cdots, t_N)^{\mathrm{T}}$, $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]^{\mathrm{T}}$ is the $N \times (M+1)$ "design" matrix, and we have defined the $N \times N$ matrix $\boldsymbol{\Gamma}$ given by

$$\boldsymbol{\Gamma} = \beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{P}^{-1}\mathbf{X}^{\mathrm{T}}.$$

We take a "coordinate-ascent" approach to maximize Eq. (2). We first define:

$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{y}, \qquad \boldsymbol{\Sigma} = (\beta\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathbf{P})^{-1}. \qquad (3)$$

Fixing λ, β, and $\{\alpha_j\}$ for all $j \neq i$, differentiating the log of Eq. (2) w.r.t α_i, equating to zero and rearranging yields the following update:

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2}\alpha_i, \qquad (4)$$

where $\gamma_i = 1 - \alpha_i\boldsymbol{\Sigma}_{ii} - \lambda(\mathbf{P}^{-1}\mathbf{K})_{ii}$. Similarly, fixing $\boldsymbol{\alpha}$ and β, differentiating w.r.t λ, and rearranging yields the following update equation for λ:

$$\lambda^{new} = \frac{\mathrm{trace}(\mathbf{\Delta K})}{\boldsymbol{\mu}'\mathbf{K}\boldsymbol{\mu}}\lambda, \tag{5}$$

where $\mathbf{\Delta} = \mathbf{P}^{-1} - \mathbf{\Sigma}$. Similarly, an update for β can be derived:

$$\beta^{new} = \frac{N}{\|\mathbf{y} - \mathbf{X}\boldsymbol{\mu}\|^2 + \mathrm{trace}(\mathbf{X\Sigma X}^{\mathrm{T}})}. \tag{6}$$

Optimization now proceeds by cycling through these equations. We initialize with $\alpha_i = 1, \forall i$, $\lambda = 1$ and $\beta = 0$; monitor the objective function at each iteration and terminate when the increase over the previous iteration is below a tolerance. Although currently we have no theoretical guarantees that the presented equations indeed increase the objective function, we have not encountered any situation where this was not the case in our experiments.

4 The RVoxM Learning Algorithm

In practice, most ($> 90\%$) α_i's tend to grow to infinity, effectively clamping the corresponding weight w_i's to zero and removing those voxels from the model. We exploit this to obtain a greedy learning algorithm for large 3-D image volumes, using two computational tricks. First, each time one of the α_i's exceeds a certain (very large) value, the corresponding voxel is pruned from the model and computations continue based on remaining voxels only, in a manner similar to the RVM algorithm [18]. Second, we use a multi-resolution approach commonly employed in image processing. We construct a pyramid representation of the training images, where each level consists of lower-resolution images computed by subsampling the images from the previous resolution. The algorithm then starts by learning the hyperparameters for the lowest resolution images, propagates them down for the initialization of the next level, and so forth until the final resolution level is reached; voxels that were pruned at the previous level remain so henceforth. Although this greedy algorithm prevents voxels from reentering once they have been removed, our experiments suggest that it works quite well in practice.

5 Using RVoxM to Make Predictions

Having learned the hyperparameters $\boldsymbol{\alpha}^*$, λ^*, and β^* from the training data, we can make predictions about t for a new input image $\mathbf{x}$ by evaluating the posterior

$$p(t|\mathbf{x}, \mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*, \beta^*) = \int_{\mathbf{w}} p(t|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \boldsymbol{\alpha}^*, \lambda^*) \mathrm{d}\mathbf{w}.$$

It can be shown that this distribution is a Gaussian with mean

$$\boldsymbol{\mu}^{\mathrm{T}}\mathbf{x} \tag{7}$$

and variance $\frac{1}{\beta^*} + \mathbf{x}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{x}$, where $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ are given by Eq. (3) in which $\boldsymbol{\alpha}$, λ, and β have been set to their optimized values α^*, λ^*, and β^*.

In the remainder, we will use the maximum a posteriori value given by Eq. (7) to predict t, which corresponds to the linear model of Eq. (1) in which the voxels' weights $\mathbf{w}$ are set to $\boldsymbol{\mu}$. In many voxels $\mu_i = 0$ (because their α_i was set to infinity) – we call the remaining voxels the "relevance voxels" as these are the only ones effectively used to predict the target variable t.

6 Experimental Results

We applied RVoxM to the problem of estimating a person's age from a brain MRI scan. This problem has attracted recent attention [1,9,11] since it provides a novel perspective for studying healthy development and aging patterns, while characterizing pathologic deviations in disease.

We used T1-weighted scans from 336 cognitively normal subjects (age range 18-93 years), available through the OASIS dataset[1]. We processed all scans with SPM8[2], using default settings, to obtain spatially aligned gray matter maps. The *gray matter density* values (tissue probabilities modulated by the Jacobian of the non-linear warp) were used as voxel-level measurements x_i in the experiment. To assess generalization accuracy, we split the data into two arbitrary[3] halves

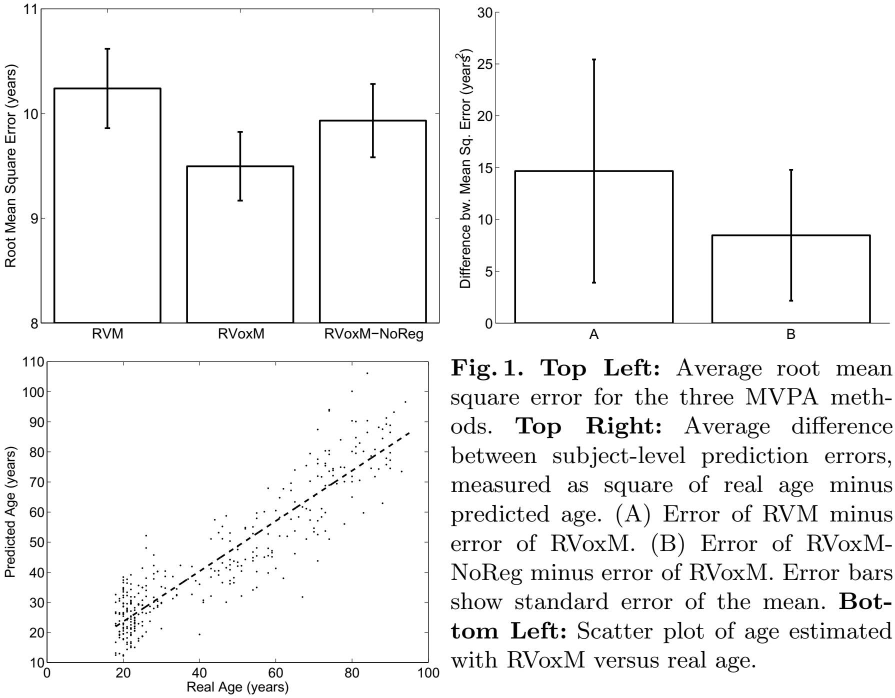

Fig. 1. Top Left: Average root mean square error for the three MVPA methods. **Top Right:** Average difference between subject-level prediction errors, measured as square of real age minus predicted age. (A) Error of RVM minus error of RVoxM. (B) Error of RVoxM-NoReg minus error of RVoxM. Error bars show standard error of the mean. **Bottom Left:** Scatter plot of age estimated with RVoxM versus real age.

[1] http://www.oasis-brains.org. 1.5T Siemens Vision scanner, $1\times1\times1.25mm^3$, MPRAGE.

[2] http://www.fil.ion.ucl.ac.uk/spm/software/spm8/

[3] Simply based on the alphabetical ordering of the anonymized filenames.

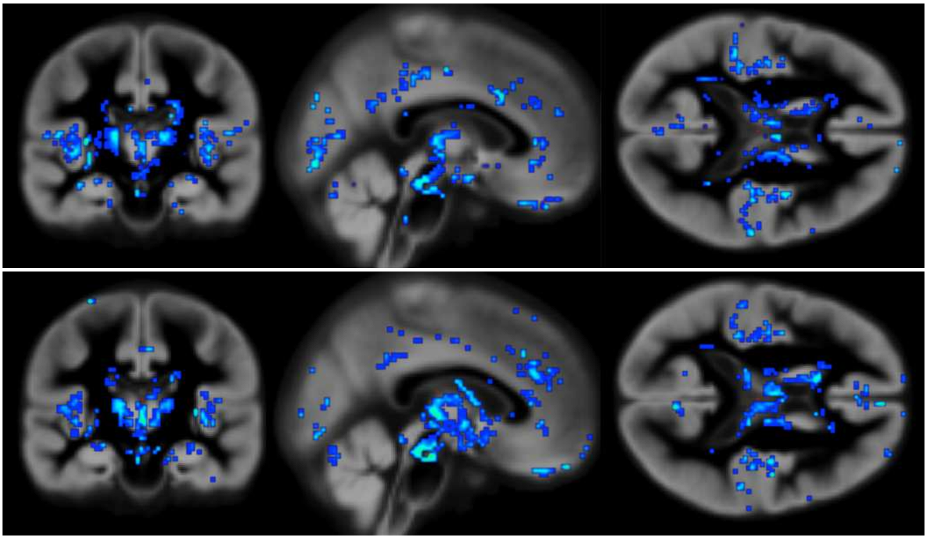

Fig. 2. Relevance voxels (in blue) for predicting age, overlaid on the average gray matter density image across all subjects. Brighter blue indicates a higher absolute value, and thus a higher relevance for prediction. Top row: Model from training on the first half of the data. Bottom row: Model from training on the second half of the data.

(age and sex matched, 43.7±23.8 years, 62.5% female). We employed each group to train the RVoxM, which was then applied to the complementary group for testing. All reported results are averages across the two training/testing sessions.

In addition to RVoxM, we used two other methods as benchmarks. The first method ("RVM") is another approach for estimating age from structural MRI [11]. It uses a principal component analysis to achieve a dimensionality-reduced representation of the images, and subsequently applies a linear RVM algorithm. We used the optimal implementation settings described in [11] and a public implementation of RVM[4]. The second benchmark ("RVoxM-NoReg") was an implementation of RVoxM with no spatial regularization, i.e., with the hyperparameter λ clamped to zero. A comparison with the latter benchmark gives us an insight into the effect of spatial regularization on the results.

Fig. 1 (top left) illustrates the root mean square error (RMSE) for the three algorithms. On average, RVoxM yields the best accuracy with a RMSE less than 9.5 years (paired t-test, $P < 0.05$); Fig. 1 (bottom left) plots the age predicted by RVoxM for each subject versus the subject's real age. Fig. 1 (top right) plots the average difference between the individual-level prediction errors (square of predicted age minus true age) obtained by RVoxM and the other two methods. On average, RVoxM achieves a statistically significantly smaller prediction error at the individual-level. RVoxM also attains the highest correlation (r-value)

[4] http://www.vectoranomaly.com/downloads/downloads.htm

between the subjects' real age and predicted age among all three methods: 0.92 for RVoxM vs. 0.90, and 0.91 for RVM and RVoxM-NoReg, respectively[5].

Fig. 2 shows $\boldsymbol{\mu}$, RVoxM's estimated voxel weights, for each of the two training sessions. Recalling that the prediction on new data is simply the linear product between $\boldsymbol{\mu}$ and the test image (Eq. (7)), the value of $\boldsymbol{\mu}$ at a specific voxel reflects the contribution of that voxel to the prediction. It can be appreciated that most voxels have a zero contribution (i.e., the model is sparse), and that the "relevance voxels" (with a non-zero contribution) occur in clusters, providing clear clues as to what parts of the gray matter are driving the age prediction process. Furthermore, the relevance voxels exhibit an overall very similar pattern across the two training sessions, providing evidence that these patterns are likely to be associated with the underlying biology and can be interpreted. We leave the interpretation of these relevance voxel patterns to future work.

7 Conclusion

In this paper, we proposed a novel Bayesian framework for image-based prediction. The proposed method yields a model where the predicted outcome is a linear combination of a small number of spatially clustered sets of voxels. We developed a computationally efficient optimization algorithm, RVoxM, to learn the properties of this model from a training data set. While RVoxM is not guaranteed to find the global optimum, our empirical results suggest that it finds a good solution in practice. Experiments on age prediction from structural brain MRI indicate that RVoxM derives excellent predictive accuracy from a small pattern of voxels that easily lends itself to neuroscientific interpretation.

Although we have used a regression model in this paper, it is straightforward to extend the technique to probabilistic classification by introducing a logistic sigmoid function [18]. In future work, we thus intend to apply RVoxM to also predict dichotomous outcomes (e.g., diagnosis), in addition to continuous ones.

Acknowledgments. Support for this research was provided by the NIH NCRR (P41-RR14075), the NIBIB (R01EB006758, R01EB013565), the NINDS (R01NS052585), the Academy of Finland (133611), the Finnish Funding Agency for Technology and Innovation (ComBrain), and MGH ECOR (2010A055663). MS received support from a KL2 Medical Research Investigator Training grant awarded via Harvard Catalyst (NIH grant 1KL2RR025757-01 and financial contributions from Harvard and its affiliations).

References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. NeuroImage 38(1), 95–113 (2007)
2. Ashburner, J., Friston, K.: Voxel-based morphometry–the methods. NeuroImage 11(6), 805–821 (2000)

[5] We note that [11] reported slightly better correlation values for RVM ($r = 0.92$), which is probably due to the increased sample size ($N = 550$) and/or different data.

3. Batmanghelich, N., Taskar, B., Davatzikos, C.: A general and unifying framework for feature construction, in image-based pattern classification. In: Prince, J., Pham, D., Myers, K. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 423–434. Springer, Heidelberg (2009)
4. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research 7, 2399–2434 (2006)
5. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
6. Cox, D., Savoy, R.: Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19(2), 261–270 (2003)
7. Cuingnet, R., Chupin, M., Benali, H., Colliot, O.: Spatial prior in SVM-based classification of brain images. In: Proceedings of SPIE, vol. 7624, p. 76241L (2010)
8. Davatzikos, C., Fan, Y., Wu, X., Shen, D., Resnick, S.: Detection of prodromal Alzheimer's disease via pattern classification of magnetic resonance imaging. Neurobiology of Aging 29(4), 514–523 (2008)
9. Dosenbach, N., Nardos, B., Cohen, A., Fair, D., Power, J., et al.: Prediction of Individual Brain Maturity Using fMRI. Science 329(5997), 1358 (2010)
10. Fan, Y., Batmanghelich, N., Clark, C., Davatzikos, C.: Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. NeuroImage 39(4), 1731–1743 (2008)
11. Franke, K., Ziegler, G., Kloppel, S., Gaser, C.: Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50(3), 883–892 (2010)
12. Kloppel, S., Stonnington, C., Chu, C., Draganski, B., Scahill, R., et al.: Automatic classification of MR scans in Alzheimer's disease. Brain 131(3), 681 (2008)
13. Li, S.: Markov random field modeling in image analysis. Springer-Verlag New York Inc., New York (2009)
14. Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., et al.: Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 51(2), 73–83 (2009)
15. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Thirion, B.: Total variation regularization for fMRI-based prediction of behaviour. Arxiv preprint arXiv:1102.1101 (2011)
16. Pereira, F., Mitchell, T., Botvinick, M.: Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45(1), S199–S209 (2009)
17. Pohl, K., Sabuncu, M.: A unified framework for MR based disease classification. In: Prince, J., Pham, D., Myers, K. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 300–313. Springer, Heidelberg (2009)
18. Tipping, M.: Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research 1, 211–244 (2001)