Regularized Tensor Factorization for Multi-Modality Medical Image Classification

Nematollah Batmanghelich, Aoyan Dong, Ben Taskar, and Christos Davatzikos

Section for Biomedical Image Analysis,
Suite 380, 3600 Market St., 19104 Philadelphia, US
{batmangh@seas,aoyan.dong@uphs,taskar@cis,christos@rad}.upenn.edu
http://www.rad.upenn.edu/sbia

Abstract. This paper presents a general discriminative dimensionality reduction framework for multi-modal image-based classification in medical imaging datasets. The major goal is to use all modalities simultaneously to transform very high dimensional image to a lower dimensional representation in a discriminative way. In addition to being discriminative, the proposed approach has the advantage of being clinically interpretable. We propose a framework based on regularized tensor decomposition. We show that different variants of tensor factorization imply various hypothesis about data. Inspired by the idea of multi-view dimensionality reduction in machine learning community, two different kinds of tensor decomposition and their implications are presented. We have validated our method on a multi-modal longitudinal brain imaging study. We compared this method with a publically available classification software based on SVM that has shown state-of-the-art classification rate in number of publications.

Keywords: Tensor factorization, Multi-view Learning, Multi-Modality, Optimization, Basis Learning, Classification.

1 Introduction

Recently, various structural (*e.g.* MRI, DTI, *etc.*) and functional (*e.g.* PET, resting state fMRI, *etc.*) imaging modalities have been utilized to develop new biomarkers for diagnosis. Multiple image modalities can provide a rich multiparametric signature that can be used to design more sensitive biomarkers [12], [10], [14]. For example, while structural MR images provide sensitive measurements for detection of atrophy in brain regions [8], FDG-PET[1] can quantify reduction of glucose metabolism in parietal lobes, the posterior cingulate, and other brain regions [5]; combination of both modalities can be very instrumental in early diagnosis of Alzheimer's disease [7].

An immediate solution to exploit multiple modalities is to concatenate all image modalities into a long vector, but learning a classifier that generalizes well in such a high dimensional space is even harder than in the uni-modality case

[1] fluorodeoxyglucose positron emission tomography.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 17–24, 2011.

because multi-modality datasets tend to be small. Therefore, dimensionality reduction plays an even more important role here. Most existing studies extract features from a few predefined areas [12]. Zhang *et al.* [14] suggested extracting features from a few pre-defined regions of interest (ROIs) and combining them into one kernel that then input to a kernel-SVM classifier. However, predefined regions might not be optimal for diagnosis on the individual level, *i.e.* classification of subjects into normal and abnormal groups. Ideally, whole image (*e.g.* brain scan) should be viewed as a large dimensional observation and relevant regions to the target variable of interest (class labels, here) should be derived from such high dimensional observation. High-dimensional pattern classification methods have been proposed for morphological analysis [6], [9] which aim to capture multivariate nonlinear relationships in the data. A critical step underlying the success of such methods is effective feature extraction and selection, *i.e.* dimensionality reduction. Batmanghelich *et al.* [2] used a constrained matrix factorization framework for dimensionality reduction while simultaneously being discriminative and representative; however, that method only works for uni-modality cases. In this paper, we propose a method inspired by the *multi-view* setting in the machine learning community [11], [1]. In the multi-view setting, there are views (sometimes in a rather abstract sense) of the data which co-occur, and there is a target variable of interest (class labels, here). The goal is to implicitly learn the target via the relationship between different views [11]. Our approach extends [2] to tensor factorization framework to handle the multi-modality case, but our formulation and optimization method is substantially different.

One could concatenate all image modalities of a subject into long columns of a matrix and simply apply [2] or a similar method. However, the advantage of extending a regularized matrix factorization to a tensor factorization framework is that because of the structure of a tensor, various factorizations can be proposed, each of which imply different hypotheses about the data. In this paper, we introduce two factorizations and explain their connotations. We derive the factorization by solving a large scale optimization problem.

2 General Framework

The novel method proposed in this paper is based on an extension of a previously presented framework for uni-modality [3], which we briefly present here for perspective. Similar to [2], the proposed method reduces the dimensionality in a discriminative way while preserving the semantics of images; hence it is clinically interpretable and produces good classification accuracy. We use regularized matrix factorization formalism for dimensionality reduction. Regularized matrix factorization decomposes a matrix into two or more matrices such that the decomposition describes the matrix as accurately as possible. Such a decomposition could be subjected to some constraints or priors. Let us assume columns of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n \cdots \mathbf{x}_N]$ represent observations (*i.e.* sample images that are vectorized), and $\mathbf{B} \in \Re^{D \times K}$ and $\mathbf{C} \in \Re^{K \times N}$ decompose the matrix such that

$\mathbf{X} \approx \mathbf{BC}$. K is the number of basis vectors, D is the number of voxels in images and N is the number of samples. The columns of matrix $\mathbf{B}$ (called $\mathbf{b}_k$) can then be viewed as basis vectors and the n^{th} column of $\mathbf{C}$ (called $\mathbf{c}_n$) contains corresponding loading coefficients or weights of the basis vectors for the n^{th} observation. The columns $\mathbf{b}_k \in \mathcal{B}$ and $\mathbf{c}_n \in \mathcal{C}$ are subjected to some constraints which define the feasible sets $\mathcal{B}$ and $\mathcal{C}$. We use variable $y_n \in \{-1(\text{abnormal}), 1(\text{healthy})\}$ to denote labels of the subjects.

An optimal basis vector ($\mathbf{b}_k$) operates as a region selector; therefore its entries (b_{jk}) must be either *on* (*i.e.* 1) or *off* (*i.e.* 0) (*i.e.* $b_{jk} \in \{0, 1\}$). Since optimizing integer values is computationally expensive, particularly for the large dimensionality characteristic of medical images, we relax this constraint to $0 \leq b_{jk} \leq 1$ which can be encoded mathematically by a combination of ℓ_∞ norm and non-negativity ($\mathbf{b} \geq 0$). Assuming that only certain structures of an anatomy are affected (*e.g.* atrophy of hippocampus in Alzheimer's disease), we can impose sparsity on the basis vectors which also make them more interpretable. The sparsity constraint can be enforced by an inequality constraint over the ℓ_1 norm of the basis vectors. These two properties constitute the feasible set for the basis vectors ($\mathcal{B}$) as follows:

$$\mathcal{B} := \{\mathbf{b} \in \Re^D : \mathbf{b} \geq \mathbf{0}, \|\mathbf{b}\|_\infty \leq 1, \|\mathbf{b}\|_1 \leq \lambda_3\}$$

where the ratio of λ_3/D encodes the ratio of sparsity of the basis vectors.

For the feasible set of coefficients ($\mathcal{C}$), we only assume non-negativity (*i.e.* $\mathcal{C} := \{\mathbf{c} : \mathbf{c} \geq \mathbf{0}\}$) because our images are non-negative but this is relaxable based on the properties of a problem.

In order to find optimal $\mathbf{B}$ and $\mathbf{C}$ matrices, we define the following constrained optimization problem:

$$\begin{aligned} \min_{\mathbf{B},\mathbf{C},\mathbf{w}\in\Re^K} \quad & \lambda_1 \mathcal{D}(\mathbf{X}; \mathbf{BC}) + \lambda_2 \sum_{n=1}^{N} \ell(y_n; f(\mathbf{x}_n; \mathbf{B}, \mathbf{w})) + \|\mathbf{w}\|_2 \\ \text{subject to: } & f(\mathbf{x}_n; \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T \mathbf{x}_n, \mathbf{w} \rangle \\ & \mathbf{b}_k \in \mathcal{B}, \quad \mathbf{c}_i \in \mathcal{C} \end{aligned} \tag{1}$$

The cost function of the optimization problem consists of two terms: 1) The *generative* term ($\mathcal{D}(\cdot;\cdot)$) encourages the decomposition, $\mathbf{BC}$, to be close to the data matrix ($\mathbf{X}$); both labeled and unlabeled data contribute to this term. 2) The *discriminative* term ($\ell(y_n; f(\mathbf{x}_n, \mathbf{B}, \mathbf{w}))$) is a *loss* function that encourages a classifier $f(\cdot)$ to produce class labels that are consistent with available labels ($\mathbf{y}$). The classifier parametrized by $\mathbf{w}$ projects each image ($\mathbf{x}_n$) on the basis vectors to produce new features ($\mathbf{v}_n = \mathbf{B}^T \mathbf{x}_n$) and produce a label. We use a linear classifier, hence $f(\mathbf{x}_n, \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T \mathbf{x}_n, \mathbf{w} \rangle$. In this paper, we set $\mathcal{D}(\mathbf{X}; \mathbf{BC}) = \|\mathbf{X} - \mathbf{BC}\|_F^2$, where λ_1 is a constant. For the loss function, we choose a hinge squared loss function: $\ell(y, \tilde{y}) = (\max\{0, 1 - y\tilde{y}\})^2$, a common choice in Support Vector Machine (SVM) literature [3].

There are three blocks in the optimization problem in Eq.(1): $\mathbf{w}, \mathbf{B}$, and $\mathbf{C}$ which is only jointly convex. In other words, if any two pairs of blocks, are fixed,

the problem is convex with respect to the remaining block. The optimization scheme starts from a random initialization of blocks, fixes two blocks, optimizes with respect to the remaining one, and repeats this process for each block. The whole process is repeated till convergence. Optimization with respect to $\mathbf{C}$ and $\mathbf{w}$ is not challenging but, due to the large-scale dimensionality of a medical image, optimization with respect to $\mathbf{B}$ requires a specialized method (see [3] for details).

3 Extension to Multi-Modality

Unlike the uni-modality case, in which each voxel stores a scalar value, in the multi-modality case, each voxel of an image is associated with an array of values. In Section 2, we stored the training data into a matrix ($\mathbf{X}$); while in multi-modality case, we need to structure the data into a tensor ($\mathbb{X}$). In fact, in the general framework (Section 2), the matrix $\mathbf{f}$ can be viewed as an order-2 tensor[2] in which the first index (rows) enumerates voxels and the second index (columns) enumerates subjects. We simply extend this matrix to an order-3 tensor in which the third index (faces) enumerates modalities. One can simply concatenate all image modalities of a subject into long columns of a matrix and simply apply [2] or a similar method. However, the advantage of extending a regularized matrix factorization to a tensor factorization framework is that various factorizations can be proposed each of which implies different hypotheses about the data because of the structure of a tensor. In this paper, we introduce two factorizations and explain their connotations (pictorially represented in Fig.1).

Our method can be viewed as *multi-view* learning [11]. In the multi-view setting, the goal is to implicitly learn about the target via the relationship between different views [11]. Depending on how to define targets, we can have different variations of the method. For example, if multiple modalities are different frequencies in spectroscopy imaging, different features extracted from diffusion tensor image (DTI), or time series in fMRI. One assumption could be that there is one hidden variable (here basis vectors: $\mathbf{B}$) that is shared across image modalities and class labels. Therefore, both class labels ($\mathbf{y}$) and data ($\mathbb{X}$) are the targets; we will refer to the method as `multi-View`$(\mathbb{X}, \mathbf{y})$.

Unlike `multi-View`$(\mathbb{X}, \mathbf{y})$, an alternative assumption could be that there is no hidden variable shared across modalities, hence every modality has its own basis vectors ($\mathbb{B}^{(v)}$), but projection on these basis vectors collaborate to predict class labels. For example, different modalities may measure quantities on non-overlapping regions of a brain (*e.g.* white matter and gray matter) each quantifying complementary features about the class labels. We refer to this variation as `multi-View`$(\mathbf{y})$. This assumption is still different than applying the uni-modality method separately because $\mathbb{B}^{(v)}$'s need to collaborate on the discriminative term.

[2] The order of a tensor is the number of indices necessary to refer unambiguously to an individual component of a tensor.

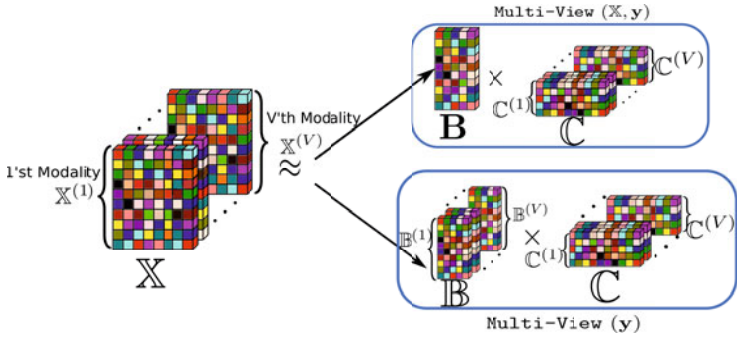

Fig. 1. The difference between the two proposed factorizations: `multi-View`$(\mathbf{y})$ versus `multi-View`$(\mathbb{X}, \mathbf{y})$. There are V modalities stored in the data tensor ($\mathbb{X}$); for `multi-View`$(\mathbf{y})$, we need to have V sets of basis vectors ($\mathbb{B}^{(1)}$,...,$\mathbb{B}^{(V)}$) and corresponding coefficients ($\mathbb{C}^{(1)}$,...,$\mathbb{C}^{(V)}$), while for `multi-View`$(\mathbb{X}, \mathbf{y})$, there is one set of basis vectors ($\mathbf{B}$) shared across modalities.

The definitions of the generative term ($\mathcal{D}(\cdot;\cdot)$) and the classifier function ($f(\cdot)$) in Eq.(1) for tensor are changed accordingly to `multi-View`$(\mathbb{X}, \mathbf{y})$ and `multi-View`$(\mathbf{y})$ (depending on the assumptions on data):

`multi-View`$(\mathbb{X}, \mathbf{y})$:	`multi-View`$(\mathbf{y})$:
$\mathcal{D}(\mathbb{X}; \mathbf{B}, \mathbb{C}) = \sum_{v=1}^{V} \|\mathbb{X}^v - \mathbf{B}\mathbb{C}^v\|_F^2$	$\mathcal{D}(\mathbb{X}; \mathbb{B}, \mathbb{C}) = \sum_{v=1}^{V} \|\mathbb{X}^v - \mathbb{B}^v\mathbb{C}^v\|_F^2$
$f(\mathbb{X}_n; \mathbf{W}, \mathbf{B}) = \sum_{v=1}^{V} \langle \mathbf{w}^v, \mathbf{B}^T\mathbb{X}_n^v \rangle$	$f(\mathbb{X}_n; \mathbf{W}, \mathbb{B}) = \sum_{v=1}^{V} \langle \mathbf{w}^v, (\mathbb{B}^v)^T\mathbb{X}_n^v \rangle$

where $\mathbb{X}$ and $\mathbb{C}$ are tensors of order-3 holding respectively all images and coefficients of the basis vectors. $\mathbb{X}^v$ and $\mathbb{C}^v$ are order-2 tensors (*i.e.* matrix) holding images and coefficients of v^{th} modality respectively. $\mathbb{X}_n$ is a order-2 tensor holding all modalities of the n^{th} subject and $\mathbb{X}_n^v$ is a order-1 tensor (*i.e.* vector) holding only v^{th} modality of the n^{th} subject. V is the number of modalities (views), $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_F$ indicate inner product and Frobenius norm respectively. $\mathbf{W}$ is a matrix holding parameters of the classifier function and $\mathbf{w}^v$ is its v^{th} column corresponding to the v^{th} modality. Notice that in `multi-View`$(\mathbf{y})$, the generative term is separable for each modality but basis matrices ($\mathbb{B}^v$'s) are coupled together through the loss function ($\ell(\cdot, \cdot)$) in Eq.(1); therefore, it is different than applying the uni-modality algorithm (Section 2) separately and concatenating extracted features later for a classifier.

4 Experiments

We acquired a subset of images from a longitudinal brain imaging study for validation of our method. The objective of this choice was to investigate the longitudinal progression of changes in brain structure (MRI) and brain function ([^{15}O]-water PET-CBF) in relation to cognitive change in cognitively normal

older adults. We used slopes of CVLT[3] score over the follow-up period as a measure of cognitive function to subdivide the entire cohort into two groups: top 20% (25 subjects) showing the highest cognitive stability (CN: cognitively normal), and bottom 20% (25 subjects) showing the most pronounced cognitive decline (CD: cognitively declining).

All T1-MR images used in this study were pre-processed according to [6] and registered to a template. Two volumetric tissue density maps [13] were formed for white matter (WM), gray matter (GM) regions. These maps quantify an expansion (or contraction) to the tissue applied by the transformation to warp the image to the template space.

Samples are divided into five folds and 4/5 of samples are used for training basis vectors (an example of which is shown in Fig.2); projections on these basis vectors are used as features and are fed to a SVM classifier.

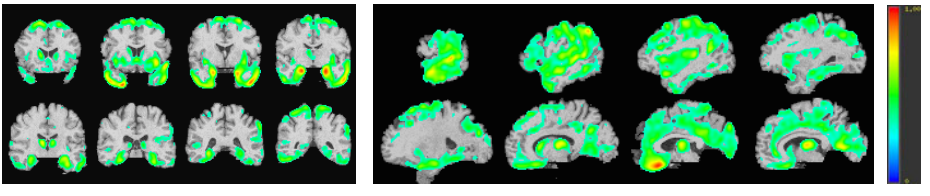

Fig. 2. Two examples of the basis vectors shown in different cuts. Left: `Multi-View`$(\mathbb{X}, \mathbf{y})$, Right: `Multi-View`$(\mathbf{y})$ ($\gamma^* = 100$; number of basis vectors is 60).

In uni-parametric dataset, the algorithm is relatively stable as long as λ's are chosen within reasonable ranges (see [3]). We set the parameters to the most frequently chosen parameters used for a uni-modality case on a totally different dataset. Numbers reported in Table 1 are produced using such parameters. Nevertheless, we performed sensitivity analysis with respect to ratio of λ_1/λ_2 and number of basis vectors, K (see Fig.3). For notational brevity, we used γ^* for ratio of λ_1/λ_2 we used for Table1. Different curves in Fig.3 denote different ratios of λ_1/λ_2. As `Multi-View`$(\mathbf{y})$ is relatively stable with respect to K and different ratios, performance of `Multi-View`$(\mathbb{X}, \mathbf{y})$ improves as K increases. Although parameters that are more inclined toward the unsupervised setting (*e.g.* $\lambda_1/\lambda_2 = 10\gamma^*$) underperform settings that are excessively discriminative (*e.g.* $\lambda_1/\lambda_2 = 0.001\gamma^*$), are more stable. Weak regularization imposed on the excessively discriminative settings can explain this observation.

Table 1 reports the average classification rates on the left-out folds for different scenarios and methods. We used a publically available software, called `COMPARE` [6], for comparison. The `COMPARE` method has been applied to many problems and has been claimed to perform very well. Its variants, *i.e.* `COMPARE` and `m-COMPARE`, are similar to `Multi-View`$(\mathbf{y})$ and `Multi-View`$(\mathbb{X}, \mathbf{y})$ respectively. For comparison, we have included `Single-View` results for each scenario in which basis vectors are extracted independently and features are concatenated

[3] California Verbal Learning Test [4].

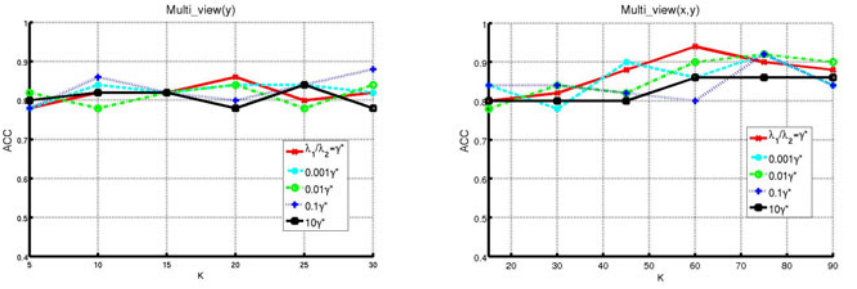

Fig. 3. Sensitivity Analysis: accuracy rates with respect to different number of basis vectors (K) for various ratios of λ_1/λ_2. Left:`Multi-View`($\mathbf{y}$). Right: `Multi-View`($\mathbb{X}, \mathbf{y}$)

and fed to the same procedure to find the best parameters for a classifier as the multi-view methods. Since results shown in the table are column-wise comparable, the highest values in the column are magnified with a bold font in each column. In general, `Multi-View`($\mathbb{X}, \mathbf{y}$) or its counterpart `m-COMPARE` perform better. In all columns, at least one of the multi-view methods outperforms the single view equivalent and the best performance is achieved by `Multi-View`($\mathbb{X}, \mathbf{y}$).

Table 1. Comparison of classification accuracy rates for different scenarios and different methods on "cognitively normal" (NC) versus "cognitively declining" (CD) subjects. Results are reported in the format: accuracy (sensitivity,specificity); with $\gamma^* = 100$; total number of basis vectors in each experiment is 60.

	NC vs. CD			
	(WM,PET)	(WM,GM)	(GM,PET)	(GM, WM, PET)
`Multi-View`($\mathbb{X}, \mathbf{y}$)	0.82 (0.84,0.8)	0.76 (0.72,0.8)	**0.84** (0.88,0.8)	**0.94** (0.88,1.0)
`Multi-View`($\mathbf{y}$)	0.86 (0.84,0.88)	0.84 (0.8,0.88)	0.78 (0.8,0.76)	0.84 (0.84,0.84)
`m-COMPARE`	**0.88** (0.8,0.96)	**0.86** (0.88,0.84)	0.8 (0.8,0.8)	0.86 (0.84,0.88)
`COMPARE`	0.78 (0.68,0.88)	0.82 (0.76,0.88)	0.82 (0.84,0.8)	0.82 (0.76,0.88)
`Single-View`	0.84 (0.8,0.88)	0.84 (0.8,0.88)	0.82 (0.84,0.8)	0.8 (0.76,0.84)

5 Conclusion

We proposed a framework that exploits all modalities in a dataset simultaneously to reduce dimensionality in a discriminative yet interpretable way. Inspired by multi-view learning, two variants of constrained tensor factorization are suggested each of which implies different hypothesis about the data. We showed that the algorithm is relatively robust with respect to choice of parameters and achieves good classification results. Computational expense of the algorithm is moderate and as future work, we plan to apply it to case for which number of modalities is large (*e.g.* HARDI data or time series).

References

1. Ando, R.K., Zhang, T.: Two-view feature generation model for semi-supervised learning. In: Proceedings of the 24th International Conference on Machine Learning, ICML 2007, pp. 25–32. ACM, New York (2007)
2. Batmanghelich, N., Taskar, B., Davatzikos, C.: A general and unifying framework for feature construction, in image-based pattern classification. Inf. Process. Med. Imaging 21, 423–434 (2009)
3. Batmanghelich, N., Ye, D.H., Pohl, K., Taskar, B., Davatzikos, C.: Disease classification and prediction via semi-supervised dimensionality reduction. In: 2011 IEEE International Symposium on Biomedical Imaging (2011)
4. Delis, D., Kramer, J., Kaplan, E., Ober, B.: California Verbal Learning Test-Research Edition. The Psychological Corporation, New York (1987)
5. Diehl, J., Grimmer, T., Drzezga, A., Riemenschneider, M., Frstl, H., Kurz, A.: Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. a pet study. Neurobiol. Aging 25(8), 1051–1056 (2004)
6. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: Compare: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging 26(1), 93–105 (2007)
7. Foster, N.L., Heidebrink, J.L., Clark, C.M., Jagust, W.J., Arnold, S.E., Barbas, N.R., DeCarli, C.S., Turner, R.S., Koeppe, R.A., Higdon, R., Minoshima, S.: Fdg-pet improves accuracy in distinguishing frontotemporal dementia and alzheimer's disease. Brain 130(Pt 10), 2616–2635 (2007)
8. Fox, N.C., Schott, J.M.: Imaging cerebral atrophy: normal ageing to alzheimer's disease. Lancet 363(9406), 392–394 (2004)
9. Golland, P., Grimson, W.E.L., Shenton, M.E., Kikinis, R.: Deformation analysis for shape based classification. In: Insana, M.F., Leahy, R.M. (eds.) IPMI 2001. LNCS, vol. 2082, pp. 517–530. Springer, Heidelberg (2001)
10. Hinrichs, C., Singh, V., Xu, G., Johnson, S.: Mkl for robust multi-modality ad classification. Med. Image Comput. Comput. Assist. Interv. 12(Pt 2), 786–794 (2009)
11. Kakade, S., Foster, D.: Multi-view regression via canonical correlation analysis, pp. 82–96 (2007)
12. Landau, S.M., Harvey, D., Madison, C.M., Reiman, E.M., Foster, N.L., Aisen, P.S., Petersen, R.C., Shaw, L.M., Trojanowski, J.Q., Jack, C.R., Weiner, M.W., Jagust, W.J., Initiative, A.D.N.: Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 75(3), 230–238 (2010)
13. Shen, D., Davatzikos, C.: Very high-resolution morphometry using mass-preserving deformations and hammer elastic registration. Neuroimage 18(1), 28–41 (2003)
14. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: The ADNI: Multimodal classification of alzheimer's disease and mild cognitive impairment. Neuroimage (January 2011)