

DISEASE CLASSIFICATION AND PREDICTION VIA SEMI-SUPERVISED DIMENSIONALITY REDUCTION

Kayhan N. Batmanghelich ^{*}, Dong H. Ye ^{*}, Kilian M. Pohl ^{*}, Ben Taskar ^{**}, Christos Davatzikos ^{*}, and ADNI [†]

^{*} Section of Biomedical Image Analysis (SBIA), Department of Radiology, University of Pennsylvania

^{**} Computer and Information Science Department, University of Pennsylvania

ABSTRACT

We present a new semi-supervised algorithm for dimensionality reduction which exploits information of unlabeled data in order to improve the accuracy of image-based disease classification based on medical images. We perform dimensionality reduction by adopting the formalism of constrained matrix decomposition of [1] to semi-supervised learning. In addition, we add a new regularization term to the objective function to better capture the affinity between labeled and unlabeled data. We apply our method to a data set consisting of medical scans of subjects classified as Normal Control (CN) and Alzheimer (AD). The unlabeled data are scans of subjects diagnosed with Mild Cognitive Impairment (MCI), which are at high risk to develop AD in the future. We measure the accuracy of our algorithm in classifying scans as AD and NC. In addition, we use the classifier to predict which subjects with MCI will converge to AD and compare those results to the diagnosis given at later follow ups. The experiments highlight that unlabeled data greatly improves the accuracy of our classifier.

Index Terms— Semi-supervised Learning, Basis Learning, Matrix factorization, Optimization, Alzheimer’s disease, Mild Cognitive Impairment (MCI)

1. INTRODUCTION

The medical imaging community frequently relies on voxel-wise image analysis to define areas of difference between groups [2] or to extract features for classification. However, this approach is not well suited for identify complex population differences, because it does not take into account the multivariate relationships in data [1, 3]. Moreover, regions showing significant group difference are not necessarily discriminative for classifying individuals. In order to overcome these limitations, high-dimensional pattern classification methods have been proposed in the recent literatures [4, 5]. A fundamental limitation in these methods with respect to medical imaging is their need for large training sets of labeled data. One way to address this issue is to train the methods using

unlabeled data, which may exist in large quantities. However, it is not clear how to exploit unlabeled data for dimensionality reduction. We will explore these topics in the subsequent sections.

Presence of unlabeled data in the learning process, also referred to as semi-supervised learning, can have interesting applications in medical imaging. For example, subjects may deviate from the normal population and may be diagnosed with a certain disease in future follow-up scans; class labels of such subjects are not very well-defined. This is the case for subjects diagnosed as Mild Cognitive Impairment (MCI) who show some impairment in their cognitive scores and have high risk to develop Alzheimer’s disease (AD) in near future. One may be interested to predict future follow-up labels (converging to AD or not) of the MCI subjects by considering them as unlabeled data. Considering MCI subjects as unlabeled data allows an algorithm to locate unlabeled subjects in the spectrum of normal vs. abnormal.

Our proposed method is combination of two techniques: the first technique proposed in [1] reduces the dimensionality in a discriminative way while preserving the semantics of images; hence it is clinically interpretable and produce good classification accuracy. The second technique incorporates information provided by unlabeled data. It computes pair-wise similarity of subjects while considering intrinsic geometry of data distribution [6]. Pair-wise similarity between subjects encodes relationship between labeled and unlabeled data and it is shown to improve classification accuracy in presence of unlabeled data [7]. Finally, the method is cast as a constrained optimization problem similar to [1] but the optimization cost function and its constraints as well as our optimizer are significantly different.

Section 2 briefly describes our general framework that we expand upon. Section 3 focuses on the added regularization term that makes use of the information latent in the unlabeled data. Section 4 presents a solution to the resulting optimization problem. The results of applying our method to a subset of ADNI data set are discussed in Section 5. Finally in Section 6, we conclude with a discussion of the method.

[†] Data used in the preparation of this article were obtained from the Alzheimer Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI)

2. GENERAL FRAMEWORK

Dimensionality reduction is typically applied to achieve a generalizable classification rate when number of samples is less than dimensionality of features. We propose to use regularized matrix factorization formalism for dimensionality reduction. This framework allows to keep the semantics of images; hence it produces interpretable results. The idea is similar to [1] but formulation and its extensions (Section 3) and other details are significantly different.

In this section, we lay out the general framework. Regularized matrix factorization decomposes a matrix into two or more matrices such that the decomposition describes the matrix as accurate as possible. Such decomposition could be subjected to some constraints or priors. Let assume columns of $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_i \cdots \mathbf{x}_N]$ represent observations (*i.e.* sample images that are vectorized), and $\mathbf{B} \in \mathbb{R}^{D \times r}$ and $\mathbf{C} \in \mathbb{R}^{r \times N}$ decompose the matrix such that $\mathbf{X} \approx \mathbf{BC}$. r is number of basis vectors which is a parameter of the algorithm and D is number of voxels of images and N is number of samples. The columns of matrix \mathbf{B} (called \mathbf{b}_k) can then be viewed as basis vectors and the i 'th column of \mathbf{C} (called \mathbf{c}_i) contains corresponding loading coefficients or weights of the basis vectors for the i 'th observation. The columns $\mathbf{b}_k \in \mathcal{B}$ and $\mathbf{c}_i \in \mathcal{C}$ are subjected to some constraints, which we denote with the feasibility sets \mathcal{B} and \mathcal{C} . We use variable $y_i \in \{-1, 0, 1\}$ to denote labels of the subjects. Healthy subjects are denoted by 1 and abnormal ones by -1 ; 0 is used simply for unlabeled subject indicating that labels are not decided for them.

In order to define the feasible sets (\mathcal{B}), we need to elaborate the requirements that our algorithm should satisfy: 1) The basis vectors must be anatomically meaningful: this means that a constructed basis vector should correspond to contiguous anatomical regions preferably in areas which are biologically related to a pathology of interest. In other words, the basis vectors should not resemble spread disjoint voxels. Sparsity of the basis vectors, *i.e.* a relatively small number of voxels with non-zeros values, encourages it to be more spatially localized. 2) The basis must be discriminative: we are interested in finding features, *i.e.* projections onto the basis vectors, that construct spatial patterns that best differentiate groups. 3) The decomposition (\mathbf{BC}) should be a good representative of data without compromising the previous properties.

In this paper, we assume that images are non-negative hence it is reasonable to impose non-negativity on \mathbf{B} and \mathbf{C} . Thus, our proposed method can be viewed as a variant of non-negative matrix factorization (NMF). NMF [8],[9] is an additive model that is known to decompose images into meaningful parts that are shared across subjects; this property is favorable for our application. Such part-based decomposition property encourages basis vectors to be similar to anatomically meaningful parts of images (*i.e.* Hippocampus, Caudate, *etc.* for brain images). We also assume that certain structures

(*e.g.* Hippocampus) of an anatomy of interest (*e.g.* brain) are affected by the abnormality (*e.g.* shrinkage of Hippocampus Alzheimer's disease); this property can be viewed as a sparsity constraint on the basis vectors which also help the basis vectors to be more interpretable. We encode these properties via non-negativity on the coefficients and combination of non-negativity and ℓ_1 and ℓ_∞ norms on the basis vectors. The ℓ_1 -norm encourages the sparsity property and combination of ℓ_∞ and non-negativity promotes part-based decomposition:

$$\begin{aligned} \mathcal{C} &:= \{\mathbf{c} \in \mathbb{R}^r : \mathbf{c} \geq 0\} \\ \mathcal{B} &:= \{\mathbf{b} \in \mathbb{R}^D : \mathbf{b} \geq 0, \|\mathbf{b}\|_\infty \leq 1, \|\mathbf{b}\|_1 \leq \lambda_3\} \end{aligned} \quad (1)$$

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

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

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{C}, \mathbf{w} \in \mathbb{R}^r} \quad & \mathcal{D}(\mathbf{X}; \mathbf{BC}) + \sum_{i \in \mathcal{L}} \ell(y_i; \langle \mathbf{B}^T \mathbf{x}_i, \mathbf{w} \rangle) + \|\mathbf{w}\|_2 \\ \text{subject to:} \quad & \mathbf{b}_k \in \mathcal{B}, \quad \mathbf{c}_i \in \mathcal{C} \end{aligned} \quad (2)$$

The cost function of the optimization problem consists of two terms: 1) *Generative* term ($\mathcal{D}(\cdot; \cdot)$) that encourages the decomposition, \mathbf{BC} , to be close to the data matrix (\mathbf{X}); both labeled and unlabeled data contribute to this term. 2) *Discriminative* term ($\ell(y_i; f(\mathbf{x}_i, \mathbf{B}, \mathbf{w}))$) 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}_i) on the basis vectors to produce new features ($\mathbf{v}_i = \mathbf{B}^T \mathbf{x}_i$) and produces a labels. In this paper, we use a linear classifier, hence $f(\mathbf{x}_i, \mathbf{B}, \mathbf{w}) = \langle \mathbf{B}^T \mathbf{x}_i, \mathbf{w} \rangle$. Only labeled data contribute to the discriminative term. Various choice are possible for $\mathcal{D}(\cdot; \cdot)$ and $\ell(\cdot; \cdot)$. In this paper, we set $\mathcal{D}(\mathbf{X}; \mathbf{BC}) = \lambda_1 \|\mathbf{X} - \mathbf{BC}\|_F^2$ where λ_1 is a constant. For the loss function, we choose a hinge squared loss function: $\ell(y, \tilde{y}) = (\max\{0, 1 - y\tilde{y}\})^2$ which is a common choice in Support Vector Machine literature. Summing over \mathcal{L} for the loss function simply indicates that the labeled subjects participate in this term.

3. SEMI-SUPERVISED REGULARIZER

In this section, we extend the method for semi-supervised learning by extending the simple ℓ_2 regularization term of \mathbf{w} in (Eqn.(2)) to a more general *Regularization* term ($\mathcal{R}(\cdot; \cdot)$) involving both basis matrix (\mathbf{B}) and \mathbf{w} vector:

$$\min_{\mathbf{b}_k \in \mathcal{B}, \mathbf{c}_i \in \mathcal{C}, \mathbf{w} \in \mathbb{R}^r} \quad \mathcal{D}(\mathbf{X}; \mathbf{BC}) + \sum_{i \in \mathcal{L}} \ell(y_i; \langle \mathbf{B}^T \mathbf{x}_i, \mathbf{w} \rangle) + \mathcal{R}(\mathbf{B}, \mathbf{w})$$

We note that our model only regularizes \mathbf{B} and \mathbf{w} but not \mathbf{C} . Our proposed regularization term is inspired by Laplacian-SVM (lapSVM) [7], and particularly the linear version of [10]. In lapSVM, samples (labeled and unlabeled) are considered as nodes of a graph. Every two nodes are connected

via an edge and there is a weight associated to the edge determining how similar the two nodes (samples) are. Properties of this graph is used to define a regularization for lapSVM [7]. Similar to lapSVM, our proposed regularization (Eqn.(3)) consists of two terms: the first term, which is a simple ℓ_2 -norm, controls the complexity of the classifier in the ambient space (here \mathbb{R}^r) and the second term is an appropriate penalty term reflecting the intrinsic structure of distribution of sample. The second term encourages the classifier to produce class label for an image that is similar to the labels of its most similar neighbors on the graph:

$$\begin{aligned} \mathcal{R}(\mathbf{B}, \mathbf{w}) &= \gamma_A \|\mathbf{w}\|_2^2 + \gamma_I \|\mathbf{w}\|_{\mathbf{Q}(\mathbf{B})}^2 \\ \mathbf{Q}(\mathbf{B}) &= \mathbf{B}^T \mathbf{X} \mathbf{L} \mathbf{X}^T \mathbf{B} \end{aligned} \quad (3)$$

where \mathbf{Q} is a $r \times r$ matrix and $\|\mathbf{z}\|_{\mathbf{Q}}^2 = \mathbf{z}^T \mathbf{Q} \mathbf{z}$; γ_A and γ_I balance relative weight between two terms and \mathbf{L} is a $N \times N$ precomputed graph Laplacian, given labeled and unlabeled data. We will shortly explain how to build a graph of labeled and unlabeled images and how to compute the graph Laplacian.

Inspired by [6], we compute the Laplacian matrix on the complete, weighted, and undirected graph with each image in the dataset being represented by one node. First, we define the distance between two images $I_i : \Omega_I \rightarrow R$ and $I_j : \Omega_J \rightarrow R$ as the weighted sum of residual and harmonic energy after registration:

$$d(I_i, I_j) = \kappa_1 \int \|I_i(x + h(x)) - I_j(x)\|_2^2 dx + \int \|\nabla h(x)\|_2 dx$$

where $h : \Omega_I \rightarrow R^3$ is the displacement field such that if $x \in \Omega_I$ then $x + h(x) \in \Omega_J$ and κ_1 is the registration parameter [6]. The images are registered by the Diffemorphic Demons [11] with coarse resolution due to its fast speed. Then, the non-negative weights \mathbf{K}_{ij} between two images are assigned by exponential of a negative distance: $\mathbf{K}_{ij} = e^{-d(I_i, I_j)/\sigma}$. A weights matrix \mathbf{K} is obtained by collecting the edge weights \mathbf{K}_{ij} for every image pair and a diagonal matrix \mathbf{T} contains the row sums for each vertex $t_{ii} = \sum_j \mathbf{K}_{ij}$. The Laplacian \mathbf{L} [12] is then constructed by $\mathbf{L} = \mathbf{T} - \mathbf{K}$.

The Laplacian \mathbf{L} encodes information relating to all pairwise relations between the images. In words, the second term in Eqn. 3 promotes smoothness of class labels on the graph. By multiplying the vector \mathbf{w} on the left and right of $\mathbf{Q}(\mathbf{B})$, (Eqn.(3)) can be written as $\mathbf{w}^T \mathbf{Q}(\mathbf{B}) \mathbf{w} = \sum_{ij} \mathbf{K}_{ij} (\mathbf{w} \mathbf{B} \mathbf{x}_i - \mathbf{w} \mathbf{B} \mathbf{x}_j)^2$.

4. OPTIMIZATION

Having all terms the cost function and the constraints, we can form the optimization problem as follows:

$$\begin{aligned} \min_{\mathbf{B}, \mathbf{C}, \mathbf{w}} \quad & \lambda_1 \|\mathbf{X} - \mathbf{B}\mathbf{C}\|_F^2 + \sum_{i \in \mathcal{L}} \ell(y_i; \langle \mathbf{B}^T \mathbf{x}_i, \mathbf{w} \rangle) \\ & + \gamma_A \|\mathbf{w}\|_2^2 + \gamma_I \|\mathbf{w}\|_{\mathbf{Q}(\mathbf{B})}^2 \\ \text{subject to:} \quad & \mathbf{b}_k \geq 0, \|\mathbf{b}_k\|_\infty \leq 1, \|\mathbf{b}_k\|_1 \leq \lambda_3 \\ & \mathbf{C} \geq \mathbf{0} \end{aligned}$$

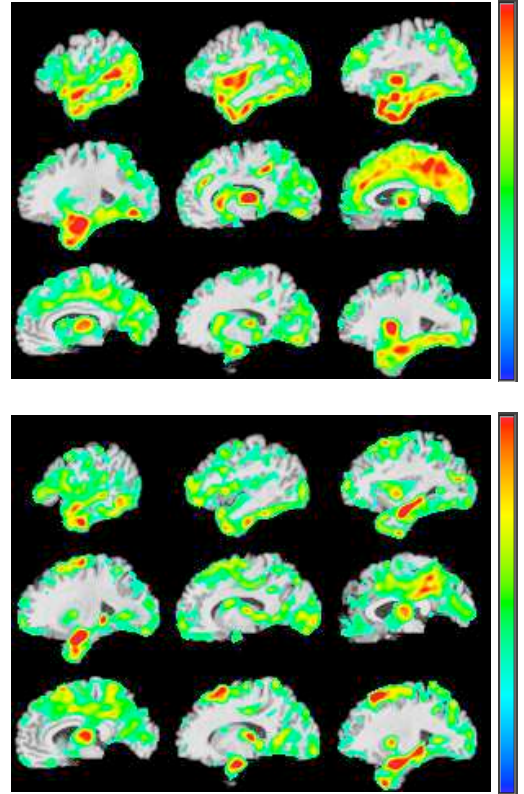


Fig. 1: Two examples of learned basis vectors shown in different sagittal cuts. The red color indicates one and blue (transparent) indicates zero.

The optimization problem is not jointly convex with respect to all blocks of variables ($\mathbf{B}, \mathbf{C}, \mathbf{w}$). However, fixing each pairs of blocks, the optimization is convex with respect the other block. Our proposed algorithm uses this property and applies a block-wise optimization scheme, *i.e.* fixing each pairs of blocks (*e.g.* fixing \mathbf{B}, \mathbf{C}) and optimizing with respect to the other (*e.g.* \mathbf{w}) and iterate till some convergence criterion is satisfied. This scheme guarantees to decrease the cost function in each iteration and, specifically for this problem, is guarantees to converge to a local minimum. Block-wise optimization with respect to \mathbf{C} and \mathbf{w} are not challenging but it is challenging with respect to \mathbf{B} because it is very high dimensional. Detail discussion of our proposed efficient method to circumvent this issue is beyond the scope of this paper; briefly, we use Spectral Projected Gradient (SPG) [13] with an efficient projection method on the feasible set (\mathcal{B}).

5. EXPERIMENTS

To evaluate the proposed method, we applied our method on a subset of Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. The dataset we used consists of structural MRI images of 54 Alzheimer's (AD), 238 Mild Cognitive

Impairment (MCI), and 63 Normal Control (CN) subjects at the baseline time point. Among 238 MCI subjects, 68 patients converted to AD in a few follow-up months (cMCI) and 170 subjects have not converted yet or may not ever convert (ncMCI). The preprocessing pipeline (bias correction, skull stripping, tissue characterization, and spatial normalization) is designed according to what is suggested in [1]; and it computes so-called RAVENS maps [14] that quantifies the density of the gray matter tissue. The observation matrix \mathbf{X} is obtained by collecting the RAVENS map for each subject.

Then, we conducted the semi-supervised basis learning method on \mathbf{X} . For computational efficiency, the basis vectors \mathbf{B} were learned only from 79 MCI subjects (unlabeled data), and 20 AD / 20 CN subjects (labeled data). The labeled subjects were divided into five folds for cross validation (4/5 for training and 1/5 for testing) and the 79 MCI subjects were shared as unlabeled data across folds. In order to investigate the affect of number of labeled data, we performed four basis learning by increasing number of revealed labels from 4 to 32; each fold has $4/5 \times (20 + 20) = 32$ AD/CN subjects and we revealed labels of AD/CN subjects as: $\{(2, 2), (4, 4), (8, 8), (16, 16)\}$. Figure 1 shows two examples of learned basis vectors in different sagittal cuts. Hippocampus, Cingulate and temporal lobe regions, which are associated with memory loss [15], are precisely highlighted in the basis.

After basis learning, features are extracted by projecting all images on the learned basis vectors. These features were fed into a supervised classifier (**SC**: Simple Logistic) and a semi-supervised classifier (**SSC**: linear Laplacian SVM [10]) to produces labels. To have a reference point for comparison, we also learned the basis without unlabeled data (supervised basis learning). Figure 2a plots accuracy rates of AD/CN with respect to number labeled data in different settings. Accuracy rates were computed on the left out labeled data and the rest of the labeled data that was not introduced during basis learning or classifier learning; in other words, the 2/3 of the data not contributed in training was used for validation. For brevity, **SF** in Figure 2a indicates Supervised Features, *i.e.* using only labeled data to learn the basis vectors, and **SSF(lap)** denotes Semi-Supervised Features, *i.e.* using labeled and unlabeled data to learn the basis vectors while the Laplacian regularization in (Eqn.(3)) is used. Figure2 shows different scenarios for classification: supervised features fed into a supervised classifier (**SF + SC**) and a semi-supervised classifier (**SF + SSC**) and compares it with semi-supervised features fed into a supervised classifier (**SSF(lap) + SC**) and a semi-supervised classifier (**SSF(lap) + SSC**). The figure shows that semi-supervised basis learning in presence of unlabeled data helps improving accuracy of detecting AD/CN: compare the blue line with the green line and the pink line with the red line. It also shows that if the semi-supervised classifier (**SSC**) is used instead of a fully supervised classifier (**SC**), we achieve the maximum performance. It also shows that presence of

unlabeled data helps improving classification rate with fewer labeled data as it saturates faster than the supervised cases.

Figure 2b shows similar results for MCI subjects. However, real labeling for ncMCI subjects are vague because they may or may not convert to AD in the future unlike cMCI who have already converted in a few follow up scans. In order to evaluate performance of the algorithm on this dataset, we reported the Area Under Curve (AUC) instead of accuracy. This figure also shows that semi-supervised basis learning improves the AUC and we can see even higher boost by using semi-supervised classifier in comparison to a fully supervised classifier.

Figure2 also includes experiments in which the Laplacian regularization is set to zero (*i.e.* $\gamma_I = 0$). Notice that the algorithm is still a semi-supervised learning because both labeled and unlabeled contribute in the generative term ($\mathcal{D}(\cdot; \cdot)$) but only the labeled data contributes the discriminative term ($\ell(\cdot; \cdot)$). This experiment is denoted by **SSF** comparing to **SSF(lap)** where the Laplacian term is in use. The purpose of this experiment is to investigate which term (the generative term or the Laplacian term) has more impact in the semi-supervised learning. In Figure2a, there is no significant difference between **SSF** and **SSF(lap)** meaning that most of gain due to presence of unlabeled data is obtained by the generative term rather than Laplacian regularization (Eqn.(3)). However Figure2b shows that adding the Laplacian regularization improves AUC slightly.

It is worth mentioning that the proposed algorithm has couple of parameters to tune (λ 's and γ 's). Due to space limitation, sensitivity analysis with respect to these parameters is not presented in this paper but fully investigated in a longer version of this paper which is currently under review for journal publication. We refer interested readers to a draft version of the journal paper that soon will be accessible through the first author homepage.

6. CONCLUSION

We presented a framework to reduce the dimension of image features in presence of unlabeled data. Constrained matrix decomposition problem was adapted for generative and discriminative basis learning and extended to semi-supervised formulation. Semi-supervised dimensionality reduction method outperforms supervised dimensionality reduction for both classification tasks considered in this paper, both in terms of classifier accuracy and area under curves (AUC). It seems that the generative term has more impact in semi-supervised learning rather than the Laplacian regularization in (Eqn.(3)); nevertheless further investigation is required to find the right balance between these two terms to exploit unlabeled data further. Future work will involve evaluating semi-supervised dimensionality reduction method for other medical image datasets and other scenarios.

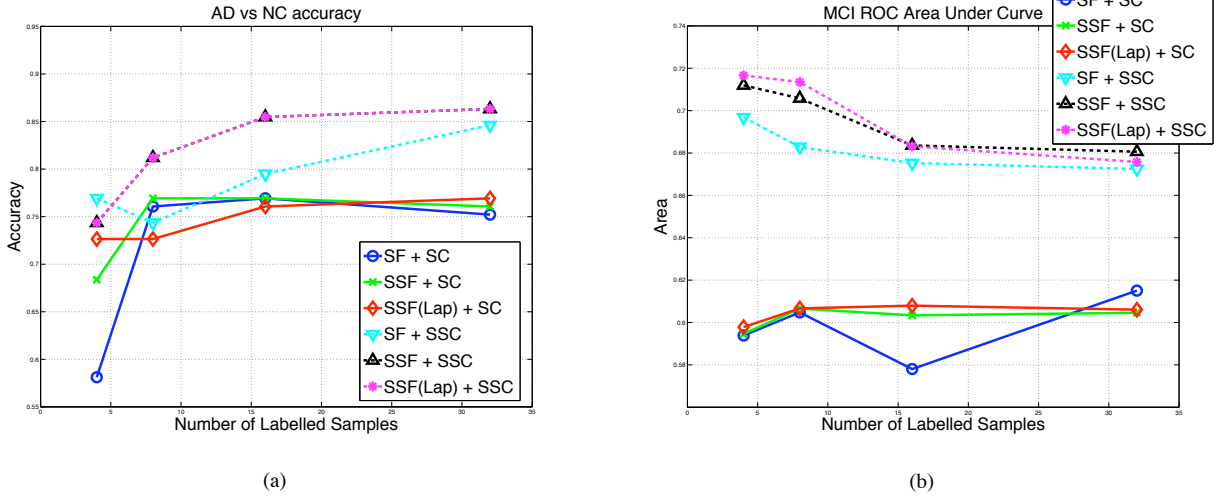


Fig. 2: For different number of labeled samples, this graph compares different measures (AUC and accuracy) for Supervised Features (SF) and Semi-Supervised Features fed into a supervised classifier (SC) and a semi-supervised classifier (SSC). **SSF(lap)** indicates $\gamma_I > 0$ while **SSF** denotes $\gamma_I = 0$: (a) represents AD-CN accuracy; (b) indicates Area Under Curve (AUC) for the MCI subjects (converters vs non-converters).

7. ACKNOWLEDGMENT

The research was supported by an ARRA supplement to NIH NCRR (P41 RR13218).

8. REFERENCES

- [1] Nematollah Batmanghelich, Ben Taskar, and Christos Davatzikos, "A general and unifying framework for feature construction, in image-based pattern classification.," *Inf Process Med Imaging*, vol. 21, pp. 423–434, 2009.
- [2] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods.," *Neuroimage*, vol. 11, no. 6 Pt 1, pp. 805–821, Jun 2000.
- [3] Melissa K Carroll, Guillermo A Cecchi, Irina Rish, Rahul Garg, and A. Ravishankar Rao, "Prediction and interpretation of distributed neural activity with sparse models.," *Neuroimage*, vol. 44, no. 1, pp. 112–122, Jan 2009.
- [4] Yong Fan, Dinggang Shen, Ruben C Gur, Raquel E Gur, and Christos Davatzikos, "Compare: classification of morphological patterns using adaptive regional elements.," *IEEE Trans Med Imaging*, vol. 26, no. 1, pp. 93–105, Jan 2007.
- [5] Polina Golland, W. Eric L. Grimson, Martha E. Shenton, and Ron Kikinis, "Deformation analysis for shape based classification," in *In IPMI*, 2001, vol. 2082, pp. 517–530.
- [6] Jihun Hamm, Dong Hye Ye, Ragini Verma, and Christos Davatzikos, "Gram: A framework for geodesic registration on anatomical manifolds.," *Med Image Anal*, vol. 14, no. 5, pp. 633–642, Oct 2010.
- [7] Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, November 2006.
- [8] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization.," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct 1999.
- [9] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts," 2003.
- [10] Sindhwani, Vikas and Niyogi, Partha, "Linear manifold regularization for large scale semi-supervised learning," in *Proc. of the 22nd ICML Workshop on Learning with Partially Classified Training Data*, 2005.
- [11] Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache, "Diffeomorphic demons: efficient non-parametric image registration.," *Neuroimage*, vol. 45, no. 1 Suppl, pp. S61–S72, Mar 2009.
- [12] Fan R. K. Chung, *Spectral Graph Theory*.
- [13] Ernesto G. Birgin, José Mario Martínez, and Marcos Raydan, "Nonmonotone spectral projected gradient methods on convex sets," Philadelphia, PA, USA, 2000, vol. 10, pp. 1196–1211, Society for Industrial and Applied Mathematics.
- [14] C. Davatzikos, A. Genc, D. Xu, and S. M. Resnick, "Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy.," *Neuroimage*, vol. 14, no. 6, pp. 1361–1369, Dec 2001.
- [15] Gael Chetelat, Beatrice Desgranges, Vincent De La Sayette, Fausto Viader, Francis Eustache, and Jean-Claude Baron, "Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment.," *Neuroreport*, vol. 13, no. 15, pp. 1939–1943, Oct 2002.