Semi-supervised Manifold Learning with Affinity Regularization for Alzheimer's Disease Identification Using Positron Emission Tomography Imaging

Shen Lu, Yong Xia*, *Member, IEEE*, Tom Weidong Cai, *Member, IEEE*, and David Dagan Feng, *Fellow, IEEE*

***Abstract*—Dementia, Alzheimer's disease (AD) in particular is a global problem and big threat to the aging population. An image based computer-aided dementia diagnosis method is needed to providing doctors help during medical image examination. Many machine learning based dementia classification methods using medical imaging have been proposed and most of them achieve accurate results. However, most of these methods make use of supervised learning requiring fully labeled image dataset, which usually is not practical in real clinical environment. Using large amount of unlabeled images can improve the dementia classification performance. In this study we propose a new semi-supervised dementia classification method based on random manifold learning with affinity regularization. Three groups of spatial features are extracted from positron emission tomography (PET) images to construct an unsupervised random forest which is then used to regularize the manifold learning objective function. The proposed method, stat-of-the-art Laplacian support vector machine (LapSVM) and supervised SVM are applied to classify AD and normal controls (NC). The experiment results show that learning with unlabeled images indeed improves the classification performance. And our method outperforms LapSVM on the same dataset.**

I. Introduction

Dementia disease is a neurodegenerative brain disorder which is usually characterized by the progressive loss of memory and cognitive impairment [1]. There are several different types of dementia categorized by cerebral metabolic patterns [2]. Among them, the most prevalent type of dementia disease is Alzheimer's disease (AD) which accounts for about 65% of all dementia cases globally [3] and the forecast for the next decades is not optimistic [4]. The decrements of cerebral metabolic activity caused by AD are mainly evident in angular gyrus and parietotemporal regions [2]. They can be detected by structural and functional brain imaging modalities such as magnetic resonance imaging (MRI) [5, 6] and positron emission tomography (PET) [3, 7]. PET provides great assistance to doctors for clinical AD diagnosis in terms of differential diagnosis, longitudinal study and early detection [3]. However, conventional clinical diagnosis for AD involves exhaustive visual examining of brain images by doctors. This method solely depends on the skill and experience of the examiners, in which case biased conclusions are likely to be drawn. Therefore, a computer-aided diagnosis method for AD is needed to assist doctors by providing the 'second opinion'. There are many studies tried to achieve this goal using various methods sourced from machine learning theories. Davatzikos et al. used voxel-based nonlinear multivariate analysis to separate AD and Frontotemporal disease (FTD) using MRI imaging [8]. In their subsequent study [9] they used similar pattern classification method but with combination of features extracted from MRI images and cerebrospinal fluid (CSF) biomarker to predict the progression from mild cognitive impairment to AD. In another study, Zhang et al. in [10] presented an approach to combine images obtained from MRI and PET together using multi-kernel learning (MKL) to classify AD and Mild Cognitive Impairment (MCI) cases. A thorough survey was done by Higdon et al. who compared the performance of several different machine learning methods on classification between FTD and AD using FDG-PET images [11]. In our previous study [12], we designed an automated classification method combining MKL and genetic algorithm (GA) to differentiate AD, FTD and normal controlled (NC) cases using FDG-PET images. In our subsequent study [13], we used infinite kernel learning (IKL) with a modified optimization constraint which helped exploit the importance of cerebral features in the AD versus NC classification task.

All the above studies make use of supervised learning (SL) paradigm. SL contains two phases: train the classifier with labeled image sample to tune the learning parameters and generate rules then test the learnt rules on new unseen examples to predict their labels. However, in the case of dementia classification, labeled data is not always available since accurate diagnosis (ground truth) is only available post mortem [3]. On the other hand, there is large amount of unlabeled or uncertain brain images. Therefore, semi-supervised learning (SSL) [14] can play an important role in dementia classification. The main difference between SL and SSL is that the dataset used to construct the classifier contains unlabeled examples. In another word, we

*Research supported in part by the National Science Foundation of China under Grants 61471297, in part by the Natural Science Foundation of Shaanxi Province, China, under Grant 2015JM6287, and in part by the Fundamental Research Funds for the Central Universities under Grant 3102014JSJ0006.

S. Lu, T. W. Cai and D. D. Feng are with the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, The University of Sydney, NSW2006, Australia (e-mail: lshe7842@uni.sydney.edu.au).

Y. Xia was with the Biomedical and Multimedia Information Technology (BMIT) Research Group, School of Information Technologies, The University of Sydney, NSW2006, Australia. He is now with the Shaanxi Key Lab of Speech & Image Information Processing (SAIIP), School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (phone: 86-29-88491533; fax: 86-29-88431518; e-mail: yxia@nwpu.edu.cn).

incorporate the information provided by large amount of unlabeled examples into the training phase of classifier. It is also suggested that the involvement of unlabeled data in machine learning process will very likely improve the overall classification performance [15]. In this paper we design and propose a novel SSL method called manifold learning with affinity regularization (MLAR) to solve the dementia classification problem. Basically, we regularize the manifold learning objective [15, 16] with affinity matrix derived from an unsupervised random forest constructed based on both labeled and unlabeled FDG-PET images. We tested the proposed method on image data obtained from Alzheimer's Disease Neurodegenerative Initiative (ADNI) cohort and compared the results with the state-of-the-art Laplacian support vector machine (LapSVM)[15] as well as supervised SVM.

II. Materials and Methods

A. Materials

We collected 145 FDG-PET images from the publicly accessible ADNI cohort (www.loni.ucla.edu/ADNI/). ADNI is a longitudinal study of a large number of elderly normal controls (NC), and older adults with AD, early MCI (eMCI) and late MCI (lMCI) collected at a number of clinical sites in the United States. All images were preprocessed by ADNI participants following standard protocol (http://www.adni-info.org/Scientists/ADNIStudyProcedures.aspx) prior to publish. The dataset we used in this study consists of 70 AD cases and 75 NCs.

We spatially normalized all FDG-PET images into the same coordinate space as the automated anatomical labeling (AAL) cortical parcellation map [17] so that spatial features can be extracted based on the AAL map. In total we extracted 286 features from each normalized image. These features contain the means and standard deviations calculated based on 116 cerebral anatomical volumes of interest (VOIs) as well as the difference between the mean voxel values of each of the 54 left-right symmetric VOI pairs.

B. Methods

Let $D = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_N\}^T$ denote a set of images where $\mathrm{x}_i = \left(x_i^1, x_i^2, ..., x_i^k\right)^T$, $x_i^k \in \mathbb{R}$, $i = 1, ..., \mathrm{N}$, N=145 represents a vector containing $k = 286$ spatial features extracted from the ith image. Let $D_L = \{\mathrm{x}_1, \mathrm{x}_2, ..., \mathrm{x}_m\}^T$ (usually, $m < N/2$) be a subset of D with corresponding image label set $L = \{l_1, l_2, ..., l_m\}^T$ where $l_i \in \{\mathrm{AD}, \mathrm{NC}\}$, $i = 1, ..., m$. We further divide D_L into a training dataset D_L^t and a testing dataset D_L^{tt}. The rest of images in D forms an unlabeled dataset D_U. Thus the SSL problem our method aims to solve is learning a prediction function f^* to predict the labels (AD or NC) for images in D_L^{tt} with small error rate by training a classifier on D_L^t with the additional information provided by $D_L + D_U = D$. We also believe that using the additional information provided by unlabeled data D_U in conjunction with D_L improves the performance of SSL on dementia classification compared to using D_L alone. In order to achieve this goal, we minimize the losses incurred on all incorrect predictions on D_L^t by the prediction function $f \in H$ during training phase. Therefore, the SSL problem can be converted to an optimization problem in the form of (1)

$$\underset{f \in H}{\text{minimize}}\ cV\left(f(D_L^t), L,\right) + \gamma I \qquad (1)$$

where V is the loss function, H is the set of functions containing the target function f^*, c and λ are constants, I is the additional information provided by both labeled and unlabeled images. The purpose of the second term involving I is to impose the smoothness conditions on the target f^* [15]. It is also well known that learning with kernel provides great generalization ability [18]. Thus, adding another penalty term involving the kernel induced by D to (1) enables us to study our problem under the kernel learning framework. Under this framework, the optimization problem (1) becomes

$$\underset{f \in H}{\text{minimize}}\ cV\left(f(D_L^t), L,\right) + \lambda \left\| f \right\|_K^2 + \gamma I \qquad (2)$$

where K is the kernel matrix defined as the function of inner product between each pair of images in D, λ is a constant. Problem in (2) can be solved by LapSVM, which is an implementation of manifold learning theory. In manifold learning the additional information I reflects the underlying intrinsic structure of D. In LapSVM the intrinsic structure is represented by the graph Laplacian matrix G under the assumption that the marginal distribution P_D from which all images in D are drawn lies on a smooth lower dimension manifold. The graph Laplacian is a data similarity measure calculated based on the data adjacency graph. LapSVM replaces I in (2) with the graph Laplacian G derived from D [15]

$$\underset{f \in H}{\text{minimize}}\ cV\left(f(D_L^t), L,\right) + \lambda \left\| f \right\|_K^2 + \gamma f^T G f. \quad (3)$$

In (3), instead of approximating the smoothness property of the underlying manifold with graph Laplacian of D, our method uses the affinity matrix [19] derived using unsupervised random forest constructed from D to model the image intrinsic structure. Compared to the graph Laplacian used in (3), the affinity matrix is not dependent on the number of nearest neighbours required by calculating graph Laplacian. Random forest is an ensemble learning method which constructs an ensemble of decision trees. The nodes in each decision tree can only be either leaf node or non-leaf node. In unsupervised random forest, the leaf nodes

of each decision tree approximate the distribution of the data clustered at them rather than predicting the class labels for the data. Data examples at each non-leaf node are split into two subsets by thresholding based on one of their features. One subset goes to the left branch of this node and the other one goes to the right. The threshold value is selected by optimizing a quality measurement of candidate splits. In our method we iteratively build each decision tree by minimizing the following unsupervised information gain [19] at each non-leaf node

$$\log\left(\left|\Lambda(\mathrm{S}_j)\right|\right) - \sum_{i\in\{L,R\}}\left|\mathrm{S}_j^i\right|\log\left(\left|\Lambda(\mathrm{S}_j^i)\right|\right)/\left|\mathrm{S}_j\right| \qquad (4)$$

where $\Lambda(\cdot)$ is the covariance operator, $|S|$ gives the cardinality of set S, S_j is the set of all data before split at node j, S_j^L is the set of all data flow to left branch of node j, and similarly, S_j^R is data flow to right branch. To solve (4) we randomly select the μ th feature $\mu \in \{1,2,...,k\}$ at current node, and then calculate

$$t = \left(\max\left\{x_i^\mu\right\} - \min\left\{x_i^\mu\right\}\right)/2,\ x_i^\mu \in \mathrm{x}_i,\ \mathrm{x}_i \in S_j \qquad (5)$$

as the threshold such that data example whose μ th feature value is less than the threshold goes to the left branch of node j and goes to the right if bigger than the threshold. Equation (4) is then evaluated based on the thresholding outcome. This procedure is repeated 50 times to find the threshold t^* that minimizes (4). After all trees are constructed, we calculate affinity matrix of each tree by $W_{ij}^t = e^{-D^t(\mathrm{x}_i,\mathrm{x}_j)}$ where $D^t\left(\mathrm{x}_i,\mathrm{x}_j\right)=1$ if training example x_i and x_j end up at the same leaf node, and 0 otherwise. At last we obtain $W = \left(\sum_{t=1}^T W^t\right)/T$ where T is the total number of trees in this ensemble. The final problem is then formulated as

$$\underset{f\in H}{\text{minimize}}\ cV\left(f(D_L^t),L,\right) + \lambda\left\|f\right\|_K^2 + \gamma f^T W f. \qquad (6)$$

The form of the target f^* in (6) bears a simple form according to the Representer theorem [15, 18]

$$f^* = \sum_{i=1}^N \alpha_i^* K\left(\mathrm{x}_i,\mathrm{x}_i^T\right) \qquad (7)$$

where $\alpha^* \in \mathbb{R}^N$ is the vector form of optimal kernel combination coefficient. Substitute (7) into (6) we obtain a new optimization problem with optimization variable α.

$$\underset{\alpha\in\mathbb{R}^N}{\text{minimize}}\ cV\left(D_L^t,L;\alpha\right) + \lambda\alpha^T K\alpha + \gamma\alpha^T K^T W K\alpha \qquad (8)$$

This problem in (8) can be solved by any standard SVM packages with modified kernel matrix to include the affinity matrix. In this paper we used LibSVM [20] to solve this problem. The summary of the MLAR algorithm is listed in TABLE I.

We randomly sampled m feature vectors together with their labels from each of the AD and NC datasets. The rest of feature vectors were treated as unlabeled. This sampling process was repeated 3 times. MLAR was then applied to these samples in a 3-fold cross validation. For each fold we used $2m/3$ feature vectors for training and the rest for testing. The constants required by MLAR algorithm were chosen empirically. We chose $T=100$, $d=5$, $\gamma_A = 1\times10^{-6}$, $\gamma_I = 1\times10^{-2}$, and $m=30$. The exact same experiment settings were also applied to LapSVM except of the absence of T and d, instead, we chose the required number of nearest neighbours $b=6$. For supervised SVM, we used the same $m=30$ as the number of labeled feature vectors belong to each of the two classes and applied 10-fold cross validation to evaluate its performance.

TABLE I. SUMMARY OF MLAR ALGORITHM

Manifold Learning with Affinity Regularization Algorithm	
Input	m labeled examples $\left\{\left(\mathrm{x}_i,l_i\right)\right\}_{i=1}^m$, $N-m$ unlabeled examples $\left\{x_i\right\}_{i=m+1}^N$, the number of trees T in the forest, the tree depth d, weights γ and λ
Output	Estimated labels for unlabeled examples $\left\{\tilde{l}_i\right\}_{i=m+1}^N$
Step 1	Construct a random forest by iteratively building decision tree with $\left\{\mathrm{x}_i\right\}_{i=1}^N$ until T trees are created. In each iteration, a decision tree grows up to d layers with (2) minimized at each non-leaf node and the affinity matrix W^t is calculated. Average across all W^t to get final affinity matrix W.
Step 2	Calculate kernel matrix K with $\left\{\mathrm{x}_i\right\}_{i=1}^N$ and solve (1) with K, W, γ, λ to obtain vector α.
Step 3	Output $\tilde{l}_i = \mathrm{sgn}\left(\sum_{j=1}^N \alpha_j K\left(\mathrm{x}_i,\mathrm{x}_j\right)\right)$ for $i = m+1,...,\mathrm{N}$.

TABLE II. PERFORMANCE OF RM AND LAPSVM IN OUR EXPERIMENT

	Error rate	**Sensitivity**	**Specificity**
MLAR	10.56%	88.89%	90%
LapSVM	24.12%	77.78%	80%
SVM	35%	71.43%	61.54%

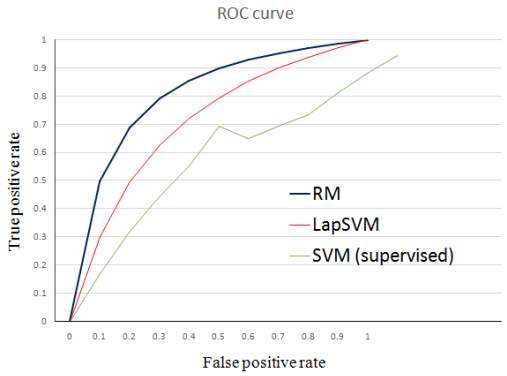

Figure 1. ROC curve for comparison among the performance of MLAR, LapSVM and supervised SVM.

C. Results

The average error rates, sensitivity and specificity of the 10-fold run of MLAR, LapSVM and supervised SVM are shown in TABLE II.

It is evident from the above table that both MLAR and LapSVM achieved better classification results than supervised SVM. This verified our belief that incorporation of information provided unlabeled data improves the dementia classification performance. Between the two semi-supervised methods, MLAR method outperformed LapSVM by a significant margin. MLAR reduced the average error rate by 13.56% down from 24.12% of LapSVM. It also increased the sensitivity and specificity of the classification by 11.11% and 10%, respectively. The ROC curve is shown in Fig. 1.

This reported performance gap between MLAR and LapSVM suggests that although both affinity matrix generated by random forest and Laplacian adjacency matrix are similarity measurement between pair of data points, affinity matrix performs better as it is obtained by solving an unsupervised random forest whose leaf nodes themselves represent a compact manifold [19] rather than simple calculation based on the nearest neighbours. In another word, after unsupervised random forest is constructed, the data examples are already well clustered on the leaf nodes. This result seems to be resonant with the work in [21] such that the unsupervised random forest construction process can be considered as unsupervised pre-training and it is able to improve the semi-supervised learning performance as well.

III. Conclusion

In this paper we proposed a dementia classification method based on a novel SSL approach to classify AD from NC cases using FDG-PET images obtained from ADNI cohort. The SSL method is a realization of manifold learning theory with the intrinsic structure representation approximated by the leaf nodes of random forest constructed on both labeled and unlabeled images. Experiment results showed that using unlabeled images significantly improved the classification accuracy. And they also showed that our method achieved higher accuracy, sensitivity and specificity than the well known LapSVM method.

References

[1] American Psychiatric Association. and American Psychiatric Association. DSM-5 Task Force., *Diagnostic and statistical manual of mental disorders : DSM-5*, 5th ed. Washington, D.C.: American Psychiatric Association, 2013.

[2] L. K. Teune, *et al.*, "Typical cerebral metabolic patterns in neurodegenerative brain diseases," *Mov Disord,* vol. 25, pp. 2395-404, Oct 30 2010.

[3] M. D. Devous, Sr., "Functional brain imaging in the dementias: role in early detection, differential diagnosis, and longitudinal studies," *European Journal of Nuclear Medicine and Molecular Imaging,* vol. 29, pp. 1685-96, Dec 2002.

[4] R. Brookmeyer, *et al.*, "Forecasting the global burden of Alzheimer's disease," *Alzheimer's & Dementia,* vol. 3, pp. 186-191, 2007.

[5] M. Bobinski, *et al.*, "The histological validation of post mortem magnetic resonance imaging-determined hippocampal volume in Alzheimer's disease," *Neuroscience,* vol. 95, pp. 721-725, 1999.

[6] N. C. Fox and J. M. Schott, "Imaging cerebral atrophy: normal ageing to Alzheimer's disease," *The Lancet,* vol. 363, pp. 392-394, 2004.

[7] S. Zhang, *et al.*, "Diagnostic accuracy of 18 F-FDG and 11 C-PIB-PET for prediction of short-term conversion to Alzheimer's disease in subjects with mild cognitive impairment," *Int J Clin Pract,* vol. 66, pp. 185-98, Feb 2012.

[8] C. Davatzikos, *et al.*, "Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI," *NeuroImage,* vol. 41, pp. 1220-1227, 2008.

[9] C. Davatzikos, *et al.*, "Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification," *Neurobiology of Aging,* vol. 32, pp. 2322.e19-2322.e27, 2011.

[10] D. Zhang, *et al.*, "Multimodal classification of Alzheimer's disease and mild cognitive impairment," *NeuroImage,* vol. 55, pp. 856-867, 2011.

[11] R. Higdon, *et al.*, "A comparison of classification methods for differentiating fronto-temporal dementia from Alzheimer's disease using FDG-PET imaging," *Statistics in Medicine,* vol. 23, pp. 315-326, 2004.

[12] Y. Xia, *et al.*, "Automated Identification of Dementia Using FDG-PET Imaging," *BioMed Research International,* vol. 2014, p. 8, 2014.

[13] S. L. Y. Xia, W. Wei, D. Feng, Y. Zhang, "Non-Sparse Infinite-Kernel Learning for Automated Identification of Alzheimer's Disease Using PET Imaging," presented at the The 13th International Conference on Control, Automation, Robotics and Vision, Singapore, Singapore, 2014.

[14] O. Chapelle, *et al.*, *Semi-Supervised Learning*: The MIT Press, 2010.

[15] M. Belkin, *et al.*, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *J. Mach. Learn. Res.,* vol. 7, pp. 2399-2434, 2006.

[16] S. Melacci and M. Belkin, "Laplacian Support Vector Machines Trained in the Primal," *J. Mach. Learn. Res.,* vol. 12, pp. 1149-1184, 2011.

[17] N. Tzourio-Mazoyer, *et al.*, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage,* vol. 15, pp. 273-89, Jan 2002.

[18] B. Schölkopf and A. J. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Cambridge, Mass.: MIT Press, 2002.

[19] A. Criminisi, *et al.*, "Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Foundations and Trends® in Computer Graphics and Vision,* vol. 7, pp. 81-227, 2011.

[20] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.,* vol. 2, pp. 1-27, 2011.

[21] D. Erhan, *et al.*, "Why Does Unsupervised Pre-training Help Deep Learning?," *J. Mach. Learn. Res.,* vol. 11, pp. 625-660, 2010.