Semi-supervised learning in MCI-to-AD conversion prediction - When is unlabeled data useful?

Elaheh Moradi and Jussi Tohka
Department of Signal Processing
Tampere University of Technology
Finland, Email: elaheh.moradi@tut.fi

Christian Gaser
Department of Psychiatry
University of Jena, Germany

Alzheimer's Disease
Neuroimaging Initiative[0]

***Abstract*—This paper investigates the use of semi-supervised learning (SSL) for predicting Alzheimers Disease (AD) conversion in Mild Cognitive Impairment (MCI) patients based on Magnetic Resonance Imaging (MRI). SSL methods differ from standard supervised learning methods in that they make use of unlabeled data - in this case data from MCI subjects whose final diagnosis is not yet known. We compare two widely used semi-supervised methods (low density separation (LDS) and semi-supervised discriminant analysis (SDA)) to the corresponding supervised methods using real and synthetic MRI data of MCI subjects. With simulated data, using SSL instead of supervised learning led to higher classification performance in certain cases, however, the applicability of semi-supervised methods depended strongly on the data distributions. With real MRI data, the SSL methods achieved significantly better classification performances over supervised methods. Moreover, even using a small number of unlabeled samples improved the AD conversion predictions.**

I. Introduction

Mild Cognitive Impairment (MCI) is a transitional stage between age-related cognitive decline and Alzheimers disease (AD). For the effective treatment of AD, it would be important to identify MCI patients with the high risk for conversion to AD. Neuroimaging data is considered to be important for the task because the progression of the AD pathology within the brain starts many years before clinical symptoms and various machine learning algorithms have been applied to construct neuroimaging biomarkers to predict MCI-to-AD conversion at an individual level, e.g., [1], [2]. However, the success of these methods has been limited so far, with a possible exception of the short-term conversion prediction [1]. One reason for this is probably the limited number of labeled data available: collecting data labels is challenging, since at the time of imaging it is not known whether an MCI subject will develop AD or not and subjects have to be followed-up for several years after the imaging to obtain a reliable clinical diagnosis.

Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning [3]. In addition to labeled data (data from MCI subjects who have been followed up and it is known if they will convert to AD or not), SSL methods make use of unlabeled data (data from MCI subjects for whom reliable future diagnosis cannot be established). While in typical SSL applications in machine learning (speech recognition, text classification, etc.) the number of available unlabeled data is expected to be huge, in our case the number of both unlabeled and labeled data is relatively small. Therefore, it is important to study when the semi-supervised learning is useful, i.e., when unlabeled data can improve the classification accuracy and what the potential bottlenecks of SSL methods are. The few SSL applications [4], [5], [6] to MRI-based MCI-to-AD conversion prediction have used a data from AD subjects and normal controls as the labeled data and tried to classify the MCI subjects into two groups (progressive and stable MCI; pMCI and sMCI). The success of these methods has been limited, the best performing method reached area under the ROC curve (AUC) of 0.73 for a short-term (15-month) conversion prediction [5], but, on the other hand, the use of unlabeled data has improved the predictions. We here set to investigate a slightly different problem, where MRIs from pMCI and sMCI subjects for whom a reliable diagnosis is available are used as labeled data. Unlabeled data are MRIs of the MCI subjects who have not been followed up for long enough (at least 3 year follow-up is expected here) or for who a reliable diagnosis cannot be assigned. We study semi-supervised learning methods for the early (up to 3 years before clinical diagnosis) detection of the MCI-to-AD conversion and compare them to relevant supervised methods with data from ADNI cohort and simulated data reminiscent of the ADNI data. We will vary the number of labeled and unlabeled data to establish bounds for the usefulness of the use of unlabeled data. With simulated data, we will also address the feature selection combined with semi-supervised learning.

II. Materials and methods

A. ADNI data

Data used in this work is obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database http://adni.loni.usc.edu/. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of

[0] Data used in preparation of this article were obtained from the Alzheimers Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

978-1-4799-4149-0/14/$31.00 ©2014 IEEE

sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. For up-to-date information, see www.adni-info.org.

We use MRIs from 404 MCI subjects, 200 AD subjects, and 231 normal controls for whom baseline MRI data (T1-weighted MP-RAGE sequence at 1.5 Tesla, typically 256 x 256 x 170 voxels with the voxel size of 1 mm x 1 mm x 1.2 mm) were available. The data from AD subjects and normal controls were only used for generating simulated data and to aid the feature selection with the classification of MCI subjects into pMCI and sMCI classes. For the diagnostic classification at baseline, 404 MCI subjects were grouped as (i) sMCI (stable MCI), if diagnosis was MCI at all available time points, but at least for 36 months (n = 115); (ii) pMCI (progressive MCI), if diagnosis was MCI at baseline but conversion to AD was reported after baseline within 1, 2 or 3 years, and without reversion to MCI or NC at any available follow-up (n = 151); (iii) uMCI (unknown MCI), if diagnosis was MCI at baseline but they are not diagnosed at the end of the project (n = 138). The MRIs were preprocessed into gray matter tissue images in the stereotactic space as described in [1], smoothed with 8-mm FWHM Gaussian kernel, resampled to 4 mm spatial resolution and masked into 29852 voxels.

B. Feature selection

Due to the high dimensionality of the data (29852 features/voxels), the feature selection is performed before machine learning analysis of the data. Because MCI is a translational stage between age-related cognitive decline and AD, we assume that the voxels that are discriminative between AD subjects and normal controls are also discriminative between pMCI and sMCI subjects. Thus, we performed the feature selection using the data from AD subjects and normal controls (without using any data from MCI subjects). The subset of voxels best separating AD subjects from controls was identified using elastic net regularized logistic regression (based on a combination of L1 (LASSO) and L2 (Ridge) regularizer) [7]. This is an embedded feature selection method that is widely applied in neuroimaging. We selected the parameter values for the regularized logistic regression using a parametric Bayesian estimate of the classification error [8], [9].

C. Simulated data generation

We generate simulated MRI data separately for both groups (pMCI and sMCI). First, a subset of voxels discriminating AD and healthy subjects were identified within MRI data by using sparse logistic regression (based on L1 (LASSO) regularizer) [7]. The analysis identified 158 voxels spread across the whole brain with the largest number of voxels in hippocampi and temporal and frontal cortices, matching well to previously observed atrophy patterns in AD. These voxels are simulated to be discriminative between pMCI and sMCI classes. Data generation process consists of the following steps:
1) We divide the ADNI data from MCI subjects randomly into two subsets in order to simulate training and testing datasets separately and to model the natural variation in the data. Data from 76 pMCI (D^p_{train}) and 58 sMCI (D^s_{train}), 75 pMCI (D^p_{test}) and 57 sMCI (D^s_{test}) subjects were used for generating simulated training and testing datasets.
2) For 158 discriminative voxels $v \in V_D$, the mean $\mu_v(G)$ and variance $\sigma^2_v(G)$ of GM image values are computed separately for each group $G = D^s_{test}, D^s_{train}, D^p_{test}, D^p_{train}$. For the non-discriminative voxels $v \in V_N$, $\mu_v(G)$ and $\sigma^2_v(G)$ are computed by pooling the data from two classes into $D_{test} = D^s_{test} \cup D^p_{test}$ and $D_{train} = D^s_{train} \cup D^p_{train}$, i.e., for these voxels $\mu_v(D^s_{test}) = \mu_v(D^p_{test})$ and $\mu_v(D^s_{train}) = \mu_v(D^p_{train})$. A simulated image representing a group G is created by, for each voxel, drawing a random number from Gaussian distribution with mean $\mu_v(G)$ and the variance $\sigma^2_0\sigma^2_v(G)$, where σ^2_0 is parameter to be varied.
3) Finally, the data is spatially smoothed by using the 3-D Gaussian filter with 5 mm isotropic FWHM to introduce a spatial dependence between the voxel values.

D. Learning algorithms

We selected to study two widely used, fairly recent SSL algorithms: low density separation (LDS) [10] and semi-supervised discriminant analysis (SDA)[11]. We combined SDA with 10 nearest neighbors method to perform the classifications as recommended in [11]. We next give a brief overview of the LDS and SDA algorithms and refer to [10], [11] for details. LDS is a two step algorithm, which first derives a graph-distance kernel for enhancing the cluster separability and then it applies transductive support vector machine (TSVM) [12] for classifier learning. Note that SSL methods applied to MCI-to-AD conversion prediction include TSVM [6] and Laplacian SVM [5]. LDS can be seen as an improved version of TSVM and related to Laplacian SVM. SDA is a SSL dimensionality reduction method that seeks to build a linear projection respecting the discriminant structure from labeled samples, such as in linear discriminant analysis (LDA), as well as the intrinsic geometric structure from both labeled and unlabeled samples. The LDA is a traditional supervised dimensionality reduction that achieves the projection vector by simultaneously maximizing the between class separability and minimizing the within-class separability of the labeled samples. However, in the case of scarce labeled samples overfitting may occur leading to inaccurate projection direction. A common way to prevent overfitting is adding a regularizer. When a set of unlabeled samples is available, SDA incorporates the information from unlabeled samples via a graph based regularization into the LDA objective function.

The support vector machine (SVM) with a RBF kernel as implemented in [13] and regularized LDA [14] were used as supervised methods in comparisons. The RBF kernel was selected instead of the linear one because its use led to better results in the preliminary testing. Even with the feature selection, data dimensionality here exceeds the number of samples and we used the regularized version of LDA with Tikhonov regularizer as described in [11], [14]. The parameters for all learning algorithms are selected via cross-validation within the training set in the case of experiments with real data. In the case of experiments with simulated data, the parameters are selected in a separate validation dataset (simulated with the parameters of training set) of a relatively large size to ensure good parameter values.

III. Experiments and Results

A. Simulated data

We generated different datasets based on ADNI MRI data as described in Sect. II.C. Since the SSL methods studied here are based on the cluster assumption, we investigated the effect of the number of unlabeled data in different data sets with different variance of the data. (The cluster assumption states that if the feature vectors are in the same cluster, they probably have the same label. This assumption clearly breaks down with higher variances.) We generated datasets with different σ_0 for this purpose. We generated 200 labeled samples (100 per class) with different number of unlabeled samples N_u ranging from 100 to 2000. We note that having N_u as large as 2000 may appear unrealistic, however, we wished to test the methods also in the case of large unlabeled dataset. We used the AUC as the performance criterion [15]. Each experiment was repeated 10 times (with a different, randomly generated simulated dataset) and we report the average AUCs across these 10 repetitions. We performed two types of experiments to address the importance of feature selection. 1) We used the knowledge of the simulated discriminative voxels and fed only the data from these 158 voxels to learning algorithms. 2) We performed the feature selection in simulated training data using elastic net regularized logistic regression as described in Sect. II.B.

The AUCs in Tables 1 and 2 indicate that the data variance was a major factor in semi-supervised learning when considering SVM-based schemes (SVM and LDS). When the data variance was not too high, adding unlabeled data improved the classification performance with LDS. However, in datasets with higher deviations adding unlabeled data degraded the performance of the classifier as the cluster assumption broke down. When the variance was high ($\sigma_0 = 1.5$), supervised method (SVM) outperformed the semi-supervised method (LDS) and adding more unlabeled data degraded the classification performance with LDS. The data variance was not a factor between SDA and LDA in a sense that semi-supervised method (SDA) was always superior to its supervised counterpart (LDA). Also, the SDA achieved its optimal performance with already relatively small number of unlabeled data ($N_u = 100$) and it did not benefit from larger numbers of unlabeled data. LDS was better of the two SSL methods with the 3 lowest variance levels, but with the highest variance SDA was better than LDS.

Comparing the AUCs in Tables 1 and 2 shows the importance of feature selection in the performance of the classifier. Not surprisingly, knowing which voxels were discriminative resulted in a better performance than using the feature selection (as we would need to do in real life). However, the AUCs sometimes improved as much as by 0.15 by knowing the important features beforehand (see, e.g., LDS, $N_u = 2000, \sigma_0 = 1.5$). The amount of improvement did not vary much between the learning algorithms, however it was clearly more important to know the discriminative features when the data variance was higher, probably indicating that the feature selection becomes more difficult when the noise level increases. Finally, application of the learning algorithms to the full data with 29852 features led to performances close to the chance level (AUC ≈ 0.5) and thus feature selection was a required step (results not shown).

TABLE I. Average AUCs, with known features. N_u is the number of unlabeled data.

σ_0	SVM	LDS	LDS	LDS	LDA	SDA	SDA	SDA
N_u	0	100	1000	2000	0	100	1000	2000
0.8	0.946	0.952	0.961	0.964	0.767	0.924	0.918	0.917
1.0	0.897	0.890	0.909	0.909	0.706	0.869	0.859	0.857
1.25	0.835	0.811	0.833	0.832	0.636	0.798	0.792	0.789
1.5	0.703	0.676	0.693	0.688	0.577	0.738	0.740	0.736

TABLE II. Average AUCs, with feature selection

σ_0	SVM	LDS	LDS	LDS	LDA	SDA	SDA	SDA
N_u	0	100	1000	2000	0	100	1000	2000
0.8	0.850	0.856	0.890	0.895	0.636	0.829	0.820	0.814
1.0	0.734	0.739	0.747	0.755	0.535	0.721	0.705	0.699
1.25	0.678	0.705	0.662	0.668	0.510	0.632	0.619	0.617
1.5	0.596	0.572	0.548	0.538	0.506	0.580	0.575	0.573

B. ADNI data

In this section, we present the experimental results for the ADNI MRI data described in Sect. II.A. while varying the number of labeled and unlabeled data used for training the classifier. We first randomly selected (without replacement) only a limited number of labeled data for training (60,100, or 140 samples, equally divided between the pMCI and sMCI classes). Then, we randomly selected (without replacement) a limited number of data from sMCI, pMCI, and uMCI subjects to be used without label information as unlabeled data (from 50 to 350 samples, with the increments of 50 samples). These random selections were repeated 100 times to create 100 different datasets per a configuration. For the evaluation of the classifier performance and estimation of the nuisance parameters for the classifiers, we computed the AUCs using two nested cross-validation loops (stratified 10-fold for each loop, inner loop for the parameter selection, outer for performance evaluation; note that the number of samples was selected so that each fold can be balanced).

Fig. 1 shows the average AUCs across 100 different samplings for the studied methods (SVM, LDS, SDA, LDA) for fixed numbers of labeled samples (indicated by different colors in Fig. 1) and with increasing number of unlabeled samples. When the number of unlabeled samples was zero, the used methods were SVM and LDA and otherwise the used methods were LDS and SDA. The feature selection within the training set (by regularized logistic regression) resulted in worse AUCs with all 4 methods than the feature selection with AD and NC data of Sect. II.B, and thus only the AUCs with the feature selection of Sect. II.B are reported. Using unlabeled data and SSLs improved the classification performance markedly, even with 50 unlabeled samples, the average AUCs always improved, on average by 0.05. The highest improvement (from 0.58 to 0.67) was with SDA compared to LDA with 60 labeled samples. In order to make statistically precise statements, we computed the p-value for unpaired AUC scores (across 100 different re-samplings of the data) with a permutation test. The improvement was always significant when comparing SSL methods (LDS and SDA) to the corresponding supervised methods (in each case $p < 0.00001$ except for the case of LDS vs. SVM with 60 labeled samples $p = 0.0045$). Thus, the use of SSL significantly improved the classification performance. The AUCs of the two SSL methods with 60 and 100 labeled samples and all available unlabeled data

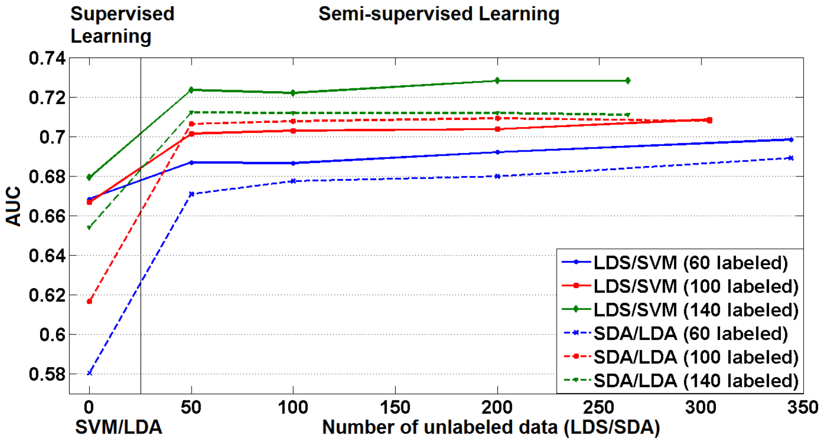

Fig. 1. The mean AUC score of LDS and SDA methods within 100 computation times with respect to different number of unlabeled data using original MRI data. When the number of unlabeled data is zero, the corresponding supervised methods (SVM and LDA) are used.

was statistically similar ($p > 0.2$) and with 140 labeled samples LDS outperformed SDA in terms of the average AUC ($p = 0.0025$). The differences between the AUCs within a fixed SSL method when the number of unlabeled data was varied were statistically not significant.

IV. Conclusion

We studied the value of unlabeled data from MCI subjects without final diagnosis in the MRI-based MCI-to-AD conversion prediction. We compared two semi-supervised learning methods, LDS and SDA, and their supervised counterparts, SVM and regularized LDA, by using ADNI MRI data and simulated data while varying the number of labeled and unlabeled samples. The use of SSL and unlabeled data significantly improved the classification performance with the ADNI data, independently on how many labeled samples were available. Importantly even a small number of unlabeled samples improved the conversion predictions. With the simulated data, the use of unlabeled data improved the classification performance in most cases, however, the improvement was smaller than with the real data and, as expected, diminished with increasing noise level. Of the two SSL methods studied, LDS had the superior performance.

Acknowledgments: Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimers Association; Alzheimers Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; And Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. This research has been also supported by the Academy of Finland under the grants 130275, 263785.

References

[1] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, A. D. N. Initiative *et al.*, "BrainAGE in mild cognitive impaired patients: predicting the conversion to alzheimers disease," *PloS ONE*, vol. 8, no. 6, p. e67346, 2013.

[2] S. F. Eskildsen, P. Coupé, D. García-Lorenzo, V. Fonov, J. C. Pruessner, and D. L. Collins, "Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the adni cohort using patterns of cortical thinning," *NeuroImage*, vol. 65, pp. 511–521, 2013.

[3] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning*. Cambridge, MA: MIT Press, 2006.

[4] K. N. Batmanghelich, D. H. Ye, K. M. Pohl, B. Taskar, and C. Davatzikos, "Disease classification and prediction via semi-supervised dimensionality reduction," in *ISBI*. IEEE, 2011, pp. 1086–1090.

[5] D. H. Ye, K. M. Pohl, and C. Davatzikos, "Semi-supervised pattern classification: Application to structural mri of alzheimer's disease," in *Pattern Recognition in NeuroImaging (PRNI)*. IEEE, 2011, pp. 1–4.

[6] R. Filipovych and C. Davatzikos, "Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI)," *NeuroImage*, vol. 55, no. 3, pp. 1109–1119, 2011.

[7] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Software*, vol. 33, no. 1, pp. 1–22, 2010.

[8] H. Huttunen, T. Manninen, and J. Tohka, "Bayesian error estimation and model selection in sparse logistic regression," in *Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[9] L. A. Dalton and E. R. Dougherty, "Bayesian minimum mean-square error estimation for classification errorpart II: The bayesian mmse error estimator for linear classification of gaussian distributions," *IEEE Trans. Signal Process*, vol. 59, pp. 130–144, 2011.

[10] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *AISTATS*, 2005, pp. 57–64.

[11] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in *ICCV*. IEEE, 2007, pp. 1–7.

[12] T. Joachims, "Transductive inference for text classification using support vector machines," in *ICML*, 1999, pp. 200–209.

[13] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Trans. Intell. Systems Tech.*, vol. 2, p. 27, 2011.

[14] J. H. Friedman, "Regularized discriminat analysis," *J. Am. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[15] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Patt. Recog.*, vol. 30, pp. 1145–59, 1997.