Domain Transfer Learning for MCI Conversion Prediction

Bo Cheng[1, 2], Daoqiang Zhang[1, 2], and Dinggang Shen[2]

[1] Dept. of Computer Science and Engineering,
Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599
{cb729,dqzhang}@nuaa.edu.cn, dgshen@med.unc.edu

Abstract. In recent studies of Alzheimer's disease (AD), it has increasing attentions in identifying mild cognitive impairment (MCI) converters (MCI-C) from MCI non-converters (MCI-NC). Note that MCI is a prodromal stage of AD, with possibility to convert to AD. Most traditional methods for MCI conversion prediction learn information only from MCI subjects (including MCI-C and MCI-NC), not from other related subjects, e.g., AD and normal controls (NC), which can actually aid the classification between MCI-C and MCI-NC. In this paper, we propose a novel domain-transfer learning method for MCI conversion prediction. Different from most existing methods, we classify MCI-C and MCI-NC with aid from the domain knowledge learned with AD and NC subjects as auxiliary domain to further improve the classification performance. Our method contains two key components: (1) the cross-domain kernel learning for transferring auxiliary domain knowledge, and (2) the adapted support vector machine (SVM) decision function construction for cross-domain and auxiliary domain knowledge fusion. Experimental results on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database show that the proposed method can significantly improve the classification performance between MCI-C and MCI-NC, with aid of domain knowledge learned from AD and NC subjects.

1   Introduction

Alzheimer's disease (AD) is the most common form of dementia in elderly people worldwide. Early diagnosis of AD is very important for possible delay of the disease. Mild cognitive impairment (MCI) is a prodromal stage of AD, which can be further categorized into MCI converters (MCI-C) and MCI non-converters (MCI-NC). The former will convert into AD in follow-up time, while the latter will not convert. Thus, accurate diagnosis of MCI converters is of great importance. Nowadays, many machine learning methods have been proposed for the classification of AD or MCI [1-3]. More recently, an increasing number of studies in AD research begin to address MCI conversion prediction, i.e., the classification between MCI-C and MCI-NC based on the baseline imaging data [2-7].

One challenge in MCI conversion prediction is that the number of MCI (including both MCI-C and MCI-NC) subjects available for training is generally very small and

thus the generalization ability of the classifier is limited. The same problem also exists in the classification between AD and NC, where the number of AD and NC subjects are also limited. Recently, to enhance the classification between AD and NC, several studies have used semi-supervised learning (SSL) methods [8] in AD diagnosis, where MCI subjects are treated as unlabeled data to aid AD classification [9-11]. Semi-supervised learning methods can efficiently utilize unlabeled samples to improve classification and regression performance, but it requires unlabeled samples and labeled samples coming from the same data distribution [8], which is usually not satisfied in practice. For example, Fig. 1 plots the distributions of AD, MCI-C, MCI-NC and NC from ADNI. As can be seen from Fig. 1, the distribution of MCI is different from that of AD and NC.

On the other hand, in the machine learning areas, a new learning methodology called transfer learning has been developed for dealing with problems involving cross-domain learning. Unlike SSL, transfer learning does not assume the auxiliary data (unlabeled data in SSL) have the same distribution as target data (labeled data in SSL), and can effectively adopt those auxiliary data from the related domain for improving classification performance, as validated in several successful applications in computer vision [12-13]. However, to the best of our knowledge, transfer learning has not been used for addressing brain disease classification problems in medical imaging, although intuitively the inclusion of additional AD and NC subjects as auxiliary data may be beneficial for the classification between MCI-C and MCI-NC. For example, as we can see from Fig. 1, although the distributions of MCI-C and MCI-NC are different from those of AD and NC, they are very related and the domain knowledge for diagnosing AD and NC may help diagnose MCI conversion.

In this paper, we propose a new transfer learning method called Domain Transfer Support Vector Machines (DTSVMs) to classify MCI-C and MCI-NC, with the aid from auxiliary AD and NC subjects. Specifically, we first construct a cross-domain kernel to transfer auxiliary domain knowledge, and then we utilize the adapted SVM to fuse cross-domain and auxiliary domain knowledge for better helping MCI-C and MCI-NC classification. We validate our method on single-modality and multi-modality biomarkers, including magnetic resonance imaging (MRI), fluorodeoxy-glucose positron emission tomography (FDG-PET), and quantification of specific proteins measured through cerebrospinal fluid (CSF), from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.
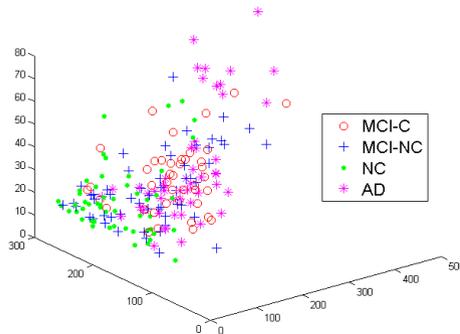

Fig. 1. Distributions of AD, NC, MCI-C, and MCI-NC subjects with CSF features

2   Domain Transfer Support Vector Machines (DTSVMs)

Transfer learning aims to apply knowledge learned from one or more related auxiliary domain to improve the performance on the target domain. In this section, we propose a new method called DTSVM for MCI-C and MCI-NC classification. There are two main steps in DTSVM, i.e., cross-domain kernel learning for transferring auxiliary domain knowledge, and the adapted SVM decision function construction for cross-domain and auxiliary domain knowledge fusion. Fig. 2 shows the flow chart of our proposed DTSVM classification method.
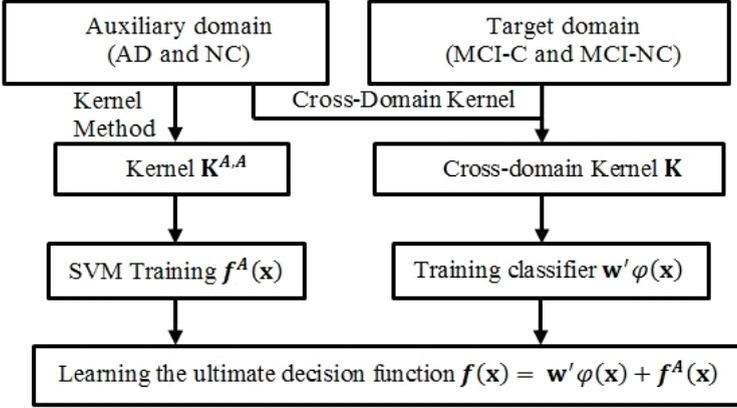

Fig. 2. Flow chart of the proposed DTSVM classification method

2.1   Cross-Domain Kernel for Transferring Knowledge of the Auxiliary Domain

Because of distribution difference between the auxiliary and the target domains, training with samples from the auxiliary domain may degrade the classification performance in another target domain. So, we cannot directly add auxiliary data to target domains for training. In cross-domain learning, it is crucial to reduce the difference between data distributions of the auxiliary and target domains. To avoid such differences, we use *cross-domain kernel learning* for transferring auxiliary domain knowledge [13]. Briefly, the auxiliary and target domains are first mapped to the Reproducing Kernel Hilbert Space (RKHS), and then the cross-domain kernel approach is used to build a new kernel function that can reduce the difference between data distributions of the auxiliary and target domains.

In the following, we formally address the problem of the cross-domain kernel approach by building a new kernel function. Firstly, we define the kernel matrices from the auxiliary domain A and from the target domain T as: $\mathbf{K}^{A,A} = k(\mathbf{x}_i^A, \mathbf{x}_j^A) \in R^{n_A \times n_A}$ and $\mathbf{K}^{T,T} = k(\mathbf{x}_i^T, \mathbf{x}_j^T) \in R^{n_T \times n_T}$, respectively, where $\mathbf{x}_i^A$ and $\mathbf{x}_j^A$ are samples in the auxiliary domain A, and n_A is the number of samples in domain A. Then, we define the cross-domain kernel matrices from the auxiliary domain to the target domain and from the target domain to the auxiliary domain as:

$\mathbf{K}^{A,T} = k(\mathbf{x}_i^A, \mathbf{x}_j^T) \in R^{n_A \times n_T}$ and $\mathbf{K}^{T,A} = k(\mathbf{x}_i^T, \mathbf{x}_j^A) \in R^{n_T \times n_A}$, respectively, where $\mathbf{x}_i^T$ and $\mathbf{x}_j^T$ are samples in the target domain T, and n_T is the number of samples in domain T. Finally, the kernel matrix $\mathbf{K} = k(\mathbf{x}_i, \mathbf{x}_j)$ is obtained as:

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^{A,A} & \mathbf{K}^{A,T} \\ \mathbf{K}^{T,A} & \mathbf{K}^{T,T} \end{bmatrix} \in R^{n \times n} \tag{1}$$

Where $n = n_T + n_A$.

2.2   Adapted SVM for Knowledge Fusion

Assume we have n_A samples and corresponding class labels in the auxiliary domain as $\{\mathbf{x}_i^A, y_i^A\}_{i=1}^{n_A}$, where $\mathbf{x}_i{}^A \in R^d$ is a sample and $y_i^A \in \{+1, -1\}$ is the class label (i.e., AD as 1 and NC as -1). Also, assume we have n_T samples and corresponding class labels as $\{\mathbf{x}_j^T, y_j^T\}_{j=1}^{n_T}$, where $\mathbf{x}_j^T \in R^d$ is a sample and $y_j^T \in \{+1, -1\}$ is the class label (i.e., MCI-C as 1 and MCI-NC as -1). For the convenience of description in the following, we can also define $y = \{y_l\}_{l=1}^n = y^A \cup y^T = \{y_i^A\}_{i=1}^{n_A} \cup \{y_j^T\}_{j=1}^{n_T}$. According to Section 2.1, we can learn the cross-domain kernel $\mathbf{K}$ from both auxiliary and target domains. In addition, since AD and NC subjects are sitting in the auxiliary domain while MCI-C and MCI-NC subjects are sitting in the target domain, we can assign the cross-domain label so that the MCI-NC label will be the same as the NC label and the MCI-C label will be the same as the AD label, considering that the learning tasks for the MCI-C vs. MCI-NC classification and the AD vs. NC classification are very related [9]. Here, we adopt the adapted SVM method in [12] to learn the cross-domain classifier $f(\mathbf{x})$, which is the ultimate decision function of our adapted SVM method. This ultimate decision function $f(\mathbf{x})$ is first learned from the cross-domain classifier $\mathbf{w}'\varphi(\mathbf{x})$ as:

$$f(\mathbf{x}) = \mathbf{w}'\varphi(\mathbf{x}) + b \tag{2}$$

Where $\varphi(\mathbf{x})$ is a kernel-induced nonlinear implicit mapping, $\mathbf{w}$ is the parameter vector of the classifier, b is the bias term, and $\mathbf{x}$ is the cross-domain data obtained by $\mathbf{x} = \mathbf{x}^A \cup \mathbf{x}^T$. Also, $\mathbf{w}'$ denotes the transpose of $\mathbf{w}$.

Because the ultimate decision function $f(\mathbf{x})$ can also be learned from the auxiliary domain classifier $f^A(\mathbf{x}^A)$, Eq. 2 can thus be written as:

$$f(\mathbf{x}) = \mathbf{w}'\varphi(\mathbf{x}) + f^A(\mathbf{x}^A) \tag{3}$$

To learn the weight vector $\mathbf{w}$ in Eq. 3, we use the following objective function, similar to the SVM:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_l$$
$$s.t. \ \zeta_l \geq 0, \ \text{with} \ y_l \mathbf{w}' \varphi(\mathbf{x}_l) + y_l f^A(\mathbf{x}_i^A) \geq 1 - \zeta_l \tag{4}$$

Where l is the l-th sample in the cross-domain $y = y^A \cup y^T$ with range of $1 \le l \le n$, ζ_l is the corresponding slack variable, and $n_T + n_A$ represents the total

number of samples in $\mathbf{y}$ as indicated above.

According to [12], we can solve this objective function (in Eq. 4) to obtain the solution for the weight vector $\mathbf{w}$. Then, we can obtain the final solution for $f(\mathbf{x})$.

3   Experiments

In this section, we evaluate the effectiveness of our proposed DTSVM method on multimodal data, including MRI, PET and CSF, from the Alzheimer's disease Neuroimaging Initiative (ADNI) database.

3.1   Experimental Settings

In our experiments, the baseline ADNI subjects with all corresponding MRI, PET, and CSF data are included, which leads to a total of 202 subjects (including 51 AD patients, 99 MCI patients, and 52 normal controls (NC)). For 99 MCI patients, it includes 43 MCI converters and 56 MCI non-converters. We use 51 AD and 52 NC subjects as auxiliary domains, and 99 MCI subjects as target domains.

The same image pre-processing as used in [1] is adopted here. First, for all structural MR images, we correct their intensity inhomogeneity by the N3 algorithm, perform skull-stripping, and remove the cerebellum. Then, we use the FSL package to segment each structural MR image into three different tissue types: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We further use an atlas warping algorithm [14] to partition each structural MR brain image into 93 ROIs. For each of the 93 ROIs, we compute GM volume in that ROI as a feature; For PET image, we use a rigid transformation to align it to its corresponding structural MR image of the same subject, and then compute the average PET value of each ROI as a feature. Accordingly, for each subject, we can acquire 93 features from the structural MR image, another 93 features from the PET image, and 3 features ($A\beta_{42}$, t-tau, and p-tau) from the CSF biomarkers.

To evaluate the performance of different classification methods, we use a 10-fold cross-validation strategy to compute the classification AUC (areas under the ROC curve), accuracy, sensitivity and specificity. Specifically, the whole set of subject samples are equally partitioned into 10 subsets, and then one subset is successively selected as the testing samples and all remaining subsets are used for training classifiers. This process is repeated 10 times. The SVM classifier training of standard SVM and DTSVM are implemented using LIBSVM toolbox [15], with a linear kernel and a default value for the parameter C (i.e., C=1). For comparison, LapSVM is also adopted in this paper. LapSVM is a typical semi-supervised learning method based on manifold hypothesis [8]. For LapSVM settings, we use linear kernel, and the graph Laplacian $\boldsymbol{L}$ with $n_T + n_A$ nodes are connected using k (i.e., k=5) nearest neighbors, and their edge weights are calculated using the Euclidean distance among samples. We use both multi-modality and single-modality biomarkers to validate our method. For combining multimodality data in DTSVM, standard SVM, and LapSVM methods, we specifically use a linear multi-kernel combination technique, with the

weights learned from the training samples through a grid search, using the range from 0 to 1 at a step size of 0.1. Also, for features of each modality, the same feature normalization scheme as used in [1] is adopted here.

3.2   Results

We compare DTSVM with LapSVM and standard SVM (i.e., SVM) for both multi-modality and single-modality cases. Table 1 shows the classification performance measures of DTSVM, LapSVM, and SVM on different modalities. Note that Table 1 shows the averaged results of 10 independent experiments. As we can see from Table 1, DTSVM can consistently achieve better results than LapSVM and SVM methods on each performance measure, which validates the efficacy of our DTSVM method on using AD and NC subjects as auxiliary domains for helping classification. Specifically, for multi-modality case, DTSVM can achieve a classification accuracy of 69.4%, which is significantly better than LapSVM and SVM that achieve only 60.8% and 63.8%, respectively. In addition, for multi-modality case, if using the leave-one-out evaluation strategy, DTSVM can achieve a classification accuracy of 70.7%.

On the other hand, AUC and ROC are further used to validate the classification performance of DTSVM, LapSVM, and SVM. Table 1 also gives the AUC values of the three methods, and Fig. 3 plots their ROC curves with respect to different modalities. Both Table 1 and Fig. 3 indicate that DTSVM is superior to LapSVM and SVM for MCI-C and MCI-NC classification. Specifically, in the multi-modality case, the AUC value is 0.736 for DTSVM, while the AUC values for LapSVM and SVM are only 0.626 and 0.683, respectively.

Table 1. Comparison of performance measures of DTSVM, LapSVM, and SVM for MCI-C vs. MCI-NC classification using different modalities. (ACC= Accuracy, SEN=Sensitivity, SPE= Specificity).

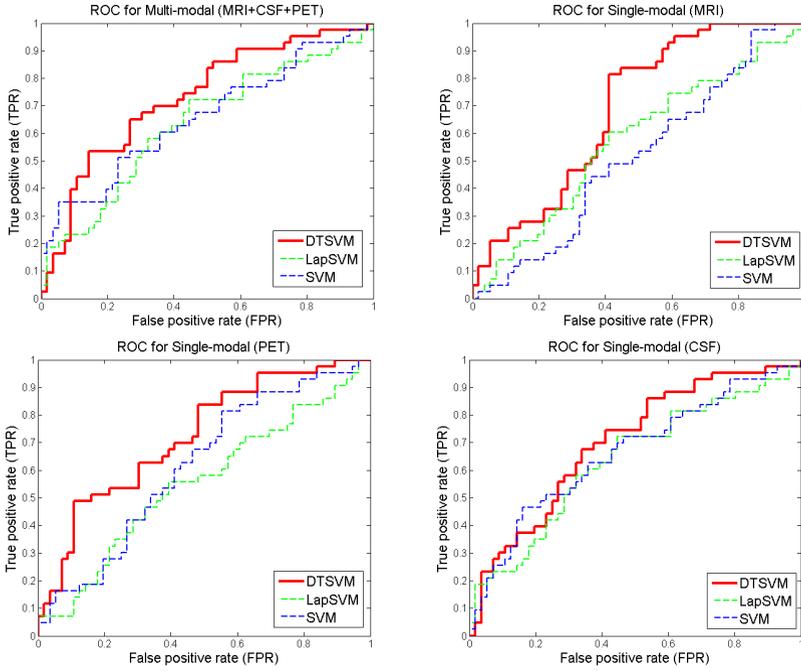| Modality	Methods	ACC %	SEN %	SPE %	AUC
MRI+CSF+PET	**DTSVM**	**69.4**	**64.3**	**73.5**	**0.736**
	LapSVM	60.8	55.3	65.0	0.626
	SVM	63.8	58.8	67.7	0.683
MRI	**DTSVM**	**63.3**	**59.8**	**66.0**	**0.700**
	LapSVM	56.8	50.6	61.5	0.563
	SVM	53.9	47.6	57.7	0.554
CSF	**DTSVM**	**66.2**	**60.3**	**70.8**	**0.701**
	LapSVM	60.8	55.3	65.0	0.626
	SVM	60.8	55.2	65.0	0.647
PET	**DTSVM**	**67.0**	**59.6**	**72.7**	**0.732**
	LapSVM	55.0	48.7	59.8	0.558
	SVM	58.0	52.1	62.5	0.612

Fig. 3. ROC curves of DTSVM, LapSVM, and SVM, using multi-modality and single-modality data, respectively
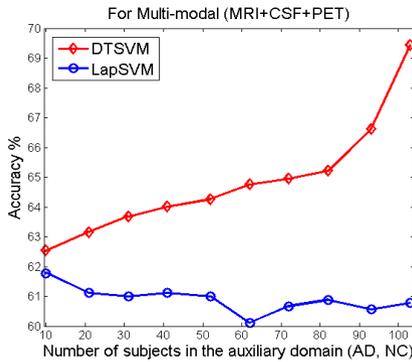

Fig. 4. Comparison of classification accuracy of DTSVM and LapSVM with respect to the use of different number of subjects in the auxiliary domain

Finally, in Fig. 4, we compare DTSVM with LapSVM for classification accuracy, with respect to the use of different number of subjects in the auxiliary domain. As we can see from Fig. 4, in most cases, the performance of DTSVM is significantly better than LapSVM as the number of subjects in the auxiliary domain increases. This validates the usefulness of adopting domain transfer learning method (i.e. DTSVM) to

learn AD and NC domain knowledge, compared to the semi-supervised learning method (i.e., LapSVM). This result further demonstrates that the distribution of MCI is different from that of AD and NC.

4   Conclusion

This paper addresses the problem of exploiting the use of auxiliary domain data (AD and NC subjects) for helping MCI-C vs. MCI-NC classification. By integrating the cross-domain kernel learning and the adapted SVM methods, we propose a Domain Transfer SVM classification method, namely DTSVM. Our method does not require the auxiliary domain data and the target domain data to be from the same distribution. Experimental results on ADNI dataset validate the efficacy of our proposed method.

References

1. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. NeuroImage 55(3), 856–867 (2011)
2. Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q.: Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of Aging 32, 2322.e19-2322.e27 (2011)
3. Cho, Y., Seong, J.K., Jeong, Y., Shin, S.Y.: Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. NeuroImage 59, 2217–2230 (2012)
4. Aksu, Y., Miller, D.J., Kesidis, G., Bigler, D.C., Yang, Q.X.: An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients. PLoS One 6, e25074 (2011)
5. Misra, C., Fan, Y., Davatzikos, C.: Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. NeuroImage 44, 1415–1422 (2009)
6. Leung, K.K., Shen, K.K., Barnes, J., Ridgway, G.R., Clarkson, M.J., Fripp, J., Salvado, O., Meriaudeau, F., Fox, N.C., Bourgeat, P., Ourselin, S.: Increasing power to predict mild cognitive impairment conversion to Alzheimer's disease using hippocampal atrophy rate and statistical shape models. Med. Image Comput. Comput. Assist. Interv. 13, 125–132 (2010)
7. Risacher, S.L., Saykin, A.J., West, J.D., Shen, L., Firpi, H.A., McDonald, B.C.: Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. Curr. Alzheimer. Res. 6, 347–361 (2009)
8. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research 7, 2399–2434 (2006)
9. Filipovych, R., Davatzikos, C.: Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). NeuroImage 55(3), 1109–1119 (2011)

10. Filipovych, R., Resnick, S.M., Davatzikos, C.: Semi-supervised cluster analysis of imaging data. NeuroImage 54(3), 2185–2197 (2011)
11. Zhang, D., Shen, D.: Semi-supervised multimodal classification of Alzheimer's disease. In: IEEE International Symposium on Biomedical Imaging, pp. 1628–1631 (2011)
12. Yang, J., Yan, R., Hauptmann, A.G.: Cross-Domain Video Concept Detection Using Adaptive SVMs. ACM Multimedia (2007)
13. Duan, L., Xu, D., Tsang, I., Luo, J.: Visual Event Recognition in Videos by Learning from Web Data. In: IEEE Int'l Conf. Computer Vision and Pattern Recognition, pp. 1959–1966 (2010)
14. Shen, D., Davatzikos, C.: HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Transactions on Medical Imaging 21, 1421–1439 (2002)
15. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)