Multiple Instance Learning for Classification of Dementia in Brain MRI

Tong Tong[1], Robin Wolz[1], Qinquan Gao[1],
Joseph V. Hajnal[2], and Daniel Rueckert[1]

[1] Department of Computing, Imperial College London, London, UK
[2] Center for the Developing Brain, Division of Imaging Sciences and Biomedical Engineering, King's College London, St. Thomas Hospital, London, UK
t.tong11@imperial.ac.uk

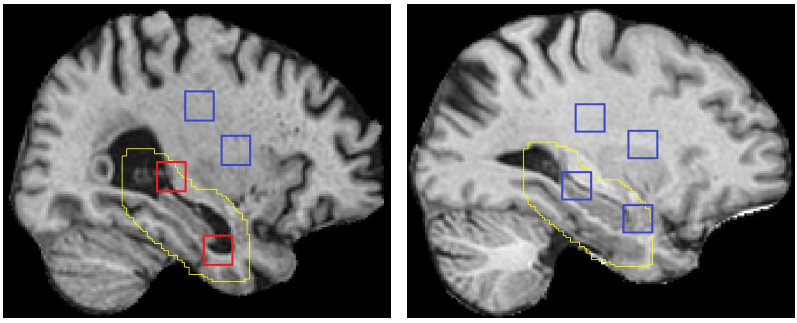
Abstract. Machine learning techniques have been widely used to support the diagnosis of neurological diseases such as dementia. Recent approaches utilize local intensity patterns within patches to derive voxelwise grading measures of disease. However, the relationships among these patches are usually ignored. In addition, there is some ambiguity in assigning disease labels to the extracted patches. Not all of the patches extracted from patients with dementia are characteristic of morphology associated with disease. In this paper, we propose to use a multiple instance learning method to address the problem of assigning training labels to the patches. In addition, a graph is built for each image to exploit the relationships among these patches, which aids the classification work. We illustrate the proposed approach in an application for the detection of Alzheimer's disease (AD): Using the baseline MR images of 834 subjects from the ADNI study, the proposed method can achieve a classification accuracy of 88.8% between AD patients and healthy controls, and 69.6% between patients with stable Mild Cognitive Impairment (MCI) and progressive MCI. These results compare favourably with state-of-the-art classification methods.

1 Introduction

Alzheimer's disease (AD) is the most common type of dementia worldwide and the prevalence of AD is predicted to quadruple in the next four decades. Early diagnosis of AD is not only crucial for future treatments currently under development but can also reduce the associated socioeconomic burden. Different imaging modalities, such as magnetic resonance imaging (MRI) and fluorodeoxyglucose positron emission tomography (FDG-PET), have been widely used to derive image-based biomarkers for AD and various classification strategies have been proposed in the context of AD [1].

The vast majority of approaches for the classification of AD are based on supervised learning. In these approaches, discriminative features are extracted from the image data and supervised classifiers are then trained to perform classification. As the dimensionality of the image data is extremely high, a feature

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 599–606, 2013.



(a) patient with AD (b) control subject

Fig. 1. Example of different bags and the region of interest used for patch extraction. (a) positive bag (patient with AD); (b) negative bag (control subject); The red boxes and green boxes represent positive patches and negative patches respectively.

selection step is necessary to avoid overtraining. To reduce the feature space and select the discriminative features, statistical approaches (e.g. t-tests) [2] or sparse regression methods (e.g. L1-regularized Lasso) [3] are often used. However, these approaches select voxel-wise features throughout the whole brain without considering the relationships among these features. To overcome this limitation, a tree-guided sparse coding method [3] and a resampling scheme [4] have been recently proposed. These approaches can select voxelwise features in meaningful brain regions, which may be related to pathology.

Although these approaches achieve promising classification accuracy, they represent the whole image as a simple feature vector, and do not consider local structural information in the image. Two recent approaches [5,6] utilize local intensity patterns within patches to capture the local structural information for AD classification. In these approaches, patches from patients with AD are used as positive samples and patches from healthy subjects are regarded as negative samples for training. However, patches are small regions in images and It is likely that some patches extracted from patients may not be characteristic of changes associated with pathology as shown in Fig. 1. Therefore, not all of patches from patients can be regarded as positive training samples. This means that there is some ambiguity in assigning training labels to patches extracted from patients. One solution to this problem is to use a weakly supervised method such as multiple instance learning (MIL) [7]. In addition, the patches from the same subject are rarely independent and often have shared information. This information across patches can convey information about the inherent structure of the images, which may be helpful for disease classification.

MIL techniques can learn classifiers from ambiguously labeled training data. They have been successfully applied to different applications in computer vision [7] and recently in medical imaging [8,9]. However, to the best of our knowledge, MIL has not been used in the context of classification of neurological diseases. In this paper, we propose to use MIL techniques to address the problem of

ambiguous patch labels. Specifically, each image is regarded as a bag; the patches extracted from the images are thus treated as inter-correlated instances in the bags. Patients and healthy subjects are treated as positive and negative bags respectively. The MIL method is then used to learn a bag-level classifier to predict the bag labels of the unseen images. In addition, a graph is constructed from each bag in order to investigate the relationships among patches and to exploit the inherent structural information of each image. The proposed method has been evaluated using 834 subjects from the ADNI study. Our experiments show that the proposed method referred to as miGraph compares favourably with state-of-the-art methods.

2 Methods

Let N indicate the number of training images. Each image is regarded as a bag and K patches are exacted from each image to form the bag. Given a training data set $\{(P_1, y_1), \cdots, (P_i, y_i), \cdots, (P_N, y_N)\}$, where $P_i = \{p_{i1}, \cdots p_{ij}, \cdots p_{iK}\}$ represents K patches extracted from the image i to form the bag and $y_i \in Y = \{1, 0\}$ is the corresponding label of the bag P_i, the goal is to learn a bag-level classifier to label unseen bags (i.e. images). Here images from patients and controls are regarded as positive and negative bags respectively. As shown in Fig. 1, if the bag contains at least one positive patch related to disease changes, the bag is labeled as positive; otherwise, the bag is labeled as negative (in this case all the patches in the bag are negative). In the following, we will show how patches are exacted from the training images to form corresponding bags. We will then introduce a graph kernel for solving the MIL problem, where a bag-level classifier is learned to predict the label of unseen bags.

2.1 Extraction of Patches

For each image, the total number of patches M is extremely high and only $K \ll M$ patches are extracted to form the corresponding bag. Ideally, the extracted K patches should be discriminative for classification. At the same time, the patches need to be representative and should reflect information about the inherent structure of the images. In order to extract discriminative patches, we assign probabilities to different patches. The assigned probabilities should represent the discriminative ability of the corresponding patches. If patches at a specific location are highly discriminative between different groups, high probabilities will be assigned to these patches. The patches with high probabilities are then extracted for classification. Various methods such as t-tests [2] or sparse coding [4] can be used to calculate the probabilities for different patches.

However, this does not mean that the more discriminative patches in the bags, the better performance the classifier achieves. The classification accuracy is also affected by the relationships among the selected patches [5,10] since these patches are used as features. If the center voxels of the extracted patches lie in a contiguous area, the patches are more likely to share common information. In this case,

these discriminative patches contain a large amount of redundant information and cannot capture information about the inherent structure of the images. To avoid large overlaps among the selected patches and reduce redundant feature information, we defined a distance threshold to select the K patches. Finally, K patches are extracted from each image according to the patch probabilities and the defined distance threshold.

2.2 Graph-Based Multiple Instance Learning

After the patch extraction step, we have N training bags with K patches. The label of the bags are known while the labels of patches in the positive bags are unknown. This multiple instance learning problem can be solved using a number of different approaches [7]. Most of these approaches typically treat patches in the bags as independent instances and neglects their relationships. The relationships among patches, however, may be beneficial for learning strong classifiers. Considering our classification problem, it is more meaningful to treat the patches extracted from the same image as inter-correlated samples. In this paper, we proposed to use a graph-based multiple instance learning method [10] for learning the bag-level classifier. In this method, a graph is constructed for each bag and the patches in each bag are treated as nodes in the corresponding graph. This requires (a) the construction of a graph for every bag and (b) a graph kernel designed to capture the similarity among graphs and (c) a bag-level classifier trained by kernel machines. In the following, we will describe these steps in details.

The construction of the graph for each bag is quite straightforward: Similar to approaches in manifold learning [11], affinity matrices are derived to construct graphs, which can capture the underlying manifold structure of the data. Let us denote a bag P_i with K patches, each patch in the bag P_i can be viewed as a node in the graph G_i. The distances between every pair of nodes are calculated using Euclidean distances and used to define the affinity matrix W^i. The affinity matrix W^i represents a graph that models the relationships among the patches in bag P_i. In the resulting graph, the weight of each edge corresponds to the dissimilarity between the corresponding pair of patches.

After mapping the bags to graphs, a graph kernel is defined to capture the similarity among graphs. Given two bags P_i and P_j which are represented as graphs with matrices W^i and W^j respectively, the graph kernel is then defined as:

$$K_G\left(P_i, P_j\right) = \frac{\sum_{a=1}^{K}\sum_{b=1}^{K} d_{ia} d_{jb} k(p_{ia}, p_{jb})}{\sum_{a=1}^{K} d_{ia} \sum_{b=1}^{K} d_{ib}} \tag{1}$$

where $d_{ia} = 1/\sum_{u=1}^{K} w_{au}^{i}$, $d_{jb} = 1/\sum_{v=1}^{K} w_{bv}^{j}$. K_G is a positive semidefinite kernel and in our case a Gaussian kernel function is used:

$$k\left(p_{ia}, p_{jb}\right) \;=\; \exp\left(-\gamma \left\|p_{ia} - p_{jb}\right\|\right) \tag{2}$$

where p_{ia} and p_{jb} are patches in bags P_i and P_j respectively. As shown in the Equations 1 and 2, both the nodes p which represent the intensities of patches and the edges w which reflect the relationships among patches are important for calculating the graph kernels K_G. Finally, the kernel K_G is normalized:

$$K_G\left(P_i, P_j\right) \;=\; \frac{K_G\left(P_i, P_j\right)}{\sqrt{K_G\left(P_i, P_i\right)}\sqrt{K_G\left(P_j, P_j\right)}} \tag{3}$$

Once the graph kernel is obtained, a classifier can be trained using various kernel machines such as kernel Fisher linear discriminant analysis (LDA), kernel principal components analysis (PCA), and support vector machine (SVM). Among them, SVM is one of the most widely used kernel machines because of its accurate classification performance [12]. In this paper, we chose SVM for our classification experiments.

3 Experiments and Results

In this paper we used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study (www.loni.ucla.edu/ADNI). All the 834 baseline images at 1.5T were included for evaluation as in [13,6]. The dataset consists of 231 Cognitively Normal (CN) subjects, 238 subjects with Stable Mild Cognitive Impairment (SMCI), 167 subjects with Progressive Mild Cognitive Impairment (PMCI) and 198 patients with AD. All images were preprocessed by the standard ADNI pipline described in [14]. After that, a non-rigid registration based on B-spline free-form deformation [15] with a final control point spacing of 2.5mm was performed to align all images to the MNI152 template space. The approach proposed in [16] was used to normalize the intensity between the subjects and the template. After preprocessing, all the images are spatial normalized and the intensities are homogeneous across the images.

To extract K patches from each image, t-tests are performed on the normalized images and p-values are calculated for each voxel. The mean p-value at each voxel is then calculated within its patch area. The probability of each patch is defined as the reciprocal of its mean p-value. The first patch is extracted when the probability of the patch is the highest. If the distances between the selected patch and its neighboring patches are smaller than the defined threshold, the probabilities of these neighboring patches are set to zeros. As a result, these neighboring patches of the previous selected patch will not be extracted. The next patch is extracted when the probability of the patch is the second highest. This selection step repeats until K patches are extracted. Patches are extracted in a region of interest (ROI) defined around the hippocampus as show in Figure 1. Experiments were performed using leave-one-out cross validation. We have compared the classification performance in different scenarios, including CN vs AD and SMCI vs PMCI. In all cases a patch size of $7 \times 7 \times 7$ is used to capture local structural information [6,17].

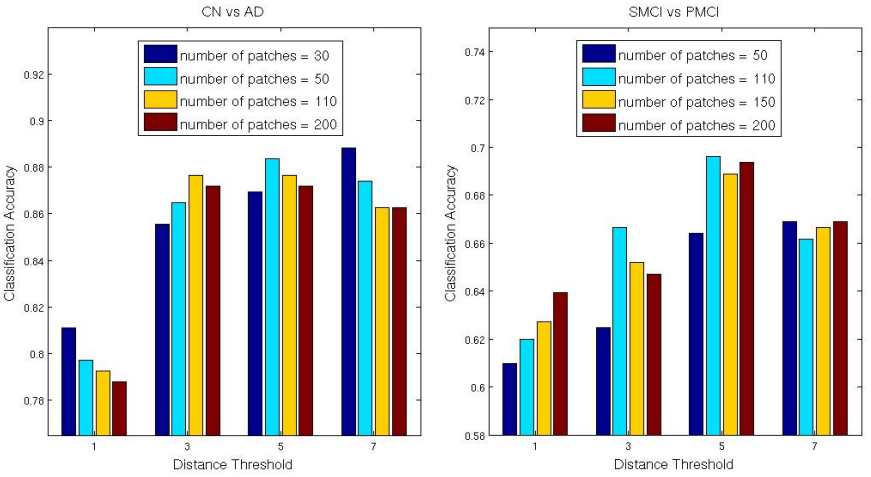


Fig. 2. Effect of the patch distance threshold and the number of the selected patches on the classification accuracy

3.1 Effect of Patch Extraction

Two key parameters in the proposed approach are the number of selected patches K and the patch distance threshold. Fig. 2 shows the classification accuracies for CN vs AD and SMCI vs PMCI when using different number of patches and various patch distance thresholds. For CN vs AD, the number of patches has less impact on the classification accuracy than the distance threshold. When a patch distance threshold of 1 is used, the extracted patches are the K patches with the smallest mean p-values of all M patches. This means that these patches are more discriminative than K patches extracted using other patch distance thresholds. However, the classification accuracy is low as shown in Fig. 2 although these patches are discriminative. The reason for this may be that the extracted patches have significant overlap with each other and contain a large amount of redundant information. As a result, they may not be able to convey information about the inherent structure of the images. Using a larger patch distance threshold (e.g. 5 or 7), the proposed method achieves a classification accuracy of 0.88 for CN vs AD.

The classification between SMCI and PMCI is far more challenging because the anatomical changes at the prodromal stage of AD disease are subtle [6]. As shown in Fig. 2, the classification of SMCI vs PMCI requires more patches (more than 100) than the classification of CN vs AD in order to achieve the best performance. When the patch distance threshold is 5, the proposed method achieves more accurate performance than those using other patch distance thresholds, which also indicates that the relationships among patches are very important for disease classification.

3.2 Comparison

The performance of the proposed method miGraph was compared with the performance of widely used linear SVM. To use the linear SVM, the same patches used in miGraph were arranged to form feature vectors and treated as input for the linear SVM. The LIBSVM toolbox [18] was used to train the classifiers. The cost parameter C was set to 1 for both methods. Table 1 shows a comparison of the classification performances using the proposed miGraph and the standard linear SVM. The results show that the proposed method achieves a significantly more accurate performance than the standard linear SVM. The improvement is gained by just replacing linear kernels with graph kernels in SVM. Although the SVM classification uses the same K patches as features, it neglects the relationships among these patches, resulting in a lower classification accuracy than the proposed method.

Table 1. Methods comparison. The classification accuracy (ACC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV) and negative predictive value (NPV) are presented in the table. In the classification of CN vs AD, the number of patches is 30 and the patch distance threshold is 7. In the classification of SMCI vs PMCI, the number of patches is 110 and the patch distance threshold is 5.

Comparison	Method	ACC	SEN	SPE	PPV	NPV
CN vs AD	Linear SVM	83.2%	73.7%	91.3%	87.9%	80.2%
	mi-Graph	88.8%	85.9%	91.3%	89.5%	88.3%
PMCI vs SMCI	Linear SVM	63.0%	56.9%	67.2%	54.9%	69.0%
	mi-Graph	69.6%	67.1%	71.4%	62.2%	75.6%

When using the baseline ADNI MR images from 834 subjects, our proposed method can achieve similar or improved results compared to approaches recently proposed in [13,6]. However, it should be noted that additional features such as age, cortical thickness or hippocampus volume were used in [13,6] to aid the classification. In the proposed method, only the MR intensities were used for classification and no segmentation was needed, which also demonstrates the effectiveness of our proposed method.

4 Conclusion and Future Work

In this work, we have proposed a novel framework for the diagnosis of subjects with AD by using multiple instance learning. The experimental results indicate that it is possible to improve the performance of a classifier by exploiting the relationships among instances. Not only the intensities of patches but also the relationships among patches affects the performance of the proposed approach. Therefore, selecting more discriminative and meaningful patches will be a focus of future work. In addition, It should be noted that the graph kernel used in the paper may not be the best choice for our classification work. A better graph kernel may be able to capture more useful structural information of the images. We will investigate the effect of using other graph kernels in future work.

References

1. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. Neuroimage 56(2), 766–781 (2011)
2. Chu, C., Hsu, A.L., Chou, K.H., Bandettini, P., Lin, C.P.: Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. NeuroImage 60, 59–70 (2011)
3. Liu, M., Zhang, D., Yap, P.-T., Shen, D.: Tree-guided sparse coding for brain disease classification. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 239–247. Springer, Heidelberg (2012)
4. Janousova, E., Vounou, M., Wolz, R., Gray, K., Rueckert, D., Montana, G.: Biomarker discovery for sparse classification of brain images in Alzheimer's disease. Annals of the BMVA 2012(2), 1–11 (2012)
5. Liu, M., Zhang, D., Shen, D.: Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Human Brain Mapping (2013)
6. Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Pruessner, J.C., Allard, M., Collins, D.L.: Scoring by nonlocal image patch estimator for early detection of Alzheimer's disease. NeuroImage: Clinical 1(1), 141–152 (2012)
7. Babenko, B.: Multiple instance learning: algorithms and applications (2008)
8. Bi, J., Liang, J.: Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure. In: IEEE CVPR, pp. 1–8 (2007)
9. Xu, Y., Zhang, J., Chang, E.I.-C., Lai, M., Tu, Z.: Context-constrained multiple instance learning for histopathology image segmentation. In: Ayache, N., Delingette, H., Golland, P., Mori, K. (eds.) MICCAI 2012, Part III. LNCS, vol. 7512, pp. 623–630. Springer, Heidelberg (2012)
10. Zhou, Z.H., Sun, Y.Y., Li, Y.F.: Multi-instance learning by treating instances as non-IID samples. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1249–1256 (2009)
11. Pless, R., Souvenir, R.: A survey of manifold learning for images. IPSJ Transactions on Computer Vision and Applications 1(0), 83–94 (2009)
12. Sanchez, A., David, V.: Advanced support vector machines and kernel methods. Neurocomputing 55(1), 5–20 (2003)
13. Wolz, R., Julkunen, V., Koikkalainen, J., Niskanen, E., Zhang, D.P., Rueckert, D., Soininen, H., Lötjönen, J.: Multi-method analysis of MRI images in early diagnostics of Alzheimer's disease. PloS one 6(10), e25446 (2011)
14. Jack Jr., C., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging 27(4), 685–691 (2008)
15. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Transactions on Medical Imaging 18(8), 712–721 (1999)
16. Nyu, L.G., Udupa, J.K.: On standardizing the MR image intensity scale. Magnetic Resonance in Medicine 42(6), 1072 (1999)
17. Liu, M., Zhang, D., Shen, D.: Ensemble sparse classification of Alzheimer's disease. Neuroimage 60(2), 1106–1116 (2012)
18. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`