Diagnosis of Alzheimer’s Disease using Multiple Instance Learning

Alexandre O.C.C. Matos
alexandre.ochoa.matos@ist.utl.pt

Institute for Systems and Robotics, IST, Lisbon, Portugal

June 2015

Abstract

Alzheimer’s disease (AD) is a burden for the contemporary society. Its early detection and accurate diagnosis may lead to an improvement in treatment. The diagnosis at pre-clinical stages or the prediction of conversion of patients with Mild Cognitive Impairment (MCI) to AD is a challenging problem, receiving attention because of the immense associated social and economic costs. This work aims to present an automatic diagnostic tool based on a recent machine learning field, the Multiple-Instance Learning (MIL) model, which differs from the more common supervised learning paradigm. In order to accomplish this, five Multi-Instance Learning algorithms - Diverse Density, CKNN, MILES, YARDS and BARTMIP - were tested, spanning the main MIL paradigms. Experiments were done both in a cross-sectional and longitudinal analysis on Fluorodeoxyglucose (FDG)-PET images retrieved from ADNI to perform binary classification between AD, MCI and Cognitively Normal (CN) patients. Also, an improved and still scarcely used method of normalisation - the Yakushev Normalisation - was implemented, confirming beneficial effects on diagnosis accuracy. Results show that, although not outperforming other supervised approaches, this paradigm is competitive with the state of the art. Furthermore, no significative differences were found between the five classifiers. Best balanced accuracies obtained for CN vs. AD, MCI vs. AD and CN vs. MCI were, respectively, 90.98%, 71.32% and 66.85%. On the other hand, Receiver Operator Characteristic (ROC) curves analysis registered areas values of 0.96, 0.77 and 0.67, respectively. What concerns the effect of longitudinal analysis, it seems to be effective only when the follow-up instant shows better accuracy. Finally, the possibility of using MIL on unregistered data was also investigated.

Keywords: Alzheimer’s Disease, Computer Aided Diagnosis, Positron Emission Tomography, Multiple Instance Learning

1. Introduction

Alzheimer’s Disease (AD) is a neurological disorder which affects mostly people aged 65 or over. It belongs to a wider category of diseases all termed as dementia, such as vascular dementia (20-30%), frontotemporal lobar degeneration (5-10%), dementia with Lewy bodies (5%) and the less studied mixed dementia and Parkinson’s disease. AD accounts for 50-75% of the cases [1].

Dementia is a syndrome related to disease in the brain, showing progressive decline in cognitive abilities. Although it can appear at any age, it is more common over the age of 65, with its prevalence doubling every 5 years after this cornerstone and its numbers doubling every twenty years [1]. Its diagnosis is usually carried out according to the Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria, which require a decline in memory and a simultaneous impairment in one cognitive function such as ability to speak coherently and understand spoken or written language or to perform motor activities. All these must interfere with daily life [2].

What concerns the AD diagnosis, it is based on the criteria published by the Alzheimer’s Association and the National Institute of Neurological Disorders and Stroke [3] in 1984. However, many developments have taken place since then. In particular, other dementia syndroms such as vascular dementia and dementia with Lewy bodies were not well described by then. Moreover, imaging techniques were not taken into account at the time and genetics of AD was poorly understood. All these factors led to the release of new criteria, published by the National Institute on Aging (NIA) and the Alzheimer’s Association in 2011. These update the ones published before, introducing new and relevant changes: they identify three stages of AD - Preclinical Alzheimer’s Disease, Mild Cognitive Impairment (MCI) and dementia due to AD - and launch biomarker tests [4].
The first stage already presents measurable changes in the brain, cerebrospinal fluid and blood, but no symptoms. It reflects the possibility, suggested by some researchers, that Alzheimer’s Disease starts up to 20 years before symptoms arise [5, 6]. The MCI stage presents the first noticeable changes in thinking, despite not affecting daily activities. Some patients show concern about changes in cognition and there is a lower performance in one or more cognitive domains such as memory, language or visuospatial skills, while maintaining their independence when performing daily tasks [7]. Finally, the probable AD patient shows interference with the ability to do usual activities, a decline in performance and cognitive impairment, which must be assessed through a history-taking from the patient and an objective cognitive exam, and encompasses at least two behavioural domains amongst memory, judgement, visuospatial abilities, language and personality. Exams such as Mini Mental State Examination (MMSE) and Clinical Dementia Rating (CDR) are often used to assess the patient mental status. Despite all the data collected from the patient, it should be emphasised that AD definite diagnosis is declared post-mortem.

AD biomarkers are relevant since cognitive measures have less statistical power and higher variability. Withal, they can act as predictors of cognitive decline and measure the changes that occur in the so called continuum of AD [8]. They have potential to be used as a clinical and prognostic tool, eventually allowing the effects of treatments to be measured. PET is a paradigmatic case of functional markers and a non-invasive nuclear medicine imaging technique. Its principle lies on the emission of gamma rays released from a radionuclide which is introduced in the body incorporated in a biological active molecule. In particular, 18F-2-fluoro-2-deoxy-D-glucose (18FDG) PET is currently one of the most used molecules. It enables a quantitative estimate of local cerebral metabolic rate of glucose consumption (CMRglc) and improves diagnosis accuracy, particularly in early AD cases [9].

Recently, an important enhancement in AD diagnosis accuracy was accomplished by Yakushev et al. with respect to reference region normalisation in statistical parametric mapping [10]. Normalisation of regional tracer uptake is a usual proceeding in PET analyses, since it allows to account for inter and intra subject variability. Another benefit lies on the reduced variability of relative CMRglc in comparison to its absolute values. Still, this step is problematic for the widespread metabolic and blood flow reduction that happen in AD. The net effects are an underestimation of metabolic reduction in diseased brains and the detection of false hypermetabolic regions in AD patients. This can easily be understood when a Cerebral Global Mean (CGM) normalisation is carried out, as in AD brains the CMRglc are abnormally low, which leads to a diminished CGM and to the aforementioned grievous effects. These can appear even in early cases of MCI [11]. The apparently hypermetabolic regions, as Yakushev denominated them, are areas usually spared in AD, such as the primary sensorimotor cortex, cerebellum, basal ganglia and brain stem. Yakushev et al. [68] tested the normalisation to a reference region named reference cluster, which is obtained through a voxelwise t-test between AD patients and controls and then a selection of the voxels with t-value above a given threshold. By doing so, the best of the areas which are apparently hypermetabolic is chosen. This process was termed by Borghammer as Yakushev Normalisation [12]. Authors such as Rodrigues and Silveira [13] and Gray [14] have exploited the benefits of this procedure.

Besides normalisation, a relevant procedure in Computer Aided Diagnosis (CAD) is feature selection, whose benefits range from a better visualisation of data or selection of discriminative features to the mitigation of the so called curse of dimensionality, which often jeopardises the classifier performance [15]. The high dimensionality of input space when there is a relatively small sample of subjects represents a drawback as it leads to increased complexity and overfitting. In the context of disease diagnosis, features should be representative of the disease process and the most insensitive as possible to noise or other artifacts [16]. Particularly, when dealing with brain image data, the number of features can easily reach hundreds of thousands, while data sets have usually under 200 subjects and not all image information is useful for classification. Some common approaches to overcome this problem are Principal Component Analysis (PCA) [17], Linear Discriminant Analysis (LDA) [18], Non-negative Matrix Factorisation (NMF), mutual information [19, 20], the Pearson correlation coefficient [20, 21], the Fisher Discriminant Ratio (FDR) [22], the absolute value of two-sample t-test statistic [23, 24, 12, 25], or Region Of Interest (ROI) parcellation [15, 26]. There are also classifiers with a built-in capacity of feature selection such as LASSO regression [27, 28], Elastic Net [29] and Adabost [30].

Concerning CAD studies applied to AD, there is a vast literature. To the best of our knowledge, the first work was published in 1990, by Kippenhan et al. [31], obtaining a remarkable 91% sensitivity and 93% specificity with PET images and using neural networks. Stoeckel et al. introduced the use of voxel intensities as features, using SPECT images [32]. Duchesne et al. [19], Silveira et al. [30] or Cuignet et al. [33] are among the relevant research
papers produced in the last ten years. All these were performed resorting to supervised methods.

Contrarily, the field of Multiple-Instance Learning (MIL) applied to medical diagnosis, and in particular AD, remains little explored. Concretely, one paper has been written about mammography [34], one about pulmonary embolism [35] and another one in the context of liver cancer [36]. With respect to Alzheimer’s Disease, only two papers, both by Tong et al., were published [24, 37].

In [24], Tong et al. developed the first MIL algorithm designed to detect AD and MCI using MRI images from the ADNI database, achieving 88.8% of accuracy between AD patients and healthy controls and 69.6% of accuracy between stable MCI and progressive MCI. In that paper, each image was segmented so that the hippocampus was used as ROI. In the following paper, the authors extended their approach to the whole brain and changed their patch extraction method, using a previously implemented penalised regression model known as elastic net [38]. The authors obtained 89% accuracy between AD and controls, 82.9% accuracy between controls and progressive MCI and 70.4% accuracy between progressive and stable MCI.

2. Multiple Instance Learning

MIL, as supervised learning, involves labeling unseen examples after a training process with labeled examples through the inference of a function that takes into account the relationship between these and their label. However, MIL differs from the latter in that each example, also known as bag, may contain more than one feature vector, known as instance. The founding work led by Dietterich et al. [39] was motivated by a drug activity prediction study concerning musk molecules, which in some conformations are detected by human nose with smell. It was assumed each molecule was a bag and each of its many conformations was an instance.

In a recent and key paper, Amores [40] formulated a comparative study in which he considered the existence of three paradigms in MIL classification according to the way information was extracted from bags:

- **Instance-Space paradigm** - The discriminative information lies on instances, that is, the classifier receives instances as input. Some of the algorithms developed under this paradigm are DD [41], EM-DD [42], MI-SVM and mi-SVM [43].

- **Bag-Space paradigm** - The classifier receives bags as input. It generally relies on the definition of a distance between bags. Some common distances are the minimal Hausdorff distance, Chamfer distance or Earth-Movers distance. CKNN [44] represents this model fairly well. MI Kernel, a bag kernel that may be used with Support Vector Machines, is also a good example, developed by Gartner [45].

- **Embedded-Space paradigm** - As in the previous paradigm, there is information extracted directly from bags but in an explicit way, by mapping the bag from the bag space to a single instance vector space. One example is the simpleMI algorithm, also called statistical kernel, in which a bag is represented by statistical measures such as maxima or minima over features in a single vector [46]. MILES, YARDIS and BARTMIP are also paradigmatic of this case.

This nomenclature is complete, in the sense that every published MIL algorithm falls into one of the paradigms.

3. Material and Methods

3.1. Neuroimaging data

The dataset used in this study contained FDG-PET images acquired from 54 patients suffering from Alzheimer’s disease, 128 with Mild Cognitive Impairment and 74 Normal Controls, at two instants: the initial visit (baseline) and the follow-up visit at 12 months. From the pool of data provided by ADNI, only those corresponding to the intersection of both time instants and satisfying some CDR restrictions were considered: 0 for normal controls, 0.5 for MCI patients and 0.5 or higher for those suffering from AD (see table 1).

<table>
<thead>
<tr>
<th>Group</th>
<th>CN</th>
<th>MCI</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients</td>
<td>74</td>
<td>128</td>
<td>54</td>
</tr>
<tr>
<td>Age</td>
<td>76.1 ± 4.7</td>
<td>74.9 ± 7.3</td>
<td>76.8 ± 6.6</td>
</tr>
<tr>
<td>Sex(%) males</td>
<td>64.9</td>
<td>65.6</td>
<td>51.9</td>
</tr>
<tr>
<td>MMSEbaseline</td>
<td>29.1 ± 1.0</td>
<td>27.2 ± 1.6</td>
<td>23.5 ± 2.0</td>
</tr>
<tr>
<td>MMSE12months</td>
<td>29.1 ± 1.2</td>
<td>26.8 ± 4.3</td>
<td>21.0 ± 4.1</td>
</tr>
<tr>
<td>CDR</td>
<td>0</td>
<td>0.5</td>
<td>0.5 or 1</td>
</tr>
</tbody>
</table>

3.2. Image preprocessing

ADNI data comes from a wide range of different FDG-PET scanners, which introduces undesired variations in image obtention. In order to obtain more uniform and similar FDG-PET data, it was submitted to pre-processing steps carried out by ADNI, in which each patient image was co-registered, averaged and unformised in terms of resolution [47].
After ADNI preprocessing steps, images from different patients were still not aligned between them, which rendered voxelwise intersubject comparison impossible. This was circumvented via a warping into the MNI152 standard space, as described in Morgado and Silveira [48].

### 3.3. Yakushev Normalisation

A SPM like approach was taken, as in Rodrigues [49]. A paired t-test was performed to provide a voxelwise computation of t-statistics. Then, in order to obtain a reference cluster of voxels, two parameters were set - a t-value threshold (t) and a minimum cluster radius (r) (see table (2)). From those voxels satisfying these constraints, only the cluster with the highest mean t-value was selected. The mean of voxel intensities inside the cluster was then used as the normalisation factor. The best parameters for each modality (CN vs. MCI, CN vs. AD and MCI vs. AD) were computed heuristically and are presented in table (2).

It should be noted that the reference clusters were obtained in a 10 fold cross-validation procedure, each corresponding to one fold. After Yakushev Normalisation, data was normalised to the [0,1] interval. This is a usual procedure, to avoid features with wide ranges dominating those with smaller ones and prevent numerical problems [50]. Also, despite Yakushev Normalisation being less effective for comparisons between controls and early stages of dementia and mild cognitive impairment, it was also applied in these cases.

<table>
<thead>
<tr>
<th>Modality</th>
<th>(t_{\text{baseline}})</th>
<th>(r_{\text{baseline}})</th>
<th>(t_{12\text{months}})</th>
<th>(r_{12\text{months}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>CN vs. MCI</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CN vs. AD</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>MCI vs. AD</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

### 3.4. Feature Selection

The first selection procedure was the application of a binary mask, since no useful information could be given by voxels outside the brain, as in [30]. Briefly, it consisted of a binary mask where each position inside the brain was set to logical true. To define the voxels inside the brain, an average brain was computed using the whole database set and a threshold set at 5% of the maximum value was set, after empirical tests (see figure 1).

Figure 1: The binary mask. Axial cut of a patient’s original image (left) and the same image after applying the binary mask (right).

Next, a ranking feature selection method was applied based on the t-value obtained from unpaired t-test that allowed to obtain patches from images in a probabilistic approach, as Tong et al.[24]. Briefly, t-tests were performed on the images and each voxel was attributed a t-value. Then, a mean t-value was calculated within a surrounding cubic volume centred at each voxel. Finally, the probability of each patch was associated with the t-statistic absolute value, corresponding to the intuition that a higher t-value means a more probable rejection of the null hypothesis. In theory, the number of patches N may reach the dimensionality of the image, which is extremely high. This may lead to dimensionality problems, which suggests the selection of only \(K \leq N\) patches. These were sorted by discriminative power, that is, in descending order of their absolute t-value. Additionally, a distance threshold calculated between the geometric centre of patches was imposed, a strategy that potentially reduces redundant information extraction [24]. Lastly, the patch size was also taken into account (see table (3)). Patches were treated as instances, and the whole images as bags with equal number of patches.

Table 3: Patch selection parameters. \(K\) - total number of patches; \(l\) - cube edge (voxel units); \(d\) - distance threshold (voxel units).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>(K)</td>
<td>[1 2 5 10 20 50 100]</td>
</tr>
<tr>
<td>(l)</td>
<td>[3 5 7 9]</td>
</tr>
<tr>
<td>(d)</td>
<td>[2 3 5 10]</td>
</tr>
</tbody>
</table>

### 3.5. Classification

1. Support Vector Machines (SVM) - Support Vector Machines are one of the most powerful tools in supervised machine learning, being widely used. They belong to the field of linear classifiers, such as the simple perceptron, but are much more flexible. SVM was originally developed by Vapnik in 1963 [51]. Its actual form is due to Boser, Guyon and Vapnik, who structurated the problem in a standard form [52]. One important improvement was the introduction of a relaxing concept - the soft margin -
2. Diverse Density (DD) - Developed by Maron [41], it is an IS paradigm algorithm which searches for the point $\vec{c}$ in instance space with the highest intersection of the positive bags minus the union of the negative bags. To accomplish so, a probabilist approach is carried out. Let $\vec{x} \in \chi$ be a point in the instance space, $\vec{c} \in \chi$ be the target concept, $\{B^+_1, B^+_2, ..., B^+_n\}$ be $n$ positive bags and $\{B^-_1, B^-_2, ..., B^-_m\}$ be $m$ negative bags, both in the training set. Then, DD is defined as:

$$DD(\vec{x}) = P(\vec{x} = \vec{c}|B^+_1, ..., B^+_n, B^-_1, ..., B^-_m)$$

(1)

As noted by Foulds [54], this is a measure of the probability that a point in instance space is the true concept, given the training bags. Applying Bayes rule and assuming some independence conditions, the maximum value of DD is given by the following maximum likelihood expression:

$$\arg\max_{\vec{x} \in \chi} \prod_{1 \leq i \leq n} P(\vec{x} = \vec{c}|B^+_i) \prod_{1 \leq j \leq m} P(\vec{x} = \vec{c}|B^-_j)$$

(2)

Maron computed the probabilities above according to the noisy-or model:

$$P(\vec{x} = \vec{c}|B^+_i) = 1 - \prod_{1 \leq j \leq n} (1 - P(B^+_ij \in \vec{c}))$$

(3)

$$P(\vec{x} = \vec{c}|B^-_j) = \prod_{1 \leq j \leq m} (1 - P(B^-_ij \in \vec{c}))$$

(4)

The expression to compute the probability that an instance belongs to a given concept was defined as:

$$P(B^-_{ij} \in \vec{c}) = \exp(- \sum_{1 \leq l \leq k} (B^-_{ijl} - \alpha_l)^2)$$

$$= \exp(-||B^-_{ij} - \vec{c}||^2)$$

(5)

where $\vec{B}_{ij}$ designates feature $l$ of instance $j$ belonging to bag $i$.

The expression above defines what Maron termed single point concept class, which assumes that every concept corresponds to a single point in the instance space and each positive bag contains at least one instance that is equal to the true concept corrupted by some Gaussian noise.

Finding in an efficient way the concept in the instance space that maximises DD is in general a difficult global optimisation problem, for the concept space is at least as big as the feature space, continuous and high dimensional [41]. Maron proposed the so called maxDD algorithm, which performs a gradient ascent based optimisation in the concept space, searching for the point which is closest to instances from positive bags and far from all negative instances. One of the drawbacks of this optimisation process is the possibility of getting stuck in local maxima. One good heuristic that diminishes this risk is to perform multiple optimisations, each started in a different random point or a priori good points, as instances belonging to positive bags. Despite not being guaranteed to reach the global maximum, this a good heuristic, even if carrying a high computation cost when dealing with a large number of instances in positive bags [41].

In order to classify a bag, one selects the distance threshold $d$ (equivalent to a probability threshold, see equation 5) that maximises the accuracy on the training set, according to a nested cross-validation procedure, a technique explained further below in this section. More specifically, the probability threshold was tuned in a search-grid between 0.1 and 1 in steps of 0.05. A bag is positive if at least one of its instances is within distance $d$ of the true concept.

It should be noted that since patches from brain images were selected in different areas of the brain, thus preventing any equivalence between features, histograms of patch voxel intensities were used to circumvent this aspect.

3. CKNN - Developed by Wang and Zucker [44] as an extension of the k-NN classifier to fit the MIL paradigm, it belongs to a lazy form of learning, where training examples are stored so that when shown a test example, these are queried [55]. It is also a BS paradigm classifier.

As in its supervised counterpart, the test example label is decided by majority vote among the labels of its k nearest neighbours. In k-NN, the Euclidean distance is a usual distance measure choice. However, under MIL, there is the necessity to define a new one, since bags are not vectorial entities, but sets of vectors. For that, Wang and Zucker applied an adapted version of Hausdorff distance:
\[ d(A, B) = \min_{a \in A} \min_{b \in B} ||a - b|| \quad (6) \]

Known as Minimal Hausdorff Distance, this is simply the minimal distance between two bags and has the advantage of being less prone to noise. Nonetheless, it was shown through the benchmark dataset Musk1, that in some cases bags from one class were incorrectly classified by the usual majority voting process. This led to the introduction of citers and references in substitution of neighbours.

The concepts of citer and reference were defined in the context of library and information science. Informally, a reference of an unseen bag may be translated as one of its R nearest neighbours. A citer of an unseen example may be defined as a bag that sees it as reference. In more formal terms, let \( n \) be the number of bags and \( BS = \{ B_1, ..., B_n \} \) the set of all bags. Then, for an example \( B \in BS \), a bag \( B_i \in BS \setminus B \) may be ranked according to its similarity to example \( B \). Let its similarity be defined by a rank number \( \text{Rank}(B_i, B) \). The C-nearest citers of \( B \) are defined based on a rank, as follows:

\[
\text{citer} s(B, C) = \{ B_i \mid \text{Rank}(B_i, B) \leq C, B_i \in BS \} \quad (7)
\]

Figure (2) shows how citers and references are assigned to bags:

<table>
<thead>
<tr>
<th></th>
<th>N=1</th>
<th>N=2</th>
<th>N=3</th>
<th>N=4</th>
<th>N=5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( b_1 )</td>
<td>( b_3 )</td>
<td>( b_2 )</td>
<td>( b_2 )</td>
<td>( b_4 )</td>
<td>( b_8 )</td>
</tr>
<tr>
<td>( b_2 )</td>
<td>( b_1 )</td>
<td>( b_4 )</td>
<td>( b_3 )</td>
<td>( b_3 )</td>
<td>( b_6 )</td>
</tr>
<tr>
<td>( b_3 )</td>
<td>( b_2 )</td>
<td>( b_1 )</td>
<td>( b_2 )</td>
<td>( b_5 )</td>
<td>( b_4 )</td>
</tr>
<tr>
<td>( b_4 )</td>
<td>( b_6 )</td>
<td>( b_2 )</td>
<td>( b_1 )</td>
<td>( b_5 )</td>
<td>( b_5 )</td>
</tr>
<tr>
<td>( b_5 )</td>
<td>( b_1 )</td>
<td>( b_2 )</td>
<td>( b_1 )</td>
<td>( b_5 )</td>
<td>( b_4 )</td>
</tr>
<tr>
<td>( b_6 )</td>
<td>( b_2 )</td>
<td>( b_3 )</td>
<td>( b_2 )</td>
<td>( b_5 )</td>
<td>( b_5 )</td>
</tr>
</tbody>
</table>

Figure 2: The nearest neighbours of 6 bags \( \{b_1, b_2, b_3, b_4, b_5, b_6 \} \). \( N \) is the nearest rank number. Let \( R \) and \( C \) be both 2. Then, for bag \( b_1 \), its R-nearest references are \( b_3 \) and \( b_2 \) and its C-nearest citers are amongst \( b_2, b_3 \) and \( b_5 \). Source: [44].

4. MILES - developed by Chen et al. [56] as a wrapper classifier, where the 1-norm SVM was used to classify and select the most important features.

In broad terms, MILES has the capability of identifying the most relevant instances and selecting those through the definition of a similarity measure between bags and instances. This is accomplished with an innovative way of bag representation, in an embedded space where each feature is associated to an instance, and thus feature selection becomes equivalent to instance selection. The embedding step makes MILES belong to vocabulary based methods.

This embedding is performed through:

\[
\tilde{M}(B_i) = [s(\vec{c}_1, B_i), s(\vec{c}_2, B_i), ..., s(\vec{c}_n, B_i)] \quad (8)
\]

where \( n \) is the dimension of the new vector \( \tilde{M} \), equal to the number of instances from all training examples and \( s \) is given by the so called most-likely cause model:

\[
s(\vec{c}_k, B_i) = \max_j \exp(-\frac{||x_{ij} - \vec{c}_k||^2}{\sigma^2}) \quad (9)
\]

The target concept points may be approximated by instances in training bags, irrespective of their label. Thus, vector \( \tilde{M} \) summarises the similarities between a bag and all training instances [57]. Desirably, if \( \vec{c}_k \) shows a higher similarity to positive bags and low similarity to negative bags, feature \( s(\vec{c}_k, B_i) \) might be useful in separating the two classes of bags [56].

MILES was originally built using 1-norm SVM, which causes the weight vector to have more null entries. Moreover, it renders the optimisation problem a linear programming one, instead of quadratic, which allows an improvement in its resolution speed [54]. However, it should be noted that any other single-instance classifier might be used inside MILES. In this work, 2-norm SVM with linear kernel was chosen.

5. YARDS - This classifier was developed by Foulds in his master thesis [54] as a modified version of MILES. Its main aim was to render the learnt instance weights an effective function over the instance space. As MILES, YARDS maps each bag into an instance-based feature space where each feature represents the similarity between this bag and a training instance and then applies some supervised classifier for classification. For the first, this similarity was given by the most-likely cause estimator which accounted only for the closest instance inside a bag (see (9)). YARDS excludes this dependency by replacing the max operator by a sum.
in the equation. For the rest, everything is identical:

\[ s(\vec{c}_k, B_i) = \sum_j \exp\left(-\frac{||\vec{x}_{ij} - \vec{c}_k||^2}{\sigma^2}\right) \] (10)

6. BARTMIP - developed by Zhang and Zhou [58] in 2008, this classifier presented a novel approach in MIL, as it was based in the first multi-instance clustering method, BAMIC. Being an ES paradigm algorithm, its vocabulary is obtained in an unsupervised way and is constituted by bags, instead of instances (the case of MILES and YARDS).

BAMIC was implemented as an adaptation of the Partitioning Around Medoids, a method with two phases, nominated the build phase and swap phase, respectively. In the first one, k examples are randomly selected as the initial medoids. Then, based on a dissimilarity matrix computed with the average Hausdorff distance, each example is assigned its closest medoid. Finally, the swap phase looks for examples inside each cluster that might lower a given cost function (the sum of average Hausdorff distance to a medoid). If such example is found, new clusters are rebuilt. This process iterates until no change occurs.

BARTMIP treats each bag as a k-dimensional feature vector whose i-th feature corresponds to the distance between the bag and the centroid of the i-th cluster and then applies some conventional supervised learner to the data. In this work, as in the original paper, 2-norm SVM with linear kernel was used, with C parameter choosen by nested cross-validation.

3.6. Longitudinal Analysis

All the algorithms and proceedings presented above were applied to cross-sectional data, that is, baseline and 12 months data separately. Besides this analysis, a longitudinal one was also tested, assembling the two time instants. This analysis was tested with SVM with patches and MILES and shared all preprocessing steps explained before, the only difference lying on the way patches were selected, so that both time instants could be considered. Hence, an average patch t-value sorting was computed for the two instants and then patches were selected from this sorted list.

3.7. Simulation of non-registered images

Image registration plays an important role in many applications, namely in medical imaging. The process is often required such that non-aligned and distorted images may be compared in a common space, the registered space. Without it, voxelwise comparison becomes almost impossible, and segmentation techniques are then required.

Despite its unquestionable value, image registration methods may become extremely time expensive when there are rotation and scaling differences [59]. This is indeed the case of ADNI data, where, besides having different scanner origins, images have intra and interpatient differences.

Bearing this drawback in mind, it was hypothesised that the MIL framework could possibly handle to some extent non-registered images. To accomplish this, it was opted for a simple and straightforward way that could capture the essence of raw images, which is the inexistence of spatial correspondence between them. Hence, an aleatory patch selection procedure was implemented, with the only restriction that at least one of the patches of each patient should belong to a discriminative area. The patches belonging to these areas were selected according to a t-statistic sorting, but being randomly selected from the best ones. The remaining patches were selected randomly from the whole brain. These procedures allowed for repetition of patches amongst the patients, thus enabling some, even if low, spatial correspondence, and were implemented in each run of cross-validation for the training set and test set.

MILES was the algorithm choosen to assess the performance of this method, in the baseline data only and for patches of size 3x3x3 with a distance threshold set to 3, for CN vs. AD.

4. Results

4.1. Classification Results

Figures (3), (4) and (5) show the best results of each modality for every classifier at the baseline and 12 months, and for SVM with patches and MILES in the longitudinal analysis, obtained with 10 runs of 10-fold cross-validation. One of the first aspects that stands out is the improvement that occurs for every classifier at the 12 months, in CN vs. AD and CN vs. MCI, while MCI vs. AD seems to have a similar performance at the two time instants. Secondly, CKNN presents the worst results for CN vs. AD at the baseline, CN vs. MCI at the baseline and 12 months, and MCI vs. AD at the baseline, even if not statistically significant. Furthermore, the best balanced accuracy attained belongs to SVM with patches in the longitudinal analysis (91.21%) in CN vs. AD. Nonetheless, longitudinal analysis did not show any significant improvement regarding the 12 months, whose results are very similar.
Figure 3: Best results for each classifier (CN vs. AD).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline b,acc</th>
<th>12 months b,acc</th>
<th>Longitudinal b,acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM with best features</td>
<td>83.02% 3.41%</td>
<td>69.49% 2.03%</td>
<td>-</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>86.74% 2.96%</td>
<td>70.14% 3.66%</td>
<td>91.26% 1.47%</td>
</tr>
<tr>
<td>CRNN</td>
<td>84.83% 2.90%</td>
<td>69.50% 3.09%</td>
<td>-</td>
</tr>
<tr>
<td>DD</td>
<td>85.30% 0.92%</td>
<td>68.61% 1.71%</td>
<td>-</td>
</tr>
<tr>
<td>MILES</td>
<td>83.00% 3.08%</td>
<td>69.00% 2.52%</td>
<td>96.10% 2.13%</td>
</tr>
<tr>
<td>YARDS</td>
<td>86.05% 2.41%</td>
<td>69.23% 1.95%</td>
<td>-</td>
</tr>
<tr>
<td>BARTMP</td>
<td>87.10% 1.83%</td>
<td>69.08% 1.87%</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 4: Best results for each classifier (CN vs. MCI).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline b,acc</th>
<th>12 months b,acc</th>
<th>Longitudinal b,acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM with best features</td>
<td>59.03% 2.74%</td>
<td>61.14% 3.24%</td>
<td>-</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>61.06% 2.10%</td>
<td>65.79% 2.43%</td>
<td>64.85% 2.54%</td>
</tr>
<tr>
<td>CRNN</td>
<td>58.40% 2.60%</td>
<td>66.05% 2.62%</td>
<td>-</td>
</tr>
<tr>
<td>DD</td>
<td>50.40% 0.7%</td>
<td>64.21% 1.9%</td>
<td>-</td>
</tr>
<tr>
<td>MILES</td>
<td>61.83% 1.13%</td>
<td>63.92% 1.32%</td>
<td>63.77% 2.26%</td>
</tr>
<tr>
<td>YARDS</td>
<td>61.13% 1.09%</td>
<td>66.83% 4.13%</td>
<td>-</td>
</tr>
<tr>
<td>BARTMP</td>
<td>63.83% 0.05%</td>
<td>65.65% 0.72%</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 5: Best results for each classifier (MCI vs. AD).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Baseline b,acc</th>
<th>12 months b,acc</th>
<th>Longitudinal b,acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM with best features</td>
<td>69.20% 2.39%</td>
<td>64.81% 3.08%</td>
<td>-</td>
</tr>
<tr>
<td>SVM with patches</td>
<td>69.60% 1.36%</td>
<td>69.79% 0.44%</td>
<td>69.36% 2.57%</td>
</tr>
<tr>
<td>CRNN</td>
<td>64.5% 1.89%</td>
<td>67.87% 3.46%</td>
<td>-</td>
</tr>
<tr>
<td>DD</td>
<td>71.1% 1.27%</td>
<td>66.2% 1.32%</td>
<td>-</td>
</tr>
<tr>
<td>MILES</td>
<td>68.8% 1.57%</td>
<td>69.0% 1.52%</td>
<td>68.10% 2.50%</td>
</tr>
<tr>
<td>YARDS</td>
<td>69.9% 1.31%</td>
<td>70.21% 5.21%</td>
<td>-</td>
</tr>
<tr>
<td>BARTMP</td>
<td>71.3% 1.25%</td>
<td>70.2% 1.70%</td>
<td>-</td>
</tr>
</tbody>
</table>

ROC analysis (not shown) confirms these results, showing no significant differences between all the classifiers in each instant of time. Furthermore, there was in general an increase in the Area Under the Curve (AUC) for the 12 months. In what concerns differences between modalities, CN vs. AD obtained the highest AUC’s, followed by MCI vs. AD and finally CN vs. MCI. The best auc’s were 0.96, 0.77 and 0.67, respectively.

4.2. Simulation of non-registered images

Results for the simulation of non-registered images show that MILES can have competitive results for the CN vs. AD case, even if worse than the alternative approach taken in this thesis. In particular, amongst the various tests performed, the one with best results was obtained with 1 random discriminative patch per patient in both training and test sets (81.12%, 3.21%).

Some general trends can also be inferred. First, when the test patients do not have patches extracted from discriminative areas, the results are far worse, namely the specificity. This behaviour is in fact part of a more general tendency, which is the increase of specificity with the number of aleatory discriminative patches in test patients. This happens at the expense of a little reduction of sensitivity. On the other hand, the increase of discriminative patches in training patients increases the sensitivity, at the expense of some decrease in the specificity. These results are not shown.

5. Conclusions

The current work exploited the use of Multiple Instance Learning, a new machine learning approach, in the automatic diagnosis of AD. CAD of AD is a hot topic for the scientific community, both for its impact in an ever ageing society and for its challenging aspects. As a starting point, it was hypothesised that this paradigm could improve diagnosis accuracy, since it introduces an ambiguity in the process of classification. In fact, people suffering from Alzheimer show brain atrophy, metabolism reduction and other artifacts only in some areas of the brain, which means that “false negative” areas do exist. MIL framework seems to tackle well this ambiguity aspect.

This approach constitutes a breakthrough in the field, since the vast majority of research has been developed in the context of supervised learning. Particularly, three main domains of MIL - IS, BS and ES - in a total of 5 classifiers, were tested and compared to two simple SVM approaches with ADNI data.

In order to accomplish this, several voxel intensity patches were extracted from FDG-PET images. For every algorithm, except one - Diverse Density -, these were directly fed to the classifier. For DD, patch intensity histograms were derived. The patch selection procedure was based on group driven voxelwise t-tests, in descent order of statistic value, thus maximising the potential discrimination between clinical groups. Several combinations of patch parameters were tested. Importantly, an enhancing preprocessing step, known as Yakushev Normalisation, was applied and revealed improved discrimination accuracy, particularly for CN vs. AD and MCI vs. AD.

Simultaneously, one less obvious but very important aspect of MIL was explored, which was its applicability for non-registered images. Indeed, it soon became evident that if this paradigm allows for the use of different patches in different patients, the use of non-registered images is just a step beyond, even if ambitious. Given the short time frame, it was decided to simulate results with different patches per patient and also with randomness included, with registered images.

Regarding the results, while not outperforming their supervised counterparts, these were competitive with the literature. CN vs. AD was the task with the best balanced accuracy, as expected. MCI
vs. AD systematically outperformed CN vs. MCI results, which indicates that patients in the MCI stage are "closer" to the normal ones. Nonetheless, both modalities show that a stricter definition of MCI stage must be established, so that an early diagnosis becomes possible. One plausible way to accomplish this is the introduction of a distinction between conversor and non-conversor MCI patients, which is indeed being assessed by some authors [24, 60]. Furthermore, it should be noted that all methods seem to be robust against patch parameters change, at least under this selection procedure, which is in disagreement with results obtained by Tong [24, 37]. A possible explanation is the different data type being analysed (PET in this thesis, MRI in Tong work), which might be less prone to redundant information carried by patches. Notably, all classifiers operated competitive results with only one patch. Even if from an anatomical point of view this seems difficult to grasp, due to patch minute dimensions, this acknowledges the fact that the most discriminative patch is probably the most important. Further research should be done under other conditions, such as randomness, different number of patches per patient and different location of patches per patient. Interestingly, results with DD, despite using histograms as instances, and CKNN, with no use of SVM, obtained very similar results to the remaining three MIL classifiers and the supervised ones. This might be a hint that improvements in classification, in the future, will probably be achieved with new feature extraction and selection procedures, not with new classifiers. In fact, most of recent research has been centered in the combination of multiple biomarkers, which together may provide more information than on their own. For instance, blood biomarkers, together with MRI or PET, might be used. In addition, one should not discard the use of longitudinal data, which might complement the use of various biomarkers. Although not showing very promising results, and sometimes contradictory ones, it is indeed a field to be further explored.

Regarding ROC curves, these revealed in general an improvement at the 12 months for CN vs. AD and CN vs. MCI, while for MCI vs. AD these were less evident. Besides, supervised SVM with best features and patches obtained often the best AUC, while YARDS and MILES showed the worse ones, a fact that did not materialise in a better or worse accuracy, respectively.

In addition, BARTMIP results set forth the use of clustering methods in the context of MIL, namely applied to AD diagnosis. The impact the cluster quality (purity and entropy) has on accuracy should be further looked into. As a matter of fact, purity and entropy variation with the number of centroids did not show any tangible effect in accuracy inside each classification task. Nonetheless, the parity between accuracy and each classification task accuracy is undeniable.

Diverse Density, even if with competitive results, is not viable due to its fastidious classification process, which requires several weeks for even few patches. EM-DD should demonstrate to be more efficient.

MILES and YARDS, both simple in their theoretical background, have the additional appeal that they are able to label training instances, even if with some false positives and (possibly) false negatives. This aspect should be further investigated and classifiers other than SVM should be tested in the context of AD. Also, the possibility of instance classification in test bags should be analysed in the future.

Concerning the simulation of non-registered images, results should be interpreted with caution, but there seems to be some applicability, even if with less accuracy than with the main approach of this thesis. This, however, is natural, since randomness was introduced, rather than a t statistic driven selection. The number of patches chosen at random from discriminant regions seems to play an important role. Future work should try new combinations of patch parameters and test this approach in non registered images. If successful, this would be a great achievement, since registration is a fastidious process. Also, it would be interesting to test a medical driven approach with registered images and aleatory patches extracted from patients under some restrictions given by the doctor.

Acknowledgements

This work was supported by Fundaç~ao para a Ciência e Tecnologia (FCT/MCTES) through the ADIAR project (PTDC/SAU-ENB/114606/2009). Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

References


[50] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification.


