Identification of Alzheimer’s Disease Using Adaboost Classifier

M. Rangini and Dr. G. Wiselin Jiji

Abstract--- Alzheimer's disease majorly affect the hippocampus region of the brain. Our first step is to segment the hippocampus region from the brain. Level set approach is used for segmenting the hippocampus region. Then after segmentation, a feature of the segmented region is calculated. Features such as intensity, gradients, curvatures, 1-D, 2-D, and 3-D Haar filters, mean filters, and standard deviation filters are estimated. Classification is done using hierarchical AdaBoost and support vector machines (SVM) classifier. Our AdaBoost and Ada-SVM segmentations compared favorably with the manual segmentations. We also evaluated how segmentation accuracy depended on the size of the training set, providing practical information for future users of this technique.

Keywords--- AdaBoost, Alzheimer's Disease, Hippocampal Segmentation, Support Vector Machines (SVMs).

I. INTRODUCTION

HIPPOCAMPAL segmentation is a key step in many medical imaging studies for statistical comparison of anatomy across populations, and for tracking group differences or changes over time. Specifically in Alzheimer’s disease, hippocampal volume and shape measures are commonly used to examine the 3-D profile of early degeneration, and detect factors that predict imminent conversion to dementia. The hippocampus is the key structure in the development of Alzheimer’s disease.

In AD, the entorhinal cortex, the CA1 field, and the subicular region are all heavily affected by cell loss, plaques, and tangles. Hippocampal volumes may vary with brain or intracranial size, and both brain and hippocampal size may be affected by acquired disease.

Alzheimer’s disease (AD) is the most common cause of dementia; its’ early and accurate diagnosis is challenging. The hippocampus is a gray matter structure of the temporal lobe known to be affected at the earliest stage of AD, even before the diagnosis can be made, at the stage of mild cognitive impairment (MCI). Hippocampal volumetry on magnetic resonance images (MRI) can thus constitute a useful diagnostic tool. Till now, hippocampal volumetry mostly relies on highly time-consuming manual segmentation, which is rater dependent, and not feasible in clinical routine.

Alzheimer's disease (AD) is the most common type of dementia, and affects over 5 million people in the United States alone. The disease is associated with the pathological accumulation of amyloid plaques and neurofibrillary tangles in the brain, and first affects memory systems, progressing to involve language, affect, executive function, and all aspects of behavior. A major therapeutic goal is to assess whether treatments delay or resist disease progression in the brain before widespread cortical and subcortical damage occurs. For this, sensitive neuroimaging measures have been sought to quantify structural changes in the brain in early AD which are automated enough to permit large-scale studies of disease and the factors that affect it.

There is no single clinical test that can be used to identify Alzheimer's disease. A comprehensive patient evaluation includes a complete health history, physical examination, neurological and mental status assessments, and other tests, including analysis of blood and urine, electrocardiogram, and an imaging exam, such as CT or MRI. While this type of evaluation may provide a diagnosis of possible or probable Alzheimer's disease with up to 90 percent accuracy, absolute confirmation requires examination of brain tissue at autopsy. Early and careful evaluation is important because many conditions can cause dementia, some of which are treatable or reversible. Potentially reversible conditions include depression, adverse drug reactions, metabolic changes, and nutritional deficiencies. The earlier an accurate diagnosis of Alzheimer's disease is made, the greater the gain in managing symptoms and allowing the person to take part in planning for their future. People with Alzheimer's disease often live for years with the disease. The duration of Alzheimer's disease from time of diagnosis can be 20 years or more. The average length of time from onset of symptoms is thought to be in the range of 4 to 8 years. No treatment is yet available that can stop Alzheimer's disease. However, the drugs donepezil (Aricept), rivastigmine (Exelon), or galantamine (Reminyl) may help delay the progression of symptoms associated with Alzheimer's disease. Also, some medicines may help control behavioral symptoms, such as sleeplessness, agitation, wandering, anxiety, and depression. Treating these behavioral symptoms often makes people with Alzheimer's more comfortable and makes their care easier.

We developments a machine learning algorithm, such as AdaBoost, have automated the feature selection process for several imaging applications. SVMs can effectively combine features for classification. AdaBoost and SVM may be used to
classify vector-valued examples, and both have been separately applied to medical image analysis before, but this paper evaluates the benefits of combining them sequentially. In our paper, we note that this paper classifies voxels in a brain MRI scan as belonging to the hippocampus versus not, but in a second step we use these classified structures to create statistical maps of systematic differences in anatomy between Alzheimer’s patients and controls. As such, although the main goal of the paper is to achieve segmentations of the hippocampus, we illustrate the use of the thesese segmentations in an application where differences between disease and normality are detected and mapped.

Among several algorithms proposed for statistical classification, AdaBoost is a meta-algorithm that sequentially selects weak classifiers (i.e., ones that do not perform perfectly when used on their own) from a candidate pool and weights each of them based on their error. A weak learner is any statistical classifier that performs better than pure chance. Each iteration of AdaBoost assigns an “importance weight” to each example; examples with a higher weight, classified incorrectly on previous iterations, will receive more attention on subsequent iterations, tuning the weak learners to the difficult examples. Testing examples with AdaBoost is therefore simply a weighted vote of the weak-learners.

SVMs, on the other hand, seek a hypersurface in the space of all features that both minimizes the error of training examples and maximizes the margin, defined as the distance between the hypersurface and the closest value in feature space, in the training data. SVMs can use any type of hypersurface by making use of the “kernel trick”.

Although SVMs have been widely used in medical imaging, AdaBoost has not. However, as AdaBoost can select informative features from a potentially very large feature pool, it is likely to offer advantages in automatically finding good features for classification. This can greatly reduce, or eliminate the need for experts to choose informative features based on knowledge of every classification problem. Instead, one just needs to define a list of possibly informative features, and AdaBoost will choose those that are actually informative.

For our classification problem, we compared three different classification techniques: 1) SVM with manually selected features (manual SVM), 2) AdaBoost, and 3) SVM with features automatically selected by AdaBoost (Ada-SVM). As AdaBoost can select features automatically, we improved the classification ability of AdaBoost and Ada-SVM by implementing them in a hierarchical decision tree framework.

As a testbed to examine segmentation performance, we trained and tested our methods on a dataset of 70 3-D volumetric T1-weighted brain MRI scans. 30 of these subjects were reserved for training, and 40 for testing. The training subjects were composed of 10 subjects with Alzheimer’s disease (AD), 10 with mild cognitive impairment (MCI), a state which carries an increased risk for conversion to AD, and 10 age-matched controls. The 40 testing subjects were composed of 20 AD and 20 controls. Due to the small number of MCI subjects available for this study, we chose to add them to the training group because it increased the variability on which to train.

Here we use the segmentation method based on level set approach. Then we classify based on the above result, we use AdaBoost inside of a new classification scheme which incorporates context information. We wish to show how AdaBoost is less effective than a combination of AdaBoost and SVM. Thus here we are focusing on the learner (AdaBoost versus Ada-SVM) and incorporates contextual information into the classification problem. Here we use a SVM classifier inside the Ada-boost. Thus our Ada-SVM works by the way that the features are selected by the Ada-boost and that features are incorporated into the SVM classifier for further classification.

II. METHODOLOGY

2.1 Subjects

We trained and tested our methods on a dataset of 70 3-D volumetric T1-weighted brain MRI scans. 30 of these subjects were reserved for training, and 40 for testing. The training subjects were composed of 10 subjects with Alzheimer’s disease (AD), 10 with mild cognitive impairment (MCI), a state which carries an increased risk for conversion to AD, and 10 age-matched controls. The 40 testing subjects were composed of 20 AD and 20 controls.

2.2 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI), nuclear magnetic resonance imaging (NMRI), or magnetic resonance tomography (MRT) is a medical imaging technique used in radiology to visualize internal structures of the body in detail. MRI makes use of the property of nuclear magnetic resonance (NMR) to image nuclei of atoms inside the body. An MRI scanner is a device in which the patient lies within a large, powerful magnet where the magnetic field is used to align the magnetization of some atomic nuclei in the body, and radio frequency fields to systematically alter the alignment of this magnetization. This causes the nuclei to produce a rotating magnetic field detectable by the scanner and this information is recorded to construct an image of the scanned area of the body. Magnetic field gradients cause nuclei at different locations to rotate at different speeds. By using gradients in different directions 2D images or 3D volumes can be obtained in any arbitrary orientation.

2.3 Level Set Formulation

The segmentation problem is expressed as the computation of a 3D surface $S(t)$ (or front) propagating in time along its normal direction. In the level set formulation the propagating front $S(t)$ is embedded as the zero level of a time-varying higher dimensional function $C(X,t)$,

$$S(t)=\{X \in \mathbb{R}^3 \mid \phi(X,t) = 0\} \tag{1}$$

The function $C$ describes a 4D surface defined by $C(X,t)d$, where $d$ is traditionally the signed distance from $X$ to the front $S$ (negative inside the object). The evolution rule for $\phi$ can be expressed as
\[
\frac{\partial V}{\partial t} + F|\Delta W| = 0 \quad (2)
\]

where \( F \) is a scalar velocity function depending on local properties of the front (like the local curvature), on external parameters related to the input data (input gradient for instance), and possibly on additional constant propagation terms.

2.4 SVM

SVMs are very popular for discrimination tasks because they can accurately combine many features to find an optimal separating hyperplane. SVMs minimize the classification error based on two constraints simultaneously. They both seek a hyperplane with a large margin, i.e., the distance from the closest example to the separating hyperplane and minimize the number of wrongly classified training examples, using slack variables. If an example is perfectly classifiable in feature space then the second constraint is not necessary. However, this is not the case in our problem, so SVMs both minimize the error on the training set and maximize the margin, increasing their generalization ability. A transformed attribute that is used to define the hyperplane is called a feature. The task of choosing the most suitable representation is known as feature selection. A set of features that describes one case (i.e., a row of predictor values) is called a vector. So the goal of SVM modeling is to find the optimal hyperplane that separates clusters of vector in such a way that cases with one category of the target variable are on one side of the plane and cases with the other category are on the other side of the plane. The vectors near the hyperplane are the support vectors.

SVM formulation

\[
\text{Min } \frac{1}{2} \| \alpha \| + C \sum_i z_i \quad (3)
\]

Subject to \( y_i (\alpha \cdot x - b) \geq 1 - z_i \)

where \( \alpha \) is the vector corresponding to the separating hyperplane, \( \frac{1}{2} \| \alpha \| \) is the margin of the hyperplane, according to the \( l_2 \)-norm, \( x \) is a vector consisting of the features, \( b \) is a scalar bias term (so the hyperplane is not forced to go through the zero point), \( z_i \) are slack variables (those classified on the wrong side of the margin of the separating hyperplane), and \( C \) is a user defined parameter controlling the tradeoff between margin and the number of slack variables. An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized. In the figure shown below, the line in the right panel is superior to the line left panel. There are many hyperplanes that might classify the data.

Figure 1: A Hyperplanes are drawn between Two Classes

\[
f(x) = \frac{z}{\sum_i |h(x)|} \quad (8)
\]

We can then see that AdaBoost explicitly minimizes the error, and implicitly maximizes the margin according to the \( l_1 \)-norm.

Figure 2: Identification of Support Vectors

To minimize (1), one can formulate the problem in its dual form (2) and maximize that problem

\[
\text{Max } (\sum_i \alpha_i \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j)
\]

Subject to \( \alpha_i \geq 0 \quad (4) \)

\[
\text{Class}(x) = \sum_i \alpha_i y_i x_i^T x + b \quad (5)
\]

Once formulated in its dual form, quadratic programming is used to find the best \( \alpha_i \) and \( b \) from (2). This formulation allows the introduction of the “kernel trick” and extends the classification ability of SVMs from variety of hyper surfaces in feature space. SVMs may be viewed as an approach to find \( \overrightarrow{w} \) and \( b \) that maximize \( P(y = + 1|y, w, b) \).

\[
P(y = + 1|\overrightarrow{w}, b) = \frac{P(w, b|y = + 1)P(y = + 1)}{P(w, b)} \quad (7)
\]

The denominator is constant, and a shape model is needed to \( P(y = + 1) \) term. Expressed in this form, SVMs may be seen as approximating the posterior distribution using a given set of features to define \( \overrightarrow{w} \) and \( b \).

2.5 AdaBoost

AdaBoost combines a set of weak learners in order to form a strong classifier in a “greedy fashion,” i.e., it always chooses the weak classifier with the lowest error, ignoring all others.

We use a decision stump as a weak learner. A decision stump, based on a given feature, classifies all examples less than a threshold as belonging to one class and greater than a threshold as another class. Formally, a decision stump consists of a feature on which the decision will be made, a separating threshold, and a Boolean saying whether positive examples are less than or greater than the threshold. A decision stump is advantageous over other weak learners because it can be calculated very quickly and there is a one-to-one correspondence between a weak learner and a feature when a decision stump is used.

AdaBoost explicitly seeks to minimize the error according to a distribution of weights, \( D_t \) at each iteration. However, if we follow the logic of and view \( \{ \alpha_i \}_{i=1}^{T+1} \) as a vector of coordinates, \( \overrightarrow{x} \), then we can rewrite \( f(x) \) as

\[
f(x) = \frac{z}{\sum_i |h(x)|} \quad (8)
\]
AdaBoost greedily selects features, it can take a complicated problem, one composed of many features, and create a sparse classification rule, one composed of only a few features. However, this is also a drawback. Due to the greedy nature of AdaBoost it can only minimize the error, and maximize the margin with respect to features that have already been selected. AdaBoost is also limited by the fact that it can only combine weak learners by adding them together.

AdaBoost approximates the Bayesian posterior distribution by incrementally adding new weak learners \((h_i(x))\) at each iteration. This is equivalent to formulating the overall classifier at time \(t\) as \(H(x) = \text{sign}[P(y = 1|\sum_{i=1}^{t} h_i(x))…h_t(x)] > 0.5\). If we let \(h_t(x)…h_1(x) = h\), we can formulate the posterior distribution as

\[
P(y = 1|h) = \frac{P(h_1|y = 1) P(y = 1)}{P(h_t)}
\]

(9)

The denominator is again a constant and \(P(y = 1)\) is a shape model which must be integrated later. In this formulation, AdaBoost also approximates the ideal Bayesian distribution after a long enough \(t\) drawing features from a very large feature pool. We could stop here and just apply an ideal Bayesian classifier to the features selected by AdaBoost.

### 2.6 Ada-SVM

SVMs globally and explicitly maximize the margin while minimizing the number of wrongly classified examples, using any desired linear or nonlinear hypersurface. This is both an advantage and a disadvantage. The advantage is that SVMs take into account each example in the entire feature space when creating the separating hypersurface. The disadvantage is that this makes them computationally intractable as the number of features becomes large.

Since AdaBoost greedily selects features, it does not have to use a large storage space to store all the features. But, since it is a greedy feature selector, given the same set of features, we expect SVM to outperform AdaBoost. We exploit this fact to design our Ada-SVM classifier. We use AdaBoost to select the features that most accurately span the classification problem, and SVMs to fuse those features together to form the final classifier.

To make AdaBoost directly compatible with SVM, one small adjustment must be made to the AdaBoost algorithm. Traditionally, AdaBoost may choose features more than once when constructing weak learners; however, having the same feature appear twice in an SVM formulation does not make sense. To overcome this, when choosing features with AdaBoost for Ada-SVM, features are chosen without replacement. In all experiments involving just AdaBoost, however, traditional AdaBoost is implemented. We implicitly take into account the Bayesian prior (shape in-formation) necessary in both models by creating a shape prior based on the LogOdds formulation. We create a signed distance map for each training subject, with negative values inside the ROI and positive values outside the ROI and then transform each of those values into the interval \((0, 1)\) by using (8), where \(I(x)\) is the intensity of voxel \(x\).

After getting a signed distance map transformed into the interval \((0, 1)\) for each subject, we then perform a voxel-by-voxel averaging in order to create one prior image that we store for both training and testing. We note that this map contains statistical information on the likely position of the target structure in the coordinate space to which all images have been aligned.

### III. Experiments

Here we have registered all brain images into the same stereotaxic space. Each subject’s brain MRI was co-registered with scaling (nine-parameter transformation) to the ICBM53 average brain template. This also allows us to define a bounding box around the training hippocampi plus some neighborhood voxels. These neighborhood voxels might contain hippocampal voxels outside the bounding box of the training set and are also necessary for computing neighborhood based features. Any voxels outside of this bounding box are definitely not hippocampus, and can therefore be bounding box is a rectangular region with corners at \((-48,-54,-44)\) and \((-1,5,17)\) for the left hippocampus and a corresponding region in the opposite hemisphere for the right hippocampus in the standard ICBM53 space. Next, we have to define our pool of candidate features from which AdaBoost will select. Our feature pool consists of information from three different image “channels”:

1) the T1-weighted image, 2) tissue classification maps of gray matter, white matter, and CSF, and 3) our Bayesian shape prior (8). From each one of these images, the following features are computed: intensity, gradients, curvatures, 1-D, 2-D, and 3-D Haar filters, mean filters, and standard deviation filters, all computed using a neighborhood kernel of size \(7\times7\times7\). Because of the large number of examples and features, we use randomization to decrease these numbers to a computationally tractable size. During each run of AdaBoost, a new set of 200,000 examples and 2500 features is randomly chosen to learn the classification rule (for either AdaBoost or Ada-SVM). These numbers were determined empirically to give optimal results.

As a final step, after segmentations are computed by either AdaBoost, Ada-SVM, or manual SVM, the binary masks are convolved with a \(3\times3\times3\) averaging kernel. Partial volume effects are removed from the resulting mask by setting voxels with a value of less than 0.5 (those with fewer than 13 neighbors) to 0 and greater than 0.5 (those with more than 13 neighbors) to 1. This is done to smooth the boundary and fill any holes.
3.1 Disease Detection

In addition to segmentation accuracy, it is also important to assess how effectively each method can differentiate disease from normal. For instance, in a study aiming to map disease effects, increases in segmentation accuracy are beneficial if they provide additional power to differentiate groups. As the effect of AD on the brain is not uniform, such studies commonly rely on mapping of group differences to identify regions that are especially susceptible to early changes, or where changes predict decline or help differentiate one type of dementia from another. We note that in reporting classification accuracy and detection of disease effects on hippocampal anatomy in groups of subjects, both of these metrics evaluate desirable characteristics of a tissue segmentation approach, but they are not necessarily causally related or even correlated. That is, a method that produces relatively better segmentation is not necessarily more discriminative and vice versa, and it is misleading to suggest that one implies the other. From a logical standpoint, there could be a bad segmentation algorithm that exaggerates the difference between AD and controls, for example, and this could be a very good discriminator. In general, this depends on whether the voxels that are misclassified by a segmentation approach are also relevant for disease classification. Ada-SVM gives a more consistent segmentation for both normal and AD subjects (the distance metrics are too prone to outliers).

IV. Conclusion and Future Work

For most of the experiments shown in this paper, both AdaBoost and Ada-SVM outperform FreeSurfer. An important subject of future study is the generalizability of the methods proposed here as a function of the MRI acquisition parameters. Even though this is interesting, this method is already useful in large scale studies for which the scanning parameters remain stable where one could segment 20 brains manually for training purposes, and then segment all the rest automatically.

In the future, we will apply both of these techniques to new datasets to examine different diseases and to rank segmentation methods for power and accuracy. It will be interesting to note if Ada-SVM more powerfully detects disease effects or segments other subcortical structures better than AdaBoost does. Although the ability to map disease effects automatically is encouraging and likely to benefit many ongoing studies, one caveat is necessary regarding the use of p-value plots to compare the effect sizes of different methods. These plots provide a clear comparison of the distribution of effect sizes in a statistical map when methodological parameters are varied, strictly speaking, many repeated large and independent samples would be required to prove that one cumulative p-value distribution differs from another on the interval. Without confirmation on multiple.

---

### Table I: These are the First 10 Features selected by Ada Boost during Ada-SVM

<table>
<thead>
<tr>
<th>Channel</th>
<th>Name</th>
<th>Neighborhood</th>
<th>Channel</th>
<th>Name</th>
<th>Neighborhood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior Image</td>
<td>Mean Filter</td>
<td>3,6,3 (3D)</td>
<td>Prior Image</td>
<td>Mean Filter</td>
<td>6,7,3 (3D)</td>
</tr>
<tr>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>6,3,3 (3D)</td>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>7,7,3 (3D)</td>
</tr>
<tr>
<td>Tissue Classification</td>
<td>Mean Filter</td>
<td>6,5,1 (3D)</td>
<td>TI-weighted Image</td>
<td>Haar Filter</td>
<td>7,7,3 (3D)</td>
</tr>
<tr>
<td>TI-weighted Image</td>
<td>Haar Filter</td>
<td>3,1,1 (3D)</td>
<td>Prior Image</td>
<td>Standard Deviation Filter</td>
<td>7,6,6</td>
</tr>
<tr>
<td>Prior Image</td>
<td>Standard Deviation Filter</td>
<td>7,6,6</td>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>5,4,5 (3D)</td>
</tr>
<tr>
<td>Tissue Classification</td>
<td>Intensity</td>
<td>n.a.</td>
<td>TI-weighted Image</td>
<td>Intensity</td>
<td>5,4,5 (3D)</td>
</tr>
<tr>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>1,3,4 (3D)</td>
<td>Prior Image</td>
<td>Mean Filter</td>
<td>6,5,1 (3D)</td>
</tr>
<tr>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>4,5,7 (3D)</td>
<td>Prior Image</td>
<td>Gradient Filter</td>
<td>5,7,2 (y)</td>
</tr>
<tr>
<td>TI-weighted Image</td>
<td>Intensity</td>
<td>n.a.</td>
<td>TI-weighted Image</td>
<td>Haar Filter</td>
<td>3,1,7 (3D)</td>
</tr>
<tr>
<td>Tissue Classification</td>
<td>Haar Filter</td>
<td>3,4,5 (3D)</td>
<td>Prior Image</td>
<td>Haar Filter</td>
<td>5,3,1 (2D)</td>
</tr>
</tbody>
</table>
samples, it may not reflect a reproducible difference between methods. FDR and its variants declare that a CDF shows evidence of a signal if it rises more than 20 times more sharply than a null distribution, so a related criterion could be developed to compare two empirical mean CDFs after multiple experiments. As simple numeric summaries sacrifice much of the power of maps, and provide a rather limited view of the differences in sensitivity among voxel-based mapping methods, additional work on CDF-based comparisons of methods seems warranted.

V. REFERENCES


