MULTIRESOLUTION WAVELET ANALYSIS AND ENSEMBLE OF CLASSIFIERS FOR EARLY DIAGNOSIS OF ALZHEIMER'S DISEASE

Genevieve Jacques[1], Jennifer L. Frymiare[2], John Kounios[2], Christopher Clark[3], and Robi Polikar[1]⋆

[1]Dept. of Electrical and Computer Engineering, Rowan University, Glassboro, New Jersey, USA
[2]Dept. of Psychology, Drexel University, Philadelphia, Pennsylvania, USA
[3]Dept. of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
⋆Corresponding author: polikar@rowan.edu

ABSTRACT

The diagnosis of Alzheimer's disease at an early stage is a major concern due to growing number of the elderly population affected, as well as the lack of a standard and effective diagnosis procedure available to community healthcare providers. Recent studies have used wavelets and other signal processing methods to analyze EEG signals in an attempt to find a non-invasive biomarker for Alzheimer's disease and had varying degrees of success. These studies have traditionally used automated classifiers such as neural networks; however the use of an ensemble of classifiers has not been previously explored and may prove to be beneficial. In this study, multiresolution wavelet analysis is performed on event related potentials of the EEG which are then used with the ensemble of classifiers based Learn++ algorithm. We describe the approach, and present our promising preliminary results.

1. INTRODUCTION

An estimated 4.5 million Americans were suffering from Alzheimer's disease (AD) as of 2000 and this number is expected to reach between 11.3 and 16 million by 2050, making it a major public health concern [1]. A further concern is the fact that an autopsy is the only tool that provides a definitive diagnosis. Clinical evaluation, the standard AD diagnostic procedure conducted at major university hospitals and research clinics, on average achieves positive predictive value of 93%, however, most patients are evaluated at community health clinics, where the expertise and accuracy of disease specific dementia remains uncertain. In fact, recently a group of Health Maintenance Organization-based physicians reported an average sensitivity of 83%, specificity of 55% and an overall accuracy of 75%, despite the benefit of a longitudinal study [2].

An effective and objective tool for early diagnosis of the disease is of course important, but to have a meaningful impact on healthcare the procedure must also be inexpensive, non-invasive and available to community physicians, who provide the first line of intervention. Several biomarkers have been linked to AD, such as the cerebrospinal fluid tau (CSF-τ) and β-amyloid, however, they have not proven to be conclusive. There is, therefore, significant need for a clinically useful, accurate, non-invasive, cost-effective and automated procedure for early diagnosis of AD. The electroencephalogram (EEG) may potentially satisfy these needs as a tool for AD diagnosis.

An EEG based technique, called the oddball paradigm, that involves the analysis of scalp recordings of auditory event related potentials (ERP) has been shown to be beneficial in detecting the changes due to mental impairment. The paradigm involves the completion of a simple task (e.g. pressing a button) by the subject when s/he hears an infrequently occurring 2 kHz (oddball) tone, presented randomly within a series of regularly occurring 1 kHz (regular) tones. In response to the oddball stimulus, the ERPs indicate a positive peak (P3 or P300) with an approximate latency of 300 ms after the stimulus. Changes in the amplitude and latency of P300 are known to be altered by neurological disorders affecting the temporal-parietal regions of the brain [3]. This includes AD where the P300 latency is prolonged and the amplitude is decreased compared to controls [4, 5]. Time-frequency analysis techniques, such as wavelets, have been used in different studies to decompose an ERP in the time-frequency plane in order to characterize its functionally relevant components [6]. However, the use of wavelets as a feature extraction tool, followed by an automated ensemble based classifier has not been explored.

The method described in this paper combines multiresolution wavelet analysis (MWA), automated classification techniques using an ensemble of classifiers approach, along with established electroencephalography (EEG) analysis, to detect the earliest stage of AD. Our goal is to analyze the ERP using MWA followed by the ensemble of classifiers based Learn++ algorithm. Our expectation in doing so is to combine wavelet transform's ability to extract features – other then just the amplitude and latency of P300 – with the superior classification and robustness of ensemble of classifiers in automated early diagnosis of the AD. Two types of wavelets along with the Learn++ algorithm have been evaluated to date on a database of the first 28 of the 80 patients planed for this study. We are interested in determining the sensitivity, specificity, and positive predictive value of this approach in distinguishing between elderly individuals with AD and cognitively normal subjects.

2. METHODOLOGY

2.1 Test Subjects and Clinical Evaluation

This study will include a total of 80 subjects, 50 of whom will be used to train the automated classification system, and the remaining 30 will be used to evaluate and validate the system's performance on previously unseen signals. Half of these subjects constitute the cognitively normal cohort, whereas the other half constitutes the AD cohort. Twenty-eight subjects, satisfying the inclusion criteria, 10 diagnosed with probable

AD and 18 cognitively normal controls, have been recruited thus far in the initial phase of this study. All subjects, control or with probable AD, are recruited from elderly patients, over 60 years of age. The subjects were verified to be free of any evidence of central nervous system neurological disease (e.g. stroke, multiple sclerosis, Parkinson's disease, etc.) by history or by exam. The use of sedative, anxiolytic or anti-depressant medications was not permitted within 48 hours of data collection. The two groups were defined by the following criteria: Cognitively normal subjects: (i) Clinical Dementia Rating (CDR)=0; (ii) Mini-Mental Scores (MMS)≥24; (iii) no indication of functional or cognitive decline during the two years prior to enrollment based on a detailed interview with the subject's knowledgeable informant or two previous annual clinical assessments. AD subjects: (i) CDR≥0.50; (ii) MMS<24; (iii) presence of functional cognitive decline over the previous 12 months based on detailed interview with a knowledgeable informant; (iv) satisfaction of NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association) criteria for probable AD [7]. All subjects received a through medical history and neurological exam at the University of Pennsylvania's Memory Disorders Clinic, in Philadelphia. Key demographic and medical items, including current medications (prescription, over-the-counter, and complementary alternative medications) are entered into their case binder and data-base. The evaluation includes standardized assessments for overall impairment, functional impairment, extrapyramidal signs, behavioral changes and depression. The clinical evaluation results constituted the gold-standard against which the proposed automated system has been compared.

2.2 Acquisition of Event Related Potentials

The ERPs were obtained using an auditory oddball paradigm while the subjects were comfortably seated in a specially designated room. The protocol originally described in [3], with slight modifications, was used in this study. Binaural audiometric thresholds were determined for each subject using a 1000 Hz tone. The evoked response stimulus was presented to both ears using stereo speakers with an amplitude level comfortable for their hearing level. The stimulus consisted of tone bursts 100 ms in duration, including 5 ms inset and offset envelopes. Tones of 1000 and 2000 Hz were presented in a random sequence with the tones occurring in 65% and 20% of the trials respectively. The remaining 15% of the trials consisted of novel sounds presented randomly. These included 60 unique environmental sounds that were recorded digitally and edited to 200 ms duration. A total of 1000 stimuli, including frequent 1000 Hz (n=650), infrequent 2000 Hz tones (n=200) and novel sounds (n=150) were delivered to each subject with an interstimulus interval of 1.0-1.3 seconds. The subjects were instructed to press a button each time they heard the 2000 Hz tone. With frequent breaks (e.g. three minutes of rest every five minutes), the data collection process lasted about 30 minutes per subject with each session proceeded by a 1 minute practice session without the novel sounds. The ERPs were recorded from 19 tin electrodes embedded in a plastic cap, using linked mastoids as reference. The electrode impedances were kept below 20kΩ to yield a good signal with the high-impedance MANSCAN amplifier system used in the study. Artifactual recordings were identified and rejected by the EEG technician. The remaining scalp potentials were amplified, digitized at 256 Hz/channel (19 channels) and stored. The ERPs that were validated by the EEG technician were preprocessed using appropriate low-pass filtering techniques and averaging.

The averaging protocol involved averaging 90~255 recordings per patient. All averages have been notched filtered at 59-61 Hz and baselined with the prestimulus interval.

2.3. Multiresolution Wavelet Analysis

Multiresolution wavelet analysis provides time localizations of spectral components in the signal thus giving a time-frequency representation. Among many time-frequency representations, the discrete wavelet transform (DWT), has become increasingly popular due to its ability to solve a diverse set of problems. For brevity, we only describe the main points here and refer the interested readers to many excellent references listed at [8]. The DWT analyzes the signal at different frequency bands with different resolutions through the decomposition of the signal (hence, multiresolution analysis). The DWT utilizes two sets of functions: scaling functions and wavelet functions, each associated with lowpass and highpass filters, respectively. Decomposition of the signal into different frequency bands is accomplished by successive highpass and lowpass filtering of the signal. The original time domain signal x(t) sampled at 256 samples/ sec. created the discrete time signal, x[n], which is passed through a halfband highpass filter, g[n], and a lowpass filter, h[n]. In terms of angular frequency, the highest frequency in the original signal is π, corresponding to the linear frequency of 128 Hz. According to Nyquist's rule, half the samples can be removed after the filtering, since the highest frequency in the signal is now π/2 radians. Therefore every other sample in the signal can be discarded. One level of decomposition is therefore given as:

$$y_{high}[k] = \sum_n x[n] \cdot g[2k - n] \quad (1)$$

$$y_{low}[k] = \sum_n x[n] \cdot h[2k - n] \quad (2)$$

where $y_{high}[k]$ and $y_{low}[k]$ are the outputs of the filters after downsampling by two. This procedure, called subband coding, is repeated for further decomposition. At each level, the filtering procedure results in half the time resolution and double the frequency resolution. Figure 1 illustrates this procedure, where $x[n]$ is the original signal to be decomposed, ↓2 indicates the downsampling operation, and H and G indicate the lowpass and highpass filters, respectively. The bandwidth of the signal at every level is marked on the figure as "B".

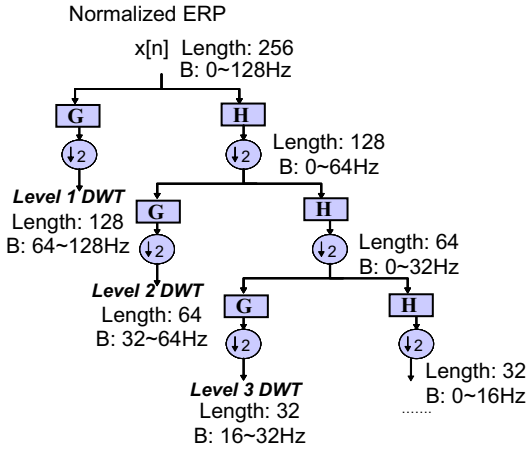

Figure 1. DWT subband coding algorithm

2.4. Automated Classification using Learn++

In automated classification applications, customarily, the distinctive features of the signals are first identified, which are then used to train the classification algorithm. The trained system is then evaluated on validation data, not seen during the training. In this study, the DWT coefficients corresponding to 0 ~ 16 Hz bandwidth were used to train the Learn++ algorithm.

Learn++, previously developed by the corresponding author [9], has two key characteristics. First, it is an ensemble of classifiers based approach, where each classifier is deliberately made relatively weak. The premise is to create diversity among the classifiers, where each classifier learns a slightly different portion of the feature space, and hence make different errors on the classification of any given instance. A suitable combination of these classifiers can then reduce the total error, conceptually similar to a lowpass filtering effect. This ensemble approach has two main advantages: not only it often provides superior classification performance over single classifier systems, but since the individual classifiers are relatively weak, it also avoids the time-costly fine-tuning of the decision boundary which may also cause overfitting.

A second key characteristic of Learn++ is its incremental learning ability from new data, as such data become available. This property will be useful later as we recruit additional patients as it will allow us to continue training without starting from scratch.

Learn++, which has its roots in the popular ensemble of classifiers algorithm AdaBoost[10], essentially generates a number of relatively weak classifiers, whose outputs are combined through a weighted majority voting scheme to obtain the final classification [9]. The classifiers (hypotheses) are generated using strategically chosen subsets of the available database. A dynamically updated distribution over the training data instances is computed such that the distribution is biased towards those instances that have not been adequately learned by the previous hypotheses. The pseudocode of the Learn++ algorithm is provided in Figure 2, where the inputs to the algorithm are: (i) a sequence of m training data instances x_i with their correct labels y_i, (ii) a supervised classification algorithm to be used as a **BaseClassifier**, and (iii) an integer T that specifies the number of classifiers (hypotheses) to be generated. Any supervised neural network whose parameters are chosen appropriately can serve as a weak classifier (such as a small multilayer perceptron, MLP, with a relatively high error goal).

For t^{th} classifier to be generated at t^{th} iteration, Learn++ starts by normalizing a weight distribution, D_t according to which training subset TR_t and test subset TE_t are drawn. This distribution is initially set to be uniform, giving equal probability to each instance to be selected into the first training subset.

At each iteration t, the weights w_t from previous iteration are normalized to give a legitimate distribution D_t (step 1). Training and test subsets are then selected according to D_t (step 2), and the weak classifier is trained with the training subset (step 3). A hypothesis h_t is obtained as the t^{th} classifier (step 4), whose error is computed on $TR_t + TE_t$ by adding the distribution weights of the misclassified instances

$$\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i). \tag{3}$$

The BaseClassifier is expected to produce at least 50% correct classification on its own training data to ensure that a meaningful performance can be obtained from each classifier.

Algorithm Learn++

Input:

- Sequence of m examples $S=[(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)]$.
- Weak learning algorithm **BaseClassifier**.
- Integer T, specifying the number of iterations.

Do for each $k=1,2,\ldots,K$:

Initialize $w_1(i) = D_1(i) = 1/m,\ \forall i,\ i = 1,2,\cdots,m$

Do for $t = 1,2,\ldots,T_k$:

1. Set $D_t = \mathbf{w}_t \Big/ \sum_{i=1}^{m} w_t(i)$ so that D_t is a distribution.
2. Draw training TR_t and testing TE_t subsets from D_t.
3. Call **BaseClassifier** to be trained with TR_t.
4. Obtain hypothesis $h_t : X \rightarrow Y$, and calculate its error $\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(i)$ on S=$TR_t + TE_t$.

   If $\varepsilon_t > ½$, discard h_t and go to step 2. Otherwise, compute normalized error as $\beta_t=\varepsilon_t / (1-\varepsilon_t)$.
5. Call weighted majority voting and obtain the composite hypothesis
$$H_t = \arg\max_{y\in Y} \sum_{t:h_t(x)=y} \log(1/\beta_t)$$
6. Compute the error of the composite hypothesis
$$E_t = \sum_{i:H_t(x_i)\neq y_i} D_t(i) = \sum_{i=1}^{m} D_t(i)\left[\left| H_t(x_i) \neq y_i \right|\right]$$
7. Set $B_t = E_t/(1-E_t)$, and update the weights:
$$w_{t+1}(i) = w_t(i) \times \begin{cases} B_t, & if\ \ H_t(x_i) = y_i \\ 1, & otherwise \end{cases}$$
$$= w_t(i) \times B_t^{1-\left[\left|H_t(x_i)\neq y_i\right|\right]}$$

Call Weighted majority voting and

Output the final hypothesis:

$$H_{final}(x) = \arg\max_{y\in Y} \sum_{t:h_t(x)=y} \log\frac{1}{\beta_t}$$

Figure 2. Learn++ algorithm pseudocode

This condition is enforced by requiring that the error computed in (3) be less than ½. If this is the case, the error is normalized to [0 1] interval

$$\beta_t = \varepsilon_t /(1-\varepsilon_t), \quad 0 < \beta_t < 1. \tag{4}$$

If the error is more than ½, the current hypothesis is discarded, and a new training subset is selected. All hypotheses generated thus far are then combined using the weighted majority voting to obtain the composite hypothesis H_t (step 5).

$$H_t = \arg\max_{y\in Y} \sum_{t:h_t(x)=y} \log(1/\beta_t) \tag{5}$$

The weighted majority voting chooses the class receiving the highest weighted vote from all hypotheses. The hypotheses with good training performance records are awarded higher voting weights. The error of the composite hypothesis is then computed as the sum of distribution weights of the instances misclassified by H_t

$$E_t = \sum_{i:H_t(x_i)\neq y_i} D_t(i) = \sum_{i=1}^{m} D_t(i)\left[\left|H_t(x_i) \neq y_i\right|\right] \tag{6}$$

where $[\![\bullet]\!]$ evaluates to 1, if the predicate holds true (step 6). The normalized composite error B_t is computed as

$$B_t = E_t /(1-E_t), \quad 0 < B_t < 1 \tag{7}$$

which is then used in updating instance weights (step 7):

$$w_{t+1}(i) = w_t(i) \times \begin{cases} B_t, & if\ H_t(x_i) = y_i \\ 1, & otherwise \end{cases} = w_t(i) \times B_t^{1-[\![H_t(x_i) \neq y_i]\!]} \tag{8}$$

The weight update rule in Eq. (8) reduces the weights of the instances that are correctly classified, making them less likely to be selected into the next training subset. The probability of misclassified instances being selected into the next training subset is therefore effectively increased at the next iteration's normalization step (step 1). The algorithm is essentially forced to focus on instances that are difficult to classify. At any point, a final hypothesis H_{final} can be obtained by combining all hypotheses generated thus far using the weighted majority voting

$$H_{final}(x) = \arg\max_{y \in Y} \sum_{t:h_t(x)=y} \log \frac{1}{\beta_t}. \tag{9}$$

3. RESULTS

Two types of wavelets were used for feature extraction: Daubechies 4 (db4) wavelet and Quadratic b-spline wavelet, both of which have been used in previous related efforts [6]. The number of coefficients varied based on the type of wavelet, 121 for b-spline and 64 for db4, though they were chosen to include all frequencies below 16 Hz, where most of the ERP information is known to reside. While 19 channels of EEGs were recorded, we have thus far analyzed the Pz electrode only, the channel that is most commonly used in oddball paradigm analysis. Five-fold cross validation was performed using the Learn++ algorithm yielding a performance of 85.3+/-5.2% for the b-splines (using 15 MLP type classifiers, all with 30 hidden layer nodes and an error goal of 0.15) and 90+/-9.2% for the db4 wavelet (5 classifiers, all with 30 hidden layer nodes and an error goal of 0.1). The sensitivities were 60% and 70%, specificities were 100% and 100%, and the positive predictive values were 100% and 80%, for b-splines and db4 wavelets, respectively. These results compare favorably to our previous efforts of using a single strong classifier, where the overall performance was in the low 80% range, sensitivities were in the 60% range, specificities were in the 90% range and the positive predictive values were in the mid 80% to mid 90% range [11].

4. DISCUSSION

These results are preliminary and based on a limited number of patients, however the use of wavelet analysis to extract features of the ERPs, followed by an automated classification system appears to be a feasible approach for early diagnosis of AD. The use of Learn++ for classification of the wavelet coefficients provides a suitable automated algorithm, and a favorable alternative to using single classifier systems. Furthermore, Learn++ possesses the ability to learn incrementally, a key asset that will be very beneficial as new data become available. It is also worth noting that Learn++ is also capable of learning new classes that may be introduced with the new data, which may also be beneficial in our future work of estimating the severity of the disease, or identifying patients with mild cognitive impairment as a third category between AD and normal.

The type of wavelet may not be too critical, as both types were able to extract much of the same information; however, the approach can be optimized for a specific wavelet.

We note that the approach using the MWA, with either single or ensemble classifiers, also compares favorably to reported community clinic correct diagnosis rates, and in some cases, even to those of the university hospital clinical evaluation rates. Additional training data is expected to help improve the generalization performance; therefore, the approach should provide a stable and effective algorithm once the remaining patients are recruited.

5. ACKNOWLEDGEMENT

This work is supported by National Institute on Aging of the National Institutes of Health under grant No P30 AG10124 - R01 AG022272 to Clark, Kounios and Polikar, and in part by the National Science Foundation grant No. ECS-0239090 to Polikar.

6. REFERENCES

[1] Alzheimer's Association, "Alzheimer's disease statistics," last accessed 09/17/04. Available at http://www.alz.org /Resources/FactSheets/ FSAlzheimerStats.pdf,

[2] A. Lim, D. Tsuang, et al." Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series," *J. American Geriatrics Soc*. vol. 47, no. 5, pp. 564-569, 1999.

[3] S. Yamaguchi, H. Tsuchiya, S. Yamagata, G. Toyoda, S. Kobayashi, "Event-related brain potentials in response to novel sounds in dementia," *Clinical Neurophysiology*, vol. 112, no. 2, pp. 195-203, 2002.

[4] J. Polich , C. Ladish, F. Bloom, "P300 assessment of early Alzheimer's disease," *EEG & Clinical Neurophysiology* , vol. 77, no. 3, pp. 179-189, 1990.

[5] J. Polich, "P300 in clinical applications," In E*lectroencephalography*, E. Niedermeyer, F. Lopez Da Silva, Ed. Philadelphia: Williams and Wilkins, 1999, pp. 1073-1091.

[6] T. Demiralp, A. Ademoglu, "Decomposition of event-related brain potentials into multiple functional components using wavelet transform," *Clinical Electroencephalography*, vol. 32, no. 3, pp. 122-138, 2001.

[7] G. McKhann, et al., "Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group to Dept. of HHS Task Force on Alzheimer's Disease," *Neurology*, vol. 34, pp. 939-944, 1984.

[8] M. Unser, editor, Gallery at wavelet.org, 09/17/2004, Available at: http://www.wavelet.org/phpBB2/gallery.php

[9] R. Polikar, L. Udpa, S. Udpa, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Trans. on Sys., Man and Cyber. (C)*, vol. 31, no. 4, pp. 497-508, 2001.

[10]Y. Freund and R. Schapire, "A decision theoretic generalization of on-line learning and an application to boosting," *Comp. and System Sci*., vol. 57, no. 1, pp. 119-139, 1997.

[11] G. Jacques, J.L. Frymiare, J. Kounios,, C. Clark, and R. Polikar, Multiresolution wavelet analysis for early diagnosis of Alzheimer's Disease, "*Proc. of 26th Int. Conf. of IEEE Eng. in Med. and Biology Soc.,"* pp. 251-254, San Francisco, CA, 2004.