Majority Vote and Decision Template Based Ensemble Classifiers Trained on Event Related Potentials for Early Diagnosis of Alzheimer's Disease

Nicholas Stepenosky[1], Deborah Green[2], John Kounios[2], Christopher M. Clark[3], and Robi Polikar[1*]

[1]Department of Electrical Engineering, Rowan University, Glassboro, NJ, USA
[2]Department of Psychology, Drexel University, Philadelphia, Pennsylvania, USA
[3]Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
*corresponding author: polikar@rowan.edu

ABSTRACT

With the rapid increase in the population of elderly individuals affected by Alzheimer's disease, the need for an accurate, inexpensive and non-intrusive diagnostic biomarker that can be made available to community healthcare providers presents itself as a major public health concern. The feasibility of EEG as such a biomarker has gained a renewed attention as several recent studies, including our previous efforts, reported promising results. In this paper we present our preliminary results on using wavelet coefficients of event related potentials along with an ensemble of classifiers combined with majority vote and decision templates.

1. INTRODUCTION

Alzheimer's disease (AD) is the most common type of dementia, a degenerative neurological disorder associated with aging. AD is a progressive brain disorder that gradually destroys a person's memory and their ability to learn, reason, make judgments, communicate and carry out daily activities. The likelihood of developing AD almost doubles every five years after the age of 65. By the age of 85, the odds of developing AD reach the alarming rate of one out of every two people. The Alzheimer's Association puts the estimated number of people affected by AD at 4.5 million, in the U.S. alone [1], with a projected number of 12 – 16 million by 2050. Alzheimer's disease has no known single cause, no cure, nor even a definitive means of diagnosis – except autopsy. Currently, AD is typically diagnosed through a clinical evaluation that involves a series of memory tests, interviews with the patient and his/her caregivers and continuous monitoring over a period of time. While such a clinical evaluation has a relatively high positive predictive value of 93%, it is only available through expert neuropsychologists at major university hospitals and/or research clinics. This level of expertise and procedures are usually very costly, and hence most patients are typically evaluated by their local community healthcare providers, where the expertise and accuracy of AD specific diagnosis remains uncertain. In 1999, a group of Health Maintenance Organization-based physicians reported an average sensitivity of 83%, specificity of 55%, and an overall accuracy of 75% [2]. Since the first line of intervention are such community clinics – at least for most people – an accurate, inexpensive, non-invasive, cost-effective, and an automated diagnostic procedure that can be made available to such clinics would be very beneficial. Furthermore, considering that most people can live up to 8 – 20 years with early intervention, a reliable early diagnosis can not only add years to patient's life, but can significantly increase the quality of life for the patient as well as their caregivers.

AD is a cortical dementia in which certain underlying processes manifest themselves on the event related potentials (ERPs) of the electroencephalogram (EEG). The EEG has not been traditionally used in AD diagnosis, however, since signals show several changes due to normal aging, coexisting medical illness, and levels of anxiety or drowsiness during the measurements as well. On the other hand, an EEG based protocol, called the oddball paradigm that involves the analysis of ERPs, has been shown to generate changes that are linked to mental impairment. In the oddball paradigm protocol, subjects are instructed to press a button when they hear an occasionally occurring oddball tone of 2 kHz within a series of regular 1 kHz tones and novel sounds. The ERPs show a positive peak called the P3 (or P300) with an approximate latency of 300 ms after the oddball stimulus. Changes in the amplitude and latency of the P300 are known to be altered by neurological disorders affecting the temporal-parietal regions of the brain [3]. Polich et al. have reported that latency and the amplitude of the P300 are in fact altered in patients with AD when compared to elderly control subjects [4, 5].

Traditional ERP analysis is performed in time domain using the amplitude and latency of the P300s. This analysis reveals only a fraction of the information available in the ERP, however, since ERPs are non-stationary signals, whose spectral content vary in time. Other studies have shown that the ERPs, and the P300 component in particular, consist of the superposition of multiple functional components, where these components extend for different, yet overlapping, time intervals in different frequency bands [6, 7]. This makes the discrete wavelet transform (DWT) an appropriate tool for the analysis of ERPs, as also shown in our earlier studies [8,9].

1-4244-0469-X/06/$20.00 ©2006 IEEE

2. METHODOLOGY

2.1. Research Subjects

This study will include a total of 80 subjects, in which half the cohort will be cognitively normal and the other diagnosed with probable AD based on clinical evaluations. Data from 48 patients – recruited to date – 25 diagnosed with probable AD and 23 cognitively normal, have been analyzed. Subjects are verified to be free of any evidence of other neurological disorders (e.g. stroke, multiple sclerosis, Parkinson's disease, etc.) by history or by exam. The two groups were defined by the following criteria: *Cognitively normal:* (i) age > 60; (ii) Clinical Dementia Rating (CDR) = 0; (iii) Mini-Mental Scores (MMS) ≥ 24; (iv) no indication of functional cognitive decline during the previous two years based on a detailed interview with the subject's knowledgeable informant or two previous annual clinical assessments. *AD subjects:* (i) age > 60; (ii) CDR ≥ 0.50; (iii) MMS< 24; (iv) presence of functional cognitive decline over the previous 12 months; (v) satisfaction of NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association) criteria for probable AD [10]. All subjects received a thorough medical history analysis, neurological exam, memory tests and standardized evaluations for several functional impairments, extrapyramidal signs for behavioral changes and depression. The clinical diagnosis was made as a result of these analyses.

2.2. Data Acquisition

The ERPs were obtained using the oddball paradigm. The protocol originally described in [3] was followed with slight modifications. The evoked response stimulus was presented to both of the subject's ears using stereo speakers with an amplitude level comfortable for their hearing. The stimulus consisted of tone bursts 100 ms in duration. Tones of 1 kHz and 2 kHz were presented in a random sequence. These tones made up 65% and 20% of the tones respectively. The remaining 15% of the trials consisted of novel sounds also presented randomly. A total of 1000 stimuli, including the standard tones of kHz (n = 650), target tones of 2 kHz (n = 200) and novel sounds (n = 150), were delivered to each subject with an inter-stimulus interval of 1.0-1.3 seconds. The subjects were instructed to press a button each time they heard the target tone of 2 kHz. When a subject responds to the target tone and presses the button, the ERPs are recorded for 1 second from 19 tin electrodes embedded in a plastic cap.

The data collection process lasted about 30 minutes per subject with each session proceeded by a 1 minute practice session without novel sounds. Artifactual recordings were identified and rejected by the EEG technician. The remaining recordings were amplified, digitized at 256 Hz/channel and stored. The saved ERPs were preprocessed using low-pass filtering and trial averaging. Averaging individual target responses (30-90 per patient) is necessary to obtain a robust P300 component. All averages have been notch filtered at 59-61 Hz and then amplitude normalized.

2.3. Multiresolution Wavelet Analysis

Multiresolution wavelet analysis provides time localizations of spectral components in a signal thus providing its time-frequency representation. The Discrete wavelet transform (DWT) has become an increasingly popular method for time-frequency analysis due to its ability to solve a diverse set of problems. It does so by decomposing a signal into different frequency bands by successive low-pass and high-pass filtering. The outputs of the high-pass filters at each level constitute the DWT coefficients at that level, while the low-pass filter outputs are further decomposed. At each successive level, the signal is analyzed at a reduced time and increased frequency resolution, hence the name multiresolution analysis. In this study we used the Daubechies wavelet with four vanishing moments as the mother wavelet and carried out the decomposition for 7 levels of detail, creating 7 frequency bands. Wavelet transforms have been well established by now, and details can be found in many excellent references listed at [11].

2.4. Features and Ensemble Classification

The signals analyzed in this study consist of the preprocessed data from the Pz electrode of the EEGs. This electrode collects data from the central parietal section of the cortex, where the P300 is known to be most prominent [12]. Furthermore, the spectral content of the P300 is known to be around 3 Hz. Therefore, the middle four of level 6 (2-4 Hz) DWT coefficients (corresponding to the 200 – 500 ms range were extracted for the analysis.

Ensemble of classifiers based approaches have recently enjoyed great attention due to their reported superiority over single classifier based systems' generalization performance. Ensemble generation techniques, such as bagging, boosting / AdaBoost, mixture of experts, along with several ensemble combination strategies, such as voting techniques, posterior probability based combinations and template matching have been proposed, analyzed and shown to be effective on a wide spectrum of applications [13]. The idea behind all ensemble based systems is that if individual classifiers are diverse, then they can make different errors, and combining these classifiers can reduce the error through averaging. Diverse classifiers can be obtained by using different data, and/or deliberately making the classifiers relatively weak.

Three such weak multiplayer perceptron type (MLP) classifiers were trained on averaged ERP responses to the target tones. All MLPs had 4 input, four hidden and two output nodes with a fairly tolerant mean square error goal of 0.1, to ensure that the MLPs are relatively weak with respect to the classification problem. These three classifiers were then combined using two different ensemble combination techniques: majority voting and decision templates. Both methods are further explained in the following sections. The generalization performances for the individual classifiers, majority voting, and decision templates were obtained through a leave-one-out cross validation scheme.

2.5. Majority Vote

Majority vote is one of the simplest and most intuitive ensemble combination techniques. Essentially, the ensemble chooses the class that is chosen by the majority of the classifiers. Let us define the decision of the tth classifier D_t as $d_{t,j} \in \{0,1\}$, t=1,…,T and j=1,…,c, where T is the number of classifiers and c is the number of classes. If tth classifier chooses class j, then $d_{t,j} = 1$, and zero, otherwise. The vote will then result in an ensemble decision for class k if:

$$\sum_{t=1}^{T} d_{t,k} = \max_{j=1}^{c} \sum_{t=1}^{T} d_{t,j} \tag{1}$$

If there is reason to believe that certain classifiers in the ensemble are “better” than the others, a weighted majority voting can also be employed.

2.6. Decision Templates

Decision templates were proposed by Kuncheva in [14], for combining continuous valued outputs of an ensemble of classifiers. The classifier outputs are typically normalized to add up to 1 using the softmax normalization: denoting the jth output of the classifier with y_j, the normalized values are

$$y'_j(\mathbf{x}) = \frac{\exp\{y_j(\mathbf{x})\}}{\sum_{k=1}^{c} \exp\{y_k(\mathbf{x})\}} \tag{2}$$

and the new y' values are used as $d_{t,j}(\mathbf{x})$, which are then interpreted as the support given by the tth classifier to the jth class. Let $\mathbf{x} \in \boldsymbol{R}^{\mathbf{n}}$ be a feature vector and W = {$\omega_1, \omega_2, \ldots, \omega_C$} be the set of class labels. Each classifier D_t in the ensemble $D = \{D_1, \ldots, D_T\}$ outputs c degrees of support for each $\mathbf{x}$. The outputs of T classifiers for a particular $\mathbf{x}$ are first organized into a *decision profile* $DP(\mathbf{x})$ as shown in Fig 1.

The column for $d_{1,j}$ to $d_{T,j}$ represents the support from classifiers D_1 to D_T for class j, and the row $d_{t,1}$ to $d_{t,C}$ is the support from classifier D_t.

The decision templates (DT) are then obtained for each class j as the average decision profile among all class j instances of the training data:

$$DT_j = \frac{1}{N_j} \sum_{\boldsymbol{X}_j \in \omega_j} DP\left(\boldsymbol{X}_j\right) \tag{3}$$

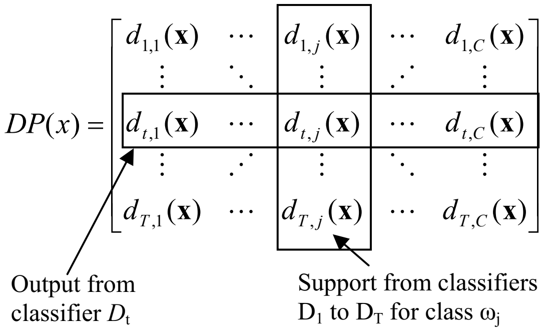
$DP(x) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & \cdots & d_{1,j}(\mathbf{x}) & \cdots & d_{1,C}(\mathbf{x}) \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ d_{t,1}(\mathbf{x}) & \cdots & d_{t,j}(\mathbf{x}) & \cdots & d_{t,C}(\mathbf{x}) \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ d_{T,1}(\mathbf{x}) & \cdots & d_{T,j}(\mathbf{x}) & \cdots & d_{T,C}(\mathbf{x}) \end{bmatrix}$
Output from classifier D_t
Support from classifiers D_1 to D_T for class ω_j

Fig. 1. Decision Profile matrix.

where N_j is the number of class j instances. Given an unlabeled instance $\mathbf{x}$, the *similarity* of its decision profile and each DT_j constitutes the support given to class j by the ensemble. The similarity measure used in this work is the squared Euclidean distance, for which the total ensemble support for class j is computed as

$$\mu_j(\mathbf{x}) = 1 - \frac{1}{T \times C} \sum_{t=1}^{T} \sum_{k=1}^{C} \left[DT_j(t,k) - d_{t,k}(\mathbf{x})\right]^2 \tag{4}$$

where $DT_j(t,k)$ is the support given by the t^{th} classifier to class k by the decision template DT_j, that is, the support given by the tth classifier to class k, averaged over class j instances. This support should ideally be high when k=j, and low otherwise. The second term $d_{t,k}(\mathbf{x})$ is the support given by the tth classifier to class k for the given instance $\mathbf{x}$. The class with the highest total support is then chosen as the ensemble decision.

In this application decision templates are created to combine individual supports given by three MLPs to each of the two classes, cognitively normal and AD. Hence, each decision profile and each decision template are 3-by-2 matrices. Detailed discussion on these and other ensemble combination rules can be found in [13].

3. RESULTS

Generalization performances and their confidence intervals are given in Table 1 for each individual classifier, $D_1 \sim D_3$, as well as the ensemble performance obtained through majority vote (MV) and the decision templates (DT). Also provided in Table 1 are the sensitivity, specificity and positive predictive values of each. All figures are average of ten trials of leave-one-out based generalization performances. Each leave-one-out performance itself is obtained by training the classifiers with 47 of the 48 instances, testing the classifiers on the remaining 48th instance, repeating the procedure 48 times changing the test instance in each case and taking the average of 48 such one-instance performances. Leave one out based generalization performance is usually considered as the best (least unbiased) estimate of the true generalization performance of the classification system.

Table 1 indicates that each classifier had an average generalization performance of 67 – 71%, which was boosted to around 76% by the majority vote or decision template. Similar performance improvements can also be seen on the sensitivity, specificity and positive predictive values.

Table 1. Results from 3 classifiers (D_1, D_2, D_3), Majority Vote (MV), and Decision Template (DT)

	Average Performance	Standard Deviation	95% Confidence Interval	Sensitivity	Specificity	Positive Predictive Value
D_1	0.7187	0.0225	0.0130	0.6760	0.7652	0.7601
D_2	0.7125	0.0165	0.0095	0.7080	0.7174	0.7322
D_3	0.6750	0.0329	0.0191	0.6360	0.7130	0.7126
MV	0.7583	0.0224	0.0130	0.7440	0.7739	0.7821
DT	0.7625	0.0176	0.0102	0.7520	0.7739	0.7841

We should note that while the difference between single classifier performances and ensemble performances are statistically significant, the difference between the majority vote and the decision templates is not. The decision templates approach is a more elegant one that considers the support given to all classes before making a decision. However, whether its additional computational overhead is justified is debatable, at least on the current dataset.

4. CONCLUSIONS & DISCUSSION

Feasibility of a diagnostic tool for early diagnosis of Alzheimer's disease is explored. An ensemble of three classifiers are combined through majority voting and decision templates, where each classifier is trained on four wavelet coefficients that characterize the P300 component at the 2-4 Hz interval. In general, results indicate that there is indeed a statistically significant performance to be gained from the ensemble combination process.

Current results on the first 48 patients (the study will eventually include 80) indicate a surprisingly promising outlook, considering that only four coefficients are used to characterize a disease for which there is still no definitive mean of diagnosis. Furthermore, all AD patients in the current cohort had MMS scores around 24, indicating that the disease is being detected at its earliest stages. Finally, the results presented in Table 1 match or exceed the current diagnosis performance at community clinics.

Our earlier results have already indicated that there is nothing significant to be gained by using a larger ensemble or stronger individual classifiers. Hence our current and future work will focus on using other ensemble generation techniques, combining features from different levels of the DWT, as well as combining data from different electrodes.

As we continuously explore alternative feature sets and classification approaches, it is hoped that we will get closer to the clinical evaluation performance figures as we obtain additional real-world data when the remaining 32 patients are recruited. The algorithm can then be easily integrated into an EEG module, and made available to community clinics to be used as a first level diagnostic screening tool for detecting the disease at the earliest stages possible.

5. ACKNOWLEDGMENTS

This work is supported by National Institute on Aging of the National Institutes of Health under grant number P30AG10124 - R01 AG022272, and by National Science Foundation under Grant No ECS-0239090.

6. REFERENCES

[1] Alzheimer's Association, "Fact Sheet About Alzheimer's Disease," last accessed 10/10/05. Available at: http://www.alz.org/Resources/FactSheets/FSADFacts.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] T. Demiralp et al. "Analysis of functional components of P300 by wavelet transform," Proc. of IEEE Eng. in Med. & Bio. Conf., vol. 20, no.4, pp. 1992-1995, 1998.

[8] G. Jacques, J.L. Frymiare, J. Kounios,, C. Clark, 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.

[9] N. Stepenosky, A. Topalis, H. Syed, D. Green, J. Kounios, C. Clark, R. Polikar, "Boosting based classification of event related potentials for early diagnosis of Alzheimer's disease," *2005 IEEE Eng. in Medicine and Biology Soc. Conf.,* Shanghai, China, 2005.

[10]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.

[11] M. Under, editor, Gallery at wavelet.org, Last accessed 10/20/05, vailable:http://www.wavelet.org/phpBB2/gallery.php

[12] B. Jansen *et al.* "An exploratory study of factors affecting single trial P300 detection," *IEEE Tran. Bio. Eng.,* vol. 51, no. 6, pp. 975–978, 2004.

[13] L. I. Kuncheva, Combining Pattern Classifiers, Methods and Algorithms. New York, NY: Wiley Interscience, 2005.

[14] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Rec.*, vol. 34, no. 2, pp. 299-314, 2001.