Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Stacked Generalization for Early Diagnosis of Alzheimer's Disease

Hardik Gandhi, Deborah Green, John Kounios, Christopher M. Clark, and Robi Polikar

Abstract –**The diagnosis of Alzheimer's disease (AD) at an early stage is a major concern due to growing number of elderly population affected by the disease, as well as the lack of a standard diagnosis procedure available to community clinics. Recent studies have used wavelets and other signal processing methods to analyze EEG signals in an attempt to find a non-invasive biomarker for AD. These studies had varying degrees of success, in part due to small cohort size. In this study, multiresolution wavelet analysis is performed on event related potentials of the EEGs of a relatively larger cohort of 44 patients. Particular emphasis was on diagnosis at the earliest stage and feasibility of implementation in a community health clinic setting. Extracted features were then used to train an ensemble of classifiers based stacked generalization approach. We describe the approach, and present our promising preliminary results.**

I. 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 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 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), β-amyloid, brain atrophy detected by MRI, etc.; however, these techniques are invasive, expensive, and/or only available at research hospitals. There is, therefore, significant need for a clinically useful, accurate, non-invasive, cost-effective and automated procedure for early diagnosis of AD.

This work is supported by National Institute on Aging of NIH under grant number P30 AG10124 - R01 AG022272, and by National Science Foundation under Grant No ECS-0239090.

H. Gandhi and R. Polikar are with the Dept. of Electrical and Computer Engineering, Rowan University, Glassboro, NJ. D. Green and J. Kounios are with the Dept. of Psychology, Drexel University, Philadelphia, PA and C.M. Clark is with Dept. of Neurology, University of Pennsylvania, Philadelphia, PA. Contact author: Robi Polikar, phone: 856-256-5372, fax: 856-256-5241, E-mail: polikar@rowan.edu,

The electroencephalogram (EEG) may potentially satisfy these needs as a tool for AD diagnosis.

A. Prior Efforts

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 a simple task (e.g., pressing a button) by the subject in response to an infrequently occurring 2 kHz (oddball) tone, presented randomly within a series of regularly occurring 1 kHz (regular) tones. Responding to the oddball tone elicits a positive peak (P300) in the ERP, 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]. The P300 component, while useful to show statistical differences between AD and normal groups, is not discriminatory enough – on its own – to be able to identify individual patients. Consider the four ERPs shown in Figure 1, obtained from the patients recruited for this study. Figure 1 (a) and (b) show a normal patient with a strong P300 component, and an AD patient with a missing P300, respectively. This is the behavior in line with statistical results obtained in earlier studies. However, Figure 1 (c) and (d) show that the exact opposite is also not uncommon: normal patients with missing P300, and AD patients with strong a P300. Hence, cognitively normal people may have delayed or absent P300; and those with AD, in particular in early stages, may still have strong P300. The inability of classical statistical approaches in individually identifying specific cases demands more sophisticated approaches for such individual identification.

Since ERPs are non-stationary signals, time-frequency analysis techniques are appropriate for their analysis. Recent studies have used wavelets to decompose an ERP in the time-frequency plane in order to characterize its functionally relevant components [6].

More recently, wavelet analysis of ERPs followed by neural network classification has been attempted for AD diagnosis, yielding promising but limited success [7, 8]. However, such studies have primarily been pilot studies with fewer patients (20 in [7], and 28 in our previous study [8]), making statistical generalization very difficult.

The novel contribution of this effort is to demonstrate the feasibility of an ensemble of classifiers based stacked generalization approach, used with a time-frequency based feature extraction method, for the earliest diagnosis of AD.

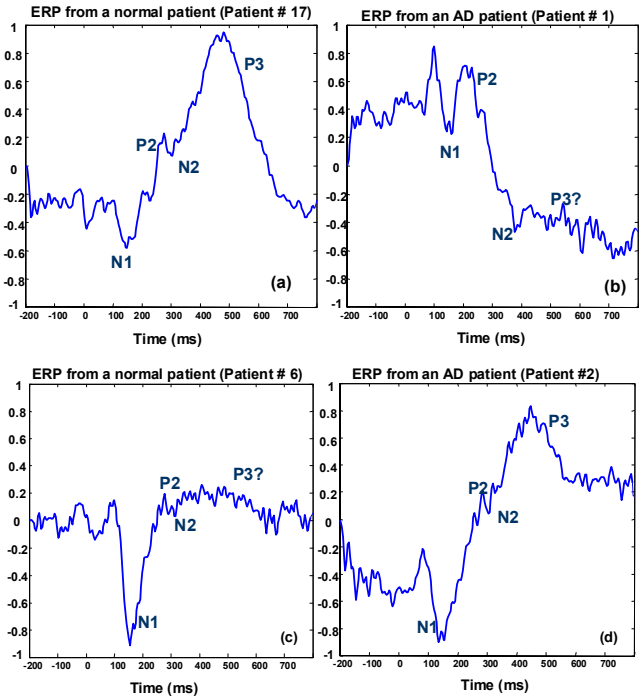

Figure 1 (a) & (b) expected P300 behavior from normal and AD patients; (c) & (d) not all ERPs follow this behavior.

Specifically, the method described in this paper combines multiresolution wavelet analysis (MWA), automated classification using an ensemble of classifiers approach, along with established EEG analysis, to detect the earliest stage of AD. Our expectation in doing so is to combine wavelet transform's ability to extract features – other than just the amplitude and latency of P300 – with the superior classification and robustness of ensemble of classifiers in automated early diagnosis of the AD. Algorithm has been evaluated on a database of the 44 of the 80 patients recruited thus far for this study.

B. Staging Dementia: Early Diagnosis

Staging of dementia is important because it alerts the caregiver and health care team, and prepares them for the next phase of illness. Dementia is typically staged in five categories: mild, moderate, severe, profound and terminal [9]**.** Clinicians assess an individual's disease stage by assessing cognition, behavior and function. Cognition includes multiple domains of which memory is one: e.g., ability to speak and comprehend others, recognize family and friends, etc. Function involves the ability to perform day-to-day activities, including personal care and social roles, while behavior includes the assessment of personality changes, and how a person acts in social situations [9].

One standard tool for AD diagnosis is the Mini-Mental State Exam (MMSE), a test for memory, language and praxis skills. It is scored on a scale of 0-30, with decreasing scores (particularly below 19) indicating increased impairment. Other tests include the Severe Impairment Battery (SIB) and the Clinical Dementia Rating (CDR) Scale, all of which are part of the NINCDS-ADRDA (National Institute of Neurological and Communicative Disorders and Stroke - Alzheimer's Disease and Related Disorders Association) criteria for probable AD [10].

II. METHODOLOGY

A. Test Subjects and Clinical Evaluation

The complete study will include a total of 80 subjects. Forty-four elderly subjects, satisfying the inclusion criteria – 22 diagnosed with probable AD and 22 cognitively normal controls – have been recruited thus far. All 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. All subjects received a thorough evaluation at the University of Pennsylvania's Memory Disorders Clinic, in Philadelphia, including a medical history analysis, neurological exam, memory tests and standardized evaluations for several functional impairments and extrapyramidal signs for behavioral changes and depression. The evaluation included standardized assessments for overall impairment, functional impairment, extra pyramidal signs, behavioral changes and depression. The clinical diagnosis was made as a result of these analyses, and constituted the gold-standard against which the proposed automated system has been compared.

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 criteria for probable AD [8]. In order to ensure that the technique developed can detect the earliest neurological effects of the disease, the AD cohort was chosen to be in the earliest stages of the disease. Hence, the average MMSE scores of the two cohort were $\mu_{AD} = 25$ (borderline between normal and mild cognitive impairment, a precursor of AD), and $\mu_{Normal} = 29$.

B. 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 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 inter-

val 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. Artifactual recordings were identified and rejected by the EEG technician. The remaining scalp potentials were amplified, digitized at 256 Hz/channel, 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 90~250 recordings per patient (per stimulus type), a necessary step to ensure robust P300 components are obtained. All averages have been notched filtered at 59-61 Hz, and baselined with the prestimulus interval.

C. Multiresolution Wavelet Analysis

Multiresolution wavelet analysis provides time localizations of spectral components in the signal, resulting in a time-frequency representation. The discrete wavelet transform (DWT) analyzes the signal at different frequency bands with different resolutions through the decomposition of the signal (hence, multiresolution analysis). A Daubechies wavelet with four vanishing moments was used for the seven level decomposition of the 256 point long preprocessed ERP signal. In our previous analyses of different frequency bands, we have found out that the 2-4 Hz range of the ERPs obtained from the Pz electrode (located in the central – parietal region of the scalp) to provide robust classification performance. This comes as no surprise, since the P300 component is strongest in the parietal region of the brain, and is known to reside in the 2-4 Hz range. Since the DWT is now a well-established technique for analysis of non-stationary signals in the time-frequency space, we do not discuss the details of the approach and refer the interested readers to many excellent references listed at [11].

D. Stacked Generalization Algorithm

Stacked generalization is an ensemble of classifiers approach for minimizing the generalization error rate of one or more classifiers [12]. Stacked generalization works by reducing the biases of the classifiers with respect to a training dataset. Stacked generalization primarily involves a set of classifiers ($C_1 \ldots C_N$) whose outputs are used as input to train a second-level classifier (C_{N+1}). The goal of the second level classifier is then to learn the correct and incorrect predictions of the first level classifiers. Essentially, C_{N+1} learns how to map the combined outputs of $C_1 \sim C_N$ to their correct classes.

The implementation of stacked generalization involves a k - fold selection process to obtain the training data for the classifier C_{N+1}. Specifically, the entire training data is first split into k subsets of (near) equal cardinality. Each classifier in the ensemble C_1 through C_N is trained k times, using $k-1$ blocks of the training data. In each case, there is one block of data not seen by any of the classifiers C_1 through C_N. The outputs of the classifiers for such instances, along with their correct labels, constitute the training data for the second level classifier C_{N+1}. Once C_{N+1} is trained, all data are pooled, and individual classifiers C_1 through C_N are trained again on the entire database, using a suitable resampling method. This algorithm is illustrated in Figure 2, where $h_i(\mathbf{x}, \boldsymbol{\theta}_i)$ indicates the classification (hypothesis) of classifier C_i for instance $\mathbf{x}$, when trained with the parameter set $\boldsymbol{\theta}_i$, which describes the training parameters of C_i.

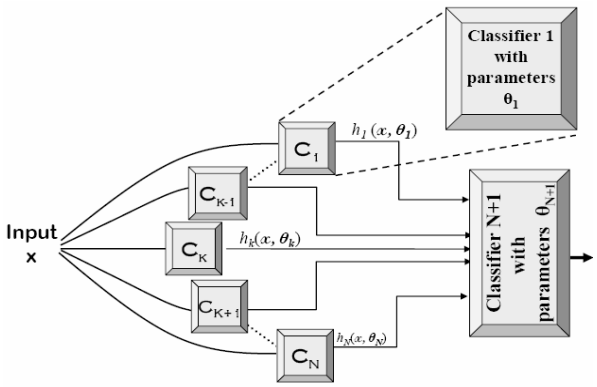

Figure 2. Stacked generalization

E. Majority Voting

One of the simplest, yet very commonly used approaches in ensemble based systems is to combine the outputs of the classifiers using a simple or weighted majority voting. Let us define the decision of the t^{th} classifier as $d_{i,j} \in \{0,1\}$, $i=1,...,N$ and $j=1,\ldots,c$, where N is the number of classifiers and c is the number of classes. If t^{th} classifier chooses class j, then $d_{t,j} = 1$, and zero, otherwise. In majority voting, the ensemble decision is chosen as the class that receives the highest number of votes, that is, we choose class J, if

$$\sum_{i=1}^{N} d_{i,J} = \max_{j=1}^{c} \sum_{i=1}^{N} d_{i,j} \qquad (1)$$

If we have reason to believe that some of the classifiers are more qualified than the others, then weighting their decisions more heavily may further improve the overall performance than that can be obtained by simple majority voting. Assigning weight w_i to classifier C_i in proportion to its estimated performance, the ensemble decision will then chose class J, if

$$\sum_{i=1}^{N} w_i d_{i,J} = \max_{j=1}^{c} \sum_{i=1}^{N} w_i d_{i,j} \qquad (2)$$

that is, if the total weighted vote received by class J is higher than the total vote received by any other class. In most applications, the weight assigned to classifier C_i is typically proportional to its performance on training data.

III. Results

Two sets of features were used along with two sets of ensemble combination schemes. The first feature set used the middle coefficients (corresponding to 200-800 ms range) from the 2-4 Hz frequency band, and the second feature set used all coefficients of the 2-4 Hz range. As mentioned earlier, the P300 is known to reside in the 2-4 Hz range (and our single classifier tests confirmed that this range provided better classification performance).

These feature sets were used to train a stacked generalization based ensemble system. All individual classifiers were multilayer perceptron type classifiers. The first level consisted of an ensemble of 5 classifiers with each having one hidden layer and five hidden layer nodes. The second layer classifier had one hidden layer with ten nodes. The error goal was 0.01. A leave-one-out cross-validation was used for training and testing to ensure the most conservative and reliable generalization performance, a standard procedure in classification applications of small databases. Of the 44 patient data, 43 were used for training all classifiers in the first level, and the remaining one was used for testing. This was repeated 44 times, each time using a different instance to test. Outputs of first level classifiers were concatenated to train the second level classifier. All data were then pooled, and the first level classifiers were retrained, again using leave-one-out validation. Final generalization performance figures were obtained as average of ten leave-one-out trials. We also repeated the ten-fold leave-one-out validation on majority vote combination of five level-1 classifiers, for comparison purposes. The individual performances of these classifiers were in the low-to-middle 60% range.

The generalization performances along with their 95% confidence intervals (averaged over ten trials) are summarized in Table 1. On average, stacked generalization using the middle coefficients of the 2-4 Hz DWT performed the best, matching or exceeding the previously reported community clinic diagnostic performances.

Table 1. Summary of diagnostic performances

Features	Classifier	Mean ± C.I.	Max
Middle coef. of 2-4 Hz	Stacked Gen.	76.4±2.4 %	84.0 %
	Majority Vote of 5 Level-1 classifiers	72.5±3.2 %	79.7 %
All coef. of 2-4 Hz	Stacked Gen.	74.2±2.9 %	81.0 %
	Majority Vote of 5 level-1 classifiers	69.4±3.9 %	78.0 %

IV. Conclusions

The application presented in this work seeks diagnostic identification of AD vs. normal patients based on their ERP recordings. Of particular interest – which makes the problem particularly challenging – is diagnosis of the disease at its earliest possible stages. Based on the results presented above, we conclude that using wavelet analysis to extract features of the ERPs, followed by ensemble based classification appears to be an effective tool for early diagnosis of AD. The stacked generalization performance was in upper 70% to lower 80% range, matching or exceeding the diagnostic performance of community clinics. This is significant, because an EEG based automated classification system is non-invasive, objective, cost-effective and can be easily implemented at community clinics, where most people get their first intervention.

We also found that the performance of stacked generalization was better than that of majority voting, but the difference was just shy of being statistically significant. We note that the individual classifiers that made up the ensemble performed around low to mid sixty percent, showing the additional improvement obtained by using an ensemble approach – whether that be majority voting or stacked generalization.

Future work will include acquiring the remaining patient data, further analysis using other electrodes, and trying other ensemble creation and combination approaches.

References

[1] Alzheimer's Assoc., "Alzheimer's disease statistics," Available at http://www.alz.org/AboutAD/statistics.asp

[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 Electroencephalography, E. Niedermeyer, F. Lopez Da Silva, Ed. Philadelphia: Williams Wilkins, pp. 1073-1091, 1999

[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] A. Petrosian, D. Prokhorov, W. Nanson and R. Schiffer, "Recurrent neural network based approach for early recognition of Alzheimer's disease in EEG," *Clinical Neurophysiology*, vol. 112, no. 8, pp. 1378-1387, 2001.

[8] G. Jacques, J. Frymiare, J. Kounios, C. Clark and R. Polikar, "Multiresolution wavelet analysis for early diagnosis of Alzheimer's disease, *Proc. of IEEE Eng. in Med. and Bio.*, vol. 1, pp. 251-254, 2004.

[9] J. Klocinski and V. Cotter, "How do we stage dementia?", University of Pennsylvania Memory Disorders Clinic Newsletter, vol.1, no.4, 2002. Available at: http://www.uphs.upenn.edu/ADC/newsletter.

[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. Unser, editor, Gallery at wavelet.org, Available at: http://www.wavelet.org/phpBB2/gallery.php

[12] D.H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241-260, 1992.