Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Ensemble Based Data Fusion for Early Diagnosis of Alzheimer's Disease

Devi Parikh[1], Nick Stepenosky[1], Apostolos Topalis[1], Deborah Green[2], John Kounios[2], Christopher Clark[3] and Robi Polikar[1*]

[1]Department of Electrical and Computer Engineering, Rowan University, Glassboro, New Jersey, USA
[2]Department of Psychology, Drexel University, Philadelphia, Pennsylvania, USA
[3]Department of Neurology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

***Abstract*—We describe an ensemble of classifiers based data fusion approach to combine information from two sources, believed to contain complimentary information, for early diagnosis of Alzheimer's disease. Specifically, we use the event related potentials recorded from the Pz and Cz electrodes of the EEG, which are further analyzed using multiresolution wavelet analysis. The proposed data fusion approach includes generating multiple classifiers trained with strategically selected subsets of the training data from each source, which are then combined through a weighted majority voting. Several factors set this study apart from similar prior efforts: we use a larger cohort, specifically target early diagnosis of the disease, use an ensemble based approach rather then a single classifier, and most importantly, we combine information from multiple sources, rather then using a single modality. We present promising results obtained from the first 35 (of 80) patients whose data are analyzed thus far.**

Keywords—**Alzheimer's disease, data fusion, wavelet analysis, oddball paradigm, ensemble system, Learn++.**

I. INTRODUCTION

A. EEG Analysis for AD diagnosis

Alzheimer's disease (AD) affects an estimated 4 million Americans, making it a major public health concern. The positive predictive value of clinical diagnosis based on neuropsychological analysis is around 93% (overall diagnostic performance around 80%) at university hospitals, however, most patients are evaluated at community clinics, where the expertise and the accuracy of disease specific dementia diagnoses is uncertain. In fact, a recent study reported that despite the advantage of longitudinal follow up, a group of HMO based physicians had a sensitivity of 83%, specificity of 55% and an overall accuracy of 75% for the clinical diagnosis of AD [1].

Several biomarkers have been linked to AD, such as the cerebrospinal fluid tau, β-amyloid, urine F2-isoprostane, brain atrophy and volume loss detected by MRI. However, none of these methods has proven to be conclusive, and even if they were, they remain primarily research hospital based tools. Consequently, there is significant need for a clinically useful, accurate, non-invasive, cost-effective and automated procedure for early diagnosis of the AD that would be available to community healthcare providers.

One such tool that is potentially feasible is the electroencephalogram (EEG). EEG analysis has not traditionally been part of a routine evaluation for AD diagnosis, however, in part due to difficulties in separating EEG changes that could be attributed to AD from those due to normal aging. An alternative EEG based technique that specifically targets the changes due to mental impairment by analyzing scalp recordings of auditory event related potentials (EPR), has been more promising, but still with inconclusive results. The protocol uses the *oddball paradigm* in which subjects are asked to respond when they hear an occasionally occurring 2 kHz (the oddball) tone, presented randomly within a series of frequently occurring 1 kHz tones. The ERPs in response to oddball tones then show a positive peak (P3 or P300, most prominently on Pz channel of the EEG covering parietal regions), with an approximate latency of 300 ms after the stimulus. Changes in the amplitude and latency of P300 are altered by neurological disorders affecting the temporal-parietal regions of the brain [2]. This includes AD, where the average P300 latency is prolonged and the amplitude decreased compared to elderly controls [3].

Traditional ERP analysis is performed either in time or frequency domain. However, this is suboptimal, since the ERP is a *time and frequency* varying signal. Despite its now mature history, studies applying time-frequency techniques, such as wavelets, to ERPs have only recently started, and mostly on non AD related studies designed specifically for P300 analysis [4,5]. Studies directly targeting AD diagnosis using wavelet analysis, have been even more rare with limited success, in part due to lack of a large study cohort (e.g., 6 in [6]); the results therefore remain largely inconclusive.

B. Ensemble Approaches for Classification & Data Fusion

Data obtained from multiple sources may carry complimentary information, suitable fusion of which, can lead to improved classification performance compared to a decision based on any of the individual data sources alone. In P300 studies, signals from the Pz electrode are usually analyzed [7], where the P300 is most prominent. However, we believe that the nearby electrodes, such as Cz and Fz, may also carry complimentary information. Hence our goal was to determine whether an appropriate combination of data from these channels may lead to improved diagnosis performance.

Traditional data fusion methods are generally based on probability theory, such as the Dempster-Schafer (DS) and its many variations. For classification and data fusion applications, ensemble approaches that use strategically trained and combined multiple classifiers constitute a new breed of algorithms that often offer an improved and more stable performance compared to their single classifier counterparts. Such approaches include bagging, Adaboost and other varia-

0-7803-8740-6/05/$20.00 ©2005 IEEE.

tions based on simpler combination schemes such as majority vote, threshold voting, averaged Bayes classifier, max/min rules, and linear combinations of posterior probabilities [8,9]. More sophisticated approaches have also been proposed, including ensemble based variations of DS, neurofuzzy systems, stacked generalization and hierarchical mixture of classifiers [10-14].

A useful addition to this list would be a general structure containing the ability to combine classifier outputs for (i) a stronger overall classifier, (ii) a classifier capable of incremental learning, and (iii) a classifier capable of data fusion.

The algorithm Learn++, described here, provides such an alternative. We had previously introduced the ensemble based Learn++ for general classification and incremental learning problems [15]. In this paper, we specifically investigate the data fusion capability of the algorithm in extracting and combining complimentary information provided from different channels of EEG recording, on a unique application to detect the earliest neurodegenerative changes of AD. We are interested in determining the sensitivity, specificity, positive and negative predictive value of this approach, in distinguishing patients with AD from cognitively normal elderly subjects.

II. METHODOLOGY

A. Test Subjects and Clinical Evaluation

This study will include a total of 80 subjects, half normal, half with AD, 50 of whom will be used to train the automated classification system, and the remaining 30 will be used to evaluate the system performance on previously unseen signals. 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 [16]. All subjects received a through 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.

B. Acquisition of Event Related Potentials

The ERPs were obtained using the auditory oddball paradigm [2]. Binaural audiometric thresholds were determined for each subject using a 1 kHz tone. Auditory stimuli were presented to both ears using stereo earphones at 60dB above each individual's auditory threshold. The stimulus consisted of tone bursts 100ms in duration, including 5ms inset and offset envelopes. Tones of 1 and 2 kHz 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 200ms duration. A total of 1000 stimuli, including frequent 1000Hz (n=650), infrequent 2000Hz 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 2 kHz tone. With frequent breaks, 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 tin electrodes embedded in a plastic cap, using linked mastoids as reference. Artifactual recordings were identified and rejected. The remaining potentials were amplified, digitized at 256Hz/channel and stored. The ERPs were then lowpass filtered, averaged (40~90 oddball tones per patient), notched filtered at 59-61Hz and baselined with the prestimulus interval. 45 subjects have been recruited so far, however, data from 10 were excluded from further analysis due to low signal to noise ratio. Of the remaining 35 patients (μ_{Age}=77) 15 were AD patients and 20 were cognitively normal individuals.

C. Multiresolution Wavelet Analysis

Since ERPs are nonstationary signals, a time-frequency technique is a natural choice. The discrete wavelet transform (DWT) has versatile properties in data compression, time-frequency localization, noise suppression, and prior successful record in analyzing EEG signals, all with modest computational expense. We have tried several types of wavelets, including Daubechies, quadratic b-spline wavelets, etc. and have found that Daubechies with 4 vanishing moments provided best overall performance [17] in prior single channel studies. Since DWT is now well-established, excellent references are readily available, such as those in [18].

D. The Learn++ Algorithm

The novelty of Learn++ is in its incremental learning capability. It can learn new information as and when new data become available, without forgetting the previously acquired knowledge and without requiring access to previous data. Specifically, Learn++ generates an ensemble of classifiers for each new database that becomes available, where the decisions of individual classifiers are combined through weighted majority voting. We recognize that data fusion and incremental learning are conceptually similar: data fusion also requires learning from additional data, albeit using a different set of features with each dataset; leading us to investigate the feasibility of Learn++ on this application. The pseudocode of Learn++ is given in Figure 1.

For each database, using a different Feature Set, FS_k, $k = 1,...,K$, submitted to Learn++, the inputs to the algorithm are (i) S_k, training data $\boldsymbol{x_i}$ along with their correct labels y_i; (ii) a supervised algorithm BaseClassifier, generating individual classifiers (henceforth, hypotheses); and (iii) an integer T_k, the number of classifiers to be generated for the k^{th} database.

Input: For each dataset from a different source

- Sequence of m_k examples $S_k=[(x_1,y_1),(x_2,y_2),\ldots,(x_{mk},y_{mk})]$.
- Weak learning algorithm **BaseClassifier**.
- Integer T_k, specifying the number of classifiers.

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

Initialize $w_1(i) = D_1(i) = 1/m_k, \quad \forall i, \quad i = 1,2,\cdots,m_k$ **(1)**

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

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

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

If $\varepsilon_t > ½$, discard h_t and go to step 2.
Otherwise, compute normalized error as $\beta_t=\varepsilon_t / (1-\varepsilon_t)$.

4. Obtain composite hypothesis from weighted majority voting

$$H_t = \arg\max_{y\in Y} \sum_{t:h_t(x)=y} \log\left(1/\beta_t\right) \quad \textbf{(4)}$$

5. Compute the error of the composite hypothesis

$$E_t = \sum_{i:H_t(x_i)\neq y_i} D_t(i) \quad \textbf{(5)}$$

6. Set $B_t = E_t/(1-E_t)$, and update the weights:

$$w_{t+1}(i) = w_t(i) \times \begin{cases} B_t, & if \;\; H_t(x_i) = y_i \\ 1, & otherwise \end{cases} \quad \textbf{(6)}$$

Call weighted majority voting and **Output** the final hypothesis.

Figure 1. Learn++ algorithm pseudocode

Each hypothesis h_t is trained on a different subset of the training data. This is achieved by initializing a set of weights for the training data, $\boldsymbol{w_t}$, and a distribution $\boldsymbol{D_t}$ obtained from $\boldsymbol{w_t}$. According to this distribution a training subset TR_t is drawn from the training data at the t^{th} iteration of the algorithm. The distribution $\boldsymbol{D_t}$ determines which instances of the training data are more likely to be selected into the training subset TR_t. Unless a priori information indicates otherwise, this distribution is initially set to be uniform, giving equal probability to each instance to be selected into the first training subset. At each subsequent iteration t, the weights previously adjusted at iteration $t-1$ are first normalized to ensure a legitimate distribution $\boldsymbol{D_t}$ (step 1). Training subset TR_t is drawn according to $\boldsymbol{D_t}$ (step 2) and the BaseClassifier is trained on TR_t. A hypothesis h_t is generated, whose error ε_t, is computed on the entire database S_k as the sum of the distribution weights of the misclassified instances (step 3).

We require that the error ε_t be less than ½ to ensure that a minimum reasonable performance can be expected from h_t. If $\varepsilon_t < 0.5$, h_t is accepted and the error is normalized to obtain the normalized error. If $\varepsilon_t \geq 0.5$ then the current hypothesis is discarded, and a new training subset is selected (return to step 2). All t hypotheses generated thus far are then combined using a voting scheme to obtain the ensemble decision, called the composite hypothesis H_t (step 4).

In the voting scheme used by Learn++, each hypothesis is assigned a weight inversely proportional to its normalized error. Hypotheses with smaller training error, indicating better performance, are given higher voting weights and thus have more say in the final decision. The error of the composite hypothesis H_t is then computed in a similar fashion (step 5) as the sum of the distribution weights of the instances misclassified by H_t. The normalized composite error B_t is obtained in step 6, which is then used for updating the distribution weights assigned to individual instances.

Equation (6) indicates that the distribution weights of the instances correctly classified by the composite hypothesis H_t are reduced by a factor of $B_t<1$. Effectively, this increases the weights of the misclassified instances making them more likely to be selected to the training subset of the next iteration. We note that this weight update rule, based on the performance of the current ensemble, facilitates incremental learning: when a new dataset is introduced, the existing ensemble is bound to misclassify instances carrying previously unlearned knowledge. The weights of these instances are therefore increased, forcing the algorithm to focus on learning novel information introduced by the new data.

For data fusion applications, voting weights for each ensemble are adjusted before final voting, based on expected or observed training performance of each data source: if reliable prior information is available about the individual feature set (e.g., we may know that Pz data are more reliable then Cz data), a higher weight can be given to classifiers trained with that feature set, otherwise, the adjustment can be based on the training performance of the ensemble on that feature set. If such a strategy is chosen, the weight of each classifier would be multiplied by the adjustment factor of the feature set to which it belongs. This adjusted weight is then used during the voting for the final hypothesis H_{final}

$$H_{final}(\mathbf{x}) = \arg\max_{y\in Y} \sum_{k=1}^{K} \sum_{t:h_t(\mathbf{x})=\mathbf{y}} \log\left(\frac{1}{\beta_t \alpha_k}\right) \quad \textbf{(7)}$$

where, α_k is the adjustment factor assigned to the ensemble trained on the k^{th} feature set. In this study, α_k was chosen as the misclassification ratio of the last H_t on S_k:

$$\alpha_k = \left(\sum_i \left[\!\left[H_{T_k}(\mathbf{x}_i) \neq y_i \right]\!\right]\right) \Big/ m_k, \; i = 1,\cdots,m_k \quad \textbf{(8)}$$

with $[\![\cdot]\!]$ evaluating to 1, if the predicate holds true.

III. Results

We have used data from three channels, Pz, Fz and Cz to be fused by Learn++. Initial results indicated that the Fz electrode (frontal region, furthest away from the parietal region), did not provide any performance improvement, and hence was later removed from analysis. The features were four DWT coefficients from each electrode (Cz and Pz), corresponding to the frequency band of 2 ~ 4 Hz and time interval of 100 ~ 500 ms after stimulus, where the P300 is known to reside. A multilayer perceptron (MLP) type network with an error goal of 0.001 and 50 hidden layer nodes was used as the BaseClassifier in Learn++. 7 classifiers were generated for each ensemble. Leave-one-out cross validation was used to estimate the true generalization performance of the algorithm, each of which was repeated 40 times and averaged in order to draw statistical conclusions about the effectiveness of data fusion.

Fig. 2 provides average classification performance over 40 trials, along with their respective 95% confidence intervals

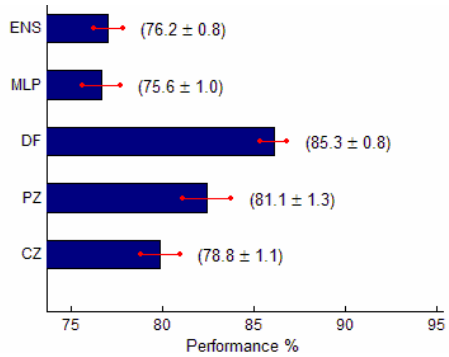

Fig. 2. Comparison of overall generalization performances

obtained by using (from bottom to top): the Cz node alone, the Pz node alone, data fusion (DF) of Cz and Pz using Learn++, a single MLP trained on the concatenated features of Cz and Pz, and an ensemble of 7 classifiers trained on the concatenated features (ENS). Several observations can me made from Figure 2: the Pz channel provided better diagnostic performance then Cz, as expected, but the difference is not significant at 95% confidence. The data fusion performance however, is better then each, and the non-overlapping confidence intervals of data fusion performance over others indicate that the improvement is statistically significant. Also, neither a single MLP, nor an ensemble, trained with concatenated Cz and Pz data performs well, indicating that concatenation is not a good data fusion approach. The sensitivity (SN), specificity (SP), positive predictive value (PPV) and the negative predictive value (NPV) for the diagnosis using data fusion are provided in Table 1.

TABLE 1: DATA FUSION DIAGNOSTIC PERFORMANCE

	Mean	95% Confidence interval width
SN	79.5%	±1.6%
SP	89.9%	±0.6%
PPV	85.5%	±0.7%
NPV	85.5%	±0.9%

Similar to the classification performance, the sensitivity, specificity, positive predictive value and negative predictive for Learn++ based data fusion were all found to be better (with statistical significance at 95% confidence) than those obtained with the individual Cz and Pz channels alone, or either of the concatenation based fusion (MLP and ENS). These results are not shown here for space considerations.

IV. DISCUSSIONS & CONCLUSIONS

Based on the results presented above, we make the following observations: (i) using wavelet analysis to extract features of the ERPs, followed by Learn++ based data fusion appears to be an effective tool for early diagnosis of AD. The approach is non-invasive, cost-effective, can be made readily available to community clinics, since EEG recording technology is well established and widely available; (ii) the approach seems to meet or exceed the current performances of community based clinical evaluations; (iii) the data fusion performance is significantly better than the individual electrodes and other classification schemes, in terms of performance, sensitivity, specificity, positive predictive value and negative predictive value; indicating that Cz electrode does carry complementary information; (iv) unlike Learn++ based data fusion, concatenation of features on their own is not effective for data fusion; (v) we have also tried several BaseClassifier architectures and error goals, and Learn++ is quite invariant to minor changes in these parameters. Therefore the approach is expected to be a stable and effective one, once the remaining patients are recruited and their signals are integrated into the knowledge base of the algorithm.

ACKNOWLEDGMENTS

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

REFERENCES

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

[2] 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.

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

[4] 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.

[5] T. Demiralp *et al*. "Analysis of functional components of P300 by wavelet transform," *Proc. of IEEE Eng. in Med. & Bio..,* vol. 20, no.4, pp. 1992-1995, 1998.

[6] S. Yagneswaran, M. Baker, A. Petrosian, "Power frequency and wavelet characteristics in differentiating between normal and Alzheimer EEG, *Proc. of IEEE Eng. in Med. and Bio.*, vol. 1, pp. 46-47, 2002.

[7] B. Jansen *et al.* "An exploratory study of factors affecting single trial P300 detection," *IEEE Tran. Bio. Eng.,* vol. 51, pp. 975–978, 2004.

[8] L.I. Kuncheva, *Combining Pattern Classifiers, Methods and Algorithms*, Hoboken, NJ: Wiley Interscience, 2004.

[9] J. Kittler, M. Hatef, R.P. Duin, J. Matas, "On combining classifiers," *IEEE Trans on Pat. Anal. Machine Intel.,* vol. 20, pp. 226-239, 1998.

[10] L.O. Jimenez, A.M. Morales, A. Creus, "Classification of hyperdimensional data based on feature and decision fusion approaches using projection pursuit, majority voting and neural networks," *IEEE Trans Geoscience and Remote Sensors*, vol. 37, no. 3, pp 1360-1366, 1999.

[11] G.J. Briem, J.A. Benediktsson, and J.R. Sveinsson, "Use of multiple classifiers in classification of data from multiple data sources," *Proc. of IEEE Geo. and Rem. Sensor Sym..* vol. 2, pp. 882-884, 2001.

[12] F.M. Alkoot, J. Kittler. "Multiple expert system design by combined feature selection and probability level fusion," *Proc of the 3rd Intl Conf on FUSION 2000,* vol. 2, pp. 9-16, 2000

[13] L.I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Tran. Patt. Anal. Machine Int.* vol. 24, no. 2, pp. 281-286, 2002.

[14] M.I. Jordan and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, no. 2, pp. 181-214, 1994.

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

[16] 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.

[17] 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.

[18] M. Under, editor, *Gallery at wavelet.org*, 04/07/2005, Available at: http://www.wavelet.org/phpBB2/gallery.php