A GENETIC APPROACH TO SELECTING THE OPTIMAL FEATURE FOR EPILEPTIC SEIZURE PREDICTION

M. D'Alessandro[1], G. Vachtsevanos[1], A. Hinson[1], R. Esteller[2,3], J. Echauz[3], B. Litt[4]
[1]Georgia Institute of Technology, Atlanta, GA, USA; [2]Universidad Simón Bolívar, Caracas, Venezuela;
[3]Intellimedix, Atlanta, GA; [4]University of Pennsylvania, Philadelphia, PA

***Abstract*- The objective of this study is to (1) develop and apply efficient algorithms to simultaneous intracranial electroencephalographic signals recorded from multiple implanted electrode sites to evaluate the spatial and temporal behavior of seizure precursors and (2) to demonstrate the utility of multiple feature and channel synergy for predicting epileptic seizures in patients with mesial temporal lobe epilepsy. Short-term seizure precursors within a 10-minute time period are investigated. The method consists of preprocessing, processing, feature selection, classification, and validation steps. The preprocessing step removes extraneous data and captures the salient signal attributes while maintaining the integrity of the signal. Processing is a three-step approach that includes first-level features extracted from the raw data, second-level features extracted from first level features, and third-level features extracted from second-level features. A genetic algorithm selects the optimal features off-line from a preselected group of features to serve as the input to the classifier.**

***Keywords*- seizure prediction, genetic algorithm, feature selection**

I. INTRODUCTION

In humans, epilepsy is the second most common neurological disorder, next to stroke, with 50 million people worldwide affected by epilepsy. Of these individuals, 25% do not respond to available therapies [1]. There is currently an explosion of interest in predicting epileptic seizures from intracranial EEG (IEEG) that has its roots in experimental and theoretical work first published in the 1970s. Many potentially useful algorithms for seizure prediction have been presented in the literature, but none that take a comprehensive approach to analyzing seizure-free (baseline) as well as preseizure (pre-ictal) periods. Most work in this area also limits analysis to the one or two electrode contacts nearest the region where electrical signs of seizure onset are first recognized, neglecting the idea that seizure precursors may evolve spatially, as well as temporally, prior to electrical seizure onset. Emphases on the seizure focus region, and the lack of adequate statistical validation, warrant studying seizure precursors from a variety of implanted electrode locations recorded simultaneously.

II. METHODOLOGY

The methodology in this research consists of the steps shown in Fig. 1.

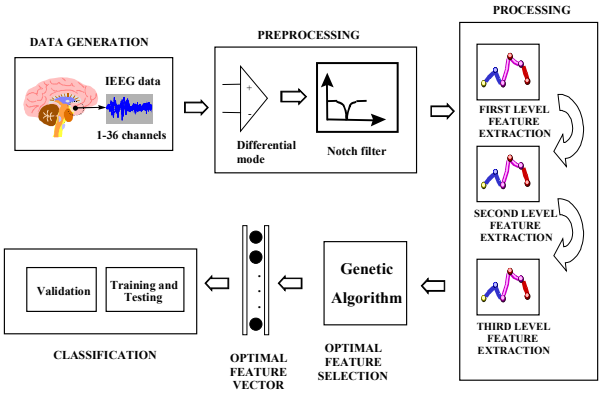

Fig. 1. Methodology for feature selection and epileptic seizure prediction.

A. Data Generation

Collaborating neurologists and neurosurgeons compiled a database of 16 patients affected by mesial temporal lobe epilepsy (MTLE) by storing both video and IEEG signals on Super VHS (SVHS) tapes. To convert the analog signal to digital for further off-line analysis, the IEEG data were copied to compact discs. There are a total of 1770 hours of data available for this research. Patients were monitored during 3- to 14-day hospital stays via video monitoring and EEG and IEEG data collection. Data were collected on a standard Nicolet 5000 video EEG acquisition system utilizing a 12 bit A/D converter and sampled at a rate of 200 Hz with bandpass filter settings of 0.1-100 Hz. Synchronization of video and EEG was achieved and stored for offline analysis of clinical onsets, asleep and awake cycles, and overall patient behavior during the stay. Both viewing the videotapes and looking at the EEG signals identified the patient's state of consciousness. The asleep/awake cycles were correlated with the IEEG data to establish a more complete database. The number of CDs per patient is dependent on the amount of data stored during the patient's pre-surgical evaluation. Each CD contains approximately 8 hours, yielding a total of 8 to 275 hours of data per patient.

In the presented research, fifteen minute records from all IEEG channels were clipped from the original raw data to address the ten minute prediction horizon. Both baseline and ictal records were created from the database. Clipping 10 minutes before the seizure onset and 5 minutes after the seizure onset created the ictal records. The 15 minute baselines were clipped according to the following criteria: 1) each record at least three hours from the onset or termination

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25OCT2001	N/A	-

Title and Subtitle A Genetic Approach to Selecting the Optimal Feature for Epileptic Seizure Prediction	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Georgia Institute of Technology Atlanta, GA	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 0949-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes Papers from the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 4	

of another seizure or any unknown activity; 2) each seizure must be a lead seizure (only the first in a cluster of seizures is used); 3) at least four brain regions are monitored. The three hour criteria is based upon recent results indicating that at least this temporal separation is required to observe seizure precursors [5]. A total of six patients were analyzed, comprising 39 preseizure/seizure records and 105 baseline records.

B. Preprocessing

All IEEG signals serve as inputs to be preprocessed. The preprocessing step captures the salient signal attributes, and maintains the integrity of the signal. To minimize the common mode artifact while maintaining the integrity of the signal, the bipolar signal is found by subtracting spatially adjacent channels to obtain the differential mode signal. In this research, the contents of one channel include all potentials at the given electrode recorded referentially. After removing the common mode artifact, a 60 Hz digital notch filter is used to remove line noise.

C. Processing

Processing is a three-step approach that includes first-level features extracted from the raw data, second level features extracted from first level features, and third-level features extracted from second-level features.

Table 1. First level features.

Feature	Equation
Curve Length	$CL(n) = \sum_{k=n}^{n+N} x(k-1) - x(k)$
Energy	$E(n) = \frac{1}{N} \sum_{k=n}^{n+N} x(k)^2$
Nonlinear Energy	$NE(n) = \sum_{k=n}^{n+N} x^2(k) - x(k-1)x(k+1)$
Spectral Entropy	$SE(m) = -\sum_{k=m}^{m+N} P(k) \log_2 P(k)$
Sixth Power	$SP(n) = \frac{1}{N} \sum_{k=n}^{n+N} x(k)^6$
Energy of Wavelet Packets (db4)	2^n wavelet packets 2^{2^n-1} possible combinations

Candidate features were selected based on criteria described in [2], expertise, observations, and our understanding of EEG signal characteristics. We desire a subset of first-level features that will extract uncorrelated information and are capable of real-time implementation. Intuitively we expect better performance using features from different domains. Based upon our research efforts, and a thorough review of the literature, the features presented in Table 1 were selected as first-level features. All first level features selected are implementable in real time. After evaluating the signals from the extracted first-level features, visual analyses of the feature plots and class-conditional probability distribution functions (PDFs) were examined. This visual analysis paved the way for selecting second-level features.

The second and third level features are identified in the third and fourth segments of the genetic algorithm chromosome shown in Figure 2.

D. Window length

To evaluate most features, it is important to maintain stationarity of the data segment. Statistical tests reveal quasi-stationarity of the EEG signal anywhere from one second (200 points) to several minutes [3]. Esteller et al. have suggested methods to optimize the window length in such experiments; however, to optimize the processing window when analyzing signals spatially and temporally would be a monumental and impractical task. For one patient for example, to optimize the window length could potentially mean 6*22 = 132 different window lengths provided we had a different window length for each first level feature and each channel. This is impractical.

Because seizures spread so quickly, a displacement as small as possible that does not provide too much variability is desired. We experimented with values ranging from 0.25 seconds to 5 seconds and observed that a displacement of 500 points and the window length to 2000 points should provide reasonable propagation resolution of seizure precursors and the ability of multi-channel analysis to effect prediction. These values are used for all tests and are in line with the definition of stationarity found in the literature and preliminary prediction results.

E. Genetic Feature Selection

Both an exhaustive search and genetic approach were considered for the feature selection stage. After a few trial runs with several patients, we found the genetic algorithm to provide optimal results by testing a maximum of 850 features compared to over 4300 features calculated when using the exhaustive search approach. The genetic algorithm applies a novel adaptive chromosome to select the best among some 4300 features to serve as inputs to a classifier. Initially, the genetic algorithm generates 48 random chromosomes and compares their performance using Fischer's Discriminant Ratio (FDR) as the objective function. A subset of approximately 70% of data are used for selecting the optimal features, while the remaining 30% of data are reserved for testing. The algorithm compares each preseizure training record with each baseline training record and takes the

average of the FDR values as the objective values. The resultant chromosomes are weighted based on their fitness values and the roulette wheel selection method is used to select surviving features. The probability of crossover remains constant at 70%, while the probability of mutation remains at 10%. A constrained crossover approach permits crossover within each subchromosome, and prohibits crossover across subchromosomes. That is, for each iteration, only one element within the channel, first level, second level, or third level subchromosomes may crossover at a time. The genetic algorithm chromosome is shown in Figure 2.

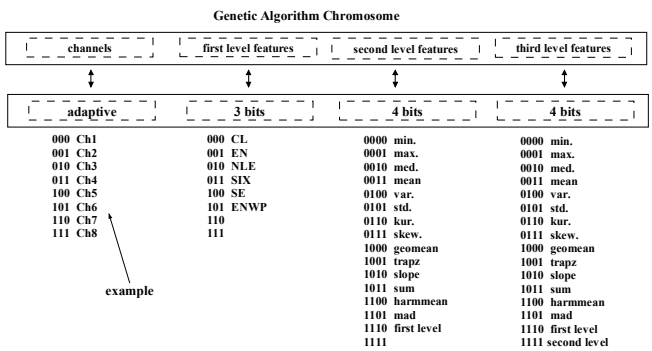

Fig. 2. Genetic algorithm chromosome.

The gene population consists of all combinations of features identified in Figure 2. The first 3 to N (variable depending on the number of channels) bits of the chromosome represent each of the channels. The next 3 bits represent the six first level features, while the next 4 bits represent the fourteen second level features and the option to choose the first level only. The last 4 bits represents the 12 third level features. The unused chromosome expressions are associated with an active penalty term if an unassigned chromosome is selected.

F. Classification and Validation

The classifier is in the developmental stages. The classifier will put the output of the feature vector into the class "preseizure" or "no preseizure". The classifier used will be determined after sufficient results are achieved, scatter plots are analyzed, and a reasonable assessment can be made regarding the best means for classifying the two data classes.

In her work, Esteller found the Probabilistic Neural Network (PNN) to most adequately provide class separability [4]. The PNN will be considered in this work; however, because the objective in this research is to predict rather than detect the UEO, the class definition used by Esteller, and perhaps the PNN itself, will not be adequate.

The literature review found that most research groups in the seizure prediction field provided limited validation in their results. Validation using split sample or "hold-out" techniques will be used in this research. To use split-sample validation, a representative sample (test set) of the data is randomly selected and is not used in any way during training.

After training, the network will be run on the test set. The resultant error will be the unbiased generalization error. In situations where the data sets are too small to justify using split-sample techniques, cross validation and bootstrapping will be considered as alternative techniques to validate the network.

III. RESULTS

The genetic algorithm was run on six patients in the database, to determine the optimal feature combination. The data analyzed included 39 preseizure and 105 baseline records. The genetic algorithm was run on each first level feature combination and each of the 32 wavelet packet energies, for a total run time of approximately 40 hours per patient. The best feature combinations for all first level feature runs were tabulated and results analyzed. The best feature combinations were patient specific. None of the six patients resulted in the same optimal feature combination. Furthermore, the focus channel was never selected as the best channel. Table 2 identifies the best feature combinations for each patient analyzed.

Table 2. Best feature combinations.

Patient	Best Channel	First level feature	Second level feature	Third level feature
3	15-16 (LIF3-4)	sixth power	minimum	trapz
4	5-6(LT5-6)	sixth power	minimum	minimum
8	10-11(RT4-5)	sixth power	minimum	sum
9	1-2(LT1-2)	energy of the wavelet packets	trapz(integrator)	minimum
15	26-27(RIT2-3)	energy of the wavelet packets	median	maximum
16	23-24(LIT3-4)	curve length	mean	mean

The genetic algorithm compared two classes: baseline data and preseizure data. A trial run was conducted with one patient using only the awake baseline records. The results revealed clear distinguishability between the preseizure and baseline records. The asleep and awake baseline records were incorporated and the genetic algorithm run again for each feature combination. Although separability between the two classes was revealed, a clear degradation in performance was observed when the asleep baseline records were included. Both asleep and awake records were included in the genetic algorithm runs for the remaining five patients.

IV. DISCUSSION

To date, most research has analyzed the focus channel since it appears that the focus channel is the channel from which the seizure generates. Only the accumulated energy has given promising results for seizure prediction when

evaluating the focus channel[5]. However, the accumulated energy feature requires the ability to distinguish between the asleep and the awake states of consciousness.

One dimensional scatter plots were created to observe the class separability and determine the need to combine features for classification. Two dimensional scatter plots revealed increased class distinguishability. Figure 3 depicts a one dimensional scatter plot for the best derived feature for one patient analyzed. Figure 4 depicts a two dimensional scatter plot for the same patient including the best derived energy of the wavelet packet feature and the best derived curve length feature. Distinguishability is evident in the one dimensional scatter plot, and improved when two derived features are used. Figure 5 displays the frequency response for the best wavelet packet for the same patient. The center frequency for this wavelet packet is around 58 hertz, in the gamma frequency band. Generally, the frequencies of interest in the IEEG signals range from 1-30 Hz, while gamma frequencies (30-60 Hz) have been observed at the cellular level and are generally stimulus dependent. This finding warrants further study and investigation of the higher frequency bands.

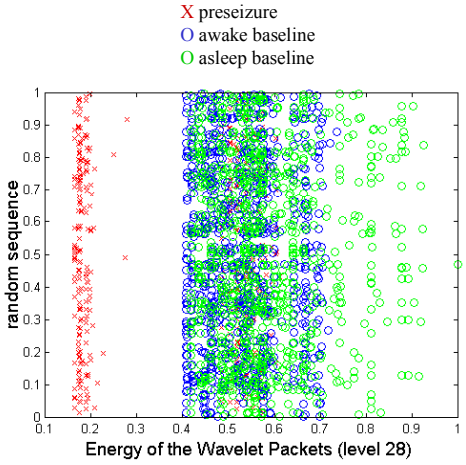

Fig. 3. One dimensional scatter plot for patient 9.

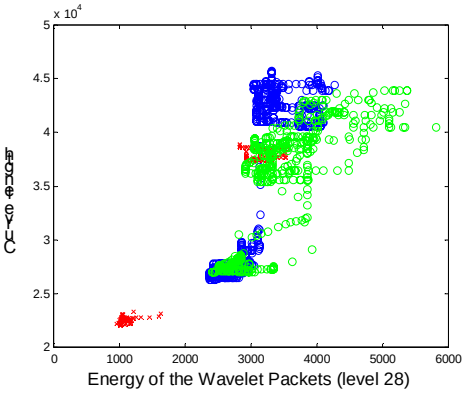

Fig. 4. One dimensional scatter plot for patient 9.

The optimal feature vector is then selected using a forward sequential approach. The neural network classifier is trained to identify inputs as preseizure or no-preseizure. The process is then validated using split sample techniques. The analysis is conducted on six patients consisting of 39 preseizure records and 105 baseline records.

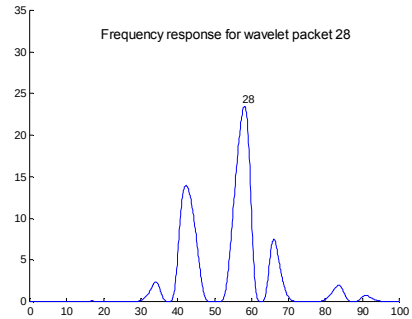

Fig. 5. Best wavelet packet energy for patient 9.

V. CONCLUSION

The optimal features selected by the genetic algorithm were different for each patient analyzed, suggesting that a patient specific predictor is necessary for prediction. The results provide additional evidence regarding the necessity to separate asleep and awake records for optimal performance. The next step is to implement a classifier and apply a forward sequential approach to select the best feature vector among the derived optimal features.

ACKNOWLEDGMENT

This work is supported by funding from the Whitaker foundation, Epilepsy Foundation, American Epilepsy Society, University of Pennsylvania Research Foundation, Jim Jacoby, and the National Institutes of Health grant #MH-62298RO1.

REFERENCES

[1] Epilepsy Foundation of America, Landover, MD, *Epilepsy Facts & Figures*, http://www.efa.org/education/facts.html. 1999.

[2] S. Blanco et al., "Stationarity of the EEG Series," *IEEE Engineering in Medicine and Biology*, vol. 14, no. 4, pp. 395-399, 1995.

[3] T. Masters, *Neural, Novel, and Hybrid Algorithms for Time Series Prediction*. New York: John Wiley and Sons, 1995.

[4] R. Esteller, *Detection of Seizure Onset in Epileptic Patients from Intracranial EEG Signals*. Ph.D. Dissertation, Georgia Institute of Technology, 2000.

[5] B. Litt et al., "Epileptic Seizures May Begin Hours in Advance of Clinical Onset: A Report of Five Patients," *Neuron*, vol. 30, 1-20, April 2001.