Influence of Combination of Criteria for Extraction of Features on the Classification of Biological Signals.

Ángel GUTIÉRREZ

Department of Computer Science, Montclair State University

Montclair, NJ 07043, USA

and

Alfredo SOMOLINOS

Departmento de Matemáticas, Universidad de Oviedo,

33007 Asturias, Spain

ABSTRACT

Expert doctors use the shape of the principal components of the Brain Stem Auditory Evoked Potential (BSAEP) signal to diagnose patients with multiple sclerosis. The diagnosis involves the estimation of the effects of the disease on the form of the waveform components of BSAEPs. Since these components are localized in time and frequency a packet wavelet decomposition of the signal is used to compress it. The information obtained by the packed wavelet can be use to feed artificial neural networks (ANN) with Radial Basis Functions for the same purpose of obtaining a diagnosis. Due to the paucity of data, the signals must be preprocessed. From the hundreds or thousands of wavelet coefficients, only eight are selected using different criteria. Those are used to train an artificial neural network with radial basis functions. We have found that if we combine some of the selection criteria to differentiate sick and healthy people, only one combination of criteria provided better results that using each criterion alone, and other combinations worked better only with some wavelet bases.

Keywords: Principal Components, Packet Wavelets, Neural Networks, Radial Basis Functions, Biological Signals and Multiple Sclerosis.

INTRODUCTION

Multiple Sclerosis is a progressive disease in which the body attacks its own central nervous system, gradually destroying myelin, the substance that surrounds nerve fibers, thereby damaging sites in the central nervous system. Symptoms vary according to the sites where this damage occurs [1]. The disease is difficult to diagnose, especially in its early stages, because the initial symptoms are usually passing and may not last long. Therefore the criteria to establish a diagnosis of clinically definite multiple sclerosis (MS) include two conditions. First, a reliable history of at least two episodes of neurological deficit, and second, objective clinical signs of lesion at more than one site within the Central Nervous System. Since the disease affects the way signals are transmitted in the brain, a recording of the reaction of the brain to external stimuli should reflect the presence of plaques, or areas of demyelization.

Thus, Brain Stem Auditory Evoked Potentials have been used by doctors in the diagnosis of MS [2]. They have a percentage success rate of 80% when diagnosing the general population, but their rate is higher, 95.7%, for determining that a person is a healthy one [3]. When asked about how they made the diagnosis, they very often find difficult to state the rules they use to reach their conclusions, because BSAEP signals can be very different from one patient to another.

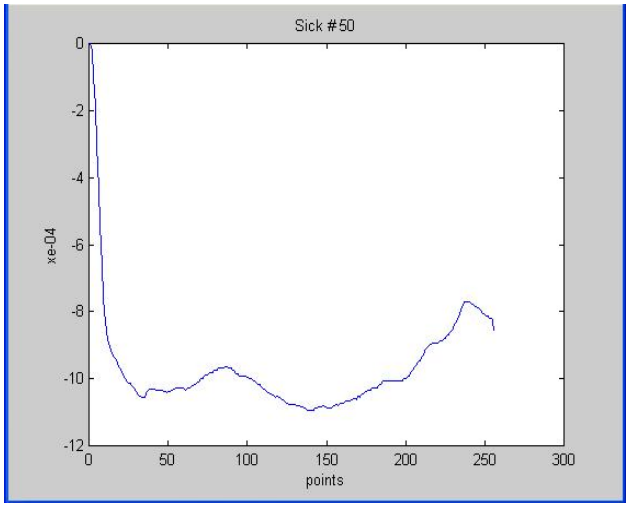

Figure 1

Figure 1 shows an example of a signal corresponding to a sick person that we have numbered as 50.

You may compare the previous signal with the one shown on Figure 2, corresponding to a different sick person, who has been cataloged as patient number 53.

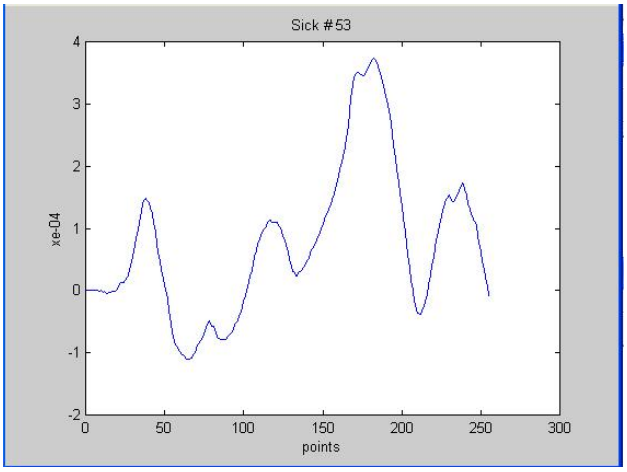

Figure 2

Doctors told us that the relevant features would involve the relative position of peaks and not their absolute value. A Fourier expansion of the signals would allow us to classify the potentials according to their frequency, but would lose the phase information. Therefore we have used the wavelet transform that can be easily implemented and it is time localized as well as frequency localized [4], [5], [6].

In this paper we describe how we have used ANN, which would learn to distinguish between normal and abnormal signals corresponding to BSAEPs. We selected the Radial Basis Function architecture for the ANN, because it has yielded the most consistent results, in the sense that changes in the wavelet basis had no major influence in the results, and we really wanted to check the influence of the criteria used in the preprocessing of the data [7].

We have a set of 197 BSAEP signals, obtained from the Hospital Ramon y Cajal, Madrid (Spain), where 123 belong to patients diagnosed with multiple sclerosis and 74 are normal signals, i.e., corresponding to healthy people. To deal with this scarcity of data to train the networks, we must limit to eight the dimension of the input space to our Radial Basis Function artificial network, which contains two hidden nodes, and a bias node. See figure 3.

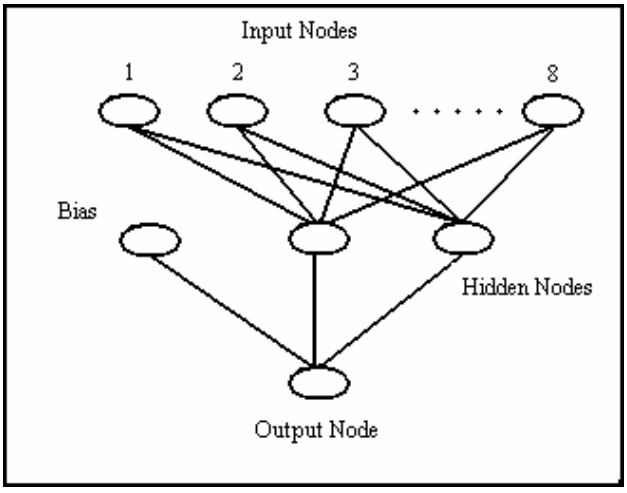

Figure 3

There are different criteria to select the 8 most significant coefficients from the wavelet transform [8], [9], [10]. Here we use a combination of pre-established criteria. The average results obtained range from 76.41% success rate, the worst case corresponding to the case of using a single criterion, to a success rate of 87.34%, for a case of combined criteria. In general the network recognizes the sick people more accurately than the healthy people. This result offers a complement to the doctors' diagnosis.

As usual we divide the signals into a training set and a validating set. With the weights assigned to the ANN by the training set, the neural network correctly classifies signals, on the validating set, which it has never seen before. Due to the shortage of data, for a given choice of wavelets and selection criteria, we trained the network removing one BSAEP signal from the training set, and then we used the left-out pattern to check the accuracy of the trained network. This was repeated until we had used each signal as a left-out one.

SELECTION CRITERIA

The selection criteria that we have used for the extraction of features are the following:

PC criterion
Compute the principal components of a set of coefficient for all the signals, and select the most significant ones. This particular criterion will be the one used to make combinations.

LC criterion
Compute the sum of the absolute value of each coefficient for all the signals, and select a preset number of those coefficients with the largest sum.

LC- criterion
Compute the sum of the absolute value of each coefficient for all the signals of only the sick people, and select a preset number of those coefficients with the largest sum.

WC criterion
For each coefficient we take as null hypothesis H_0: The distributions of the sample corresponding to the normal signals and the sample corresponding to the abnormal signals are identical. We use the Wilcoxon rank sum non-parametric test to find the probability of obtaining the two samples when the null hypothesis is true. We sort the coefficients by the value of this probability, and select those coefficients with the lowest values.

KS criterion
In this case we use the Kolmorof-Smirnov test to obtain the most significant coefficients

PREPROCESSING OF DATA

Our original signals were of different time-length. After truncation to a common time, the number of data ranged from 350 to 490. We made the signals of a uniform length of 256 data

points, using spline interpolation. All signal manipulations, before submission to any of the ANN, have been done using MATLAB 5.3, including the spline, statistics, neural networks and wavelet toolkits.

From the many wavelet bases offered in MATLAB, we have used: all biorthogonal bases (bior11- bior68) bases, all Coiflets bases (coif1-coif5), the first 10 Daubechies bases (db1-db10) and the 7 first Symlets bases (sym2- sym8).

EMPIRICAL RESULTS

We run the training of the ANN with 8 inputs obtained in the following way:

Simple PC criterion
Select the best 8 coefficients using the PC criterion.

Simple LC criterion
Select the best 8 coefficients using the LC criterion

Combined LC and PC criteria
Select 30 coefficients using the LC criterion, and then apply the PC criterion to reduce the number to 8.

Simple LC- criterion
Select the best 8 coefficients using the LC- criterion.

Combined LC- and PC criteria
Select 30 coefficients using the LC- criterion, and then apply the PC criterion to reduce the number to 8.

Simple KS criterion
Select the best 8 coefficients using the KS criterion.

Combined KS and PC criteria
Select 30 coefficients using the KS criterion, and then apply the PC criterion to reduce the number to 8.

Simple WC criterion
Select the best 8 coefficients using the WC criterion.

Combined WC and PC criteria
Select 30 coefficients using the WC criterion, and then apply the PC criterion to reduce the number to 8.

Table 1 summarizes the average percentage of success rate in the diagnosis of sick and healthy people for the different mentioned criteria and the different families of wavelet decomposition bases.

Looking at the table, we can see that the combination of the WC and PC criteria gave better results than just using any one of them, except for the biorthogonal family of bases, for which the KS criteria gave better overall results. Unfortunately in this case, there was no consistency on obtaining better results for the sick people than for the general population, as happened when the LC & PC, LC- & PC and WC & PC combined criteria were used.

	Biorxx	Coiflets	Daub	Symlets
PC	85.77	86.01	85.80	85.86
LC	85.22	83.83	85.65	85.79
LC & PC	85.53	85.49	85.65	85.64
LC-	85.32	83.21	85.75	84.83
LC- & PC	85.49	85.59	85.49	85.71
KS	87.05	85.80	85.62	86.31
KS & PC	86.15	87.15	84.92	86.16
WC	76.41	81.76	80.88	80.02
WC & PC	86.49	87.15	86.37	87.34

Table 1 Success Rate for Diagnosis of MS

It should also be noted that the KS criterion had a higher standard deviation than the same combined criteria, when considering each family of wavelet bases. See Table 2, which shows the standard deviation of all the criteria for the different bases

	Biorxx	Coiflets	Daub	Symlets
PC	1.89	0.00	0.36	0.39
LC	2.41	2.02	1.85	0.99
LC & PC	0.99	0.00	0.25	0.25
LC-	1.61	1.49	1.49	0.63
LC- & PC	1.17	0.23	0.35	0.28
KS	2.05	1.62	2.33	1.82
KS & PC	2.44	2.18	2.15	1.15
WC	6.27	4.42	6.76	3.68
WC & PC	1.90	0.23	1.20	1.81

Table 2 Standard Deviation per Family of Bases

Excluding the KS-PC combination, all the other combinations gave better results, when comparing them with the results of using the LC, LC- or WC criterion alone, for all families of wavelet bases. As for the best family, the Symlets were the ones with the higher average success rate, except in the case of the mentioned KS & PC combination.

In the same Table 2, we can also see that the combination LC & PC and LC- & PC were the ones that gave the least oscillation on the percentage of success rate for all wavelet bases. The same two combinations were also the ones that gave best success rates in diagnosing the sick people, as it is shown in table 3, where the success rate shown corresponds only to the sick people for the same criteria and the same family of wavelet decomposition bases.

	Biorxx	Coiflets	Daub	Symlets
PC	89.32	87.96	87.96	88.03
LC	87.32	87.48	87.80	87.57
LC & PC	90.19	89.75	90.32	90.47
LC-	87.66	85.85	87.89	86.41
LC- & PC	90.08	89.92	90.08	90.59
KS	88.45	86.18	86.83	87.34
KS & PC	89.27	90.24	88.94	90.01
WC	77.20	83.40	82.20	81.00
WC & PC	89.81	90.41	89.27	90.36

Table 3 Success Rate for Diagnosing Sick People

CONCLUSIONS

The higher success rates for a general diagnosis were found mostly in the combination of the Wilcoxon and the principal components criteria. When we look at the specific diagnosis of sick people, we see that the combined criteria always exceed the success rate of any of the individual ones, except in the case of the KS & PC combination. This same combination of criteria, when considering the Biorxx and the Coiflets families, was also the exception to the fact that all the combined criteria give more homogeneous general results within a particular wavelet basis family.

This special behavior of the KS & PC combination of criteria, for these two families, also appears when we look at the success rate for diagnosing sick people. The two highest success rates were found on the wavelet basis Bior39, 93.50%, and on the wavelet basis Coif4, 92.68%. But for the Biorxx family, the standard deviation on the population of success rates for sick people has a value of 3.52 and for the Coiflets family the value is 4.10. These are values far greater than those found on the same diagnosis, for other combination of criteria and other wavelet families.

Since the doctor(s) recognize better the healthy people, we could adjust the design of the ANNs to respond even better to detecting sick people. We are working on this approach. We are also working to see how we could use independent component analysis [11] to select the most discriminating coefficients before we submit the data to the ANNs, or use this criterion as a new one on a combination of criteria similar to the ones we had done on this paper.

Although we have developed a procedure of combining different criteria to better extract features for the diagnosis of multiple sclerosis, it can be applied to other signals with similar characteristics, especially when there is a limited number of data available.

REFERENCES

[1] A. Blinowska, J. Verroust and D. Malapert, "Bayesian statistics as applied to multiple sclerosis diagnosis by evoked potentials", **Electromyogr. Clin. Neurophysiol**., Madrid, 32(1-2), 1992, pp.17-25.

[2] C.M. Poser, D.W. Paty, L. Scheinberg, W.I. MacDonald, F.A. Davies, G.C. Ebers, K.P. Johnson, W.A. Sibley, D.H. Silberberg and W.W. Tourtellotte, "New diagnosis criteria for multiple sclerosis: Guidelines for research protocols", **Ann-Neurol**., 13, 1983, pp. 227-231.

[3]. J. Fernandez Plaza, A. Canales, J. Dorronsoro, V. Lopez Martinez and J. Siguenza, "Reconocimiento de potenciales evocados de tronco mediante redes neuronales", in **Segundas Jornadas Nacionales de Informatica de la Salud,** Madrid, 1993, pp. 233-239.

[4] Charles K. Chui, **Wavelets: A Mathematical Tool for Signal Analysis**, Philadelphia, SIAM, 1997.

[5] J. Raz and B. Turetzky, "Wavelet Models of Event-Related Potentials", in **Wavelets in Medicine and Biology**, A. Aldroubi and M.Unser eds. CRC Press 1996. pp. 571-590.

[6] M. Akay, **Detection and Estimation Methods for Biomedical Signals,** New York, Academic Press Inc., 1996.

[7] C. Fernandez-Garcia, A. Gutierrez and A. Somolinos, "Diagnosis of Multiple Sclerosis using Radial Basis Functions", **Proceedings of the IASTED International Conference on Modelling and Simulation**, Pittsburgh, PA, May 13-16, 1998, pp. 3-6.

[8] C. Fernandez-Garcia, A. Gutierrez and A. Somolinos, "Selection of wavelet coefficients for neural networks used in Medical Diagnosis", **Proc. of the Summer Computer Simulations Conference in Arlington**, VA, (SCS, San Diego, 1997).

[9] C. Fernandez-Garcia, A. Gutierrez and A. Somolinos, "The Role of Non Linearity in Neural Networks used for Medical Diagnosis", **Proceedings of the 1999 Health Sciences Simulation Conference,** S. Francisco, CA, January 17-20, 1999, pp. 200-205.

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, "**Numerical Recipes in C, The Art of Scientific Computing**", 2nd Ed., Cambridge, Cambridge University Press, 1994.

[11] Mark Wachowiak, Renata Smolíková, Georgia D. Tourassi, and Adel S. Elmaghraby, "Separation of Cardiac Artifacts from EMG Signals with Independent Component Analysis", **BIOSIGNAL 2002**, Brno, Czech Republic, June 26-28, 2002, pag.23.