Influence of the Training Methods in the Diagnosis of Multiple Sclerosis Using Radial Basis Functions Artificial Neural Networks

Angel GUTIÉRREZ
Department of Computer Science, Montclair State University
Montclair, NJ 07043, U.S.A.

ABSTRACT

The data available in the average clinical study of a disease is very often small. This is one of the main obstacles in the application of neural networks to the classification of biological signals used for diagnosing diseases. A rule of thumb states that the number of parameters (weights) that can be used for training a neural network should be around 15% of the available data, to avoid overlearning. This condition puts a limit on the dimension of the input space.

Different authors have used different approaches to solve this problem, like eliminating redundancy in the data, preprocessing the data to find centers for the radial basis functions, or extracting a small number of features that were used as inputs. It is clear that the classification would be better the more features we could feed into the network.

The approach utilized in this paper is incrementing the number of training elements with randomly expanding training sets. This way the number of original signals does not constraint the dimension of the input set in the radial basis network. Then we train the network using the method that minimizes the error function using the gradient descent algorithm and the method that uses the particle swarm optimization technique.

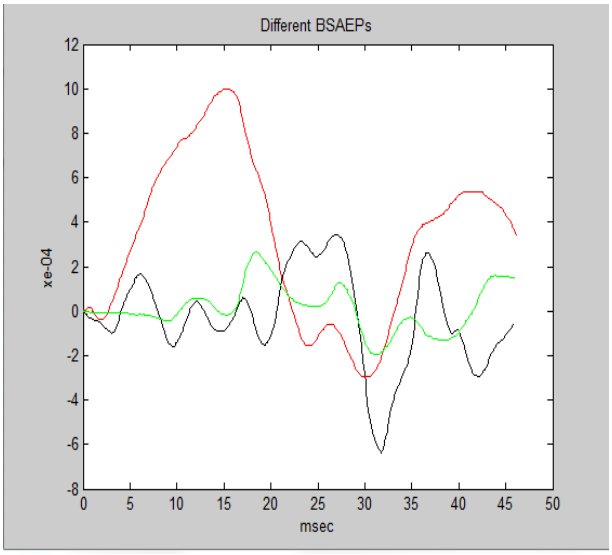
A comparison between the two methods showed that for the same number of iterations on both methods, the particle swarm optimization was faster, it was learning to recognize only the sick people. On the other hand, the gradient method was not as good in general better at identifying those people.

Keywords: Neural Networks, Radial Basis Functions, Particle Swarm Optimization, Signal Processing, Wavelets, Health Sciences, Multiple Sclerosis.

1. INTRODUCTION

Doctors utilize BSAEP to diagnose patients with multiple sclerosis. MS can reveal, among other symptoms, a decrease of the wave V amplitude, an increase on absolute latencies and interpeak interval latencies I-III, I-V, III-V. But the border between pathological and normal values sometimes is not well defined [1]. Doctors very often find it difficult to state the rules they use to reach their conclusions, and their success rate is higher for healthy people than for sick people. It should be noted that the biological signals studied in this paper are Brain Stem Auditory Evoked Potentials (BSAEP) for the diagnosis of Multiple Sclerosis, the techniques that we applied to them could be easily applied to study any time series related to the evolution of biological parameters. For instance, they could easily be used dealing with VEP, ECG's, EEG or EMG's potentials, [2] - [8].

The relevant features in a BSAEP would involve the relative position of peaks and not their absolute value. Figure 1 shows the BSAEP of one of the healthy people, who is called healthy # 25, one of the sick people, called sick # 3., and another one of a patient called sick #6.



Green: Healthy # 25, Black: Sick # 3, Red: Sick # 6

Fig.1. Different BSAEP signals

The BSAEP of a sick people and a healthy one could look sometimes very similar, see Figures 2 and 3. But other times the shape of the signal is completely different for sick people, as Figure 3 and 4 shows.

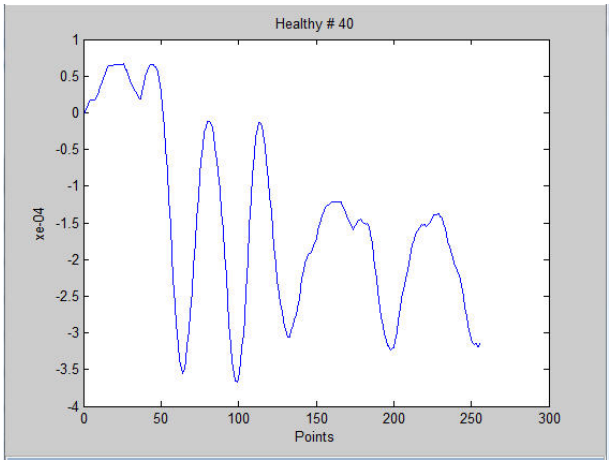


Figure 2: Healthy # 40

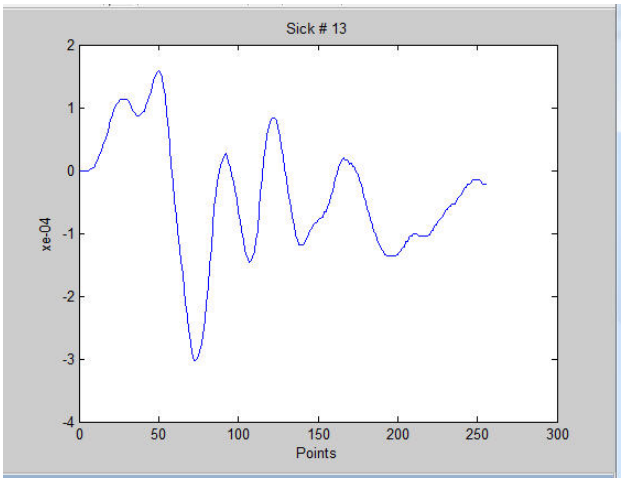


Figure 3: Sick patient # 13

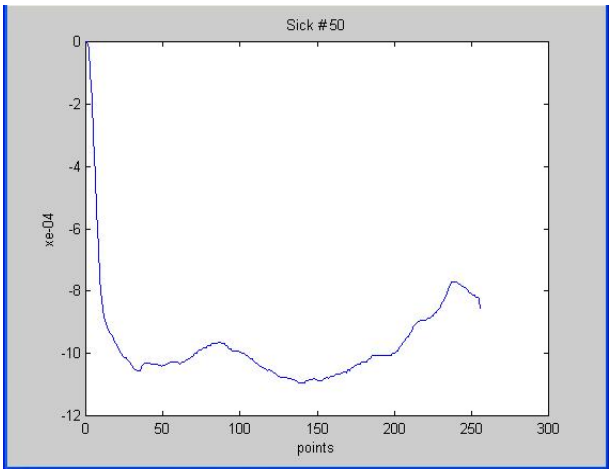


Figure 4: Sick patient # 50

We work with an artificial neural network, that uses radial basis functions, but is trained using two different methods. One method tries to minimize the error function using the gradient descent algorithm with decreasing learning rates, by locating where the gradient is equal to zero [9]. The other method uses the particle swarm optimization technique [10], [11], [12].

But before we feed the signals to the neural networks, we preprocessed and compress them. The preprocessing begins by using the same time interval for all signals [13]. Then we normalize them [14]. Since we need to identify the instants in time at which certain events occur, we use wavelet transforms for the compression [15]. We assume that they will be able to capture the amplitude and relative position of the peaks of the signals, information that doctors use for their diagnosis.

For this compression we use all the 37 different wavelet transforms found in MATLAB that allow us to capture the decrease of the wave V amplitude, and an increase of interpeak interval latencies.

Once they have been compressed, the author selected a small number of the most significative features, according to the Kolmogorov-Smirnov statistical criteria, following the ideas found in a previous paper [16]. These selected coefficients were then used as inputs. It is quite clear that the classification would be better the more features we could feed into the network.

We have a set of 193 BAEP signals, obtained from the Hospital Ramon y Cajal, Madrid (Spain), where 70 are normal signals, i.e., corresponding to healthy people, and 123 belong to patients diagnosed with multiple sclerosis. Small samples impose a limit on the number of parameters that can be learned by neural networks. In this paper we first increment the number of training elements, using randomly expanded training sets [17] and we use them to train the radial function network, following the ideas on [18], [19].

Clustering algorithms were used previously to find centers and radii for the radial basis functions [20], [21]. The availability to generate an arbitrary number of samples removes not only the need to find centers and radii, but also the constraint that the number of original signals places on the dimension of the input set of the network. For each neuron we can determine the coordinates of the center (the same number as the inputs), the radius and the output weight. Thus, an n input network, with m radial functions, would require the fitting of m*(n + 2) parameters. This implies that the computing time will be in the order of m*n, but it will also depend upon the number of iterations performed in the training. So we still must select, from the hundreds of wavelet coefficients, only a handful of them and they must be the coefficients that contain the most significant features [22]. We use these networks with different kinds of wavelets and the Kolmogorov-Smirnov test as the criteria for the selection of 25 input coefficients. Our hidden nodes consists of 4 radial basis functions.

Once the radial basis function has been trained with each method, we tested them and recorded our results. The process was repeated seventeen times, and we obtained the mean and standard deviation of all the cases. As a result, we could see that for the same number of iterations on both methods, the particle swarm optimization was faster, but tended to recognize mostly the sick people On the other hand, the gradient method was in general better at recognizing the healthy people.

2. PRE-PROCESSING OF DATA

Expert doctors use the shape of the principal components of the Brain Stem Auditory Evoked Potential (BSAEP) signal to

determine if a person is sick or healthy. This suggests that the wavelet transform of the BSAEP could be used to capture the features that determine if a person is sick or healthy with the help of a neural network.

We have a set of 193 BSAEP signals. The signals were taken from 84 people with multiple sclerosis, using the left and/or right hemisphere and from 35 healthy volunteers, using both hemispheres. The signals for the sick people, where obtained from the Hospital Ramon y Cajal, Madrid (Spain). These signals where acquired from people that complied with the criteria needed to establish a diagnosis of clinically definite multiple sclerosis (MS): A reliable history of at least two episodes of neurologic deficit, and 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 existence of the disease. Thus doctors can diagnose the disease using BSAEP.

When doctors diagnose this disease they often find it difficult to state the rules they use to reach their conclusions. We aim to help them with the diagnosis using an artificial neural network with radial basis functions in the hidden nodes

In order to work with the signals, we digitized them using a scanner, and restricted them to a common time (the minimum of all of them). Then we generated analog signals using cubic splines. Finally we selected 512 equidistant points, from the analog signals. After this process was done, we applied all the discrete wavelet transforms found in MATLAB to the set of 512 points already obtained. Since we need to identify the instants in time at which certain events occur, we use wavelet transforms because we assume that they will be able to capture the amplitude and relative position of the peaks of the signals, information that doctors use for their diagnosis.

It is impossible to feed the coefficients supplied by the wavelet transforms directly into a neural network. It is clear that the more features we could feed into the neural network, the better the classification would be.

Therefore we increment the number of training elements, using randomly expanded training sets [17]. We generate 579 new signals, with the same proportion of sick and healthy people as in the original set, i.e. 369 for sick people and 210 for the healthy ones. In fact for For each of the two clusters corresponding to sick and healthy people, an estimation of the values for the elements in the probability density function, $f_{kME}(z)$, also denoted as $N_k(\mathbf{U}, R)$, k =1,2 Eq.(1), that maximized the differential entropy for that cluster, were computed.

$$\mathbf{N_k}(\mathbf{U}, \mathbf{R}) = \frac{1}{(\sqrt{2\pi})^{n+1} |\boldsymbol{R_k}|^{\frac{1}{2}}} \, e^{-\frac{1}{2}(z-U_k)^T R_k^{-1}(z-U_k)} \qquad (1)$$

Here z denotes an input-output data vector, U_k is the mean vector of the cluster k, R_k is the covariance matrix of the same cluster, $|R_k|$ is its determinant, and T denotes the operation that performs the vector transpose operation. We represent the estimation of the mean vector as $\hat{U}_k$ and of the covariance matrix as $\hat{R}_k$, where a diagonal load was added to insure its invertibility

With this information, data were drawn for each cluster using the formula given in Eq. (2)

$$Z^i = \hat{U}_k + \hat{L}_k s^i \qquad (2)$$

where s^i is an independently identically distributed (i.i.d.) vector sequence drawn from N(0,1), and $\hat{L}_k$ is the Cholesky lower triangular matrix from the decomposition of $\hat{R}_k$.

Then from the hundreds of wavelet coefficients, we select 25 coefficients using the Kolmogorov-Smirnov test. .

3. NEURAL NETWORK ARCHITECTURE

The radial basis function network architecture used for this work can be seen in Fig. 5. There are *n* input nodes in the fanout layer, *m* nodes and a bias in the hidden layer, and one output node.

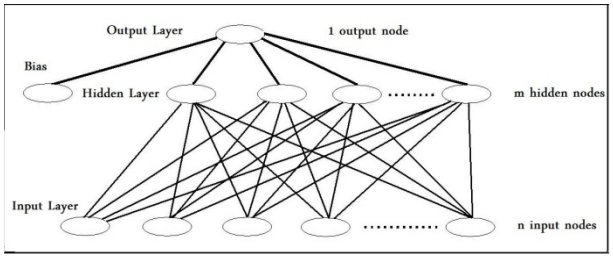


Figure 5: Radial basis function neural network

The value of *n* is 25, as the number of most significant coefficients selected. As for *m* we used 4, so the total number of free parameters is 105, well within the range of the 15% to 20% of the number of training elements.

The network was trained using the 37 different wavelet bases offered in MATLAB: all biorthogonal bases (bior11- bior68), all Coiflets bases (coif1-coif5), the first 10 Daubechies bases (db1-db10) and the 7 first Symlets bases (sym2- sym8).

The input-output space of our data requires that all the values of every coefficient on our sample, are normalized, with mean zero, and standard deviation of one. This avoids the problem that the output values, being far greater than any of *n* inputs in the case of sick people, could dominate the making of the partitions and in doing so, defeat the purpose of the algorithm. The mean value for each coefficient, and the corresponding standard deviation should be kept, to be utilized for the normalization of any future input vector that needs to be tested.

For the method of minimizing the error function using the gradient descent algorithm, each training process consisted of 10,000 random presentations, beginning with different random values. In this case the learning rates η(k) for the centers, the radii and the weights were given by the linear function

$$\eta(k) = \eta_0 + (\eta_1 - \eta_0) * \frac{k}{NPR} \qquad (3)$$

where *k* is the iteration step, NPR is the number of presentations, η_0 is the initial learning rate, set at 0.001, and η_1 is the final rate, set at 0.08. These values for the initial and final learning rate for both the hidden and input layers were known to be acceptable.

Wavelet	Average	St. Deviation
bior11	56%	0.01
bior13	29%	0.07
bior15	59%	0.16
bior22	85%	0.03
bior24	51%	0.10
bior26	95%	0.00
bior28	13%	0.03
bior31	1%	0.00
bior33	55%	0.13
bior35	23%	0.06
bio37	63%	0.10
bior39	61%	0.01
bior44	93%	0.00
bior55	19%	0.02
bior68	34%	0.11
coif1	27%	0.07
coif2	32%	0.08
coif3	35%	0.09
coif4	20%	0.05
coif5	95%	0.23
db1	12%	0.03
db2	49%	0.12
db3	80%	0.04
db4	65%	0.16
db5	6%	0.01
db6	31%	0.07
db7	35%	0.03
db8	93%	0.23
db9	33%	0.08
db10	92%	0.22
sym2	8%	0.02
sym3	60%	0.03
sym4	25%	0.06
sym5	20%	0.05
sym6	55%	0.07
sym7	72%	0.02
sym8	69%	0.17

Table 1: Gradient Method

Wavelet	Average	St. Deviation
bior11	100%	0.00
bior13	100%	0.00
bior15	70%	0.26
bior22	99%	0.00
bior24	99%	0.00
bior26	98%	0.01
bior28	53%	0.29
bior31	100%	0.00
bior33	100%	0.00
bior35	100%	0.00
bio37	100%	0.01
bior39	99%	0.00
bior44	100%	0.00
bior55	100%	0.00
bior68	100%	0.00
coif1	100%	0.00
coif2	100%	0.00
coif3	100%	0.00
coif4	100%	0.00
coif5	91%	0.06
db1	100%	0.00
db2	78%	0.42
db3	98%	0.01
db4	99%	0.01
db5	100%	0.00
db6	100%	0.00
db7	100%	0.00
db8	100%	0.00
db9	100%	0.00
db10	100%	0.00
sym2	19%	0.25
sym3	100%	0.00
sym4	100%	0.00
sym5	100%	0.00
sym6	100%	0.00
sym7	100%	0.00
sym8	98%	0.01

Table 2: Particle Swarm Optimization

For the method of minimizing the error using the particle swarm optimization we used 10 particles, and each of them updates its position and its velocity 1000 times.

Once the artificial neural networks were trained, we checked the results with our original set of data, and recorded the general rate of success and the corresponding rates for sick and healthy people. We repeated the process 17 times for each method and for each wavelet.

4. EMPIRICAL RESULTS

After all the training had occurred, we took the average and standard deviation of all the samples. Tables 1 and 2 shows the results for diagnosis for sick people using the gradient algorithm and the particle swarm optimization algorithms for training, respectively. Both tables have the same structure. Each row corresponds to the success rates for a particular wavelet basis whose name appears in the first column. The second column reflects the general success rate for recognizing the sick people and the third column of the table shows the standard deviation corresponding to the sample of trainings.

Looking at table 1, it is worth noting that in the case of the particle swarm optimization, although the average is very high in 33 of the wavelet decomposition, the values of the standard deviation are very high for the other 4 cases, with values of 0.25 (sym2), 0.26 (Biorthogonal 15), 02.9 (Biorthogonal 28) and 0.42 (Daubechies 2). On the other hand, in table 2, there are only 4 cases with very high average, and all of them, except one, have the highest values of the standard deviation, although not as high as in table 1. In fact, these standard deviations are 0.22 (Daubechies 10), 0.23 (Coiflet 5) and 0.23 (Daubechies 10).

It is worth noting that two wavelet that performed poorly according to the results in table, 2, (Biorthogonal 28 and Symlet 2) also performed poorly,, with averages of 13% and 8%, as shown in table 1

There are other samples that were computed, but due to the lack of space they are not shown. In the conclusion some of their properties will be discussed.

5. CONCLUSIONS

Radial basis function networks had been used to diagnose Multiple Sclerosis. They provide an automatic, fast and reliable way to discriminate the signals from sick and healthy people. But it seems that the results differ according to the method used for the training of the neural network. But since this was the result of only one specific network architecture, with a specific method of expanding the training set, further investigation is needed to determine if this result is similar when we use a different artificial network, and/or expand the set of training elements applying a different technique, and/or we use a different random generator that MATLAB supplied.

To answer these questions we should first allow to modify the number of hidden nodes. This will increment the number of centers and radii, and it will constraint the number of input nodes. We could probably assume that the first coefficients that discriminate more are enough to convey most of the $\hat{R}_k$ information, and selecting a larger number does not enhance the learning of the network. But on the other hand, we should be careful when the number of hidden nodes is so large that the number of input nodes goes below 8.. It seems that in this case we will not be able to capture enough discriminating features of the input space. [19].

Another point to highlight is that table 1 has a great variety of averages, and standard deviations. Table 2 shows small differences in the averages, and most of them with very low or zero standard deviation. But in this last case, the standard deviations reach big values while using some wavelet, as mentioned above.

For future research, we could compare these results to those obtained by using a different statistical discriminating criterion, like the largest sum of the absolute value of the coefficients, the principal components analysis, the Wilcoxon rank sum test, or Shannon's entropy.. We could also apply the expanding of the training set according to [17] using the original values of the 512 points instead of the coefficients of the discrete wavelet transform. Or we could increment the number of training elements using white noise applied to the original signals. We could also change the number of hidden nodes, with the corresponding variation of the number of input nodes, to avoid overlearning. With respect to the artificial neural network that we have used, we could investigate if the removal of the bias hidden node would affect the result.

We could also use a margin based feature selection criterion and apply it to measure the quality of sets of extracted features [23]. Another possibility is to pass a message between the different particles at various level of training. Finally we could select the even or odd values in the set of original data when they are expanded using cubic splines. This will generate twice as many numbers of start data for the randomly generated expanded training set. Of course we could use a combination of all these approaches to compare the results with those obtained in this paper.

In conclusion we can say that our findings are a good sign that artificial neural networks with radial basis functions could be used to help doctors when they are diagnosing cases of multiple sclerosis

6. REFERENCES

[1] 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.

[2] J.A. Sigüenza, S. González, J.R. Dorronsoro and Vicente López. ``Automatic Classification of Visual Evoked Potentials by Feedforward Neural Nerworks", Artificial Neural Networks, **Proc. of the International Conference on Artificial Neural Networks.** (North-Holand Elsevier T. Kohonen et al Edi.), 1991, pp. 1117 - 1120.

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

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

[5] A. Subasi , M Yilmaz and H. R. Ozcalik, "Classification of EMG signals using wavelet neural network", **Journal of Neuroscience Methods** 156, 2006, pp. 360–367.

[6] 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.

[7] Holdaway, R.M., et alii, "Classification of somatosensory-evoked potentials recorded from patients with severe head injuries", **IEEE Engn. Medicine and Biology**, 9, 1990, pp. 43-49.

[8] Freeman, D.T. "Computer recognition of brain stem auditory evoked potential wave V by a neural network",**Ann Otol Rhinol Laryngol**. 101 (9), 1992, pp. 782-790.

[9] C. M. Bishop, **Neural Networks for Pattern Recognition,** Oxford University Press, 1999

[10] J. Kennedy and R. Eberhart "Particle Swarm Optimization", **Proceedings of IEEE International Conference on Neural Networks (IV),** 1995, pp. 1942–1948.

[11] M. Meissner,, M. Schmuker and G. Schneider, "Optimized Particle Swarm Optimization (OPSO) and its application to artificial neural network training". **BMC Bioinformatics**, **7,** 2006, pp. 125.

[12] M.E.H Pedersen and A.J. Chipperfield, "Simplifying particle swarm optimization". **Applied Soft Computing**, 10 (2), 2010, pp. 618–628.

[13] A. Gutiérrez & A. Somolinos, "Influence of Wavelet Boundary Conditions on the Classification of Biological Signals", **Proceedings of the IEEE 26th Annual Northeast Bioengineering Conference**, Storrs, CT, April 8-9, 2000, pp. 25-26,

[14] C. Fernández-García, A. Gutiérrez and A. Somolinos, Diagnosis of multiple sclerosis using radial basis functions, **Proceedings of the IASTED98, International Conference on Modeling and Simulation**, Pittsburgh, PA, 1998, pp. 3-6.

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

[16] A. Gutiérrez & C. Fernández, "Using a Combination of Artificial Neural Networks for the Diagnosis of Multiple Sclerosis", **Proceedings of the 5th. International Symposium on Bio- and Medical Informatics and Cybernetics**, Orlando, Florida, July 19th- 22nd, 2011, pp. 146-151.

[17] G.N. Karystinos, and D. A Pados, On Overfitting, Generalization, and Randomly Expanded Training Sets, **IEEE Trans. Neural Networks**, 11(5), 2000, pp. 1050-1057.

[18] A. Gutiérrez, "Processing Brain Stem Auditory Evoked Potential for Improving Diagnosis of Multiple Sclerosis", **Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition (AIPR -09)**, Orlando, Florida, July 13-16, 2009, pp. 193-197.

[19] A. Gutiérrez, "Diagnosis of Multiple Sclerosis Using Brain Stem Auditory Evoked Potentials"**, Proceedings of the 13th World Multi-Conference on Systemics, Cybernetics and Informatics**, Orlando, Florida, July 10-13, Vol. III, 2009, pp. 45-50.

[20] A. Gutiérrez & A. Somolinos, "Preprocessing of Brain Stem Auditory Evoked Potentials for Diagnosing Multiple Sclerosis", **Proceedings of the IASTED International Conference on Advances in Computer Science and Technology**, Puerto Vallarta, Mexico, January 23-25, 2006, pp. 196-201.

[21] A. Gutiérrez, Carlos Fernández & A. Somolinos, "Clustering Algorithms for Preprocessing of Data in Diagnosis of Multiple Sclerosis using Radial Basis Functions", **Proceedings of the IASTED International Conference on Modelling and Simulation**, Philadelphia, PA, May 5-8, 1999, pp. 123-127,

[22] A. Gutiérrez and A. Somolinos, Extracting Features for Brain Stem Auditory Potential Signals**, Proceedings of the 8th World Multi-Conference on Systemics,. Cybernetics and Informatics**, Orlando, FL, 2004, pp.140-144.

[23] R.Gilad-Bachrach, A. Navot , and N. Tishby, Margin Based Feature Selection- Theory and Algorithms, **Proceedings of the Twenty-first International Conference on Machine Learning**, Banff, Alberta, Canada, 2004, pp. 43-50.

[24] B**. J. Frey and D. Dueck**, "Clustering by Passing Messages Between Data Points", **SCIENCE**, **315**, 2007, pp. 972-976

[25] V. Elser, I. Rankenburg, and P. Thibault, "Searching with iterated maps". **Proceedings of the National Academy of Sciences USA**, **104**, 2007, pp. 18-42