Evaluation of ANN and SVM for the classification and prediction of patients with diabetic neuropathy

Antonio SalgadoCastillo[1], Tahimy González Rubio [1]
[1]Computer Science Department, [1]Biomedical Engineering Department, Universidad de Oriente, Santiago de Cuba, Cuba. {asalgado@csd, tgonzalez@fie} .uo.edu.cu

Abstract. Diabetic neuropathy is a disease that affects a large proportion of the world population, so that its prevention and early detection is of vital importance at the present time. In this paper we evaluate an ANN and SVM designed using MatLab 7.9.0.529 for the classification and prediction of patients with diabetic neuropathy, using Pulse Waves Sequences of Blood Volume. Efficiency was evaluated taking into account the algorithms and training time as well as effectiveness in classification and prediction. Considering 40 cases in the process of learning and 18 in the validation, the best classification results were obtained with the ANN for an 88.88% effective with the Gradient descent learning algorithm with adaptive learning rate, and the SVM was obtained 72.22% success rate using the Quadratic programming algorithm. In predicting both methods were 100% effective.

Keywords: *SVM, ANN, Pulse Waves, Diabetic Neuropathy*

1 Introduction

Research in the field of neuroscience and artificial intelligence at present, is of vital importance because of the large number of patients with diseases such as diabetes and diabetic neuropathy. In the past 20 years the increase of population with diabetes and neuropathic diseases led the medical community to formulate new methods or tests for early diagnosis of them, all based on the principle that they are the least invasive and painful for the patient, with a high level of confidence, less expensive and reproducible in any laboratory. So it is necessary to create computational tools, able of early diagnosis of the diseases mentioned above, using in this case the Pulse Waves Sequences (PWS) [5], obtained through ANGIODIN PD3000 [3] and whose technique to capture these waves is based in the process of photoplethysmography [1], which is not invasive or harmful to human body.

It should be noted that there are several techniques that allow the classification of the signals obtained, among them we can mention the Artificial Neural Networks (ANNs) [1], [4], and Support Vector Machines (SVMs) [2]. This paper makes an evaluation of ANN and SVM for classification of PWSs of patients with diabetic neuropathy, in order to measure their efficiency and effectiveness, taking into account the following indicators: algorithms, training time and effectiveness in classification and prediction. For this, an ANN training algorithms Gradient Descent with Momentum (TRAINGDM) and Gradient Descent with Learning Rate Variable (TRAINGDA) platform is designed on Matlab 7.9.0.529, and a SVM with Quadratic Programming

algorithms (QP) and Sequential Minimal Optimization (SMO). For the training process the signals of 40 patients were coded: 30 sick and 10 healthy subjects. The proposed tools according to the PWS of a patient, allow classification in healthy or sick, and predict whether they will remain stable or is likely to have the disease under study. For validation tests it used 18 patients (13 sick and 5 healthy).

2 Methodology

The methodology is reflected in the Figure1 showing the phases involved in the development of the research. First, the signal to be processed was coded, then proceeds to training with both techniques and uses them for classification and prediction.

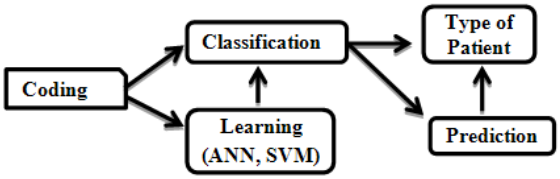

Figure 1. Scheme describing the developmental stages of the computational proposal tool.

ANNs are a tool for statistical modeling, geared toward the recognition of patterns (classification and prediction). An ANN is composed of a large number of highly interconnected processing elements (neurons or nodes) working to solve specific problems. The nodes or neurons are grouped in layers and distinguish among input layer, hidden layer and output layer. Thus, the network supports p input values to provide q output values , depending on the connections between neurons, as well as final values adopted by the weights [4]. In supervised training networks including backpropagation, the training data are constituted by several pairs of training patterns of input and output. Knowing the output implies that the training benefits from the supervision of a teacher.
A SVM "learns" the decision surface of two different kinds of entry points. As a classifier of two classes, the description given by the vector data support is able to form a decision boundary around the domain of the training data with little or no knowledge of those outside the decision boundary. The data are mapped by means of a kernel function to a feature space in a higher dimensional space, where it looks for the maximum separation between classes [7]. This feature border, when brought back to the input space, can separate the data into two classes, each forming a cluster.

2.1 Coding

The methodology for coding the signal before starting the learning process of ANN and SVM are described in the diagram shown in Figure 2.

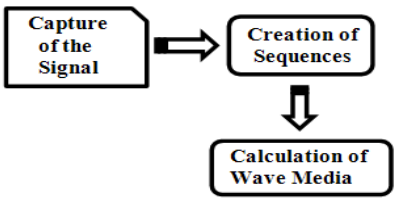

Figure 2. General outline that describes the stages of encoding Waves Blood Volume Pulse.

2.1.1 Signal capture

The Pulse Wave from which the process of classification begins is obtained through the Cuban electronic device ANGIODIN PD3000, where the signal is just the spectrum of PWS, a magnitude biophysics (usually measured in m/s) with pre-clinical and clinical significant value as shown in Figure 3a. Catching the signal is performed until the patient reaches a state called basal. When the signal is captured, each pulse wave applies a mathematical model [6], [9], implemented by another computational module, which returns 12 values that characterize each wave (see Figure 3b).

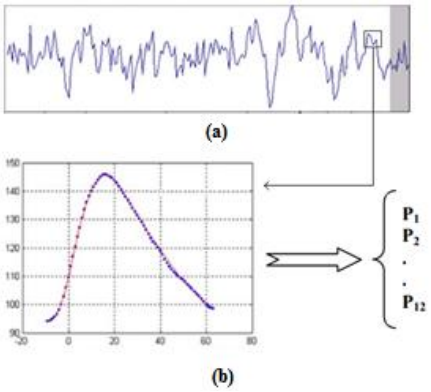

Figure 3. a) Pulse wave signal. b) Wave extracted with 12 trait or characteristic values returned by the math model.

2.1.2 Creating the Wave Pulse Sequences

At this stage, each signal is partitioned into 22 groups of sequences because the obtained average of the number of waves in each signal is 110 (Figure 4), where each sequence has 5 consecutive waves. This process is performed with a completion with zero to signals which do not reach the 110 waves and eliminating signals exceeding this value, guaranteeing the same size as the characteristic vector that serves as input to the ANN and SVM.

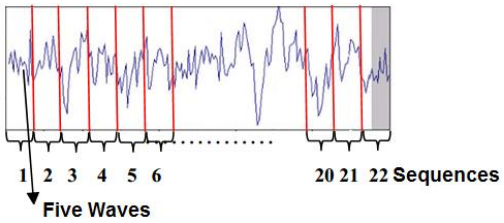

Figure 4. Partitioned signal in the 22 sequences' groups of Pulse Wave Blood Volume.

2.1.3 Calculation of Medium Wave

In this process for each sequence (which consists of 5 waves) the root mean square (1) is calculated, with the aim of extracting a characteristic vector: of every 5 waves a representative wave or feature is obtained. These 22 mean vectors will form the features of a signal wave. Importantly, each vector is represented by 12 features, forming a eigenvector of 264 elements.

$$O_{Media}(i) = \sum_{i=1}^{n}\sum_{j=1}^{m}(V_j(i)/m) \qquad (1)$$

Where n represents the amount features of each vector (n= 12), m the amount vector to be scanned (m = 5) and the calculated average of 12 features.

2.2 Design of ANN and SVM

2.2.1 Training and learning of ANN

The structure which is designed for operation of the network is which has three layers [10]. The first layer has 264 neurons that represent the values of the characteristic vector of each signal obtained in the previous process of coding. The second layer has 150 neurons, and the last layer 1 neuron.
For training the ANN the transfer functions, sigmoidal (tansig) and linear (purelin) were used.

$$\text{tansig: } f(n) = \frac{e^n - e^{-n}}{e^n + e^{-n}} \qquad f(n) = 1 - \left(f(n)\right)^2 \quad f'(n) = (1 - a^2) \qquad (2)$$

$$\text{purelin: } \qquad f(n) = n \qquad f'(n) = 1 \qquad (3)$$

For this work, the learning rate was initialized at $5*10^{-2}$ and error which must converge the algorithm to ensure the effectiveness of the neural network designed is 10^{-6}.

2.2.2 Training and learning of SVM

For training the SVM kernel linear function was used:

$$K(x, x_i) = (x^T x_i) \tag{4}$$

2.3 Prediction

Once the training process is done for both methods, the tools are ready to classify a new patient. A given input vector, if the output is a value near to 0 then the patient is healthy, and if approaches 1 the patient is sick.
After the ANN and the SVM are classifying properly the prediction process proceeds, which is performed only if the patient is classified as healthy, it is important to say that all patients used in this training were performed on the capture of several signals at a remote time interval and it was found that each signal obtained for the same patient, has a very close distance to the other, it is important that doctors make patients prone prediction according to the degree of similarity to other patients who are studied, it means, if the results of the patient seems to indicate that a patient is healthy, after a certain time was classified as sick, then according to this similarity doctors classify the new patient prone, but in the conditions mentioned above this is not true, because the diagnostic tests give results only at advanced stages, so we predict some of these pathologies early in the risk groups existing at present.
To achieve this prediction we have drawn the following diagram (see Figure 6) showing the methodological process used.

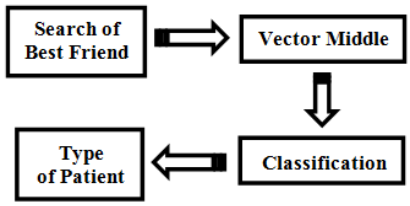

Figure 6. General scheme describing the steps of predicting the pulse waves.

2.3.1 Search of Best Friend

This process leads to the search for the two nearest neighbors to the patient being analyzed, using the Euclidean Distance (5), the search is performed on data from patients stored, as a basis on which we have knowledge of their pathological status. After finding the two nearest neighbors the Middle Vector process is done.

$$d_{ij}(z, v) = \sqrt{\sum_{k=1}^{n} (z_{kj} - v_{ki})^2} \tag{5}$$

2.3.2 Vector Middle

The Middle Vector process is responsible for calculating an average signal (centric vector) between the two nearest neighbors and the patient referred to the study, where the average signal would be a possible vector close to the signal that could show this patient and would give us near information signal about the patient in the future.

2.3.3 Ranking

In this process the classification of the calculated average vector is carried out: in stable or prone, having these outputs the possibility of analyzing two fundamental cases raises. The first case (see Figure 7) is that the two nearest neighbors are found healthy patients and the average vector calculated to be simulated belongs to the class of the stables, then we expect stability in the patient (stable), in the second case (see Figure 7) one of the two nearest neighbors is a sick patient and simulated the middle vector belongs to the class of sick patients then we expect a possible change in the patient. The explanation here is that the patient under study is classified as healthy because the nearest neighbor is healthy that is more representative than the nearest neighbor who is ill at the time of classification, but as expected to signal the patient approaches the middle vector, which was classified as sick, the need arises to give a patient undergoing follow-up study as a prone patient.

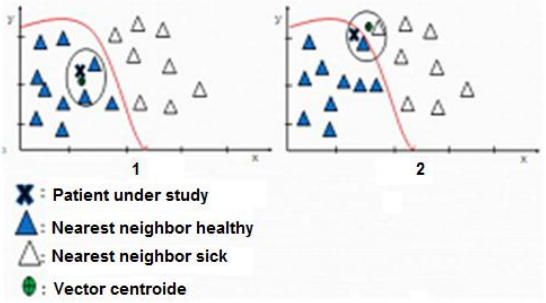

Figure 7: Most important cases in the prediction process.

3 Discussions of results

3.1 Training using ANN and SVM

The training phase of ANN and SVM was performed with 40 signals, of which 30 belong to sick patients and 10 to healthy patients. Table 1 shows the general data of the training of both learning machines and the time response of these for the classification of a new patient. The first method exposed is TRAINGDM where the slowness of convergence is observed, which was reached in 21027 iterations, in a time of 18:53 minutes, converging to the desired value (10^{-6}). The second method is TRAINGDA that after 18797 iterations in a time of 14:09 minutes converges to the value 10^{-6} minimizing the error, which was the main objective of the training. The

third method is the SMO which completes the training in 3:22 seconds, very similar to QP method that took 3:15 seconds. The speed of response that is the delay time to classify a new patient, for the four methods was 1:70 seconds.
From these results we conclude that in our research to the classification of pulse wave signals, TRAINGDA methods for ANN and QP for SVM are the most recommended.

Table 1. Duration of the training and response rates with SVM and ANN.

	Algorithms	Training time (in hours)	Iterations	Answer of time in the validation (in hours)
ANN	TRAINGDM	00:18:53:00	21027	00:00:01:70
	TRAINGDA	00:14:09:00	18797	00:00:01:70
SVM	SMO	00:00:03:22	-	00:00:01:70
	QP	00:00:03:15	-	00:00:01:70

3.2 Validation using ANN and SVM

To validate the effectiveness of the techniques under study it was simulated with 18 signals of which 5 belong to healthy patients and 13 to sick patients. The results returned by the ANN and SVM are shown in Table 2.
As shown, the ANN and SVM approached the values of the PWS to 1 in sick patients while that of the healthy to 0, which shows that the training was successful. However with the TRAINGDA method the ANN achieved 16 patients correctly in classification while working with the TRAINGDM method it achieved 15 patients. The SVM with QP and SMO methods classified correctly only 13 patients. In Table 2 the results correspond to the values misclassified by both techniques.

Table 2. Values returned by the ANN and SVM in the validation process.

ANN				SVM			
TRAINGDA		TRAINGDM		QP		SMO	
Class 1 Sick=1	Class 2 Healthy=0	Class 1 Sick=1	Class 2 Healthy=0	Class 1 Sick=1	Class 2 Healthy=0	Class 1 Sick=1	Class 2 Healthy=0
1.0269	-	0.9662	-	1	-	1	-
0.6859	-	0.6765	-	1	-	1	-
1.7412	-	1.6081	-	1	-	1	-
0.7483	-	0.7441	-	1	-	1	-
0.7486	-	0.8104	-	1	-	1	-
1.0789	-	1.0713	-	1	-	1	-
0.9992	-	0.9995	-	1	-	1	-
1.179	-	1.156	-	1	-	1	-
1.2254	-	1.2208	-	1	-	1	-
0.7017	-	0.7090	-	1	-	1	-
1.0283	-	0.8693	-	1	-	1	-

0.9377	-	0.7932	-	1	-	1	-
0.9515	-	0.8809	-	1	-	1	-
-	0.4555	-	0.3041	-	***1***	-	***1***
-	0.1102	-	0.2147	-	***1***	-	***1***
-	0.4271	-	***0.5038***	-	***1***	-	***1***
-	***0.8928***	-	***0.9703***	-	***1***	-	***1***
-	***0.9999***	-	***1.0541***	-	***1***	-	***1***

Table 3 shows the percentages of effectiveness obtained for each method used. TRAINGDA was achieved with the 5 patients classified healthy 3 well, and all patients successfully, achieving a 88.88% total effective.These percentages may be given by the low number of patients included in this first proposal, however when you consider that for the training there were used 30 patients, in the classification 10 patients, and all patients were correctly classified, for this group we have 100% effectiveness, unlike healthy cases used 10 in training and validation correctly, classified only 3, obtaining a 60% success rate for this group. With the TRAINGDM method for sick patients yielded the same results, but of 5 patients classified as healthy, only 2 correctly for a 40% success rate in this group and a 83.33% effectiveness. Respect to SVM classified the 13 patients sick with a 100% success rate for this group, but in relation to patients, it is scored 0% effective with QP and SMO methods, generally yielded a 72.22% effective with the SVM.

Table 3. Results of the validation process in percentages of effectiveness.

		Results in (%) of patients classified correctly		
	Algorithms	**13 Sick patients**	**5 Healthy patients**	**TOTAL**
ANN	TRAINGDA	100 %	60 %	***88.88 %***
	TRAINGDM	100 %	40 %	83.33 %
SVM	QP	100 %	0 %	***72.22 %***
	SMO	100 %	0 %	72.22 %

3.3 Prediction using ANN and SVM

Table 4 shows the results of the prediction, it is observed that there are values returned by the SVM and ANN that do not correspond with the class to which they belong, at this point it is important to clarify in the case of healthy patients, their behavior can be predicted as stable, which means that their PWS is most similar to a sequence feature of a healthy patient, or can be predicted as a likely meaning that their PWS, although it does not become that of a sick patient , has similar features, but not most. In the case of sick patients these can be classified only in prone, because having the disease its PWS has mostly similar features to a diseased patient. In Table 4 the shaded results correspond to the values that do not belong to the same class in

which they appear, it does not mean that there was a bad prediction as explained above. As can be seen, had obtained for both techniques a 100% effective for evaluating the effectiveness of the prediction can only use ill patients, who constitute the group with diagnosed disease.

Table 4. Values returned by the ANN and SVM in the prediction process.

ANN		**SVM**	
TRAINGDA		**QP**	
Class 1 (Prone=1)	**Class 2 (Stable=0)**	**Class 1 (Prone=1)**	**Class 2 (Stable=0)**
1.0161	-	1	-
0.9191	-	1	-
0.9376	-	1	-
0.9182	-	1	-
0.9189	-	1	-
1.0284	-	1	-
1.0162	-	1	-
1.1126	-	1	-
1.1114	-	1	-
0.9160	-	1	-
1.0324	-	1	-
0.9930	-	1	-
1.0311	-	1	-
-	0.4825	-	***1***
-	0.3480	-	***1***
-	***0.8262***	-	***1***
-	***0.9592***	-	***1***
-	***1.041***	-	***1***

4 Conclusions

Considering the parameters in Table 6 and the analysis developed we arrive at the following conclusions: of the four training methods used in research, the best results taking into account the training time, response time of classification, and the percent effective in the classification were obtained with the TRAINGDA method for ANN, and QP method for SVM.

The prediction by both methods showed the same results, although the SVM training took less than with the ANN, this is not significant comparing with the classification results that show a clear superiority of the ANN over the SVM in terms of effectiveness, which is the main objective of both techniques. This allows us to

conclude on the relevance of using ANN as effective classification technique for PWS in diabetic patients.

Table 6. Comparison between ANN and SVM as the most important indicators for this study.

	Algorithms	Training time (in hours)	Answer of time in the validation (in hours)	Classification	Prediction
ANN	TRAINGDA	00:14:09:00	00:00:01:70	88.88 %	100 %
SVM	QP	00:00:03:15	00:00:01:70	72.22 %	100 %

References

1. Allen J, Photoplethysmography and it application in clinical physiological measurement. Topical Review. Phyfiol. Meas. 28(2007) R1- R39.

2. Betancourt GA, Las Máquinas de Soporte Vectorial (SVMs). Facultad de Ingeniería Eléctrica, Universidad Tecnológica. Abril 2005.

3. Cuadra M, Corzo A, Pascau A, Ferrer O, García JC, Hernández D., Implementation of a medical device for the diagnosis of vascular diseases. Its introduction on the clinical Practice. TELEC'2000 International Conference, Santiago de Cuba,Cuba.

4. Fausett, LV, Fundamentals of Neural Networks: Architecture, Algorithms, and applications. Prentice Hall; US Ed edition (9 April 1999). ISBN-13: 978-0133198492.

5. Loukogeorgakis S, Dawson R, Phillips N, Martyn CN, Greenwald SE., Validation of adevice to measure arterial pulse wave velocity by photoplethysmographic method. Physiol. Meas. 23 (2002) 581–596.

6. Milanés D, Solución a un Modelo Matemático de Programación no Lineal, utilizado en el contorno de la Onda de Volumen de Pulso. *Degree. Thesis*: june, 2010, Universidad de Oriente.

7. Nocedal, J; Wright, SJ. (2006). Numerical Optimization (2nd ed.). Berlin, New York: Springer – Verlag. p. 449. ISBN 978-0-387-30303-1 .

8. Parrella F, Online Support Vector Regression. Department of Information Science, University of Genoa, Italy. Junio 2007.

9. Pascau A, Fernández Britto JE Allen J., Relación de nuevos modelos conceptuales y matemáticos para el contorno de la Onda de Volumen de Pulso Arterial. Revista cubana de investigaciones biomédicas 2011.

10. Rifkin R, Everything Old is New Again: a Fresh Look at Historical Approaches in Machine Learning, *Ph.D. thesis*: 18, 2002.