AUTOMATYKA/ AUTOMATICS • 2012 • Vol. 16 • No. 1

http://dx.doi.org/10.7494/automat.2012.16.1.89

Wiesław Wajs*

Neural Network and Artificial Immune Algorithms for the Classification of Medical Data Series

1. Introduction

We intent to demonstrate that the immune system concept can be used as a computational tool for data classification and prediction. The immune system has several useful ideas from the viewpoint of data manipulation. The immune network theory hypothesizes the activities of the immune cells, the emergence of memory, and the discrimination between own cells, knows as self, and external invaders, knows as non-self. The immune system has an internal image of all existing pathogens, infections non-self, to which it might be exposed to during its lifetime. The artificial immune model consists of a set of cells, named antibodies, interconnected by links with associated connection strengths. There are several immune network models presented in the literature [3]. The immune network models are based upon a set of differential equations to describe the dynamics of the network cells and molecules [5, 13, 9]. The interactions between different types of elements lead to the network connectivity pattern and dynamics. In the model proposed by Varela and Coutinho (1991), it is possible to stress three characteristics of the immune network. The first one is the structure. The structure describes the types of interactions among the network components, represented by matrices of connectivity. The second is the dynamics that accounts for the variation in time of the concentrations and affinities of its cells. The third one is metadynamics, a property related to the conti-nuous production of novel antibodies and the death of non-stimulated or self-reactive cells.

The main characteristic of the immune network theory is the definition of an individual's molecular identity, which emerges from a network organization followed by learning of the molecular composition of the environment where the system develops.

2. Description of Medical Problem

Initial stabilization of the infant state is a difficult task. Analysing huge amounts of data requires experience. The decision process can be verified by comparing it to the model

* AGH University of Science and Technology, Krakow, Poland

of respiratory insufficiency progress carried out by clinical staff. The diagnosis is based on theoretical and empirical knowledge. The proper selection of input parameters is crucial for the accuracy of the model prediction. A short characteristic of parameters to be used as input data for a neural network algorithm is given below. Blood gas values are given as the input parameters of each network forecasting value of: pH, PaO2, PaCO2, and HCO3 (Figs 1–3).

Arterial blood pH means the degree of acidity or alkalinity. pH is a symbol for the degree of acidity or alkalinity of a solution. It is the logarithm of the reciprocal of the hydrogen-ion concentration in gram equivalents per liter of solution. Arterial oxygen pressure PaO2 means pressure of O2, interval from 40 to 60 mmHg. Arterial carbon dioxide pressure PaCO2 means the pressure of CO2, interval from 35 to 55 mmHg. Serum bicarbonate concentration HCO3 describes the concentration of HCO3 in the blood, interval from 21 to 25 mmol/l. We use pH to determine acidosis or alkalosis:

If (pH>= 7.35 and pH =< 7.45) then Normal
If (pH< 7.35) then Acidosis
If (pH> 7.45) then Alkalosis.

We use PaCO2 to determine the respiratory effect:

If (PaCO2<35) then Tends toward Alkalosis,
If (PaCO2>45) then Tends toward Acidosis,
If (PaCO2 >= 35 and PaCO2 =< 45) then Normal.

If (PaCO2<35 and pH < 7.4) then Metabolic Acidosis,
If (PaCO2<35 and pH > 7.4) then Respiratory Alkalosis,
If (PaCO2>45 and pH < 7.4) then Respiratory Acidosis,
If (PaCO2>45 and pH > 7.4) then Metabolic Alkalosis with

Respiratory Compensation,

If (PaCO2 > 35 and PaCO2 < 45 and Normal pH) then Normal Arterial Blood Gases,
If (PaCO2 > 35 and PaCO2 < 45 and Abnormal pH) then Uncompensated Metabolic Acidosis or Alkalosis.

We use HCO3 to verify the metabolic effect:

If (High PaCO2 and Low HCO3) then Acidosis,
If (Low PaCO2 and High HCO3) then Alkalosis.

We assume a metabolic cause when the respiratory one is ruled out:

If (High pH and High PaCO2) then Metabolic,
If (High pH and Low PaCO2) then Respiratory,
If (Low pH and Low PaCO2) then Metabolic,
If (Low pH and High PaCO2) then Respiratory.

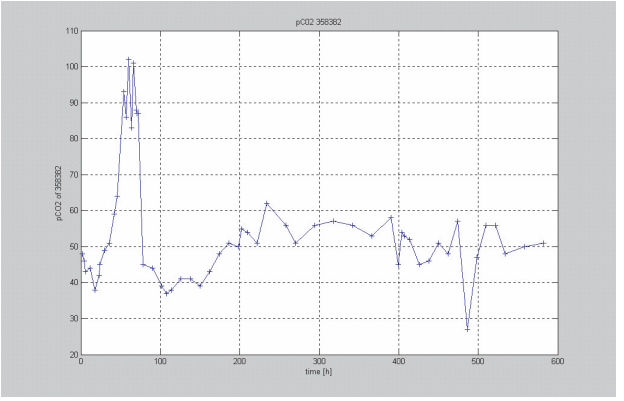

Fig. 1. Function of PaCO2 from time 0 to time 600 h. Infant is No. 358382 in the database

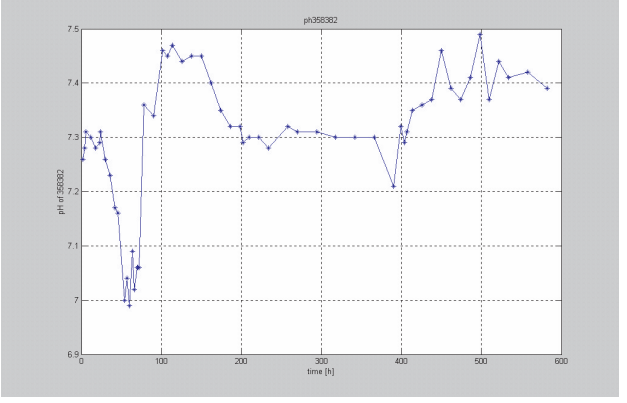

Fig. 2. Function of pH from time 0 to time 600 h. Infant is No. 358382 in the database

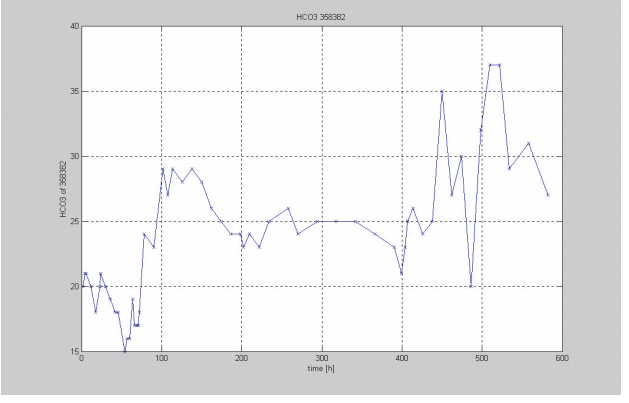

Fig. 3. Function of HCO3 for 0 hrs to 600 hrs duration. Infant is No. 358382 in the database

3. Multidimensional Data Classification

Clustering is useful in several exploratory pattern analyses, grouping, decision making, and machine-learning tasks including data mining [8]. The data clustering approach is implemented in association with hierarchical clustering and graph theoretical techniques. Based upon a set of unlabeled samples {x1, x2, …, xn}, where xi is described by L variables, a network is constructed to answer such questions as [2]: Is there a great amount of redundancy within the data set? Is there any group or subgroup intrinsic to the data? How many groups are there within the data set? What is the structure of these groups? How can we generate decision rules to classify novel samples?

We assume a problem-dependent set of L measurements to characterize a molecular configuration as a point s in a shape – space S. A point in an L – dimensional space, called shape space, specifies the set of features necessary to determine the antibody-antibody and antigen-antibody interactions. This shape set of features that defines either antibody or an antigen, is represented as an L – dimensioned vector. The possible interactions within the Artificial Immune Network are represented in the form of a connectivity graph. The Artificial Immune Network [1] is an edge – weighted graph, composed of a set of nodes, called antibodies, and the set of node pairs called edges with an assigned number called connection strength, associated with each connection edge.

A minimal spanning tree [7, 14] is used to define the number of network clusters, which will be equal to the number of higher peaks in the bar graph plus one, indicating large variations in the minimax distances between cells. The number of clusters can be measured as the number of valleys of the respective histogram. The dendrogram allows us not only to define the number of clusters but also to identify the antibody (node) belonging to each cluster. A dendrogram is defined as a rooted weighted tree where all terminal nodes are at the same path length from the root [15].

The training process of the artificial immune network consists of two phases. The first phase is learning of the artificial immune network. The training data set is comprised of blood parameters (pH, PaO2, PaCO2, and HCO3) in the form of time series of the same length. The second phase is the test. We test the network's generalization abilities using input vectors from the test dataset. The network interprets elements of the training dataset in the form of time series as antigens. An antigen stimulating particular antibody creates an artificial immune network. The time series form allows one to calculate the antigen – antibody affinity. We used an algorithm presented in the paper [3]. The following notation is adopted: $\xi = 30\%$ – percentage of the mature antibodies to be selected, $\sigma d = 1.2$ – natural death threshold, $\sigma s = 0.2$ or 0.01 – suppression threshold. The important parameter, which has a major influence on the immune network structure, is σs. The value of the suppression threshold influences the generalization abilities of the artificial immune network. The greater value of the suppression threshold causes the cells for which the distance is smaller than the suppression threshold to be removed from the network structure. The elimination process realized for those cells improves the generalization abilities of the network. The results of

the training process are shown in Figure 4. The number of valleys over the graph bar should be associated with the number of clusters into which time series of the training dataset is classified. The smaller value of suppression threshold results in a higher number of clusters – classes into which data from the training set are classified. The smaller value of the suppression threshold forces the network to create more and more subtle divisions, but at the same time reduces its generalization abilities, see Figure 5.

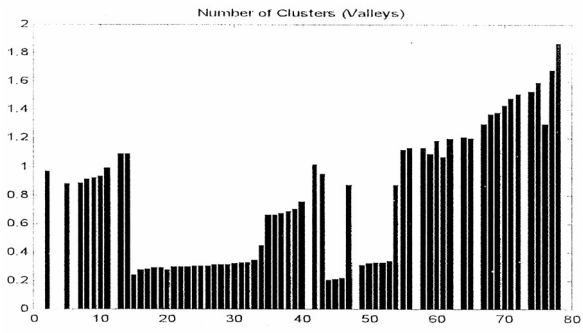

Fig. 4. Bar graph, number of clusters for $\sigma s = 0.2$

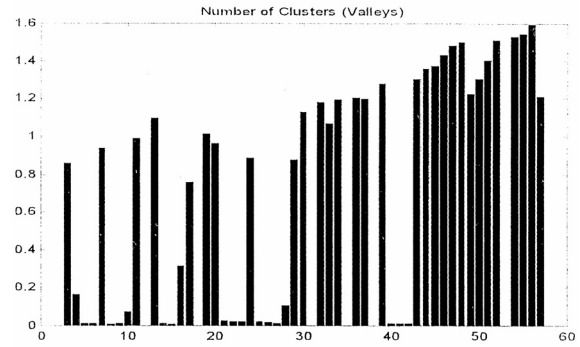

Fig. 5. Bar graph, number of clusters for $\sigma s = 0.01$

4. Prediction Results

Blood gas values and treatment parameters are recorded during a few days of hospitalisation. Neural networks learning process consists of two data sets. The first one contains the data introduced to the first layer of neurones. The second set contains the expected data. The analysis of the input data characteristics revealed that the data should be normalised before entering the first neurone layer. Each sequence of values of the separate input parameter recorded during several days is normalised in a simple way. All input values are

divided by the maximum value. All input parameters are subject to polynomial approximation. The blood gas values can be measured four times a day, or less. The learning vectors are determined with the step 0.1 [h]. Unsatisfactory results are caused by disturbances in the function the representing values of the learning data set and an inappropriate data step.

An original data value is recorded once, or a few times a day. The intermediate values are calculated by the polynomial approximation method.

The second group of input data consists of values that have an influence on blood gas values. Those input data are referred to as medical treatment parameters. There are: birth weight, gestational age, A/A (alveolar-arterial ratio), PDA (patent ductus arteriosus), mean saturation of arterial hemoglobin, standard deviation of saturation of arterial hemoglobin, percent of the time duration for which the saturation of arterial hemoglobin is less then 85%, and the mean pulse rate.

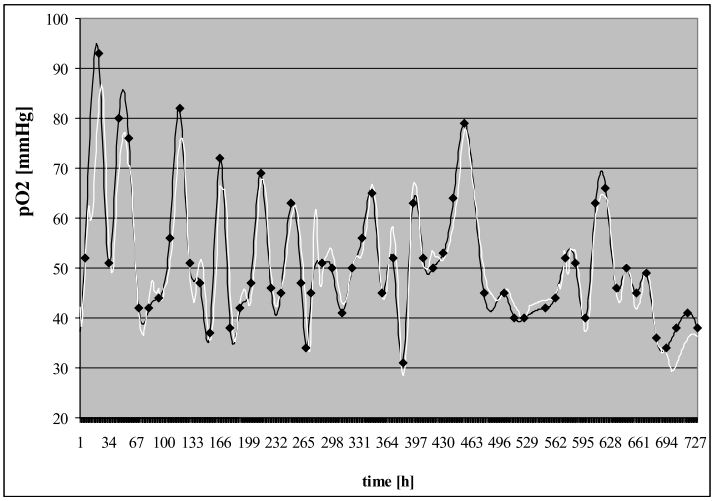

Fig. 6. PaO2 forecast results. Black line depicts the approximated values, the white line depicts predicted values, sign * depicts source value

The values of the other parameters are represented by the setting of the respirator, and there is no need for their approximation. Up to fourteen sequences representing input parameters have been created and each of the sequence has been normalized to improve neural networks learning process of each of them.

It can be seen that results are not quite satisfactory. It is a result of too large a time step and a poor selection of input parameters. The results could be improved by adding more input parameters that influence the blood gases nature. The parameter value can be calculated for a shorter time step. The learning process can be improved thanks to the fact that the decreasing step implies more data vectors. In this case there are about 2500 data vectors. The correctness for selecting a patient's data should be also considered. There is a tendency

to increase the prognostic error at the end of the approximated sequence that is related to the nature of approximation methods implemented for the time series. We achieved good prognoses with an error below assumed the value for HCO3 and PaCO2, PaO2. In the case of pH the prognosis is under expectation. This factor is strongly related to PaCO2 and HCO3. It can be concluded that we have received a good prognosis for two or three parameters. Table 1 contains the results of the simulation of neural network predictions for a 6 hour period of time.

Table 1
Percentage of good results prediction for pH, PaCO2, PaO2, HCO3
by different neural network structures: #Input-#Hidden-#Output]

Neural Network Structure	pH %	PaCO2 %	PaO2 %	HCO3 %
5-3-1	74.96	90.94	68.35	86.94
5-5-1	79.26	89.86	67.59	87.71
5-10-1	79.87	90.94	67.12	86.18
5-15-1	80.81	89.56	68.51	88.02
5-20-1	78.96	90.32	66.82	89.40
8-3-1	77.27	82.64	61.44	86.33
8-5-1	76.80	88.94	68.20	82.80
8-10-1	77.42	88.17	65.44	83.10
11-3-1	75.62	88.02	60.52	81.11
11-5-1	73.27	87.71	66.05	84.02
11-10-1	78.34	87.56	60.22	80.18
14-3-1	64.82	86.79	62.06	80.49
14-5-1	75.73	86.33	60.52	81.87
14-10-1	76.04	97.25	61.75	76.96

One of the problems occurring during neural network training is called overfitting. The error on the training set is driven to a very small value, but when new data are presented to the network the error is large. One method for improving network generalization is to use a network that is just large enough to provide an adequate fit. The larger the network used, the more complex function the network can create [12].

6. Conclusion

Respiratory problems are the most prominent in the pathology of newborns hospitalised in the Neonatal Intensive Care Unit (NICU). Respiratory insufficiency is the leading cause of hospitalisation and mortality as well. Blood gases have been analysed in the context of actual respiratory setting. Arterial blood gases (ABG) are good indicators of these problems.

Accurate forecasting of ABG alternations caused by different factors would be of great value in newborn intensive care. A neural network structure yields a collection of values that describe the prognosis for pH, PaCO2, PaO2, and HCO3. Blood gas values are forecasted with an error. We assume to receive the error less than or equal to the acceptable value.

References

[1] De Castro L.N., Von Zuben F.J., *An Evolutionary Immune Network for Data Clustering*. Proc. of the IEEE SBRN, 2000, 84–89.
[2] De Castro L.N., Von Zuben F.J., *The Clonal Selection Algorithm with Engineering Applications*. GECCO'00 Workshop Proceedings, 2000, 36–37.
[3] De Castro L.N., Von Zuben F.J., *Artificial Immune Systems*. Part I – Basic Theory and Applications, Technical Report – RTDCA 01/99, 1999.
[4] Wajs W., *Predicting of Dynamic Medical Data Series Using Neural Network Method*. 15th Triennial World Congress of the IFAC b'02 Barcelona 2002.
[5] Jerne N.K., *Towards a Network Theory of the Immune System*. Ann. Immue nol. Ins. Pasteur, 125C, 1974, 373–389.
[6] Jerne N.K., *Clonal Selection in a Lymphocye Network*. In G.M. Edelman (Ed.). Cellular Selection and Regulation in the Immune Response (p. 39). Raven Press, New York 1974.
[7] Leclerc B., *Minimum Spanning Tree for Tree Matrices, Abridgements and Adjustments*. Journal of Classification, 12, 1995, 207–241.
[8] Zahn C.T., *Graph-theoretical Methods for Detecting and Describing Gestalt Clusters*. IEEE Transactins on Computers, C-20(1), 1971, 68–86.
[9] Varela F.J., Coutinho A., *Second Generation Immune Networks*. Immunology Today, 12(5), 1991, 159–166.
[10] Chryssolouris G.C., Lee M., Ramsey A., *Confidence Interval Prediction for Neural Network Models*. IEEE Trans. on Neural Networks, vol. 7, No 1, Jan. 1996.
[11] Graves S.C., Redfield C.H., *Equipment selection and task assignment for multi product assembly system design*. Int. J. Flexib. Manufacturing Sys., vol. 1, 1988, 31–50.
[12] Lippman R.P., *An introduction to computing with neural nets*. IEEE ASSP Magazine, 1987, 2–22.
[13] Bonna C.A., Kohler H., *Immune Networks*. Annales of the New York Academy of Sciences, 418, 1983.
[14] Carrol J.D., *Minimax Length Links' of a Dissimilarity Matrix and Minimum Spanning Trees*. Psychometrica, 60(3), 1995, 371–374.
[15] Lapointe F-J., Legendre P., *The Generation Tests for Dendrograms: A Comparative Evaluation*. Journal of Classification, 8, 1991, 177–200.