Srđan Ribar
Teaching Assistant
University of Belgrade
Faculty of Mechanical Engineering

Miroslav Dramićanin
Associate Research Professor
Institute of Nuclear Sciences "Vinča"

Tatjana Dramićanin
Assistant Research Professor
Institute of Nuclear Sciences "Vinča"

Lidija Matija
Associate Research Professor
University of Belgrade
Faculty of Mechanical Engineering

Classification of Breast Cancer Luminescence Data Using Self-Organizing Mapping Neural Network

Self-organizing mapping neural networks are applied in the analysis of breast cancer luminescence data. Data consist of three dimensional vectors presenting normal and malignant human tissue. The possibility of such data classificaton in two groups (normal and malignant tissue) is analyzed.The network performed sucessfull classification.

Keywords : *self-organizing mapping, breast cancer, tissue luminescence*

1. INTRODUCTION

Breast cancer is one of the most common malignant tumors among women in the world and the most frequent cause of death, exceeding cardiovascular diseases and lung cancer in Serbia. It is now a well-established fact that early detection of cancer can play a significant role in its treatment, making possible improvement in the quality of patient`s life and increase in survival rates. If diagnosed early, breast cancer is one of the most treatable forms of cancer. Recently, modern luminescence techniques are applied for the diagnostics of breast cancer [1]. Huge efforts are made to establish non invasive diagnostic method [2]. In this paper we examine potential of artificial neural networks for breast tissue classification based on data input form luminescence measurements.

This paper is concerned with neural network data analysis. Neural networks are structures which perform various types of input-output mappings.

Artificial neural networks (ANN) is the software engineering discipline concerned with parallel, distributed, adaptive information processing systems that develop information processing capabilities in response to exposure to an information environment. Medical applications of ANNs are mostly based on their ability to handle classification problems including classifications of illnesses or to estimate prognosis [3,4]. There are a few studies which examined the use of ANN to both predict nodal involvement and investigate its predictive applicability to the long-term prognosis of patients with breast cancer [5]. Seven inputs were analyzed by the ANN: the patient's age, the tumor size and grade, percentage ER- and PgR –expressing cells, and expression levels of *h-mtsl 1* and *nm*23 [6]. When *h-mtsl 1* and *nm23* are considered as sole inputs, it was found out that genes do have an important effect on prediction statistics. It was concluded by authors that the specificity parameter attained in this case is equal to that achieved by all seven parameters when analyzed (90%), suggesting that the analysis of the various expression levels of both *h-mtsl 1* and *nm23* is perhaps more effective in predicting node-negative status.

Neural networks are highly adaptive structures with numerous adaptible coefficients which have to be set on specific values. This adaptation is called a training process. The training process is applied to train the network off line to perform specific mapping that is required. During the training process a representative set of data are presented to network. The training process is finished when a preset network mapping accuracy is reached.

Most types of networks need exact correct output network value for each input value during training process [2]. It means that output data set has to be given as desired output value for each input data. In that way for one set of input data network may be trained to perform arbitrary input-output mapping. After trained process is finished network input-output mapping is memorised and network may be tested with test data set. Test data set usually have values different from training data set. That happens when noise is presented in data set or when training and test data sets are defined on continous domains [2]. Since network is trained on specific training data set for arbitrary new (not memorised) inputs, network outputs are closest to memorised output. It means that networks have: a) filtering properties in presence of noise, b) property of input data classification into predefined subsets. This is important network property which is commonly used. After the training process is finished, network is ready to perform required mapping with predefined accuracy.

Specific types of networks can be trained with no desired output value.

It has been shown that SOM is one of the best known artificial neural network algorithms [7]. A SOM is learning algorithm, which represents high-dimensional data in a low-dimensional form without loosing any of the 'essence' data. Also, it is organized data based on similarity by putting entities geometrically close to each other.

The self-organizing map is an unusual mapping network in that the mapping it approximates is defined implicitly, not explicitly. Graphic representation as one

Received: Jun 2006, Accepted: September 2006
Correspondence to: Srdjan Ribar
Faculty of Mechanical Engineering,
Kraljice Marije 16, 11120 Belgrade 35, Serbia and Montenegro
E-mail: sribar@mas.bg.ac.yu

of major features depends upon the fact that the weight vector in each of the processing elements of the network corresponds to a specific feeler mechanism position in the feeler mechanism universe. After it is trained, the self-organizing map neural network implicitly defines a continuous topological embedding map of the rectangular neural network array into the topological space of the feeler mechanism universe. Naturally, this map operates in the opposite direction from the feeler mapping that translates the position of the pointer into a corresponding neural network activation. Two attributes which characterize the self-organizing map are: first, the number density of the weights in the target space is approximately proportional to the probability density used for selecting the example. Secondly, the mapping from the rectangular grid of processing elements into the target space must be topologically continuous.

Such networks perform self organized mapping (SOM) [8]. Their input and output space are necessary of the same dimension. Such network tends to perform input-output mapping in a manner that position of input data in input space should correspond to output data in output space. Since input and output space dimension is the same, after training coordinates of inputs and corresponding outputs are as close as possible. The number of elements in input and output sets is not the same. Input set has commonly extremely more elements than output set. So, more than one element from input space is mapped to one output element which is closest to it. The number of output elements is arbitrary and depends on the type of mapping which has to be obtained. So, this type of network can perform either exact input-output mapping or data classification. In the first case the number of output elements is close to the number of input elements and in the other presents the number of classes for input data classification. SOM input data are training and test set only with no desired network output.

Data present quantification of luminescence excitation- emission matrices obtained from normal and malignant human breast tissues [1]. Data consist of threedimensional vectors. Since normal and malignant human tissues have different structure, this property is noticed in vectors arrangement. So, each pattern is presented by three component vector. The high accuracy of such differentiation of normal and malignant tissue is confirmed by histopathology analyses of set of patterns which seems to be nowadays only valid diagnostic method.

This high accuracy vector presentation of tissue patterns could be very promising. As a matter of fact, the whole set of vectors could be divided in to two groups: one would represent normal and other malignant tissue. A tool which separates vector input space into two groups is analyzed in this paper: self-organizing mapping neural network (SOM) is applied to an input set of vectors. Input vector set consists of two subsets: traininig set and contol set. Training set is composed of vectors with statistical parameters the same as normal and malignant tissue have. Test set consists of vectors obtained by measurements of normal and malignant tissue.

The point of this is to analyse if it is possible to train the network by training set. Such network should be able to divide vectors obtained by measurements (test set) into two groups - with normal and malignant tissue. The result of network mapping would be compared with histopathology analysis of tissue. If network mapping is satisfactory, this method could be helpful as diagnostic method. It means that surgical intervention as a diagnostic method could be avoided.

2. MATERIALS AND METHODS

In this work we used the results of the *in vitro* synchronous scanning method - SLS (constant wavelength mode) of human skin specimens obtained after clinical surgery. Normal and malignant tissue status is confirmed by histopathology and dermoscopy. Synchronous luminescence spectra were recorded at room temperature using a Perkin Elmer LS45 Fluorescence Spectrophotometer. A Xenon discharge lamp was used for illumination, while a gated red-sensitive R928 photomultiplier operating up to 900 nm was used for detection. Data were recorded at a 200 nm/min scan rate and excitation wavelenghts from 330 to 550 nm with 0.5 nm increnments, averaging 4 scans to obtain the spectra. [9]

This work we are reporting concerns the classification of experimental data by SOM neural network.

3. INPUT DATA

Input data consist of training and test set data. Training set contains 2000 vectors with properties of normal and malignant tissue: $x_{in}(i) \in R^n$, $i = 1, 2000$, $n = 3$. Input data are normalized, fig.1. Test set contains 200 vectors obtained from normal and malignant tissue.

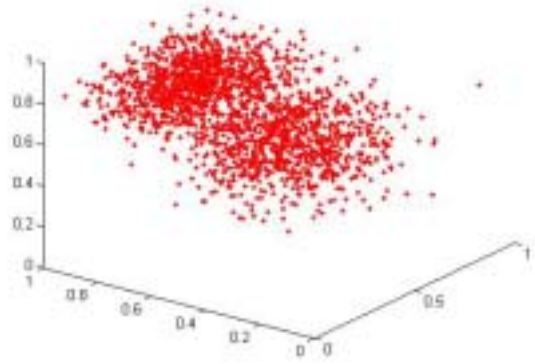

Figure 1. Training set

It is clear that training data could not be separated into two subsets, fig.1.

4. NETWORK STRUCTURE

The first step in this classification is a choice of a network. Self-organizing map (SOM) as topology network was a promising choice. Self-organizing

mapping is a one-layered network very similar to Kohonen's network [8]. It means that one network input activates only one node in output layer (with nonzero output) called 'winner'.If nodes are in linear distribution, each of them have two adjacent nodes, called 'neighbours '. The number of neighbours rises in the case of 2D distribution of nodes, fig.2. (nodes are represented by circles and lateral connections by lines).

Dimension of nodes distribution is not limited and is equal to input vector dimension only. All nodes are laterally connected to their neighbours (lines at fig.2). The only difference between Kohonen's network and SOM is neighbouring function. This function moves the 'winner' closer to its input and influences, so adjacent elements become sensitive to similar input. Such influence is performed by distance function which determines number of elements which become sensitive

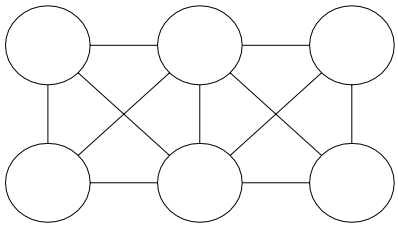

Figure 2. 2D-SOM structure

to similar input. It means that there are no desired ouput data of network. Each node is connected to all of its neighbours. Each input vector is input to all nodes. SOM property is that dimensions of input and output space are equal. Since input space dimension is $n = 3$, and SOM tends to map to the same dimension output space, nodes should be arranged before training in 3D space. Output nodes are positioned in space after training near input vectors as close as possible. Their position depends on density of input vectors. The number of output nodes has to be set before network training. For any control set input, the only active node will be the node nearest to that input.

Since a mapping tool is selforganizing map, its structure depends of input set.

The main task of such a network is to form outputs: $x_{out}(j) \in R^n$, $j \in N$ which should be a close to inputs as possible. After training process each input vector from test set will activate one output element closest to it. This specific property of SOM was the main reason for its implementation solving this problem. At this moment the number of output elements j has to be chosen. That number depends on the complexity of input vectors separation into disjunctive sets. Input space should be in this case divided into two subsets. If it is possible, two output elements are enough. Testing network with $j = 2$ leads to enormous mapping error. So, in this case it is adopted much greater value: $j = 125$. These elements form at the beginning of training process three-dimensional matrix structure in output space: $5 \times 5 \times 5$. At the beginning of training process all elements are jointed in the centre and training process will spread them through output space. Their density in the output space depends of density of input elements. So their matrix position at the beginning of training process is irrelevant to their final position after enough number of training epochs. It is only necessary that dimension of output elements is the same dimension as input elements eg. threedimensional.

3.1 Network training

Mapping problem was then analyzed by SOM. All that is necessary for SOM network training are input training vectors since network's goal is to get the same vectors on its output. So, input data are presented to the network several times through traning epochs. The number of training epochs is experimentally variable and has increased till mapping error became constant and sufficiently small. Output elements are positioned in the centre of output space at the beginning as it was mentioned earlier. The results of training after just one training epoch are in fig.3. It can be considered that output elements tend to spread through output space like input elements of training set (fig.1). In fig.3. test data are presented (malignant tissue by dash dots and normal by black dots) as well as lateral connections between network elements (lines).

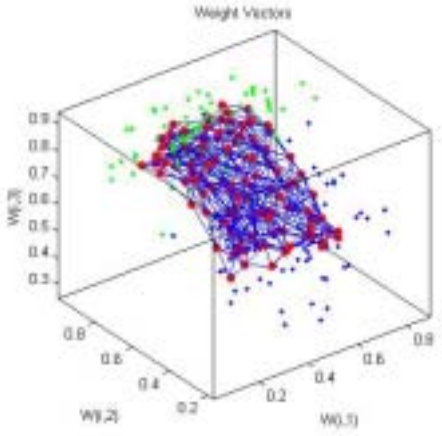

Figure 3. Training results after 1 epoch

During further training elements are spread through output space. It can be observed after 10 training epochs, fig.4, comparing to training result after 1 epoch, fig.3. (min value of $w(i,1)$ axe). So output elements are meshed with input elements tending to cover them all.

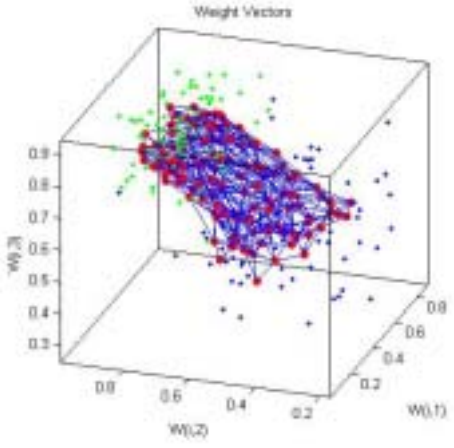

Figure 4. Training results after 10 epochs

This process continues as mapping error is getting smaller. As it is given in fig.5. after 50 epochs network outputs are spread through output space trying to map input elements. Density of output depends on density of input elements, so single inputs far away from pile, far left and far right in fig.1., could not be mapped well.

3.2 Mapping accuracy

SOM network accuracy depends mainly on input data distribution. Mapping accuracy is high if input elements can be divided into subsets. After a certain training period (500, 1000, 5000 epochs) mapping accuracy was tested. Tests were obtained applying test set data as an network input and mapping accuracy is defined as follows: each test set element activates one output element. Input elements that are near each other will activate the same output element. As test set presents normal and malignant tissue satisfactory data classification means that one output element must not be activated by input data of two different groups. In that case such input data are considered as classified incorrectly and mapping element is incorrect too.

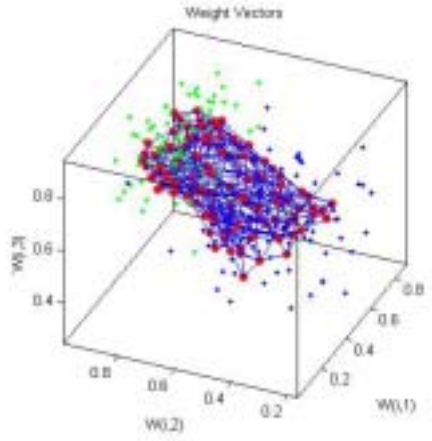

Figure 5. Training results after 50 epochs

A set of test data is presented in two figures: fig.6 and fig.7. Black dispersed dots correspond to normal tissue data and dashed concentrated dots to malignant tissue. Test data can not be divided into two disjunctive subsets although they are not meshed like training set data, fig.1.

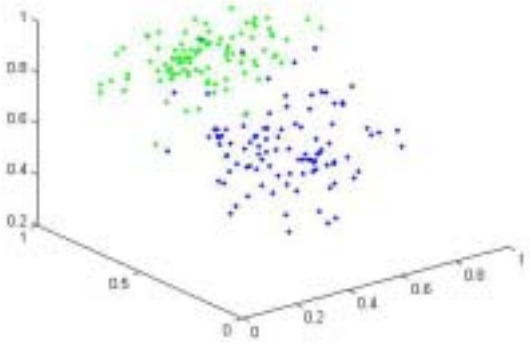

Figure 6. Test data

Mapping accuracy was tested after certain training period (1, 10,..,50, 100, 200, 500, 1000, 5000 epochs).

Test data mapping results are presented in table 1.

Mapping accuracy was tested with test data set. It means that test set is an input set. It contains results of measurements of 100 normal and 100 malignant tissue patterns presented as threedimensional vectors. As it is clearly seen after 1000 epochs minimum incorrect mapping elements was obtained. Further training did not increase network accuracy (same accuracy was obained with no. of epochs > 5000 which is not presented here).

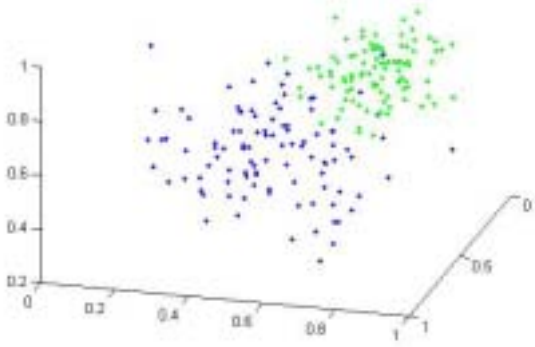

Figure 7. Test data

Increasing number of mapping elements made no benefit in network accuracy. The same input data were presented to SOM network with higher dimension: $10\times10\times10$ but with no influence to network accuracy. Network accuracy is also invariant to the choice of distance function during training, which was also tested.

Table 1. Mapping results of test data

No. of epochs	No.of incorrect mapping elements (%)	No. of incorrect selected data (%)
1	4 (3.2)	12 (6)
10	5 (4)	15 (7.5)
50	7 (5.6)	22 (11)
100	5 (4)	11 (5.5)
200	5 (4)	13 (6.5)
500	4 (3.2)	13 (4)
1000	4 (3.2)	9 (4.5)
5000	5 (4)	12 (6)

4. CONCLUSION

Breast cancer luminescence data of normal and malignant human tissue are analyzed in this paper. These data (in the form of threedimensional vectors) presented input of self-organizing map. The task was to train the network to classify inputs into two distinguished groups. It is expected that each group should contain only data of the same kind: normal or malignant tissue data. Such mapping should also prove that numerical presentation of tissue is valid. Mapping obtained by SOM is satisfactory. The results over 5000

training epochs are presented, table 1. It can be noticed that mapping error is <10%. The number of network elements that perform unsatisfactory mapping is 4% as 6% of input data are incorrectly selected. It is important to observe that best mapping is performed after only 500 epochs of training- 3.2% of incorrect network elements and 4% incorrect selected data. Further training did not improve mapping quality. These mapping errors are caused by test data arrangement in input space. Such test data set can not be mapped by SOM with zero-error. Test data that present different inputs (normal and malignant tissue) are meshed in input space, so it is not possible to separate them with zero-error.

To decrease mapping error, it is necessary to improve input data separation before presenting them to SOM. This means that additional information regarding input data should be known and presented to SOM. In that case input data should be clearly separated in input space and network should perform better. This might arise the input dimension of network (dimension >3) and disable simple 3D presentation network performance but also diminish network error. If input data set could be divided into two subsets, zero-error might be achieved. In that case not only self organizing map as a classificator could be applied.

ACKNOWLEDGMENT

The author is very grateful to Željko Ratkaj from University of Belgrade, Faculty of Mechanical Engineering, Belgrade, Serbia, for useful suggestions for the first step of applying neural network.

REFERENCES

[1] Dramićanin, T., Dramićanin, M., Dimitrijević, B.: Excitation - emission and synchronous luminescence spectroscopy of normal and malignant breast tissue, XLIV Conference of Serbian Chemical Society, pp. 70, 2006.

[2] Markley, M., Lo, J., Tourassi, G., Floyd, C.: Self-organizing map for cluster analysis of a breast cancer database, Artif Intell Med, pp. 113-127, 2003.

[3] Astion, M.L., Wilding, P.: The application of backpropagation neural networks to problems in pathology and laboratory medicine, Arch. Pathol. Lab.Med. 116, pp. 995-1001,1992.

[4] Truong, H. at al.: Neural networks as an aid in the diagnosis of lymphocyte-rich effusions, Anal. Quant. Cytol. Histol.17: pp. 48-54, 1995.

[5] Naguib, R.N.G, at al.: The detection of nodal metastasis in breast cancer using neural network techniques, Physiological Measurement, 17, pp. 297-303,1996.

[6] Naguib, R.N.G., at al.: Prediction of nodal metastasis and prognosis in breast cancer: a neural model, Anticancer Research, 17, pp. 2735-2742,1997.

[7] Kohonen, T.: *Self-Organization and Associative Memory*, Springer-Verlag, Berlin, 1988.

[8] Kohonen, T., Kangas, J.A., Laaksonen, J.T.: Variants of self-organizing maps, IEEE Trans. on Nerual Networks, I, pp.93-99, 1990.

[9] Dramićanin, T., Bandić, J., Dimitrijević, B., Dramićanin, M.D.: Synchronous Luminiscence Spectroscopy of Melanoma, in the Book of Abstracts of International Conference on Physics of Optical Materials and Devices, ICOM 2006, Herceg Novi, Montenegro, August 31-Septembre 2, p. 124. 2006.

КЛАСИФИКАЦИЈА ПОДАТАКА ДОБИЈЕНИХ ЛУМИНЕСЦЕНЦИЈОМ РАКА ДОЈКЕ ПОМОЋУ САМО-ОРГАНИЗУЈУЋЕ НЕУРОНСКЕ МРЕЖЕ

Срђан Рибар, Мирослав Драмићанин, Татјана Драмићанин, Лидија Матија

Примењена је самоорганизујућа неуронска мрежа при анализи података луминесценције рака дојке. Улазни подаци су тродимензионални вектори који представљају нормално и малигно хумано ткиво. Анализирана је могућност класификације података у две групе. Мрежа је задовољавајуће обавила класификацију улазних података.