The self-organizing maps of Kohonen in the medical classification

Manel ZRIBI ;Younes BOUJELBENE
Faculty of Economic and Management
URECA
Sfax University- Tunisia
manel0306@yahoo.fr
boujelbene.younes@yahoo.fr

Inès ABDELKAFI. Rochdi FEKI
Higher schoolar of trade
URED
Sfax University- Tunisia
rochdi.feki@escs.rnu.tn
inesabdelkafi@yahoo.fr

Abstract: **In Tunisia, breast cancer is the most common cancer among women; it presents the leading cause of female mortality in the age group 35 to 55 years. This paper uses a neural approach based on Kohonen self-organizing maps to perform a classification of tumors (benign and malignant) using a sample of Tunisian women. Empirical results demonstrate the relevance of the approach and show that neural networks are an important decision support technique for detecting the presence of cancerous tissue in the breast and the classification of tumors.**

Key words: Classification; breast cancer; Kohonen self-organizing maps

I. INTRODUCTION

Over the last decade, research related to the detection of breast cancer has been increased. Many studies have been oriented to detecting the presence of cancerous tissue in the breast and the classification of tumors (Subhash et al., 2003; Fei et al., 2003 ; Madden 2002 ; De Santo et al., 2003 ;Karnan, 2008,…). The approaches used are from the field of statistics, probability and connectionism.

In our work we will be interested in self-organizing maps of Kohonen (SOM) neural approach is one that consists of a classification strategy based on principle of competition and cooperation.

Their ability to automatically discovery categories in a set of data is of such networks effective classification techniques (Lambert, 1999).

The Kohonen selforganizing maps have been implemented and have shown their worth in various fields.

The purpose of this paper is to build a technical decision support for better detection of the presence of cancerous tissue in the breast and a classification of tumors (malignant and benign) using the SOM based on a sample of Tunisian women.

The rest of this article is organized as follows: the first section recalls the main research and the main techniques used for detecting the presence of cancerous tissue in the breast and the classification of tumors. The second section presents the SOM and their advantages over conventional classification techniques. The third section explains the results and their interpretation.

II. METHOD OF CLASSIFICATION THE BREAST CANCER

In Tunisia, breast cancer is the most common cancer among women; it presents the leading cause of female mortality in the age group 35 to 55 years. Obtaining prognostic factors at diagnosis allows treatment is considered graduated tailored for each patient.

Many classification techniques have been used in order to group patients with similar biological and which are likely to share certain factors etiopathological.

The constitution of these classes is important since it allows, first, to clear the disease process and therefore establish the proper térapeutique approach and also identify the overall prognosis of this disease. In the literature, many studies have been related to breast cancer medical diagnosis:

- Subhash et al. (2003) propose a K-means approach for performing a classification in the context of screening for breast cancer. This is a nonparametric classifier and has been adapted and tested on two samples of patients from databases WDBC (Wisconsin Diagnosis Breast Cancer) and WBC (Wisconsin Breast Cancer).
- Fei et al. (2003) propose a method for separating large margins for a classification in the context of screening for breast cancer, for projecting the vectors of characteristics a number of patients in a space of large dimensions. This technique leads to interesting results in the case where the classes are linearly separable.
- De Santo et al., (2003) have proposed an approach based on an expert system for multiple

recognition of microcalcifications from a database of mammographic images.

More this techniques, the Artificial Neural Networks (ANN) have become a tool increasingly used in screening for breast cancer:

- Cheng et al. (1994), used two ANN models (the first is a back-propagation network and the second uses radial basis function), their approach led to acceptable results but requires a long execution time causing over-learning.
- Baeg et Kehtarnavaz (2000),have proposed an back-propagation networks with three layers on the basis of two characteristics of image texture to classify abnormalities in digitized mammograms, were considered good results.
- Guo et Nandi (2006), used a multilayer perceptron (MLP) with back-propagation algorithm to train a network parameters from the WDBC data base.
- Verma (2008), used the ANN with Soft cluster neural network algorithm for Digital Database of Screening Mammography

In this work, we use the neural approach based on self organizing maps of Kohonen to clear a classification of tumors (benign and malignant) based on a sample of patients in Tunis

III. THE SELF-ORGANIZING MAPS OF KOHONEN

The Kohonen self-organizing maps, were introduced by Teuvo Kohonen in the 80.

This is a neural model which is a technique for data visualization. This method has any characteristics: the number of classes is set by the user, and it is adapted to large data, for projecting a multidimensional input space in an output space, generally one-dimensional or two-dimensional.
Were competitive and dynamic networks to elect a winner neuron and the foster (Kohonen 1993, Kohonen 1995, Cottrell et al. 1998).

The self-organizing maps Kohonen, using the unsupervised learning algorithm (Kohonen, 1995 ; Cottrell et al., 2003), Were often used the context of data analysis (Blayo et Demartines, 1991 ; Oja et Kaski,1999).

Specifically, the applications were for classification problems.

The combination is effected by reference to a concept of neighborhood among classes for the examples in a close neighbors (within a class or related classes) have similar characteristics. In this section, we propose to present briefly the technique with a focus on certain elements that make the originality of the method.

Basically, self organizing maps Kohonen correspond to ANN composed of two layers, one input layer and one output layer:
The input layer receives information it constitutes the input network. This information generally relates to a number of examples and provides a series of n variables; the number of variables defined as the number of neurons in the input layer. There the vector of observed values of the variables in example i.
The output layer consists of a number of cells arranged in a rectangular grid, this is a "map".Each neuron i on the input layer is connected to each neuron j of the output layer. At each connection is assigned a synaptic weight w_{ij} The classification method using Kohonen maps belong to the family of algorithms leading to Sheet as the methods of aggregation around mobile centers, or cloud dynamics (Lebart, 1999).

The learning Kohonen algorithm is an iterative procedure that can be seen as a stochastic algorithm for classification, which groups the observations in classes, while maintaining the space topology of observations.

This is done so that two examples presented at the input of the system that are affected in the same class or in neighboring classes on the map, have similarities in their respective characteristics.

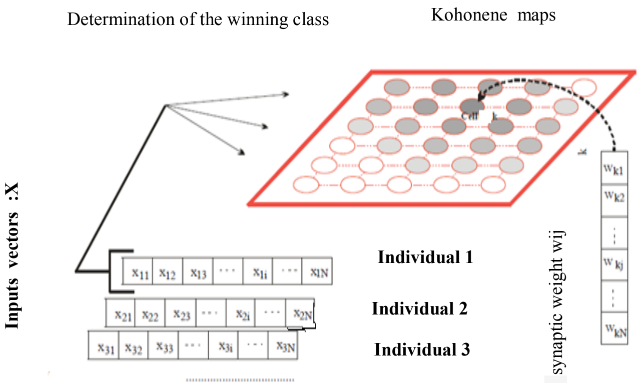

Determination of the winning class
Kohonene maps
Cell k
Inputs vectors :X
x_{11} x_{12} x_{13} ··· x_{1i} ··· x_{1N}
x_{21} x_{22} x_{23} ··· x_{2i} ··· x_{2N}
x_{31} x_{32} x_{33} ··· x_{3i} ··· x_{3N}
Individual 1
Individual 2
Individual 3
synaptic weight wij
w_{k1}
w_{k2}
w_{kj}
w_{kN}

Figure 1. The Kohonen algorithm

The Kohonen algorithm is an iterative process whose principle can be summarized as follows:

(i) The first step is an initialization step random weights wij. For each class j of the map is associated a weight vector code, regarded as the representative of this class.
(ii) The second step is to choose a random example and present its feature vector X to the network input to determine the gaining unit j* .This is the class that this example will be used is determined such as the partner weight vector is nearest the vector of characteristics shown at the entrances .This is competitive learning; the code vector to winning class its update in the direction of the vector subject. This adjustment also concerns all the weight vectors of other cells in the map according to their proximity to the winning class.
(iii) The same procedure is subsequently made for another example randomly selected in the database, and so until convergence.

Classification by the Kohonen algorithm tenders many advantages (Kohonen 1993, Kohonen 1995, Cottrell

1998): originality to propose a visualization of the neighborhood structure between classes and the Kohonen algorithm is designed to optimize use of visual properties of the algorithm, i.e. The concept of neighborhood between classes (Blayo 1991, Cottrell 1995, Rousset 1999).

After finished with training, each vector in the database is assigned to a cell on the map, and each cell is represented by its weight vector. This vector sums most of the information in all the features of people hat are grouped together. For this reason the analysis will focus to a great extent on the structure of code vectors achieved.

IV. Experimental study

A) Data Description

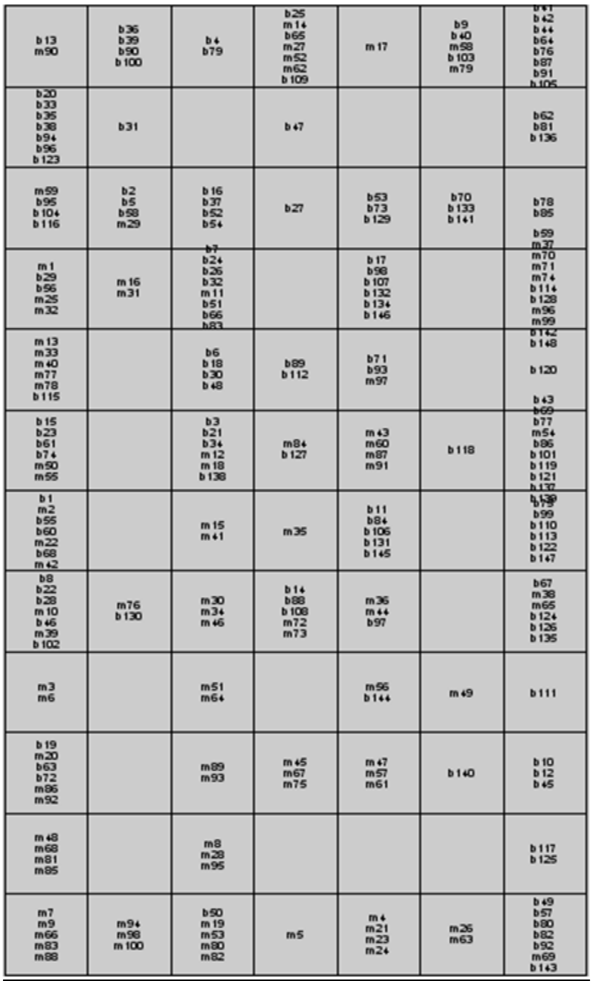
Our sample consists of 247 Tunisian women with the disease of breast cancer; these data were collected from university CHU in cooperation with specialists in the field for the identification a typology of malignant breast tumor (m) or benign (b).

There is no single factor responsible for the emergence of breast cancer. Empirical studies on the detection breast cancer have revealed several risk factors.

In our application, we selected 18 risk factors for type epidemiological, clinical and paraclinical: Origin(X1) , Marital Status(X2), antecedent Personal(X3), Antecedent family(X4), age at menarche (X5), age of first full-term pregnancy(X6), Breast feeding(X7), Ménopausal status(X8),history of oral contraceptive(X9), skin types(X10), state of the nipple(X11), sign of inflammation(X12), Ganglion(X13), mammography(X14), tu mor size(X15), nodules(X16), Metastasis(X17) ,anapath (malignant (m) or benign (b))(X18).

B) Results and Interpretation

The learning of a Kohonen map that includes 7 × 12 = 84 classes by using the software Matlab version 7.11. This size of the map is most appropriate for our sample by the process proposed by Vesanto et al. (2000), for determining the number of lines and the number of columns (optimal map) in a Kohonen map for a particular purpose. The results is in Figure 2.



b13 m90	b36 b39 b90 b100	b4 b79	b25 m14 b65 m27 m52 m62 b109	m17	b9 b40 m58 b103 m79	b41 b42 b44 b64 b76 b87 b91 b105
b20 b33 b35 b38 b94 b96 b123	b31		b47			b62 b81 b136
m59 b95 b104 b116	b2 b5 b58 m29	b16 b37 b52 b54	b27	b53 b73 b129	b70 b133 b141	b78 b85 b59 m37
m1 b29 b56 m25 m32	m16 m31	b7 b24 b26 b32 m11 b51 b66 b83		b17 b98 b107 b132 b134 b146		m70 m71 m74 b114 b128 m96 m99
m13 m33 m40 m77 m78 b115		b6 b18 b30 b48	b89 b112	b71 b93 m97		b142 b148 b120 b43 b69
b15 b23 b61 b74 m50 m55		b3 b21 b34 m12 m18 b138	m84 b127	m43 m60 m87 m91	b118	b77 m54 b86 b101 b119 b121 b137 b139
b1 m2 b55 b60 m22 b68 m42		m15 m41	m35	b11 b84 b106 b131 b145		b75 b99 b110 b113 b122 b147
b8 b22 b28 m10 b46 m39 b102	m76 b130	m30 m34 m46	b14 b88 b108 m72 m73	m36 m44 b97		b67 m38 m65 b124 b126 b135
m3 m6		m51 m64		m56 b144	m49	b111
b19 m20 b63 b72 m86 m92		m89 m93	m45 m67 m75	m47 m57 m61	b140	b10 b12 b45
m48 m68 m81 m85		m8 m28 m95				b117 b125
m7 m9 m66 m83 m88	m94 m98 m100	b50 m19 m53 m80 m82	m5	m4 m21 m23 m24	m26 m63	b49 b57 b80 b82 b92 m69 b143

Figure 2 : Classification with a self-organizing map of Kohonen

An initial review of the map obtained allows a first idea about the way in which the combination was effected, and shows two the assembly areas of patients by their tumor type: 'm' is a malignant tumor and 'b' the benign tumor.

The first area, at the top left and right of the map, includes the majority of patients with benign tumors (b).

The second area, the center of the map to the bottom, which defines the area in patients with malignant tumor (m).

Each group of patients classified in one cell involves these patients have similar prognostic factors.

There are also some wrong classifications which appear in either area, and that can be assigned to situations where it is difficult for some patients, to make a decision as to their affiliation with one of two groups based only on the own variables.

This can be viewed from distances also referred to as U-Matrix. This is a map that uses the code vectors to illustrate the distances between the different boxes of the Kohonen map.

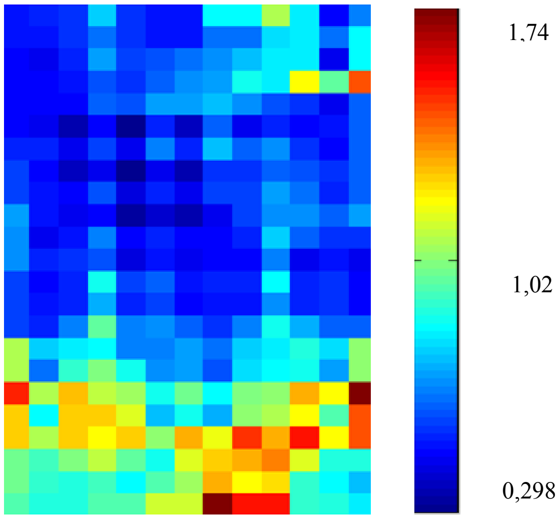


Figure 3 : Distance map (U-Matrix)

In our situation, the first area which is mainly related to patients with a benign tumor of the blue color. Thereby affirming about a great similarity of patients characteristics subject. However, if one moves to the bottom from the center of the map, we see that the color in several places turns to red. This fact suggests that which patterns of patients with malignant tumors are more heterogeneous.

It is can also use the code vectors to generate a map of weight used for each variable: the maps variable. In this map, a red color is associated with a high value of the cell relates and a blue color indicates a low value.

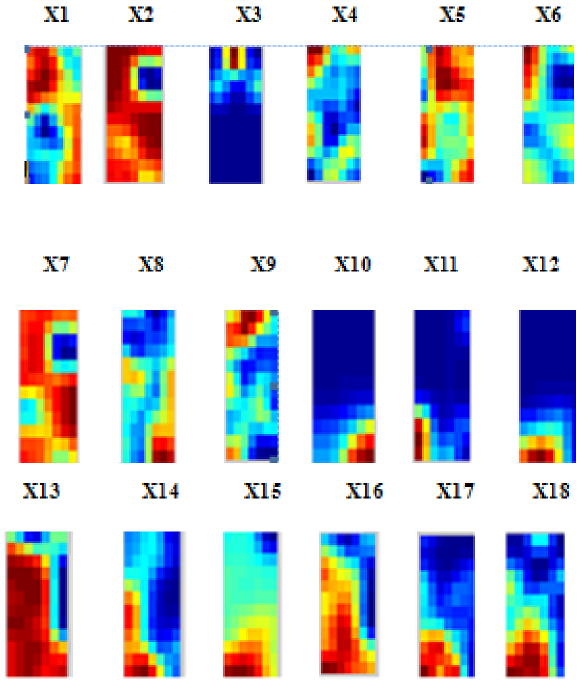


Figure 4. Maps variables

If two cards for two variables are similar, it can be concluded if a correlation between it. So, we can see that variables the nipple(X11)and sign of inflammation(X12) are highly correlated.

The same comment also applies to variables mammography(X14), metastasis(X17) and anapath (X18).

The use of code vectors may provide additional elements of analysis with a more comprehensive view of the map. In this vein, Cottrell et Rousset (1997), et Rousset (1999), Feki (2008), have made a combination of code vectors by using a hierarchical clustering with Ward distance. This process leads to the production of new classes called 'macro classes' or 'super classes'.

In our application, this process allowed us to group the 84 classes of the optimal map achieved in the four macro classes in Figure 5.

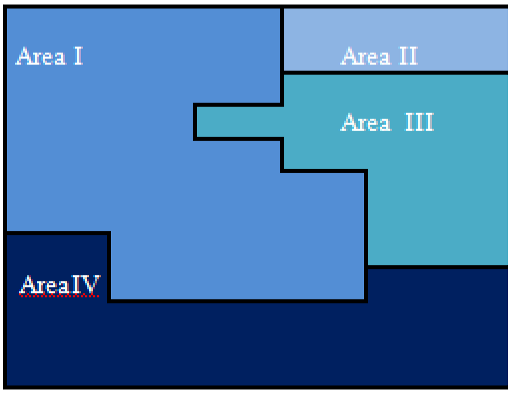


Figure 5: Macro classes map

In each area, patients with one hand, some degree of overall similarity in relation to different risk factors. Moreover, the characteristics of these patients differ significantly from those of patients in the other three areas. It is as possible to characterize the factor that has dominated in all four areas obtained.

Seranno Cinca (1996) demonstrated in this work that it is can identify the most critical variable in assigning a patient to a particular class on the map by studying the synaptic weights. The process is to identify the variable which represents the greatest weight (in absolute value).The results with our data shows that each area is characterized by a leading variable that distinguishes it reports by other macros classes:

The analysis of four area sensed in our application brings up the (or) characteristic (s) that characterize (s) patients are screened. We see that the variable "tumor size" has the feature which corresponds to the weight of most important (in absolute value) followed by the variable "age at first pregnancy" and "nodules".

These latter three variables with the variable "state of the nipple" feature (with similar weights) area 2. The characteristic "tumor size" also dominates the area followed by three variables "nodules" state of the nipple "and" signs of inflammation ". Finally, the variable "tumoral size" and "Nature of skin" is the main features of area 4.

V. CONCLUSION

In this paper, we presented a classification of 247 patients with the disease breast cancer with the Kohonen self-organizing maps. Building on 18 risk factors for type epidemiological, clinical and para -clinical, it was possible to identify some important results show the relevance of the approach adopted and the contribution of this neuronal technique over the classical techniques of classification .
Specifically, the use of this method allowed us to distinguish patients with a malignant tumor and the benign, the results showed that tumor size, the nature of nodule and type of skin are the main features of classification.

Even if the results do not exactly coincide with conclusions drawn by several experts in the field, the resulting map can be very useful for differen allowing better understanding of the main features of each type of patient. Secondly it few serve as a tool for decision support for doctors allowing them, after presentation of the unique features of the patient ,to place it in the appro

REFERENCES

[1] Asoke Hong .G, Nandi K, 2006 : Breast cancer diagnosis using genetic programming generated feature , Elsevier Pattern recognition, Vol 39, pp 980-987.

[2] Baeg, S. and Kehtarnavaz, N., 2000 : Texture based classification of mass abnormalities in mammograms, Proc. of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS), Houston, TX, vol. 1, pp. 163-168.

[3] Cheng S.N.C, Chan H.P, He1vie M.A, Goodsitt M.M, Adler D.D, St. Clair D.C,1994 : Classification of mass and non-mass regions on mammograms using artificia1 neural networks, Imaging Sei, Technol

[4] Cottrell M. and Rousset P., 1997 : A powerful Tool for Analyzing andRepresenting Multidimensional Quantitative and Qualitative Data. InProceedings of IWANN'97, pages 861-871, Springer Verlag, Berlin.

[5] Cottrell M., Fort J.C., Pagès G., 1998 : Theorical aspects of theSOMalgorithm, Neuro Computing, 21, pages 119-138.

[6] Cottrell, M., Fort, J.C., Pagès, G., 1995. Two or three things that we know about the Kohonen algorithm, in Proc of ESANN'94, M. Verleysen Ed., D Facto,Bruxelles, p.235-244.

[7] Cottrell,M., SmaïI I., Letrémy P., Rousset P. : Cartes auto-organisées pour l'analyse exploratoire de données et la visualisation , SAMOS-MATISSE (UMR 8595), Univ. Paris 1.

[8] Feki R. et Khoufi W., 2008 : Caractéristiques des entreprises défaillantes à l'aide des cartes auto organisatrices de Kohonen (Cas des PMI Tunisiennes), La revue comptable et financière N1.

[9] Fei Sha, Lawrence K., Saul, Daniel D., Lee, 2003 : Multiplicative updates for nonnegative quadratic programming in support vector machines. Advances in Neural Information Processing Systems 15, Sebastian and K. Obermayer, Eds. Cambridge, MA: MIT Press.

[10] Kohonen T., 1990 : Learning Vector quantization , IJCNN , San Diego, pp. 223-226.

[11] Kohonen T, 1988 : Self-Organization and Associative Memory, Second Edition, Springer-Verlag, New York.

[12] Lebart L., 1999 : Qualité de l'information dans les enquêtes : QUESTII ¨ O´, vol. 23, 2, p. 321-342

[13] Madden Michael G., 2002 : Evaluation of the Performance of the Markov Blanket Bayesian Classifier Algorithm. CoRR, c s . LG/0211003.

[14] Serrano Cinca, C., 1996 : Self Organizing Neural Networks for Financial Diagnosis, Decision Support Systems, 17.

[15] Subhash C. B., Sikha B., Kuhu P., and Nikhil R, 2003 : Breast cancer detection using rank nearest neighbor classification rules. Pattern Recognition, 36:25 34.

[16] Thiria S. , Ambroise C., Sèze G. et Badran F., 2000: Hierarchical clustering of self-organizing maps for cloud classification , Neurocomputing, Volume 30, Issues 1–4, January 2000, Pages 47–52

[17] Verma B., 2008 : Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms, In: Artificial Intelligence in Medicine, pp. 4267-79.

[18] Vesanto J. and Alhoniemi E., 2000 : Clustering of the Self-Organizing Map, IEEE transactions on neural networks, 11(3).