IJST EXPLORE International Journal of Science and Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-7318

International Journal of Science and Technology Volume 3 No. 2, February, 2014

Visualization of Breast Cancer Data by SOM Component Planes

P.Venkatesan.[1], M.Mullai[2]
[1]Department of Statistics,NIRT(Indian Council for Medical Research),Chennai
[2]Department of Mathematics, Ethiraj College For Women,Chennai

ABSTRACT

The component plane presentation of integrated self-organized maps(SOM) is a powerful artificial intelligence tool for analysis of large, complex, biological databases. This approach allows the display of multi-dimensional SOM outputs disease data bases in multiple sample specific presentation providing distinct advantages in visual inspection of biological significance of features clustered in each unit.

This paper attempts to analyse the behavior of the various attributes constituting breast cancer and the correlation between them through their component planes. This additional information obtained will, definitely, complement the results obtained through other techniques.

Keywords: *Medical diagnosis, Artificial intelligence (AI),Artificial Neural network(ANN), Self Organizing Map(SOM), Best Matching Unit(BMU),Component Plane(CP).*

I. INTRODUCTION

Artificial Neural Network (ANN) is a powerful AI technique playing a vital role in the medical field. It has emerged as one of the indispensible tools in clinical diagnosis for its ability to explore medical data.ANN is capable of handling data exploration efficiently and effectively and hence enhances the decision making in the medical field when used for classification, clustering, pattern recognition and prediction[Akay et al 1994,Bishop 1995,Pattichis et al 1995].

Supervised and unsupervised are two of the modes in which ANN performs. supervised learning necessitates the presence of a desired output result for each input while training the network where as desired output is not available for unsupervised learning. Self organizing map(SOM), introduced by Teuvo Kohonen(2001), is one of the best known unsupervised, competitive, natural learning algorithm[Kohonen T,1995].

Data mining is ‘ non-trivial extraction of implicit, unknown and potentially useful information from data’ (Frawley et al, 1992). Its aim is to search through the data for its hidden features and other relations. Scatter plot is one of the traditional techniques to identify the dependencies or the relations between the variables. But this technique does not seem to be practical where variables are more in number. Visualization technique based on component planes (Miguel A et al,2007) helps to find the correlating attributes.

II. SELF ORGANISING MAP

A self-organizing map consists of components called nodes or neurons [Kohonen T, 1982]. Weight vector of the same dimension as the input data vectors is associated with each node in the map space. The usual arrangement is a regular spaced nodes in a hexagonal or rectangular grid [Ultsch&Siemon, 1989]. SOM describes a mapping from a higher dimensional input space to a lower dimensional map space, usually two dimensional. SOM combines the features of vector quantization and vector projection.

During the learning process, weight vector for each node is found. In each iteration, the most similar node called the best matching unit (BMU) is found by a similarity measure. Learning takes place with an objective of grouping similar nodes of the BMU and updates their corresponding weight.

SOM is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, called a map[Lipp,1987].Self organizing maps are different from other artificial neural networks in the sense that they use neighbourhood function to preserve the topological properties of the input space.SOM is organized so that similar data are mapped on to the same node or to the neighbouring nodes and there by similar input patterns are spatially clustered and get organized themselves.

SOM Process

Competitive process identifies the BMU, **Co-operative process** locates the neurons closer to BMU using a neighbourhood function, most preferably Gaussian

$$\mathbf{h_{j,i(x)}(n) \;=\; exp\;\; [- d^{2}_{j,I} \,/\;\; 2\,\sigma(n)]}$$

having the property of being symmetric about the maximum point, monotonically decreasing with time (a unique feature of SOM). Here, d_{ij} is the lateral distance between the BMU and excited neuron. σ explains the width of the neighbourhood, which shrinks with time. **Adaption process** enables the excited neurons to increment weight vector by suitable adjustment

$$\mathbf{w_j(n+1) \;=\; w_j(n) \;+\;\; n\; h_{ji\;(x)}\,(x - w_j(n))}$$

Self organizing of the neurons and fine tuning of feature map take place during the adaptation process. Topological ordering, density matching and feature selection are the distinct properties of SOM making it the most effective tool.

Component Planes

Visualization of data by component planes is one among the various visualization techniques using SOM grid(Marcos et al 2004,Miguel A et al,2007).They show the emerging patterns of data distribution on SOM grid(Kohonen,2001), and detect dependencies among variables and contribution of each one to the SOM grid. This depends on code book vectors and hence a more comprehensive method. Each component plane exhibits the behavior of an attribute constituting the data using colour coding. With the help of component planes, data set can be explored for interpretation of patterns and association of the attributes and hence gain knowledge about the data set.Kaski et al (1996) applied both U-matrix and CP to classify countries by their socioeconomic global indexes. Winter et al(1994) used CP to analyze the racial segregation issue of South Africa.

Data base

Breast cancer data obtained from UCI -Machine Learning Repository is used for the purpose . Data consists of details of 699 individuals tested for breast cancer based on 32 real valued features contributing the disease attributes and nine important features, *viz*, radius, texture, perimeter , area, smoothness, compactness, concavity, concave points and symmetry. The component planes for the attributes are obtained for both actual and normalized data with Gaussian neighborhood and bubble kernel. SOM_PAK (version 3.1) has been used in matlab (version 7.9.0.529-R2009b) for the purpose (Vesanto .J etal,1999). Correlation co-efficient (given within brackets) are obtained with the help of viscovery SOMine 5 software. Map is trained using batch algorithm.

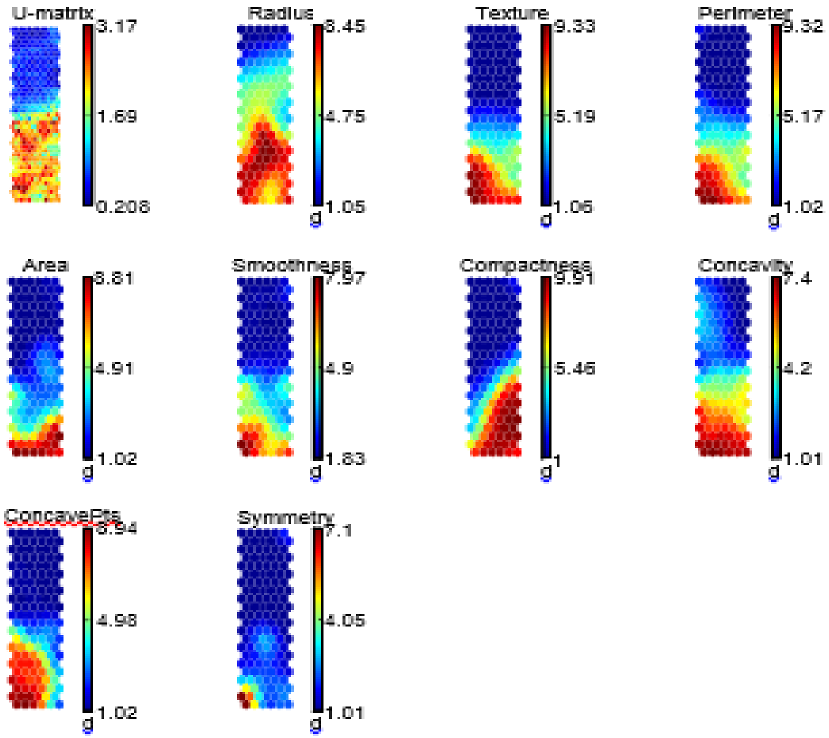

Figure-1: Component Planes(Actual Data) with Gaussian neighbourhood

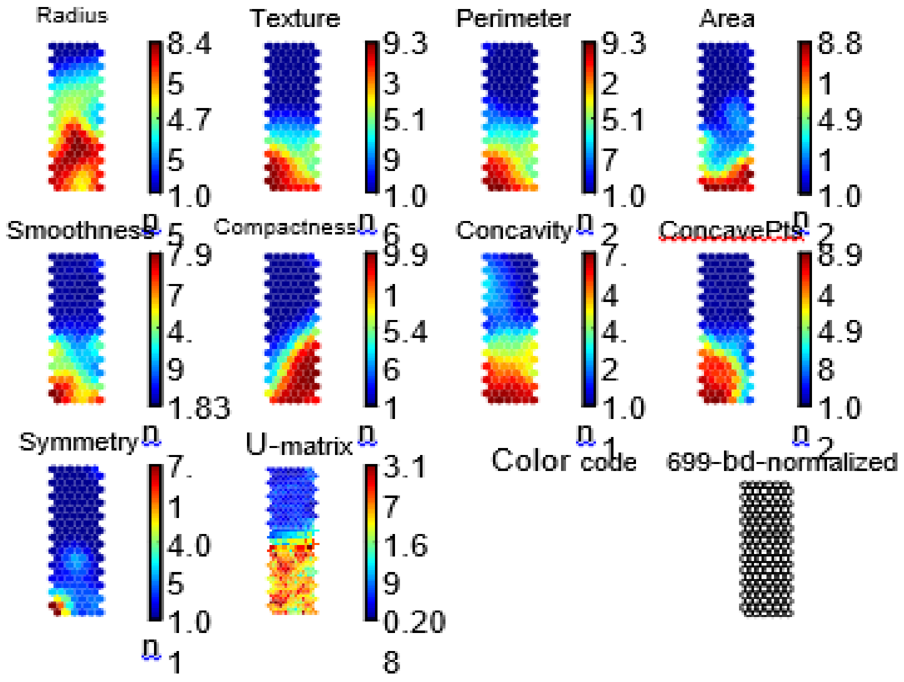

Figure-2.Component Planes (Normalized Data) with Gaussian neighbourhood

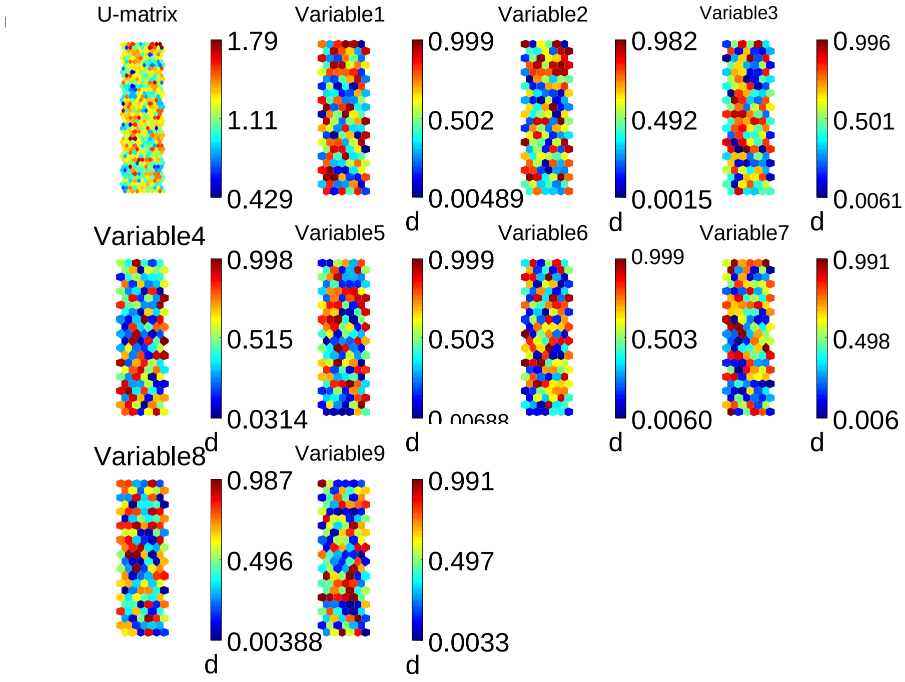

Figure-3 Component Planes(Normalized Data) with Bubble neighbourhood

Final quantization error : 2.489 ; Final topographic error : 0.033

Observation

Figure 1. shows a U-Matrix and nine CPs for each of nine attributes where high values are primarily located below the first third of the map. U-matrix has three short clusters in the lower part. This suggests that these areas could be associated with radius, texture, perimeter and compactness. Moreover, while comparing the U-matrix and CP of radius, both of them have void space in similar position which might suggest that radius seem to dominate the decision. At this point, one can make two inferences that data set does not have a similarity or the U-matrix could not separate the dataset in a proper manner.

For the CPs, high value is associated with red colour and low values with blue colour. Some visible patterns can be observed from the component planes. CPs for texture and perimeter have a very similar colour pattern indicating a strong correlation between them.(0.6568, the highest correlation among the attributes).Both of them have deep dark left bottom corner an identical nature of their CPs.

It is also observed that the attribute ‘symmetry’ contributes very less to the prediction and is negatively correlated (-0.3469) to compactness as their CPS do not have similar colour coding of the regions. From figure-2 it is inferred that CPs pertaining to the actual data and normalized data look identical with the Gaussian neighbourhood. CPs pertaining to bubble neighbourhood (fig.3) looks totally different and does not help us to draw inferences as the case in Gaussian neighbourhood.Performance of Gaussian function can be attributed to its nature such as uni-modal, monotonic decreasing with increasing neighbourhood (necessary condition for convergence), translation invariance (independent of location of winning neuron), attainment of maximum value at the winning neuron and dependence on neighbourhood (the necessary condition for co-operation among neighbouring neurons) while bubble neighbourhood uses step function.
The quantization error assesses the vector quantization properties. Quantization being 2.489 implies the fact that more neurons represent, stability, better input. The simplest measure of topology preservation, the topographic error, being 0.33 indicates the better quality of the map.

III. CONCLUSION

Comparison of U-matrix with the component planes show the contribution of the attributes to the data set and comparison between the component planes show their correlations visually, which is an interesting and important aspect. Moreover, the attributes whose contribution is negligible (revealed by their respective component planes) can be ignored and attributes contributing almost equivalently can help to remove the redundancy in data, there by number of attributes can, further, be reduced enabling the reduction of data size, a welcoming aspect indeed.

REFERENCES

[1]. Bishop,C.M(1995).Neural Networks for pattern recognition, Oxford University Press, NEW YORK

[2]. Brause, R.W.,(2001). Medical analysis and diagnosis by neural networks. *Proceedings of Second International Symposium on Medical data Analysis*. Oct 08-09, Springer-Verlag, LONDON, UK.,pp: 1-13

[3]. Frawley.W.J,Piatetsky-Shapiro& Matheus(1992). Knowledge discovery in_databases –an overview *AI Magazine* 13: 57-70

[4]. Kasi .S ,Kohonen.T(1996) Exploratory Data Analysis by the self Organizing Map: *Proceedings of the Third International conference on Neural Networks in the capital Markets* 498-507

[5]. Kohonen, T. (1982) Self-organizing formation of topologically correct feature__maps, *Biological Cyberbetics*. Volume 43, 1982, pp:59-96

[6]. Kohonen, T. (1995) Self-Organizing Maps, *Springer Series in Information Sciences*, Vol. 30, Springer, Berlin, Heidelberg, NEW YORK, 1995

[7]. Kohonen, T. (1997) Self-Organizing Maps__,2nd edn.,Springer Verlag

[8]. Kohonen, T. (2001) Self-Organizing Maps, *Springer Series in Information Sciences*, Vol. 30, Springer, Berlin, Heidelberg, NEW YORK, 2001

[9]. Marcos A,Antonio M,Jose SM(2004) Visulization of geospatial data by component planes and U-Matrix *Proceedings of geoinormation*.75-89

[10]. Miguel A ,Andres P(2007) Improving the correlation hunting in a large quantity of_SOM Component *planes ICANN,Part II,LNCS* 4669.379-388

[11]. Pattichi C.S(1995*) IEEE Tranc Biomed Eng* 42: 486-496

[12]. Ultsch A &Siemon H. P(1989) Exploratory Data Analysis: Using Kohonen Networks on Transputers, *Research Report No. 329, University of Dortmund*, 1989

[13]. Vesanto J(1999) SOM based visualization methods. *Intelligent Data Analysis*,3:111-126,

[14]. Vesanto, J., Ahola, J (1999).: Hunting for Correlations in Data Using the Self-Organizing Map. In: *Proceeding of the International ICSC Congress on*

Computational Intelligence Methods and Applications, pp. 279–285

[15]. Vesanto .J,Himberg .J,Alhoniemi.E,Parhankangas,J (2000) SOM Toolbox for Matlab 5, Report A57,April 2000,ISBN 951-22-4951-0

[16]. Winter,K,Hewitson.B (1994). Self organizing maps-applications to cencus data.In Hewitson.B,Crane.R,editors,*Neural nets:applications in geography*.kluwer.