Clustering of Disease Data base using Self Organizing Maps and Logical Inferences

Dr.P.Venkatesan[1] and M. Mullai[2]

[1]*Department of Statistics,NIRT(ICMR),Chennai*
[2]*Department of Mathematics, Ethiraj College For Women,Chennai*
venkaticmr@gmail.com [1], *mmullai5@gmail.com* [2]

Abstract

Disease classification requires an expertise in handling the uncertainty. ANNs emerge as a powerful tool in this regard. ANNs have featured in a wide range of applications with promising results in biomedical sciences. The self-organized maps (SOM) use unsupervised learning to produce low dimensional discretized representation of the input space. SOMs are different from other neural networks in the sense that they use neighborhood function to preserve the topological properties of the input space. This paper compares Kohanen's SOM network with other clustering method. The SOM gives faster and accurate results in clustering the data. The results were presented and compared.

Keywords: Medical diagnosis, Artificial intelligence (AI), Neural network, Self Organizing Map(SOM), Best Matching Unit(BMU),Tuberculosis (TB).

1. Introduction

Artificial Intelligence (AI) is the branch of computer science, which intends to make computers more intelligent. AI may probably be the single most successful technology in the last decades, which has been widely used in a variety of applications in diversified areas. AI, as technique mimic the processing way of information in human brain, has emerged as promising methods in dealing with non-linear and complex relations (Victor Alves *et al.,* 2003). The ability to learn, tolerance to data noises and capability of modeling incomplete data has made it unique analyzing approaches in many scientific procedures.

Expert or knowledge based systems are the common type of AI in routine clinical use. They contain medical knowledge usually about a very specifically defined task and are able to handle the data effectively to arrive at an appropriate result (Brause, 2001). There are many different types of clinical task ,such as generating alerts and remainders, diagnostic assistance, therapy planning, image recognition and interpretation to which expert systems can be applied. One of the most successful areas in which expert systems are applied is in the clinical laboratory.

Artificial Neural Network (ANN) is a powerful AI technique possessing the ability to learn a set of data and build weight matrices to denote learning pattern. ANN can learn from experiences to improve the performance and has the ability to adopt itself to changes in the environment. They can be very effective in handling incomplete information or noisy data and in situations where specified rules cannot be set to solve a problem.

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. Back Probagation Neural Network (BPNN), Radial Basis Function Network (RBFNN), Probilistic Neural Network (PNN) and Generalized Regression Neural Network(GRNN) are a few of the supervised learning type.

2. Self-Organising Map

Self organizing map (SOM), introduced by Teuvo Kohonen (2001), is one of the best known unsupervised natural learning algorithm(Kohonen, 1995). It is a feed forward network that has been used for medical image segmentation (Victor Alves *et al.,* 2003). It projects a high dimensional space into a low dimensional space. This type of architecture is fundamentally different in arrangement of neurons and motivation consisting of a two dimensional array of nodes. Each node is associated with a weight vector $\{w_j\}$,of the same dimension

, and a position in the map space. An unsupervised, iterative learning process finds the weight vector for each node. In all iteration, the most similar node called the best matching unit (BMU) was found by a similarity measure. A training process considers the neighboring 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 (Lippman, 1987). Self-organizing maps are different from other artificial neural networks in the sense that they use neighborhood function to preserve the topological properties of the input space. SOMs are useful in visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. Training builds the map using input examples. It is a competitive process, also called vector quantization. Mapping automatically classifies a new input vector.

A self-organizing map consists of components called nodes or neurons (Kohonen, 1982). Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid (Ultsch &Siemon, 1989). The self-organizing map describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to find with the closest weight vector to the vector taken from the data space. Once the closest node is located, it is assigned the values from the vector taken from the data space.

3. Process

Competition, Co-operation and Synaptic adaption are the processes involved in the formation of SOM. In the competition process, neurons with largest discriminant function $W_j^T X$ is declared the winner. Co-operation process locates the excited neurons with the help of a topological neighborhood function, most preferably, the Gaussian function

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

having the property of being symmetric about the maximum point, onotonically decreasing with time(a unique feature of SOM) Adaption process enables the excited neurons to increment weight vector and there by increment their discriminant functions.

$$w_j(n+1) = w_j(n) + \Delta w_j$$

$$\Delta w_j = n\, h_{ji\,(x)}(x - w_j(n))$$

Properties such as topological ordering, variations in mapping and feature selection make SOM the most effective tool.

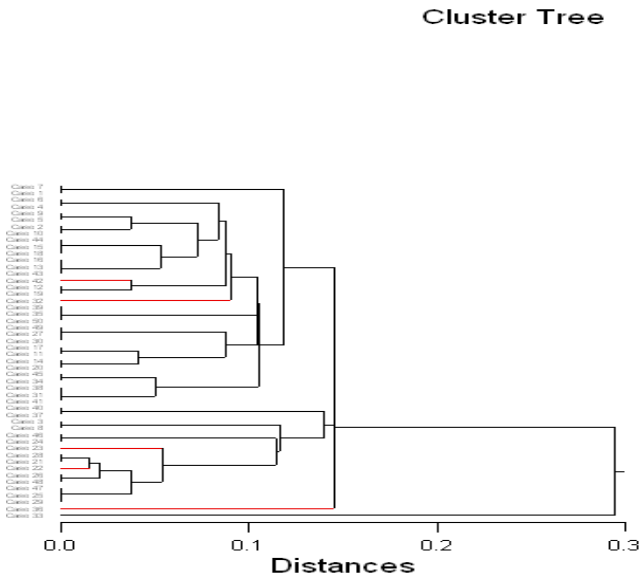
4. Database

For empirical comparisons, we have considered a data consists of 50 individuals (10 individuals from 4 categories of patients namely TB, TB lymphadenitis, Asthma and COPD and 10 healthy individuals). The spectral wave length in the mid infrared regions was considered within the range of 400-4000.Using the feature extraction technique namely principal component analysis(PCA), components which accounted for 99 percent variations were chosen for the clustering using SOM and other algorithms.

5. Clustering

Clustering of numerical data forms the basis of many classifications and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behaviour.

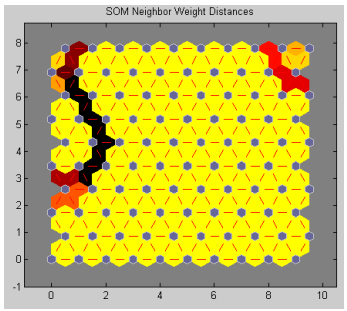
Fig.1. *Hierarchial clustering*



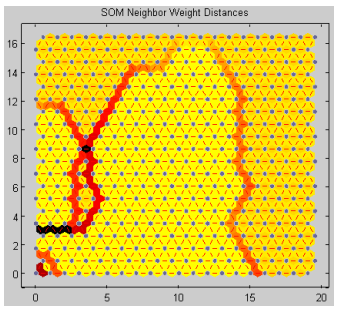
6. Hierarchial clustering

As a first step, traditional hierarchial clustering method (single linkage) is used to cluster the data (Spath, 1980). Single linkage tends to give the best results, compared to complete and median linkage. Here, order of the branching point projection on the horizontal axis corresponds to the sequence of clustering.

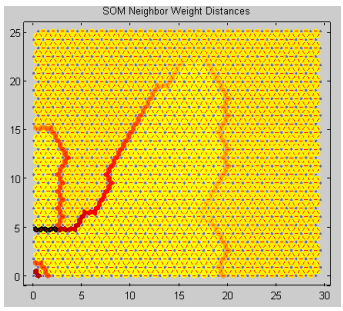
Fig.2. *SOM clustering in 10x10, 20x20, 30x30 neurons*



10 x 10

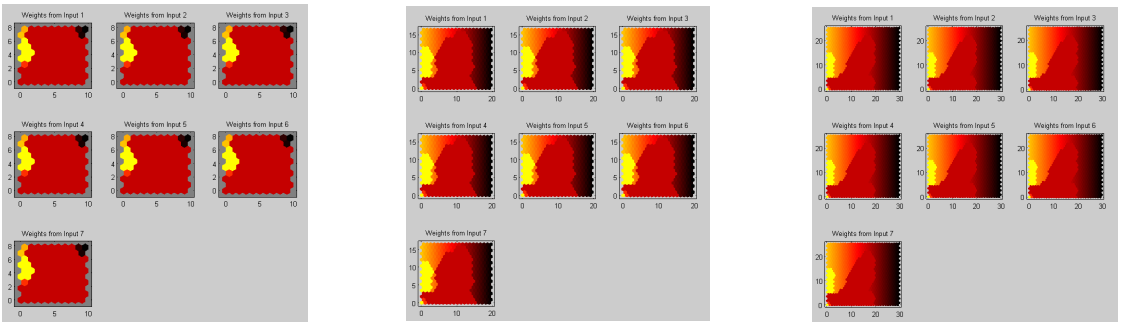


20 x 20



30 x 30

Fig.3. *SOM weights from input in the 10x10 20x20 30x30 Hexagonal grid*



10 x 10 20 x 20 30 x 30

7. SOM Clustering

Next, data is clustered using SOM. Two-dimensional SOMs with hexagonal (default topology of the SOM) organizations of the centroids have been obtained using the matlab software. Each hexagon represents a node in the SOM. Light colours depict the closely spaced nodes. Darker colours indicate nodes that are more distant. Thus, groups of light colours can be considered as clusters and the dark parts as boundary regions. Clusters can be further highlighted by plotting appropriate labels on the map. Clustering done on 20 x 20 neurons gives a better picture.that the inputs are highly correlated.Following figure 2,3,4 shows a weight plane for each element of the input vector. They are the visualization of the weights that connect each input to the neuron (darker colours represent larger weights).As the connection patterns of the input are very similar, it can be assumed.

8. Conclusion

In bench markings and practical problems, not only the ultimate accuracies but the speed of the computation, too, should be compared.SOM which is similar to K-means clustering (Hartigan & Wong, 1979], is not only faster than the hierarchical clustering but little richer. With K-means, one can choose the number of clusters to fit the data into (Gordon, 1994). For SOM one can choose the shape and size of a network of clusters to fit the data into. This technique proved to be a better tool for handling data for the sake of clustering.

9. References

1. Brause, R.W.,(2001). Medical analysis and diagnosis by neural networks.Proceeings of Second International Symposium on Medical data Analysis. Oct 08-09, Springer-Verlag, Lonon, UK.,pp: 1-13
2. Gordon A.D, (1994). Identifying Genuine Clusters in a Classification, Computational Statistics and Data Analysis 18 ,pp: 561-581
3. Hartigan J. A & Wong M. A, (1979). A K-means Clustering Algorithm, Applied Statistics,28,pp: 100-108
4. Kohonen, T. (1982) Self-organizing formation of topologically correct feature maps, Biological Cyberbetics. Volume 43, 1982, pp:59-96
5. Kohonen, T. (1995) Self-Organizing Maps, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995
6. Lippman R.P (1987) An Introduction to Computing with Neural Nets, IEEE, ASSP Magazine, April 1987, pp: 4-22
7. Spath H (1980) Cluster Analysis Algorithms,Chichester, UK, 1980
8. Ultsch A &Siemon H. P(1989) Exploratory Data Analysis: Using Kohonen Networks on Transputers, Research Report No. 329, University of Dortmund, 1989

9. Victor Alves, Paulo Novais, Luis Nelas, Moreira Maia & Victor Ribeiro (2003) Case based reasoning versus artificial neural networks in medical diagnosis. Proceedings of IASTED International Conference Artificial Intelligence and Applications. pp: 1-5