IJIREEICE

ISSN (Online) 2321-2004
ISSN (Print) 2321-5526

International Journal of Innovative Research in
Electrical, Electronics, Instrumentation and Control Engineering

Vol. 6, Issue 10, October 2018

Instigating Self Organizing Map with Linear Neurons for Effective Gaussian Clustering

Priyanka.D[1], **Dr.S.Prema**[2]

M. Phil Research Scholar, Department of Computer Science (PG),

K. S. Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu, India[1]

Associate Professor, Department of Computer Science (PG),

K. S. Rangasamy College of Arts and Science (Autonomous), Tiruchengode, Tamilnadu, India[2]

Abstract: The Self-Organizing Map (SOM) is an excellent tool in experimental phase of data mining. It projects input space on models of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large, to simplify quantitative analysis of the map and the data, similar units need to be grouped, i.e., clustered. Self-organizing maps are known for its clustering, visualization and classification capabilities. SOM is a popular unsupervised artificial neural network algorithm to produce a low dimensional, discredited representation of the input space of the training samples, called a feature map. Such a map retains standard features of the input data. The Expectation Maximization (EM) algorithm was proposed to calculate the Maximum Likelihood Estimator (MLE) in the occurrence of missing observations. An EM algorithm that yields topology preserving maps of data based on probabilistic mixture models. Compared to other mixture model approaches to self-organizing maps, the function our algorithm maximizes has a clear interpretation; it sums data log-likelihood and a penalty term that enforces self-organization. Finally allows just handling of missing data and learning of mixtures of self-organizing maps. The proposed SOM and EM algorithm is implemented with MATLAB.

Keywords: Self-Organizing Map (SOM), Neural Networks, Feature Map, Clustering, Unsupervised Learning, Mixture Models, EM Algorithm

I. INTRODUCTION

Data Mining is the process of extracting or mining knowledge or analyzing large amounts of data in order to discover patterns and other information. It is usually performed on databases, which store data in a structured format. By "Mining" large amounts of data, hidden information can be exposed and used for other purposes. The data mining step may relate with the user or a knowledge base. The exciting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. According to this view, data mining is only one step in the entire process, even though an essential one since it uncovers hidden patterns for evaluation.

The data mining is a knowledge discovery process. However, in industry, in media, and in the database research location, the term data mining is becoming more popular than the longer term of knowledge discovery in databases. Data mining is the process of learning interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. A data mining task can be stated in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives to interactively connect with the mining system. Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have achieved.

Various techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Data mining is a relatively young discipline with wide and various applications; there is still a nontrivial gap between general principles of data mining and application-specific, effective data mining tools. In this section, we examine several application domains. For addressing all these issues, this paper proposes SOM and EM algorithm with the help of clustering is implemented with MATLAB. Let us discuss about the existing problems and the proposed work carried on in a brief manner by the following processes.

DOI 10.17148/IJIREEICE.2018.6104

II. LITERATURE REVIEW

A. Generalized Self-Organizing Maps for Automatic Determination of the Number of Clusters and Their Multiprototypes in Cluster Analysis Marian B. Gorzałczany et.al [6] presents a generalization of self-organizing maps with 1-D neighborhoods (neuron chains) that can be effectively applied to complex cluster analysis problems. The results of extensive experiments demonstrate the generally low sensitivity of our approach to changes in a broad range in control parameters. The application of an earlier version of our approach to WWW document clustering.

B. Topology Based Clustering Using Polar Self-Organizing Map: Lu Xuet.al [5] proposed clustering method provides a visual representation as polar self-organizing map (PolSOM) allows the characteristics of clusters to be presented as a 2-D polar map in terms of the data feature and value. Theoretical analysis and extensive simulations show the effectiveness and clustering superiority of the algorithm.

C. Selectable and Unselectable Sets of Neurons in Recurrent Neural Networks with Saturated Piecewise Linear Transfer Function: Lei Zhang et.al [4] concepts of selectable and unselectable sets are proposed to describe some interesting dynamical properties of a class of recurrent neural networks (RNNs) with saturated piecewise linear transfer function. A set of neurons is said to be selectable if it can be co-unsaturated at a stable symmetry point by some external input. A set of neurons is said to be unselectable if it is not selectable, i.e., such set of neurons can never be co-unsaturated at any stable symmetry point regardless. To illustrate the importance of these new concepts, the application of this class of RNNs for group selection was studied. The relationships among groups and selectable sets, as well as the external inputs, were explored.

D. Distribution Based Cluster Structure Selection: Zhiwen Yuet.al [10] investigates the problem of how to select the suitable cluster structures in the ensemble which will be summarized to a more representative cluster structure. Specifically, the cluster structure is first represented by a mixture of Gaussian distributions, the limitations of which are estimated using the expectation–maximization algorithm. DSCE is effective in elucidating the cluster structure of different real-world datasets. In the future, we will perform distribution-based cluster structure ensemble (DSCE) on more real-world applications.

E. Bayesian Cluster Enumeration Criterion for Unsupervised Learning: Freweyni K. Teklehaymanotet.al [3] proposes a two-step cluster enumeration algorithm. First, a model-based unsupervised learning algorithm partitions the data according to a given set of applicant models. Subsequently, the number of clusters is determined as the one associated with the model for which the proposed Bayesian Information Criterion (BIC) is maximal. Simulation results indicate that the penalty term of the proposed criterion has a curvature point at the true number of clusters which is created due to the change in the trend of the curve at that point. The analysed summary of existing work is given in table 1.

Table I: Summary of Literature Review

S. No	Title	Author, Publisher and Year	Working Platform	Objective	Future Scope
1.	Generalized Self-Organizing Maps for Automatic Determination of the Number of Clusters and Their Multiprototypes in Cluster Analysis	Marian B. Gorzałczany et.al [6] IEEE [2018]	Multi-prototypes in Cluster Analysis	Automatic Determination of the Number of Clusters	These features enable the network to automatically determine the number of clusters in a given data set and to generate a collection of multiprototypes.
2.	Topology Based Clustering Using Polar Self-Organizing Map	Lu Xu et.al [5] IEEE [2015]	Topology Based Clustering	Polar Self-Organizing Map	Theoretical analysis and extensive simulations show the effectiveness and clustering superiority of the algorithm.
3.	Selectable and Unselectable Sets of Neurons in Recurrent Neural Networks with Saturated Piecewise Linear Transfer Function	Lei Zhang et.al [4] IEEE [2011]	Recurrent Neural Networks	Selectable and Unselectable Sets of Neurons	Selectable and unselectable sets in this class of RNNs were established by rigorous mathematical analysis.

4.	Distribution Based Cluster Structure Selection	Zhiwen Yu et.al [10] IEEE [2017]	EM Algorithm	Cluster Structure Selection	A unified cluster structure from multiple cluster structures obtained from different datasets.
5.	Bayesian Cluster Enumeration Criterion for Unsupervised Learning	Freweyni K. Teklehamanot et.al [3] IEEE [2018]	Bayesian Cluster	Unsupervised Learning	The estimation of the number of data clusters in unsupervised learning problems.

III. IDENTIFIED PROBLEM FROM EXISTING SYSTEM

During the connection of each data from SOM, some problems are sorted out in this existing system are,

- That every SOM has different similarities among the sample vectors.
- Sometimes, it is very difficult to acquire right data so this is a limiting feature to the use of SOMs often referred to as missing data.
- The dimensions of the data increases, dimension reduction imagining techniques become more important, but unfortunately then time to compute them also increases.

These identified problems are solved with the help of implementing a new algorithm named EM (Expectation Maximization) to calculate the maximum likelihood estimator (MLE) in the proposed thesis work. The proposed algorithm is implemented with MATLAB.

IV. PROPOSED WORK: SELF-ORGANIZING MAP FOR CLUSTERING AND GAUSSIAN MIXTURE MODELS

This provides an overview of the techniques which acts as the base for the proposed algorithm.

- Self-Organizing Map
- Gaussian Mixture Models
- Artificial Neural Networks

These are the three core areas which this research focuses on. These three machine learning techniques are explained subsequently.

A. Self-Organizing Map: Self-Organizing Map provides a proficient platform for the exploratory data analysis. The input data can be projected into grid structure and easily visualized and properties can be explored in significant manner. The Self-Organizing Map can be deployed for the Gaussian distribution. The input data can be projected into vectors. The exploration of the vectors topology and distribution can be learned easily which will end up in effective clustering. Self-Organizing Map (SOM) is a parts located within a rigid area such as a grid. Different kinds of the grids can be deployed based on the need of the problem. It is a visualization method to represent higher dimensional data is usually a 1-D, 2-D or 3-D manner. Two most used two dimensional grids in SOMs are rectangular and hexagonal grid. Three dimensional topologies can be in form of a cylinder or toroid shapes.

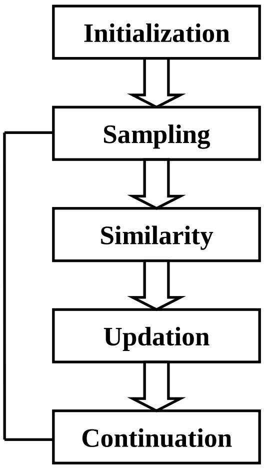

Figure 1: Steps Involved in SOM

DOI 10.17148/IJIREEICE.2018.6104

B. Gaussian Mixture Models: Mixture models represent the presence of the sub-populations in the overall populations. The sub-populations are exempted to replicate or resemble the overall population and act as a candidate for representing overall model. Data can either be multimodel or unimodel. Multimodel has more than one core points where the algorithm can track or retrieve multiple features at a time. In unimodel variant one core point is focused and retrieved. Encapsulation of the models has to be done very carefully.
Expectation Maximization (EM) Algorithm for Gaussian Mixture Models: The Expectation Maximization (EM) algorithm was proposed to calculate the maximum likelihood estimator (MLE) in the presence of missing observations.

C. Artificial Neural Networks: A neural network is an interconnected gathering of simple processing elements, units or nodes, whose functionality is loosely based on the animal neuron. The processing ability of the network is stored in the inter unit connection strengths, or weights, obtained by a process of reworking to, or learning from, a set of training patterns. Human and animal brains are highly complex, nonlinear and parallel systems, consisting of billions of neurons integrated into numerous neural networks. Neural networks within a brain are massively parallel distributed processing system suitable for storing knowledge in forms of past experiences and making it available for future use. They are particularly suitable for the class of problems where it is difficult to propose an analytical solution convenient for algorithmic implementation.

V. ALGORITHM IMPLEMENTATION

A. The Operation for SOM Algorithm

- Initialization - Random values are chosen for the initial weight vectors $w_j(0)$. The weight vectors of all the neurons must be distinct and the magnitude of these weights is preferably a small value.

- Sampling - A sample x is drawn from the input space with a certain probability, this vector x represents the activation pattern applied to the lattice. The dimension of x is kept equal to m.

- Similarity Matching - The best matching winning neuron i(x) is found at time step n by using the minimum Euclidian distance criterion.

$$i(x)=\arg\min_j \|x - w_j\| \text{ where, } j=1, 2, \ldots, l$$

- Updation - The synaptic weight vectors of all the neurons are adjusted using the following formula,

$$w_j(n+1) = w_j(n) + □(n)\, h_{ji(x)}(n)\, (x(n) – wj(n))$$

where, □□(n) learning rate, $h_{ji(x)},(n)$ neighborhood function around the winner neuron i(x), □(n) and $h_{ji(x)}(n)$ could both be dynamically varied for obtaining optimal results.

- Continuation - Step 2 is continued until some noteworthy changes are found in the feature map.

B. Expectation Maximization (EM) Algorithm

$$\text{Goal: } \hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \overset{\text{log of expectation of } p(x|z)}{\log}[\textstyle\sum_z p(x, z \mid \theta)] \qquad f(E[x]) \geq E[f(x)]$$

1. E-Step: compute

$$E_{z|x,\theta^{(t)}}[\log(p(x, z \mid \theta))] = \overset{\text{expectation of log of } p(x|z)}{\textstyle\sum_z} \log(p(x, z \mid \theta))p\left(z \mid x, \theta^{(t)}\right)$$

2. M-Step: solve

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \textstyle\sum_z \log\left(p(x, z \mid \theta)\right) p(z \mid x, \theta^{(t)})$$

C. Measures in Clustering

- Minkowski Distance

  $d(x,y) = L_q(x,y) = \sqrt[q]{\sum_{i=1}^{n}(x_i - y_i)^q}$

- Manhattan Distance

  $d(x,y) = L_1 = \sum_{i=1}^{n}|x_i - y_i|$

- Euclidean Distance

  $d(x,y) = L_2 = \sqrt{\sum_{i=2}^{n}(x_i - y_i)^2}$

VI. RESULT AND DISCUSSION

A. Significant Aspects of the SOM

Factor 1: Supervised Learning

- In supervised learning, a chosen output result for each input vector is required.
- The supervised learning type, such as the multi-layer perceptron, uses the target result to monitor the formation of the neural parameters.

Factor 2: Unsupervised Learning

- An unsupervised learning algorithm which is used to visualize high dimensional data sets.
- Unsupervised learning is a technique of finding a structure in unlabelled data.

Factor 3: Neural Networks

- Neural network models which belong to the category of competitive networks.
- A neural network is an interconnected gathering of simple processing elements, units or nodes.

Factor 4: Clustering

- Clustering groups based of their mutual similarities.
- Clustering achieves high within cluster similarity and low inter-cluster similarity.

B. SOM Components

The self-organization process involves four major components,

- Initialization - All the connection weights are initialized with small accidental values.
- Competition - For each of the input patterns, a discriminant function is computed by the neuron which provides the basis for competition. The neuron with the smallest value of discriminant is declared the ‘winner’.
- Cooperation - The winning neuron determines the three-dimensional location of a topological neighborhood of excited neurons, thereby providing the basis for cooperation among neighboring neurons.
- Adaptation - The motivated neurons decrease their individual values of the discriminant function in relation to the input pattern through suitable adjustment of the associated connection weights, such that the response of the winning neuron to the successive application of a similar input pattern is enhanced. The discriminant function is defined as the squared Euclidean distance among the input vector x and the weight vector wj for each neuron j.

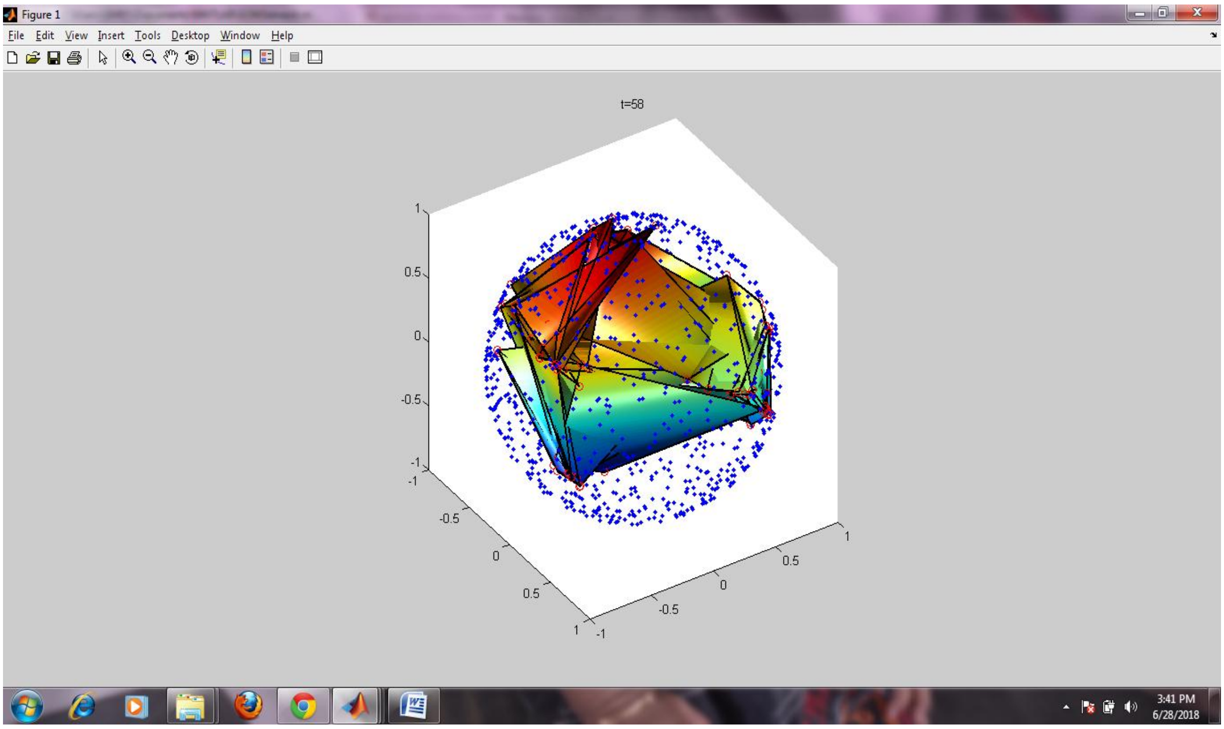

Figure 2: SOM 3-D MESH for the Gaussian Mixture Models: Initialization

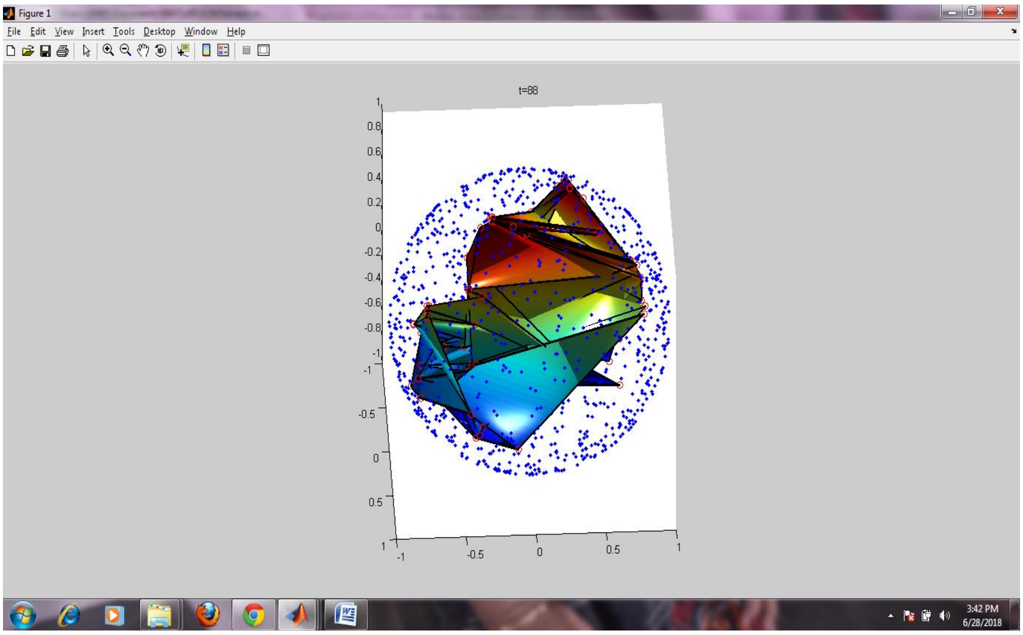

Figure 3: SOM 3-D MESH for the Gaussian Mixture Models: Competition and Cooperation

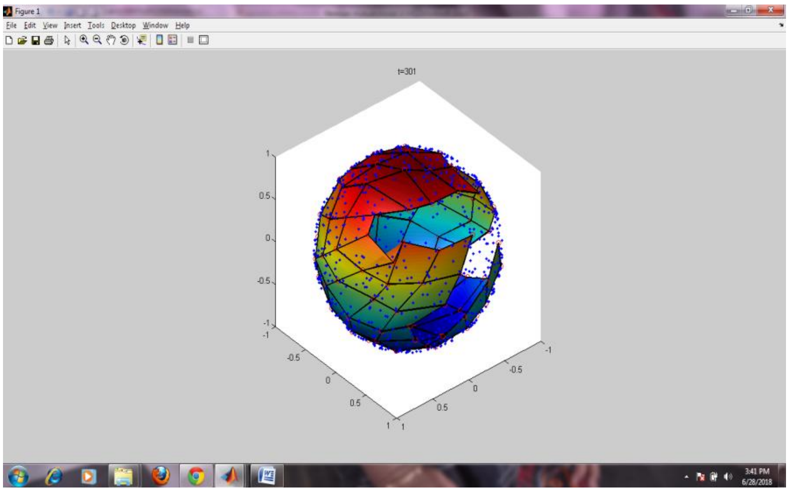

Figure 4: SOM 3-D MESH for the Gaussian Mixture Models: Neighbor Updating and Synaptic Adaptation

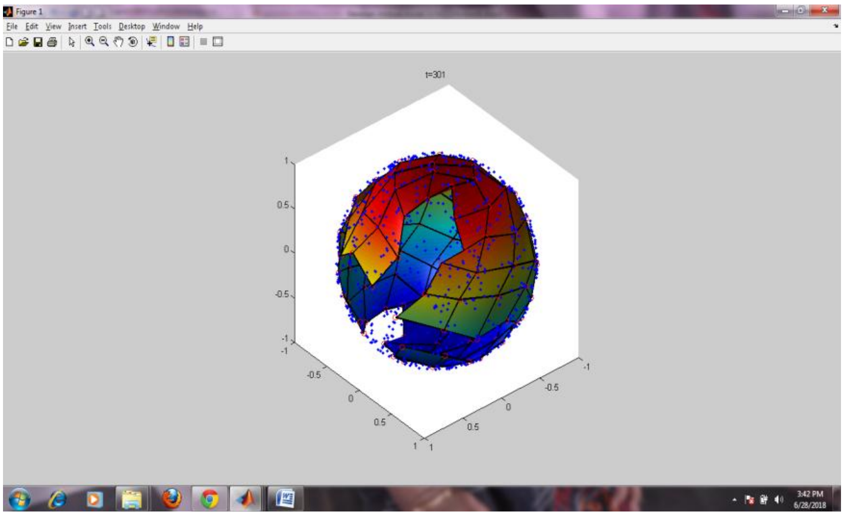

Figure 5: SOM 3-D MESH for the Gaussian Mixture Models: Continuation

DOI 10.17148/IJIREEICE.2018.6104

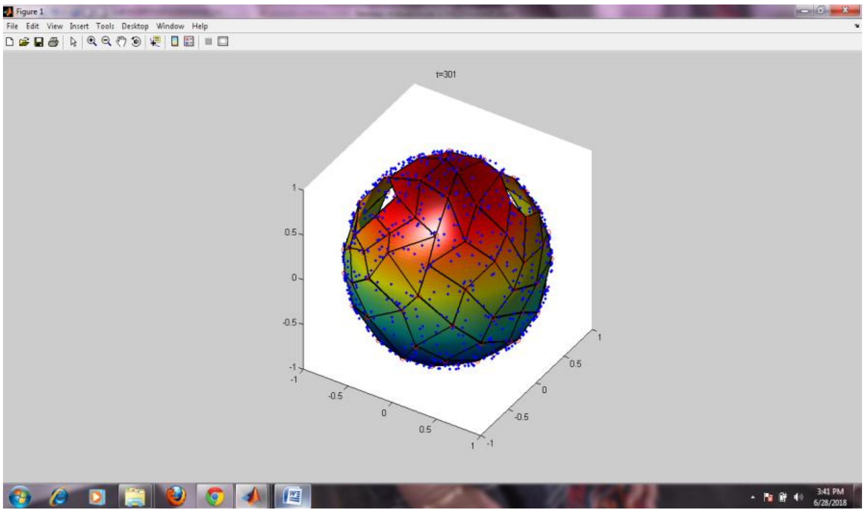

Figure 6: SOM 3-D MESH for the Gaussian Mixture Models: Components Map

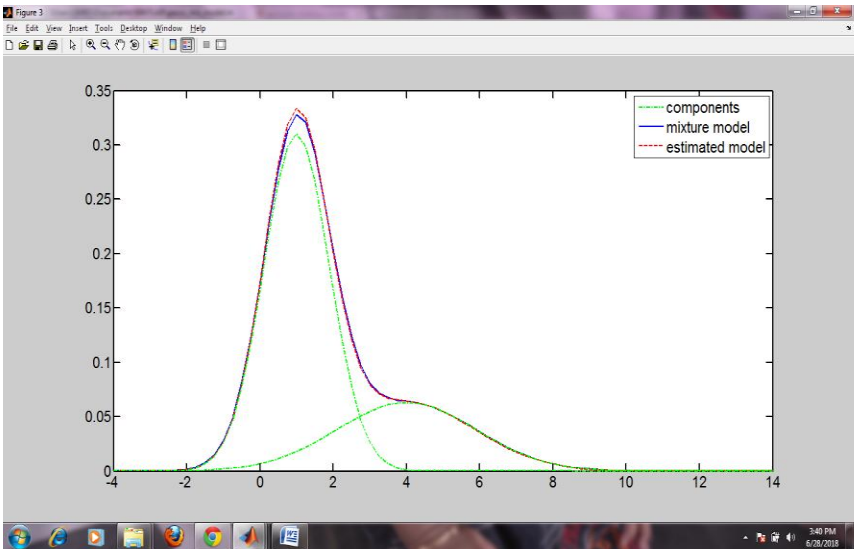

Figure 7: Comparison of the MSE in Mixture Model and Estimated Model

C. Summary: This research has offered an algorithm SOM, which is used to visualize high dimensional data sets and Clustering tool of high-dimensional and complex data. Self-organized map (SOM), as a particular neural network model has found its inspiration in self-organizing and biological systems. Neural networks of neurons with lateral communication of neurons topologically organized as self-organizing maps. The proficient points regarding SOM are,

- A self-organizing map (SOM) is a grid of neurons which adjust to the topological shape of a dataset, allowing us to visualize large datasets and identify potential clusters.
- An SOM learns the shape of a dataset by repeated moving its neurons closer to the data points. Distinct groups of neurons may thus reflect basic clusters in the data.
- SOMs are the best for datasets with continuous variables, and it should be repeated to check for consistency. Resulting clusters should also be validated.

CONCLUSION

The Self-Organizing Map (SOM) can be used to detect features inherent to the problem and thus has also been called Self-Organizing Feature Map (SOFM). It provides a topology conserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping from high dimensional space onto a plane. The property of topology conserving means that the mapping preserves the relative distance between the points. Points that are near each other in the input space are mapped to immediate map units in the SOM. This research has offered an algorithm SOM, which is used to visualize high dimensional datasets, clustering tool of high-dimensional and complex data. SOM as a particular neural network model has found its inspiration in self-organizing and biological systems. Neural networks of neurons with lateral communication of neurons topologically organized as self-organizing maps. It is concluded that the proposed method performs well when compared with the existing methods.

FUTURE WORK

High dimensionality of the data leads to loose potentially incurred knowledge. In order to overcome the problem in large dimensional data the SOM based method is proposed in this research. When the fuzzy and intuitionistic fuzzy methods are incorporated with the SOM in data analysis phase then it could yield better performance which would be the future scope of the research.

REFERENCES

[1]. Bei Zhao, Yanfei Zhong, Ailong Ma and Liangpei Zhang. “A Spatial Gaussian Mixture Model for Optical Remote Sensing Image Clustering.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol.9, No.12, pp.5748-5759. IEEE, December 2016.
[2]. Ezequiel López-Rubio. “Improving the Quality of Self-Organizing Maps by Self-Intersection Avoidance.” IEEE Transactions on Neural Networks and Learning Systems, Vol.24, No.8, pp.1253-1265. IEEE, August 2013.
[3]. Freweyni K. Teklehaymanot, Michael Muma and Abdelhak M. Zoubir. “Bayesian Cluster Enumeration Criterion for Unsupervised Learning.” IEEE Transactions on Signal Processing, Vol.66, No.20, pp.5392-5406. IEEE, October 2018.
[4]. Lei Zhang and Zhang Yi. “Selectable and Unselectable Sets of Neurons in Recurrent Neural Networks with Saturated Piecewise Linear Transfer Function.” IEEE Transactions on Neural Networks, Vol.22, No.7, pp.1021-1031. IEEE, July 2011.
[5]. Lu Xu, Tommy W. S. Chow and Eden W. M. Ma. “Topology-Based Clustering Using Polar Self-Organizing Map.” IEEE Transactions on Neural Networks and Learning Systems, Vol.26, No.4, pp.798-808. IEEE, April 2015.
[6]. Marian B. Gorzałczany and Filip Rudzinski. “Generalized Self-Organizing Maps for Automatic Determination of the Number of Clusters and Their Multiprototypes in Cluster Analysis.” IEEE Transactions on Neural Networks and Learning Systems, Vol.29, No.7, pp.2833-2845. IEEE, July 2018.
[7]. Punit Rathore, James C. Bezdek, Sarah M. Erfani, Sutharshan Rajasegarar and Marimuthu Palaniswami. “Ensemble Fuzzy Clustering Using Cumulative Aggregation on Random Projections.” IEEE Transactions on Fuzzy Systems, Vol.26, No.3, pp.1510-1524. IEEE, June 2018.
[8]. Richard Rzeszutek, Dimitrios Androutsos and Matthew Kyan. “Self-Organizing Maps for Topic Trend Discovery.” IEEE Signal Processing Letters, Vol.17, No.6, pp.2530-2538. IEEE, June 2010.
[9]. Sarwar Tapan and Dianhui Wang. “A Further Study on Mining DNA Motifs Using Fuzzy Self-Organizing Maps.” IEEE Transactions on Neural Networks and Learning Systems, Vol.27, No.1, pp.113-124. IEEE, January 2016.
[10]. Zhiwen Yu, Xianjun Zhu, Hau-San Wong, Jane You, Jun Zhang and Guoqiang Han. “Distribution-Based Cluster Structure Selection.” IEEE Transactions on Cybernetics, Vol.47, No.11, pp.3554-3567. IEEE, November 2017.

BIOGRAPHIES

Dr. S. Prema, currently working as an Associate Professor in K. S. Rangasamy College of Arts & Science has received Ph.D., from the Bharathiar University in 2015. She has been involved in the teaching for the past 13 years. She secured the 1st Rank in B.Sc under Periyar University, Salem. She has totally 52 publications and one of her research paper entitled “An NLP based Approach for Facilitating Efficient Web Search Results using BSDS” received the **Best Paper Award**. Her papers are cited at various publications (IEEE Xplore, Elsevier, Springer and International Conference Proceedings). She has h-index value: 5, i-10 index: 3, Citations: 125 and her profile is listed in Marquis Who is Who in World, International Biography Center London, UK, 2011. She has been awarded “Innovative Research & Dedicated Women Academician Award” at International Awards & Honors Convocation 2018 Conducted by The Society of Innovative Educationalist & Scientific Research Professional, Malaysia. She has produced 5 M.Phil scholars and currently guiding 4 M.Phil and 1 Ph.D Scholars for doing their research. She has been involved in generating funds for R&D.

Ms. Priyanka .D is pursuing M.Phil (Computer Science) in K. S. Rangasamy College of Arts and Science (Autonomous), Tamilnadu, India. She completed MCA degree under Anna University, Chennai in 2017. She secured the 3rd Rank in BCA under Periyar University, Salem in 2013. She has attended 1 workshop related to Android and 1 seminar related to Network. Her areas of interest are Network and Data Mining.

DOI 10.17148/IJIREEICE.2018.6104