IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 4 (Sep. - Oct. 2013), PP 06-12
www.iosrjournals.org

Impulsion of Mining Paradigm with Density Based Clustering of Multi Dimensional Spatial Data

R.Dinesh Sunder, Bobby Lukose
(MPhil Scholar, PG & Research Department of Computer Science, Hindusthan College of Arts & Science, Coimbatore, Tamil Nadu, India)
(Associate Professor, PG & Research Department of Comp Science, Hindusthan College of Arts & Science, Coimbatore, Tamil Nadu, India)

Abstract : *Mining knowledge from large amounts of spatial data is known as spatial data mining. It becomes a highly demanding field because huge amounts of spatial data have been collected in various applications ranging from geo-spatial data to bio-medical knowledge. The amount of spatial data being collected is increasing exponentially. So, it far exceeded human's ability to analyze. Recently, clustering has been recognized as a primary data mining method for knowledge discovery in spatial database. The development of clustering algorithms has received a lot of attention in the last few years and new clustering algorithms are proposed. DBSCAN is a pioneer density based clustering algorithm. It can find out the clusters of different shapes and sizes from the large amount of data containing noise and outliers. This paper shows the results of analyzing the properties of density based clustering characteristics of three clustering algorithms namely DBSCAN, k-means and SOM using synthetic two dimensional spatial data sets.*

Keywords: *Clustering, DBSCAN, K-Means, SOM, SOFM*

I. INTRODUCTION

Clustering is considered as one of the important techniques in data mining and is an active research topic for the researchers. The objective of clustering is to partition a set of objects into clusters such that objects within a group are more similar to one another than patterns in different clusters. So far, numerous useful clustering algorithms have been developed for large databases, such as K-MEANS [4], CLARANS [6], BIRCH [10], CURE [3], DBSCAN [2], OPTICS [1], STING [9] and CLIQUE [5]. These algorithms can be divided into several categories. Three prominent categories are partitioning, hierarchical and density-based. All these algorithms try to challenge the Clustering problems treating huge amount of data in large databases. However, none of them are the most effective.

In density-based clustering algorithms, which are designed to discover clusters of arbitrary shape in databases with noise, a cluster is defined as a high-density region partitioned by low-density regions in data space. Density Based Spatial Clustering of Applications with Noise (DBSCAN) [2] is a typical density-based clustering algorithm. In this paper, we analyze the properties of density based clustering characteristics of three clustering algorithms namely DBSCAN, k-means and SOM.

II. DBSCAN ALGORITHM

Density-Based Spatial Clustering and Application with Noise (DBSCAN) was a clustering algorithm based on density. It did clustering through growing high density area, and it can find any shape of clustering. The idea of it was,

- ε-neighbor: the neighbors in ε semi diameter of an object
- Kernel object: certain number (*MinP*) of neighbors in ε semi diameter
- To a object set *D*, if object *p* is the ε-neighbor of *q*, and *q* is kernel object, then *p* can get "direct density reachable" from *q*.
- To a ε, *p* can get "direct density reachable" from *q*; *D* contains *Minp* objects; if a series object p1,p2,...,pn, p1=qn then p_{i+1} can get "direct density reachable" from Pi, P$_i$€D, 1<i<n.
- To ε and *MinP*, if there exist a object *o*(*o*€*D*) , *p* and *q* can get "direct density reachable" from *o*, *p*and *q* are density connected.z

III. EXPLANATION OF DBSCAN STEPS

DBSCAN requires two parameters: epsilon (eps) and minimum points (minPts). It starts with an arbitrary starting point that has not been visited. It then finds all the neighbor points within distance eps of the starting point. The possibility of outcomes are..,

- If the number of neighbors is greater than or equal to minPts, a cluster is formed. The starting point and its neighbors are added to this cluster and the starting point is marked as visited. The algorithm then repeats the evaluation process for all the neighbors recursively.
- If the number of neighbors is less than minPts, the point is marked as noise.
- If a cluster is fully expanded (all points within reach are visited) then the algorithm proceeds to iterate through the remaining unvisited points in the dataset.

IV. RESEARCH BACKGROUND AND RELATED WORKS

The company is using a legacy application for their day to day works. Though it helps in tracking the work progress of various aspects like labor, item, construction and accounting, there still remains some ambiguity. They feel complexity in executing certain process. This ambiguity makes them to turn towards a prolific package of software that prevails over the ease of construction.

The continuing development of enterprise resource planning (ERP) systems has been considered by many researchers and practitioners as one of the major IT innovations in this decade. ERP solutions seek to integrate and streamline business processes and their associated information and work flows. What makes this technology more appealing to organizations is its increasing capability to integrate with the most advanced electronic and mobile commerce technologies. However, research in the ERP area is still lacking and the gap in the ERP literature is huge. Attempts to fill this gap by proposing a novel taxonomy for ERP research. Also presents the current status with some major themes of ERP research relating to ERP adoption, technical aspects of ERP and ERP in IS curricula. The discussion presented on these issues should be of value to researchers and practitioners. Notable issues to be handled are depicted below

4.1 Intrinsic complexity in achieving results:

As a customized work flow there are 'N' levels of approvals for each transaction and there remain multiple hierarchies for master records, since then the need of categorized view and supplementary filtering options for various data sets cannot be retrieved.

Since there are portals for buyers and vendors, there remains difficulty in tracking the payments phase by phase and grouping by dimensions.

Due to multiple projects running simultaneously the civil works carried out at different stages requires stream lined supply of materials. So there comes the necessity of knowing the consumption Vs estimated material cost which leads to know the actual profit of the business. Provision for vendor rating based on various factors such as supply methods, timed delivery, quality in supply etc., will help in quoting and tender evaluation. This remains one of the major issues lacking in the existing system.

Maintaining stock management has been a challenging task in the construction industry and it resides the way of representing the stock quantity and value by segregating the materials based on its usage, but the existing system provides only the flat cost for the stock and not the moving average because of this the possibility of tracking the consumed and estimated comparison will not provide the accuracy in result.

4.2 Implication of poor data selection.

According to the survey (as on January 2013) made in the construction industries around Coimbatore, with respect to addressing the common mistakes that lead to poor selection of data are emphasized below.

- The requirement analysis data for material cost and labor cost gets varied before and after estimation, it means at the time of estimation and once the production kicks off.
- If tried to get the variation data and reason behind the cost, then there occurs the scenarios for different dimensions to be analyzed. This may consume adequate man power and to fine tune the results there requires several stages of verification for data correction.
- The dimension factors will then lead to integrate with other modules too. For example, if involved in finding the facts that influence the delay of production works, the company has to step into stock management which in turn directs them to find the vendor evaluation in purchase management which in turn tends to go for accounts management for payment history and finally to funds flow and so on.

4.3 Fact finding results:

As per the research made on technical aspects on ERP system in companies that are follow systemized process, It is find to be more negative replies towards the progression which means the success ratio is below average. Few of them are depicted below.

- 65% of company's executives believe that ERP will not be flexible to their practical scenarios when facing emergency issues. It means the data correction or rollback the process is not so easy but which is needed in certain situations.

- 30% of management faculties feel if concentrated on the data flow in ERP for budget plans, then they could miss their completion dates by wise margin.
- 25% of organizations adopting ERP systems are facing significant resistance from staff and 10% of organizations also encountered resistance from managers.
- End level users provide their suggestion in terms of limitation if further customization is required over different dimensions of data entry.

V. PROPOSED WORKS

During the last years some researchers have studied the topic of critical success factors in ERP implementations, out of which 'training' is cited as one of the most ones. Up to this moment, there is not enough research on the management and operationalization of critical success factors within ERP implementation projects. This technical research report proposes a framework for monitoring and evaluating training in ERP implementation projects. In order to develop a set of metrics for such monitoring and evaluating tasks, we have used the Goals/Questions/Metrics (GQM) approach. The GQM approach is a mechanism for defining and interpreting operational, measurable goals. Because of its intuitive nature the approach has gained widespread appeal. As a result, we propose a GQM preliminary plan with different metrics to monitor, control and evaluate training while implementing an ERP system. We also propose a three dimensional framework to interpret the metrics defined.

5.1 Implementation through GQM:

The GQM approach is a mechanism that provides a framework for developing a metrics program. It was developed at the University of Maryland as a mechanism for formalizing the tasks of characterization, planning, construction, analysis, learning and feedback. GQM does not provide specific goals but rather a framework for stating measurement goals and refining them into questions to provide a specification for the data needed to help achieve the goals. The GQM method contains four phases: planning phase, definition phase, data collection phase and interpretation phase.

The definition phase is the second phase of the GQM process and concerns all activities that should be performed to formally define a measurement program. One of the most important outcomes of this phase is the GQM plan. A GQM plan or GQM model documents the refinement of a precisely specified measurement goal via a set of questions into a set of metrics. Thus, a GQM plan documents which metrics are used to achieve a measurement goal and why these are used - the questions provide the rationale underlying the selection of the metrics. The definition phase has three important steps:

- Define measurement goals.
- Define Questions
- Define metrics.

5.2 K-Means Algorithm:

The naive k-means algorithm partitions the dataset into 'k' subsets such that all records, from now on referred to as points, in a given subset "belong" to the same center. Also the points in a given subset are closer to that center than to any other center. The algorithm keeps track of the centroids of the subsets, and proceeds in simple iterations. The initial partitioning is randomly generated, that is, we randomly initialize the centroids to some points in the region of the space. In each iteration step, a new set of centroids is generated using the existing set of centroids following two very simple steps. Let us denote the set of centroids after the ith iteration by C(i).

The following operations are performed in the steps:

- Partition the points based on the centroids C(i), that is, find the centroids to which each of the points in the dataset belongs. The points are partitioned based on the Euclidean distance from the centroids.
- Set a new centroid c(i+1) € C (i+1) to be the mean of all the points that are closest to c(i+1) € C(i) The new location of the centroid in a particular partition is referred to as the new location of the old centroid.

The algorithm is said to have converged when recomputing the partitions does not result in a change in the partitioning. In the terminology that we are using, the algorithm has converged completely when C(i) and C(i – 1) are identical. For configurations where no point is equidistant to more than one center, the above convergence condition can always be reached. This convergence property along with its simplicity adds to the attractiveness of the kmeans algorithm. The k-means needs to perform a large number of "nearest-neighbour" queries for the points in the dataset. If the data is 'd' dimensional and there are 'N' points in the dataset, the cost of a single iteration is O(kdN). As one would have to run several iterations, it is generally not feasible to run the naïve k-means algorithm for large number of points.

Sometimes the convergence of the centroids (i.e. C(i) and C(i+1) being identical) takes several iterations. Also in the last several iterations, the centroids move very little. As running the expensive iterations so many more times might not be efficient, we need a measure of convergence of the centroids so that we stop the iterations when the convergence criteria are met. Distortion is the most widely accepted measure.

5.3 SOM Algorithm:

A Self-Organizing Map (SOM) or self-organizing feature map (SOFM) is a neural network approach that uses competitive unsupervised learning. Learning is based on the concept that the behavior of a node should impact only those nodes and arcs near it. Weights are initially assigned randomly and adjusted during the learning process to produce better results. During this learning process, hidden features or patterns in the data are uncovered and the weights are adjusted accordingly. The model was first described by the Finnish professor Teuvo Kohonen and is thus sometimes referred to as a Kohonen map.

The self-organizing map is a single layer feed forward network where the output syntaxes are arranged in low dimensional (usually 2D or 3D) grid. Each input is connected to all output neurons. There is a weight vector attached to every neuron with the same dimensionality as the input vectors. The goal of the learning in the self organizing map is to associate different parts of the SOM lattice to respond similarly to certain input patterns.

5.4 Variables used:

s is the current iteration

λ is the iteration limit

t is the index of the target input data vector in the input data set $\mathbf{D}$

$\mathbf{D}(\mathbf{t})$ is a target input data vector

v is the index of the node in the map

$\mathbf{W_v}$ is the current weight vector of node v

u is the index of the best matching unit (BMU) in the map

$\Theta(u, v, s)$ is a restraint due to distance from BMU, usually called the neighbourhood function

$\alpha(s)$ is a learning restraint due to iteration progress

5.5 Algorithm Working:

STEP-1: Randomize the map's nodes' weight vectors

STEP-2: Grab an input vector $\mathbf{D}(\mathbf{t})$

STEP-3: Traverse each node in the map

STEP-3.1: Use the Euclidean distance formula to find the similarity between the input vector and the map's node's weight vector

STEP-3.2: Track the node that produces the smallest distance (this node is the best matching unit, BMU)

STEP-4: Update the nodes in the neighborhood of the BMU (including the BMU itself) by pulling them closer to the input vector

$\mathbf{W_v}(s + 1) = \mathbf{W_v}(s) + \Theta(u, v, s)\, \alpha(s)(\mathbf{D}(t) - \mathbf{W_v}(s))$

STEP-5: Increase s and repeat from step 2 while $s < \lambda$

Table-1: SOM Algorithm applied with 3 weight vectors at different time intervals.

Applying the SOM Algorithm

Data sample utilized

time (t)	1	2	3	4	5	6	D(t)	η(t)	Weights Updated
1	C						1	0.5	C, A
2		B					1	0.5	B, A
3			A				1	0.5	A, B, C
4				B			1	0.5	B, A
5					A		1	0.5	A, B, C
6						C	1	0.5	C, A
7	C						0	0.25	C
8		B					0	0.25	B
9			C				0	0.25	C
10				B			0	0.25	B
11					B		0	0.25	B
12						A	0	0.25	A
13	C						0	0.1	C
14		B					0	0.1	B
15			C				0	0.1	C
16				B			0	0.1	B
17					B		0	0.1	B
18						A	0	0.1	A

35

'winning' output node

5.6Training SOM Algorithm:

Initially, the weights and learning rate are set. The input vectors to be clustered are presented to the network. Once the input vectors are given, based on the initial weights, the winner unit is calculated either by Euclidean distance method or sum of products method. Based on the winner unit selection, the weights are updated for that particular winner unit. An epoch is said to be completed once all the input vectors are presented to the network. By updating the learning rate, several epochs of training may be performed.

A two dimensional Kohonen Self Organizing Feature Map network is shown in Fig- 1 which is given below

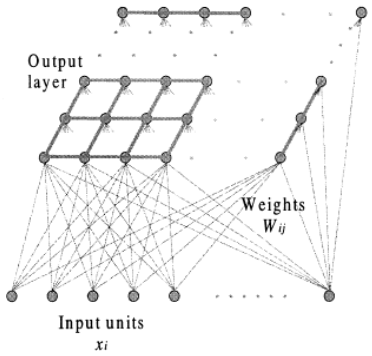


Fig-1: The SOM Network

VI. EVALUATION AND RESULTS

6.1 The Test Database:

To evaluate the performance of the clustering algorithms, two dimensional spatial data sets were used and the properties of density based clustering characteristics of the clustering algorithms were evaluated. The first type of data sets was prepared from the guideline images of some of the main reference papers of DBSCAN algorithm. So that data was handled from image format. The Fig- 2 shows the type of spatial data used for testing the algorithms

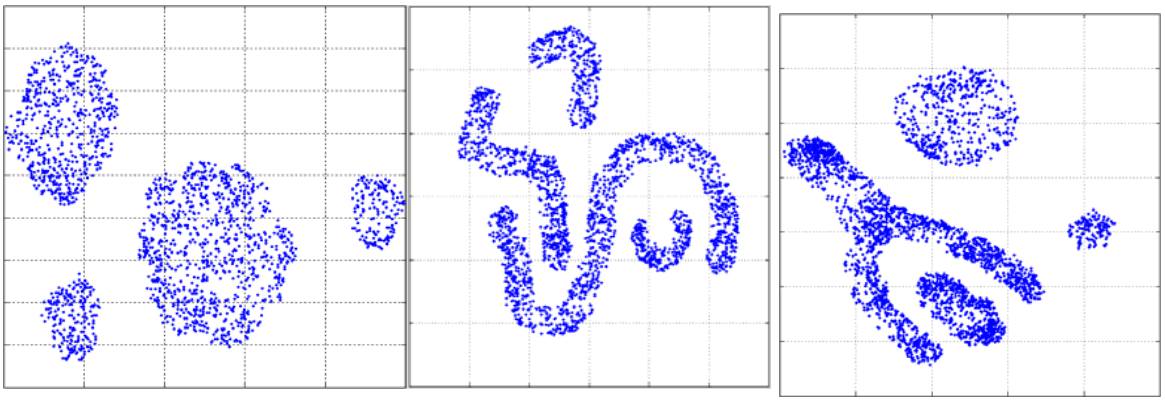
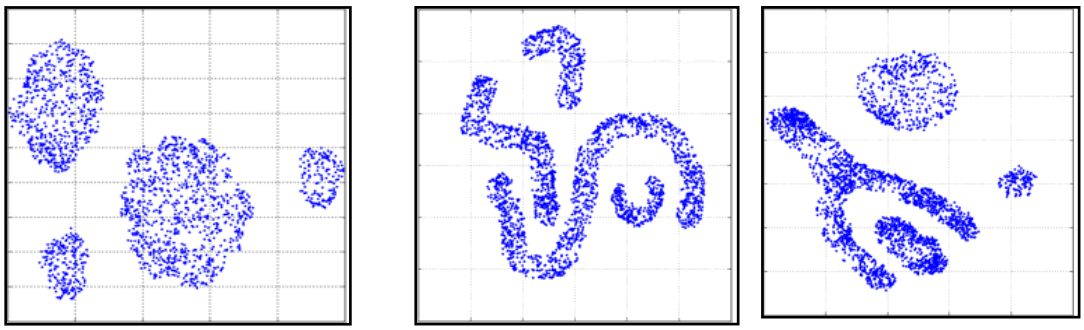
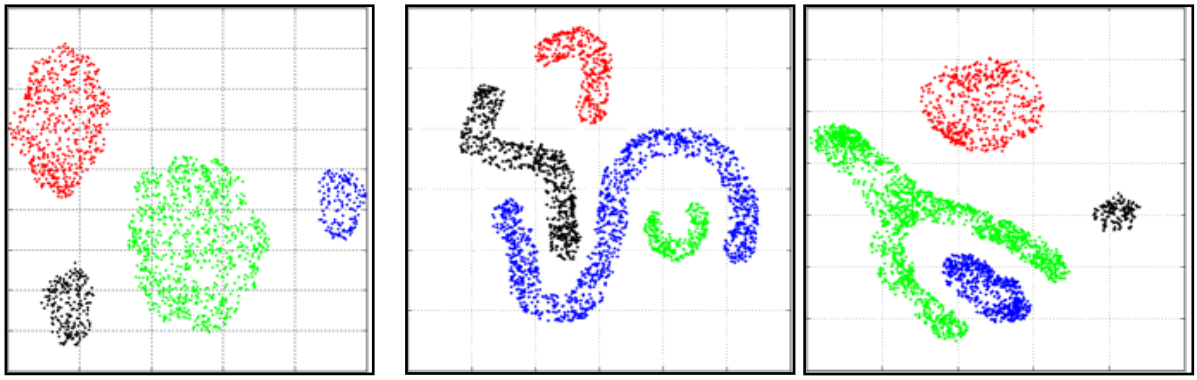


Fig-2: Spatial data used for testing.

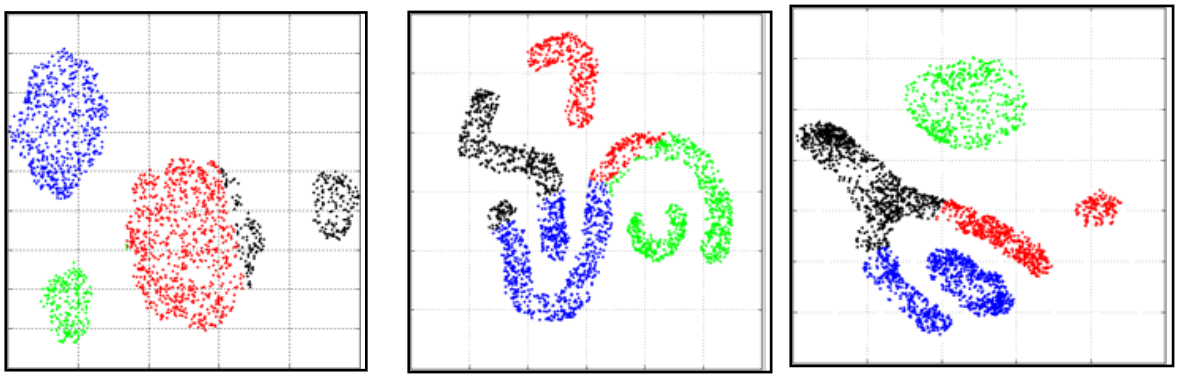
The Fig-3 shows the properties of density based clustering characteristics of DBSCAN and SOM.

Original Data

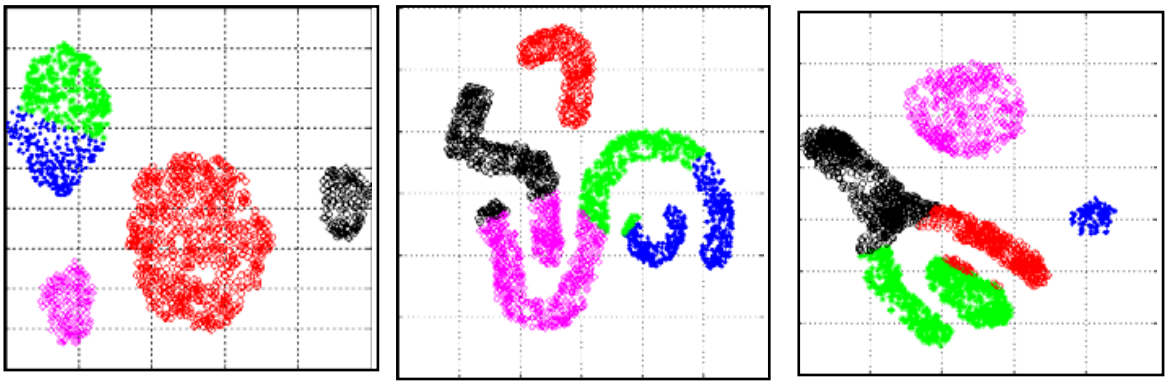




Clustering by DBSCAN



Clustering by SOM



Clustering by K-means

Fig-3: Clustering characteristics of DBSCAN, SOM and K-means clustering algorithm.

From the plotted results, it is noted that DBSCAN performs better for spatial data sets and produces the correct set of clusters compared to SOM and k-means algorithms. DBSCAN responds well to spatial data sets

VII. CONCLUSION

The Clustering algorithms are attractive for the task of class identification in spatial databases. This paper evaluated the efficiency of clustering algorithms namely DBSCAN, k-means and SOM for a synthetic, two dimensional spatial data sets. The implementation was carried out using MATLAB 6.5. Among the three algorithms DBSCAN responds well to the spatial data sets and produces the same set of clusters as the original data.

REFERENCES

Books:

[1] Kaufman L. and Rousseeuw P. J (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley & Sons.

[2] Ankerst M., Markus M. B., Kriegel H., Sander J(1999), "OPTICS: Ordering Points To Identify the Clustering Structure", Proc.ACM SIGMOD'99 Int. Conf. On Management of Data, Philadelphia, PA, pp.49-60.

[3] Guha S, Rastogi R, Shim K (1998), "CURE: An efficient clustering algorithm for large databases", In: SIGMOD Conference, pp.73~84.

[4] Ester M., Kriegel H., Sander J., Xiaowei Xu (1996), "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", KDD'96, Portland, OR, pp.226-231.

[5] Wang W., Yang J., Muntz R(1997), "STING: A statistical information grid approach to spatial data mining", In: Proc. of the 23rd VLDB Conf. Athens, pp.186~195.

[6] Raymond T. Ng and Jiawei Han (2002), "CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 5.

[7] Rakesh A., Johanners G., Dimitrios G., Prabhakar R(1999), "Automatic subspace clustering of high dimensional data for data mining applications", In: Proc. of the ACM SIGMOD, pp.94~105.

Thesis:

[8] http://genome.tugraz.at/MedicalInformatics2/SOM.pdf

[9] http://www.cs.bham.ac.uk/~jxb/NN/l17.pdf

[10] http://lib.tkk.fi/Diss/2002/isbn951226093X/article4.pdf

[11] http://www.cs.hmc.edu/~kpang/nn/som.html

[12] http://www2.tku.edu.tw/~tkjse/5-1/5-1-5.pdf

[13] http://users.ics.aalto.fi/mikkok/thesis/book/node12.html

[14] http://kanaya.naist.jp/SOM/

[15] http://www.isegi.unl.pt/fbacao/papers/The%20self-organizing%20ma%20the%20Geo-SOM.pdf