Research article Available online www.ijsrr.org ISSN: 2279–0543

International Journal of Scientific Research and Reviews

Inference & Learning of Mining Pattern with Multi-Dimensional Clusters

S.Kowsalya[1*] and Reshmi.S[2]

[1] Department of Computer Science and Application, Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India

[2]Department of Computer Science and Application, Sri Krishna Arts & Science College, Coimbatore, Tamil Nadu, India

ABSTRACT

Mining knowledge from large volume of unstructured data is referred as spatial data mining. It turned out as highly demanding scope of field, the reason being huge volume of spatial data have been collected in and around various applications ranging between geo-spatial data and bio-medical knowledge. The volume of spatial data that was collected or the existing is found to be increasing gradually. This exceeds human's ability to analyses. Recently, clustering technology has been recognized as a prominent data mining method for knowledge discovery in spatial database. The developing clustering algorithms have received lot of good reviews and attention in the past years and on the other hand there were some new clustering algorithms are proposed. DBSCAN is proposed to be a pioneer density based clustering algorithm. It can identify the clusters with different shapes and size among the large volume of data involving noise and outliers. In this paper I have shown the results of analysing the properties and attributes of density based clustering. I have also emphasized on 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.

***Corresponding author**

Mrs. S.Kowsalya

Assistant Professor

Department of Computer Science and Application,

Sri Krishna Arts And Science College,

Coimbatore - 641008, Tamil Nadu, India.

Email: kowsalaya.selvaraj@gmail.comMob.No:+91-9003958767

INTRODUCTION

Clustering concepts are revolving as one among the vital techniques in data mining trends. It is an active and most wanted research topic for the present day researchers. The objective of clustering we have focused in this paper is to partition a set of objects into clusters in such a way that objects within a specific group are more likely similar to one another than patterns available in different clusters. In near past days, numerous clustering algorithms have been developed for large volume of databases stating as effective utilized, such as K-MEANS, CLARANS, BIRCH, CURE, DBSCAN, OPTICS, STING and CLIQUE. The algorithms mentioned above shall be divided into different categories. Three notable categories are partitioning, hierarchical and density-based. All these algorithms obviously try to challenge the clustering problems taking huge amount of data in account. Eventually, none among them are resulted as most effective.

In density-based clustering algorithms, the macro level focus that we proposed in this paper is designed to discover clusters in an arbitrary shape from the databases with noise. A cluster is obviously defined as a high-density region or a defined area that got partitioned by low-density regions in data space. Density Based Spatial Clustering of Applications with Noise (DBSCAN) is most likely behaves as a density-based clustering algorithm in terms of results. In this paper, we have analyzed and phrased about the properties exists in density based clustering, characteristics of three clustering algorithms (DBSCAN, K-means and SOM).

DBSCAN ALGORITHM

Density-Based Spatial Clustering and Application with Noise (DBSCAN) is termed and referred as one of the clustering algorithm based on density. It works by clustering through keep growing high density area, and it can also find any shape of clustering. The idea of its functioning is elaborated as follows

- ε-neighbor: the neighbors in ε shall be taken as a semi diameter of an object. Provided the object is placed in a wide range.
- Kernel object: certain number (MinP) of neighbors in ε semi diameter not exceeding the boundaries.
- To the 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, then it means D contains Minp objects; Similarly if a series object p1,p2,...,pn, p1=qn then obviously pi+1 can get "direct density reachable" from Pi, Pi€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.

EXPLANATION OF DBSCAN STEPS

There are two parameters to be used for DBSCAN algorithm: epsilon (eps) and a minimum point (minPts). The initial trigger shall be an arbitrary starting point provided it should not be visited. All the neighbor points will then be identified within the distance 'eps' from starting point. The possible outcomes are as follows

Case-1: When number of neighbors greater than or equal to minPts, then a cluster is formed. In such case, the starting point along with its neighbors shall be added to this cluster and the starting point is marked as visited. This step in algorithm then gets iterates the evaluation process for all the other neighbors' recursively.

Case-2: When number of neighbors less than minPts, then the point shall be marked as noise.

Case-3: When a cluster is completely expanded (meaning that all the points within reach are visited) then the algorithm proceeds to iterate through the remaining unvisited points in the dataset.

RESEARCH BACKGROUND AND RELATED WORKS

We have chosen a construction and infrastructure based concern to evaluate and implement my research works. Even though the existing system helps them in tracking the work in progress of different portfolios such as labor, material, construction and accounting, they still face some ambiguity. The operations team feels complexity in executing certain process. This ambiguity makes the management to turn towards a prolific package of software forcefully that prevails over the ease of construction and its dependent activities.

The present generation of enterprise resource planning (ERP) systems has been focused by many researchers and practitioners as a major IT innovation. ERP solution's expectation is to integrate and streamline the business processes and their information with work flows. The main appealing thing for the organizations in this technology is its increasing capability with integration of most advanced electronic and mobile commerce technologies. However, research in the ERP area is very few and the gap remains huge. Many attempts were made to fill this gap by proposing an enhanced novel taxonomy

for ERP research. The current status is presented with some different themes of ERP in research relating to ERP adoption, functional and technical aspects of ERP in IS curricula. The outcomes of discussions presented on these topics should be of value added to researchers and practitioners. We have depicted in below sections about the handling of such notable issues.

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.

Implication of poor data selection.

According to the survey (as on January 2018) made in the construction industries in and around the Coimbatore city, we were able to identify the common mistakes that lead to poor selection of data are emphasized below.

The data being prepared as a part of requirement analysis for material cost and labour cost gets varied before and after estimation, in other words the data are not consistent and keep varying between the time of estimation and the production kicks off.

Upon trying to get the variation data and reason behind the cost, then there occur the scenarios for different dimensions to be analyzed. It seems that there is a need of adequate consumption of man power and to fine tune the results there requires several stages of verification and validation for data correction.

The third notable finding is on the dimension factors that are badly recommended to integrate with other modules. In case of finding the facts that influence the delay of production works, the company has to step its focus into stock management which in turn directs them to find the vendor evaluation in purchase management and then tends to go for accounts management for payment history and finally to funds flow and this keep continuing.

Fact finding results

Based on the research made on ERP system's technical aspects, it was noted that though the companies follow a systemized process, we received more negative replies towards the progression and this means the success ratio is below average. Few of such examples are depicted below.

- Over 65% of company's peer level executives believe that ERP system and its approaches will not be flexible to their practical scenarios in their routine activities when facing emergency issues. This means the data correction or rollback process is not so convenient easy perhaps is needed in certain situations.
- Around 30% of the middle management faculties feel if there is enough effort being put on the data flow in ERP system for budget plans, and then there are high possibilities that they could miss their completion target dates by a wide margin.
- Close to 25% of organizations whoever adopting an ERP systems are facing significant measures of resistance from their staff and over 10% of organizations too encountered resistance from their managers.
- The end level users provide their suggestion in view of limitation if further customization is required over different dimensions of data entry.

PROPOSED WORKS

During the last few years there were many researchers who have studied the topic of critical success factors during implementation of an ERP system, out of which the 'training' is cited as one of the most prominent ones. As on date, there are no enough researches undertaken on the core management activities and operationalization of critical success & failure factors in ERP implementation projects. This technical research report in particular proposes a framework for analyzing, monitoring and evaluating the training in ERP implementation projects. In order to develop a predefined set of metrics for such analyzing, monitoring and evaluating tasks, we have used the most familiar and effective approach named Goals/Questions/Metrics (GQM) approach. This approach is a mechanism for defining and interpreting the operational and measurable goals. Due to its intuitive nature, this approach has

already gained a widespread appeal across the research industry. Hence based on the evaluation of the results and considering the advantages, we propose a GQM approach to follow the preliminary plan with different metrics to analyses, monitor, control and evaluate the training while implementing an ERP system. In addition, we propose a three dimensional framework to interpret the metrics defined in this approach.

Implementation through GQM

The GQM approach is a kind of mechanism that provides a defined framework for developing a metrics program. It was developed in the University of Maryland as a mechanism for formalizing the project tasks of dynamic characterization, planning, construction, analysis, learning and user 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.

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 it 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 re-computing 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 k-means algorithm. The k-means needs to perform a large number of "nearest-neighbor" 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.

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 TeuvoKohonen 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.

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 D.

'D(t)' is a target input data vector.

'V' is the index of the node in the map.

'WV' 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.

Algorithm Working

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

STEP-2: Grab an input vector D(t).

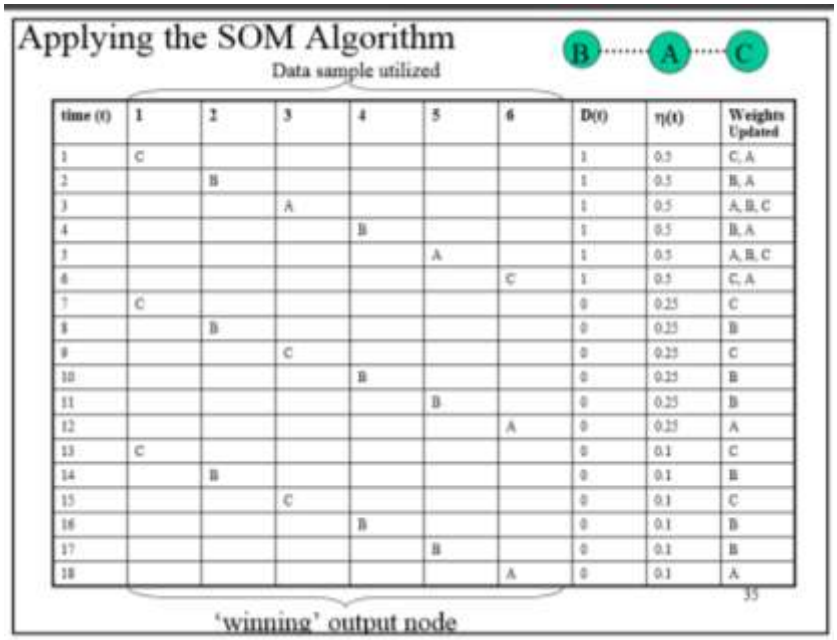
STEP-3: Traverse each node in the map.

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. $Wv(s + 1) = Wv(S) + \Theta(U, V, S)\ \alpha(S)(D(t) - Wv(S))$

STEP-5: Increase S and repeat from step 2 while .



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

Figure-1. SOM algorithm applied with 3 weight vectors at different time intervals

Training 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 figure-2 which is given below

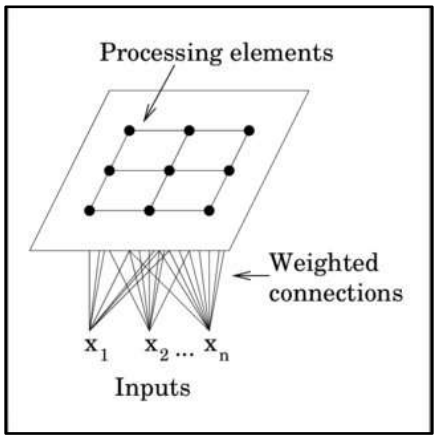


Figure-2. The SOM Network

EVALUATION AND RESULTS

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 figure 2 shows the type of spatial data used for testing the algorithms

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.

CONCLUSION

The Clustering algorithms are attractive for the task of class identification in spatial databases. This paperevaluated 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

1. Kaufman L. and Rousseeuw P. J "Finding Groups in Data: An Introduction to Cluster Analysis", 1990;127 – 246.
2. Ankerst M., Markus M. B., Kriegel H., Sander J, "OPTICS: Ordering Points To Identify the Clustering Structure", Proc.ACM SIGMOD'99 Int. Conf. On Management of Data, Philadelphia, PA, 1999; 49-60.
3. Guha S, Rastogi R, Shim K, "CURE: An efficient clustering algorithm for large databases", In: SIGMOD Conference, 1998;73-84.
4. Ester M., Kriegel H., Sander J., XiaoweiXu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", KDD'96, Portland, OR,1996;226-231.
5. Wang W., Yang J., Muntz R, "STING: A statistical information grid approach to spatial data mining", In: Proc. of the 23rd VLDB Conf. Athens 1997; 186-195.
6. Raymond T. Ng and Jiawei Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining", IEEE Transactions on Knowledge and Data Engineering, 2002; 14: 5.
7. Rakesh A., Johanners G., Dimitrios G., Prabhakar R, "Automatic subspace clustering of high dimensional data for data mining applications", In: Proc. of the ACM SIGMOD, 1999; 94-105.