KSOMKM: An Efficient Approach for High Dimensional Dataset Clustering

Momotaz Begum and Md. Nasim Akthar
Department of Computer Science & Engineering,
Dhaka University of Engineering & Technology, Gazipur, Bangladesh.
E-mail: momotaz.2k3@gmail.com.

***Abstract*—The process which was used for grouping the similar elements or occurring closely is called cluster. Nowadays cluster analysis is one of the major data analysis techniques. On the other hand many important problems involve clustering for large datasets. KSOM and k-means is one of the most popular partitioning clustering algorithms that are widely used. The original k-means algorithm is computationally expensive and the number of clusters K, to be specified before the algorithm is applied. The other thing is, it is quite sensitive to initial centroids. When more number of dimensions is added then K-Means fails to give optimum result. For this "Curse of High Dimensionality" problem is occurred. Here we propose that Kohonen Self Organizing Map (KSOM) is used to define number of clusters and then load based initial centroid K-Means algorithm (KSOMKM) is used to find out the more accurate number of cluster for High Dimensional Dataset. Finally the Kohonen Self Organizing Map (KSOM) with Load based K-Means algorithm (KSOMKM) is tested on different datasets. There are an IRIS data set, Diabetes dataset, Thyroid, Blood pressure dataset. Its performance is compared with other clustering algorithm for number of iteration, quantization errors and topographic errors.**

***Index Terms*—curse of dimensionality, data mining, high-dimensional datasets and Kohonen Self Organizing Map (KSOM).**

I. Introduction

Data mining is the process of extracting useful information from a collection of data's in a large data-base. It is used in many applications such as pattern recognition, medical purpose, web documentation, business purposes, and scientific purposes and so on. Classification is a procedure where data objects are assigned to predefine classes. Clustering is an unsupervised classification of data items, or feature vectors into groups. The term unsupervised here means that clustering of data objects doesn't depend on predefines classes.

K-means is the most exoteric partitioning clustering method for its efficiency and simplicity in clustering large Data sets. K-means is among the commonly used partitioning based clustering method that tries to find a specified number of clusters (k), represented by their centroids, by minimizing the sum of square error function. It is very simple and fast but is very sensitive to initial positions of cluster centers. Usually only a small number of dimensions are relevant to certain clusters; the irrelevant one may produce noise and mask the real clusters to be discovered. Moreover when dimensionality increases, data usually become increasingly sparse, this can affect the quality of clustering [1].

Most of the time due to noise and outliers associated with original data the traditional K-means algorithm does not work well for high dimensional data and results may not be accurate. Also the computational complexity increases rapidly as the dimension increases. Moreover the exact number of clusters cannot be determined and that is very sensitive to initial centroids. Hence to improve the performance we proposed KSOM with load based initial centriods K-Means algorithm (KSOMKM), basically SOM for dimension reduction and determining the number of clusters after that load based K-Means Algorithm for better cluster. It is found out that the approach gives better accuracy and better performance in terms of speed. Below we gave a brief description of the proposed experiments.

The remainder of this paper is organized as follows. The Section II presents an overview of related works on high dimensional dataset by KSOM with different K-means clustering algorithms. Section III describes the process of experiment. We also discussed some related algorithm i. e the original k-means algorithm, KSOM and for better selection of initial centroids load based k-means algorithm in Section IV. The Section V shows the experimental result in two ways that is for multivariate data set (Diabetes, Thyroid, Pima-Indian-diabetes and Blood pressure) for calculating average time and the other is high dimensional dataset (IRIS) in different criteria with various popular clustering algorithm. Lastly we describe the conclusion and future work.

II. Related Work

The selection of initial centroids is strongly dependable in the original k-means algorithm. Different selection of centroids leads to different resulting clusters. So the correctness and excellence of clusters mainly depends on initial selection points. Over the last few years many authors have been proposed to progress the quality and accuracy of k-means in several methods [2].

An enhanced algorithm to put data points for the suitable cluster is proposed. Although initial centroid was

Manuscript received May 10, 2013; revise July 19, 2013.

doi: 10.12720/ijoee.1.2.102-107

selected randomly in that method; for this, the method also contains the same problem i.e sensitive to the initial seed points and doesn't make correct clustering results [3].

For high dimensional data set a hybridized K-Means clustering approach is proposed where PCA, Canonical Variate Analysis (CVA) and Genetic Algorithm (GA) was used for dimensional reduction and for finding the initial centroids, a new method was employed, finding the mean of all the data sets divided in to k different sets in ascending order [4].

A new algorithm that avoided the random selection of initial centriods as well as improves performance of k-means is proposed [5]. But accuracy is low. Divisive Correlation Clustering Algorithm (DCCA) proposed the cluster of data set without taking the initial centroids and the value of the desired number of clusters k [6]. But the time complexity of this algorithm is also high. Self-Organizing Map (SOM) for clustering purpose is proposed [7].

Many works have been done for overlapping clusters. To overcome this problem the hybridized self organizing map is proposed [8]. On the other hand gives a brief idea on dimensionality reduction and the various techniques that can be applied for dimensional reduction with improved initial center by the distance from the origin is proposed [9].

III. Process of the Experiment

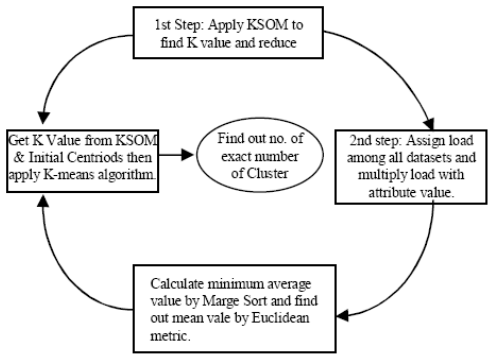

Figure 1. Flowchart of the proposed work.

IV. Related Algorithms

A. Original K-Means Algorithm

This section describes the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases: the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner. Eventually, a situation will be reached where the centroids do not move anymore. This signifies the convergence criterion for clustering [10].The pseducode of Orginal K-means algorithm is as follows:

Input:
D = d1, d2,.....dn // set of n data items.
k = the number of desired clusters,
Output: a set of k clusters.
Steps:
1. Randomly choose k objects from n as the initial cluster centers.
2. Repeat
3. Assign each object from n to one of k clusters to which the object is most similar based on the mean value of the objects in the cluster.
4. Update the cluster means by taking the mean value of the objects for each of k cluster.
5. Until no change in cluster means/ min error E is reached.

B. KOHONEN Self-Organizing Map (KSOM)

Kohonen Self-Organizing Maps (or just Self-Organizing Maps, or SOMs for short), are a type of neural network. They were developed in 1982 by Tuevo Kohonen, a professor emeritus of the Academy of Finland [11]. The KSOM neural network is basically a single-layer feed forward network. This network contains two layers of nodes. The input layer acts as a distribution layer. When an input pattern is presented, each unit in the 1st layer takes on the value of the corresponding entry in the input pattern. The 2nd layer units then sum their inputs and compete to find a winning unit. By observing the input patterns, KSOM reorganizes those by clustering similar patterns into groups. In addition to these unique neural network based clustering algorithms for information science applications; prior research in neural networks has strongly suggested the Kohonen self-organizing feature map (SOM) as a good candidate for clustering textual documents .The topology of the Kohonen SOM net-work is shown in Fig. 2.

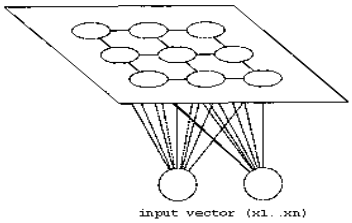

Figure 2. Simple KSOM map.

Thus, the input layer and each node of the mapping layer can be represented as a vector which contains the number of features of the input. The mapping nodes are initialized with random numbers. The ``best'' mapping node (BMU) is defined as that with the smallest Euclidean distance between the mapping node vector and the input vector. The input thus maps to a given mapping node. The value of the mapping node vector is then adjusted to reduce the Euclidean distance. In addition, all of the

neighboring nodes of the best node are adjusted proportionally. In this way, the multi-dimensional (in terms of features) input nodes are mapped to a two-dimensional output grid. After all of the input is processed (usually after hundreds or thousands of repeated presentations), the result should be a spatial organization of the input data organized into clusters of similar (neighboring) regions [12].

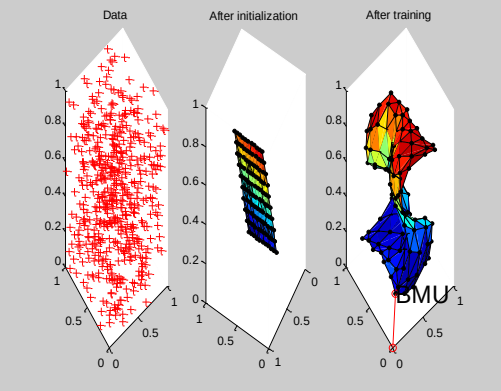

Figure 3. Best matching unit (BMU) selection

As mentioned earlier, the original K-Means algorithm does not work well for high dimensions. And we have seen some of its weakness, such as sensitive to initialization, unknown number of clusters and complexity problem. So to overcome its entire problem, we used the Kohonen self-organization map. Basically we apply KSOM on the dataset for reducing the dimension. The KSOM, not only it reduces the dimension but it gives us a clear confirmation on the number of clusters. We then applied the load based system for data points to obtain the initial centroid and finally the data's are grouped into cluster using the K-Means algorithm. The above mentioned process is called KSOMKM.

C. Load Based Initial Centroid K-Means Algorithm

A method is used to find out the initial centroid of the dataset. A sorting algorithm is applied to determine the score of each data point and divided into k subsets where k is the number of clusters which is given from the KSOM. Finally the nearest value of mean from each subset is taken as initial centroid. For example a data set D is consists of n number of data such as d1, d2, d3... dn. Each data point of this set may contain multiple attributes such as di may contain attributes a1, a2, a3,....., am, where m is the number of attributes. In case of multidimensional attributes we set a load with each attribute by manually between-p one to ten. After multiplying the load factor with each attribute we sum the values and make an average by dividing the total with m. The entire set of data points are then sorted using Merge Sort. The sorted list of data points are then divided into k subsets. The nearest possible value of mean from each dataset becomes the initial centroids of the cluster to be constructed [13]. The Pesudocode of load based initial centriod k-means algorithm is as follows:

Input:
D = d1, d2.....dn // set of n data items
L // set of load for data points.
K // number of desired clusters from KSOM
Output: A set of k initial centroids and resultant clusters.
Steps:
1. Calculate the average score of each data point;
i) di=a1,a2,a3,..,am
ii) sl=(l1*a1+l2*a2+l3*a3+ ..... +lm*am)
iii) sl(avg)=sl/m //
where sl=sum of load, a = the attributes value, m = number of attributes and l = load to multiply to ensure fair distribution of cluster by manually between (1-10).
2. For minimum average sort the data;
3. Now split the data into k clusters from KSOM;
4. Calculate the mean value of the each subset by Euclidean metric;
5. Take the nearest possible data point of the mean as the initial centroid for each data subsets;

The above described method for finding initial centroids of the clusters is more meaningful than the original k-means where centroids are selected randomly. The algorithm converges earlier than the original k-means. It also gives exact initial centriods for high dimensional data set. In addition it can be clustered very accurately and smoothly. The Pseudo code for the algorithm of KSOMKM is as follows:

Input: The Pseudo code for the algorithm of full method is as follows
//Set up input layer, total number of input neurons, I = Input P* Input Q.
// Set up output layer. Total number of output neurons, J = Output P * Output Q
//Initialize connection weights (randomize) between input layer neurons and output layer neurons, Wij
//Set initial learning rate parameter, α (between 0.2 to 0.5)
//Set total number of iterations, T
//Start with iteration t = 0.

Output: A set of K resultant clusters.
//Step 1: Apply KSOM to reduce the dimension of the dataset and find clusters.
1. The network is fully connected that is all nodes in input layer are connected to all nodes in output layer.
2. Apply the first input of the KSOM and Compute the winning neuron (jc) in the output layer which is the minimum Euclidean distance .The Euclidean distance as follows:

$$\|E(j)\| = \sqrt{\sum_{i=0}^{i=I-1} (w_{ij} - p_i)^2}$$

And find the minimum distance *E j*
Which is the winner neuron, jc?
3. Update weight for each connection i.e. for all neurons j and for all i:

$$w_{ij}(new) = w_{ij}(old) + \Delta w_{ij}(new)$$

4. Update learning rate α such that: $\alpha_t = \alpha_0(1 - \frac{t}{T})$

5. Repeat Steps 6 to 8 until t=T
6. Repeat with next pattern chosen randomly
(Do Steps 3-6)[18]

//step2: Find the initial centriod

Describe in above.

//Step 3: Apply the K-means clustering algorithm with the initial centriods and no. of cluster from KSOM.

Input:

//Initial centriod from step2.

K // number of desired clusters from KSOM

Output: A set of exact clusters

7. For each data point, find the nearest cluster centriod from list *Center* that is closest and assign that data point to the corresponding cluster.
8. Update the cluster centriod in each cluster of the data points, which are assigned to that cluster.
9. Repeat the steps 7 and 8 UNTIL the convergence criteria are met.

V. Experimental Result

A. Experiment 1: For Multivariate Data Set

The multivariate data set, taken from the UCI repository of machine learning databases, such as Diabetes, Thyroid, Pima-Indians-diabetes and Blood Pressure that is used to test the time complexity. This same data set is given as input to the standard k-means algorithm, load based k-means algorithm and KSOM with load based k-means algorithm i.e KSOMKM. The experimental results are shown in Table I.

TABLE I. TABLE I: Performance Comparison of the Algorithm

Dataset	Algorithm	Average time taken(ms)
Diabetes	Standard K means	0.0481
	Load based K means	0.0313
	KSOMKM	0.0256
Thyroid	Standard K means	0.0781
	Load based K means	0.0748
	KSOMKM	0.0648
Pima-Indians-diabetes	Standard K means	0.0854
	Load based K means	0.0812
	KSOMKM	0.0746
Blood Pressure	Standard K means	0.0944
	Load based K means	0.0922
	KSOMKM	0.0871

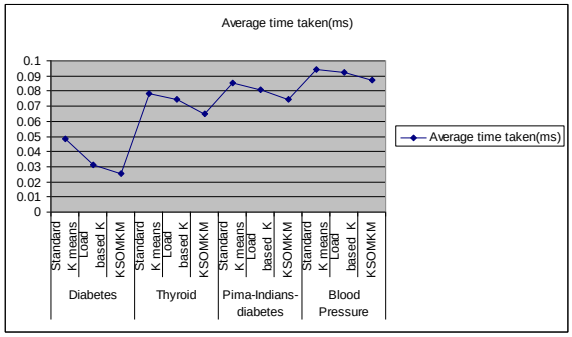

Figure 4. Comparison of the time complexity for different datasets.

B. Experiment 2: For High Dimensional Dataset

The experimental analysis is performed on an IRIS data set which we have taken from UCI Repository of Machine Learning Databases. The data set contains 5 dimensions of three types of flower species setosa, virsicolor and virginica. Here, the U-matrix of the map is shown by SOM toolbox in MATLAB. The IRIS data set also has labels associated with the data samples. Actually, the data set consists of 50 samples of three species of Iris-flowers (a total of 150 samples) such that the measurements are width and height of sepal and petal leaves.

The 'U-matrix' shows distances between neighboring units and thus visualizes the cluster structure of the map. The label associated with each sample is the species information: 'Setosa', 'Versicolor' or 'Virginica' [14].

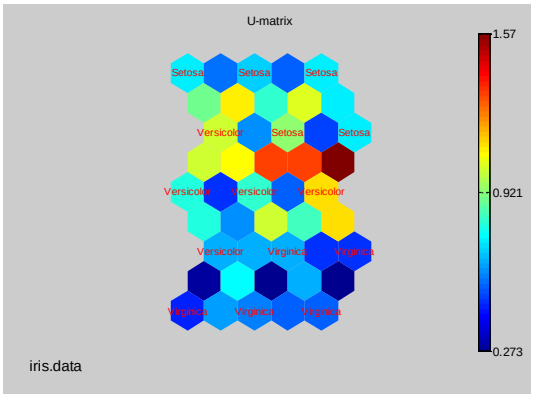

Figure 5. Visualization of clusters for IRIS data.

The variable values have been demoralized to the original range and scale. The component planes ('PetalL', 'PetalW', 'SepalL' and SepalW') show what kind of values the prototype vectors of the map unit have.

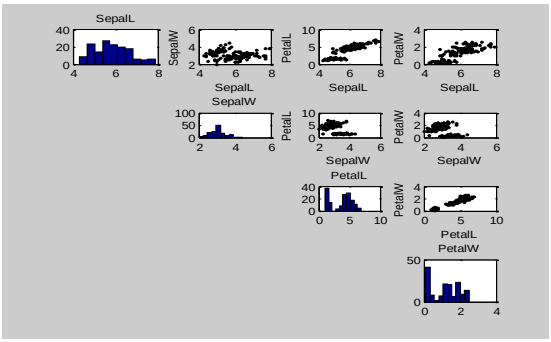

Figure 6. The histograms and scatter plots of the four clusters.

Here the data set and map prototypes are plotted again, but information of the cluster is shown using color: red for the first cluster, green for the second, blue for the third and gray for the last. When the information is added this way, the visualization becomes harder and harder to understand.

The cluster number has been found out by applying the KSOM. After getting k value then it is applied in load based initial centroid K-Means algorithm. The result is compared with some of the well-known high dimensional clustering algorithm and utilized to have better performance in terms of topographic error, quantization error and number of iteration.

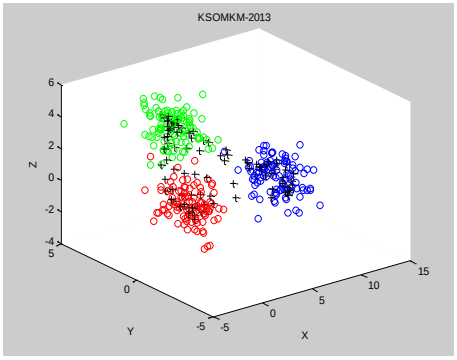

Figure 7. Visualization of Clusters using KSOMKM.

TABLE II. PERFORMANCE COMPARISON OF THE OTHER AL-GORITHM FOR HIGH DIMENSIONAL DATASETS

Algorithm	Topographic Error	Quantization Error	No. of iteration
KSOMKM	0.0139	0.391	6
PCA	0.1331	0.74	7
Sammon Mapping	0.2689	0.73	8
Hybridized k-Means Algorithm with PCA	0.129	0.77	6
CVA	0.252	0.67	8

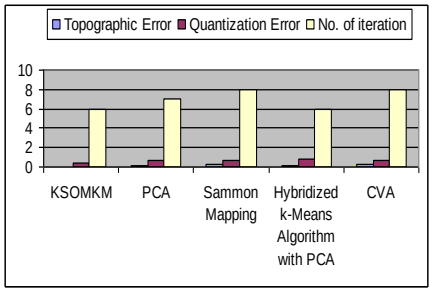

Figure 8. Comparison of others algorithm for different criteria.

The above result shows that the proposed algorithm KSOMKM shows a better performance in terms of some errors and number of iteration when compared with the earlier proposed hybridized K Means clustering algorithm using PCA, original PCA, CVA and Summon mapping taken on the experimental data.

VI. CONCLUSION AND FUTURE WORK

In this paper we have applied the Kohonen self-organization map (KSOM) for dimension reduction and determining the number of clusters. After that for initial centriod we used load based k-means algorithm (KM) .And we have got initial centriods .Finally we added k value which is from KSOM and initial centriods from load based system. The KSOMKM is used to find out exact number of clusters. The standard K-Means algorithm faced some problem such as unknown number of clusters and the sensitivity to initial centroid also it is unable to use high dimensional dataset. The proposed KSOMKM algorithm can be applied to find out the cluster very easily for high dimensional dataset .The cluster number can be identified more accurately in compared with other method. One limitation of this process is that it takes long time to provide the correct cluster number. A research issue on this point remains open. Further research can be done to use a more accurate method i. e to find correct number of cluster using better optimization technique for good accuracy, time complexity and quantization error.

REFERENCES

[1] B. A. Shboul and S. H. Myaeng, "Initializing K-means by using genetic algorithm," *World Academy of Science, Engineering and Technology*, vol. 54, pp. 114-117, 2009.

[2] R. Dash, D. Mishra, A. K. Rath, and M. Acharya , "A hybridized k-means clustering algorithm for high dimensional dataset," *International Journal of Engineering, Science and Technology*, vol. 2, no. 2, pp. 59-66, 2010.

[3] A. M. Fahim, A. M. Salem, F. A. Torkey, M. A. Ramadan, and G. Saake, "An efficient K-means with good ini-tial starting points," *Georgian Electronic Scientific Journal: Computer Science and Telecommunications*, vol. 2, no. 19, pp. 47-57, 2009.

[4] H. S. Behera, R. B. Lingdoh, and D. Kodamasingh, "An improved hybridized k-means clustering algorithm (IHKMCA) for high dimensional dataset and its performance analysis," *International Journal of Computer science and Engineering*, vol. 3, no. 2, pp. 1183-1190, 2011.

[5] C. Zhang and S. Xia, "K-means clustering algorithm with improved initial center," in *Proc. Second International Workshop on Knowledge Discovery and Data Mining*, 2009, pp. 790-792.

[6] A. Bhattacharya and R. K. De, "Divisive correlation clustering algorithm (DCCA) for grouping of genes: Detecting varying patterns in expression profiles," *Bioinformatics*, vol. 24, pp. 1359-1366, 2008.

[7] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586-600, May 2000.

[8] M. N. M. Sap and E. Moheb, "Hybrid self organizing map for overlapping clusters," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, pp. 11-20, 2011

[9] G. Bohling, "Dimension reduction and cluster analysis," *EECS 833*, 6 March, 2010.

[10] M. Yedla, S. R. Pathakota, and T. M. Srinivasa, "Enhancing K means algorithm with improved initial center," *International Journal of Computer Science and Information Technologies*, vol. 1, no. 2, pp. 121-125, 2011.

[11] J. Vesanto, "SOM-based data visualization methods," *Intell. Data Analysis*, vol. 3, no. 2, pp. 111-126, 1999.

[12] Self-organizing map. [Online]. Available: en.wikipedia.org/wiki/Self-organizing map

[13] M. S. Mahmud, M. M. Rahman, and M. N. Akhtar, "Improvement of K-meansclustering algorithm with better initial centroids based on weighted average," in *Proc. International Conference on Electrical and Computer Engineering*, December 2012.

[14] Self-organizing map in matlab: the SOM toolbox. [Online]. Available: http://cda.psych.uiuc.edu/martinez/edatoolbox/Docs/toolbox2paper.pdf

Momotaz Begum, was born at Norsingdi, Bangladesh dated 30.06.1984. She obtained her Bachelor of Science in Engineering degree from Department of Computer Science and Engineering (CSE) of Dhaka University of Engineering and Technology (DUET), Gazipur-1700, Bangladesh in 2008.At present she is taking M.Sc. in Engineering, Department of Computer Science and Engineering (CSE) from the same University. She has been serving as ASSISTANT PROFESSOR, Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur-1700, Bangladesh. Field of interest: Advanced Database System, Software Engineering, Artificial Neural Network, Data Mining, and Data clustering. Ms. Begum is the member of International Association of Computer Science and Information Technology (IACSIT).

Md. Nasim Akthar was born at Rajshahi, Bangladesh dated 10.01.1973.He obtained his Bachelor of Science in Engineering and Master's degree from Department of Computer Science and Engineering (CSE) of National Technical University of Ukraine, Kiev, Ukraine in 1996 and 1998 respectively. In 2010 he obtained his PhD from Moscow State Academy of Fine Chemical Technology, Russia. He has been serving as ASSOCIATE PROFESSOR and HEAD, Department of Computer Science and Engineering (CSE), Dhaka University of Engineering & Technology (DUET), Gazipur-1700, Bangladesh. Field of interest: Data ware house mining, Software Engineering, Artificial Neural Network, Distributed search techniques and Data clustering.