Evaluating Cluster Quality Using Modified Density Subspace Clustering Approach

V.Kavitha[1], Dr.M.Punithavalli[2],

[1]Research Scholar, Dept of Computer Science, Karpagam University Coimbatore, India.
[2]Director,Dept of MCA, Sri Ramakrishna College of Engineering Coimbatore, India.

Abstract

Clustering of the time series data faced with curse of dimensionality, where real world data consist of many dimensions. Finding the clusters in feature space, subspace clustering is a growing task. Density based approach to identify clusters in dimensional point sets. Density subspace clustering is a method to detect the density-connected clusters in all subspaces of high dimensional data for clustering time series data streams Multidimensional data clustering evaluation can be done through a density-based approach. In this approach proposed, Density subspace clustering algorithm is used to find best cluster result from the dataset. Density subspace clustering algorithm selects the P set of attributes from the dataset. Then apply the density clustering for selected attributes from the dataset .From the resultant cluster calculate the intra and inter cluster distance. Measuring the similarity between data objects in sparse and high-dimensional data in the dataset, Plays a very important role in the success or failure of a clustering method. Evaluate the similarity between data points and consequently formulate new criterion functions for clustering .Improve the accuracy and evaluate the similarity between the data points in the clustering,. The proposed algorithm also concentrates the Density Divergence Problem (Outlier). Proposed system clustering results compared them with existing clustering algorithms in terms of the Execution time, Cluster Quality analysis. Experimental results show that proposed system improves clustering quality result, and less time than the existing clustering methods.

Keywords— Density Subspace Clustering, Intra Cluster, Inter Cluster, Outlier, Hierarchical Clustering.

I. INTRODUCTION

Clustering is one of the most important tasks in data mining process for discovering similar groups and identifying interesting distributions and patterns in the underlying data. Clustering of the time series DataStream problem is about partitioning a given data set into groups such that the data points in a cluster are more similar to each other than points in different clusters .Cluster analysis [3] aims at Identifying the groups of similar objects and helps to discover distribution of patterns, finding the interesting correlations in large data sets. It has been subject of wide research since it arises in many application domains in engineering, business and social sciences. Especially, in the last years the availability of huge transactional and experimental data sets and the arising requirements for data mining created needs for clustering algorithms that scale and can be applied in diverse domains.

The clustering of the times series data streams becomes one of the important problem in data mining domain. Most of the traditional algorithms will not support the fast arrival of large amount of data stream. In several traditional algorithms, a few one-pass clustering algorithms have been proposed for the data stream problem. These methods address the scalability issues of the clustering problem and evolving the data in the result and do not address the following issues:

(1)The quality of the clusters is poor when the data evolves considerably over time.

(2) A data stream clustering algorithm requires much greater functionality to discovering accurate result and exploring clusters over different portions of the stream.

Charu C. Aggarwal [5] proposed micro-clustering phase in the online statistical data collection portion of the algorithm. This method is not dependent on any user input such as the time horizon or the required granularity of the clustering process. The aim of the method is to maintain statistics based on sufficiently high level granularity used by the online components such as horizon-specific macro-clustering as well as evolution analysis.

The clustering of the time series data stream and incremental models are requiring a decisions before all the data are available in the dataset. The models are not identical to find the best clustering result.

Finding clusters in the feature space, subspace clustering is an emergent task. Clustering with dissimilarity measure is robust method to handle large amount of data and able to estimate the number of clusters automatically by avoid overlap. Density subspace clustering is a method to detect the density-connected clusters in all subspaces of high dimensional data. In our proposed approach Density subspace clustering algorithm is used to find best cluster result from the dataset. Density subspace clustering algorithm selects the P set of attributes from the dataset. Then apply the density clustering for selected attributes from the dataset .From the resultant cluster calculate the intra and inter cluster distance. In this method finds the best cluster distance, repeat the steps until all the attributes in the dataset are clustered among the clustered result in the dataset and finally finds the best cluster result.

Clustering methods are used to support estimates a data distribution for newly attracted data and their ability to generate cluster boundaries of arbitrary shape, size and efficiently. Density based clustering for measuring dynamic dissimilarity measure based on the dynamical system was associated with Density estimating functions. Hypothetical basics of the system proposed measure are developed and applied to construct a different clustering method that can efficiently partition of the whole data space in the dataset. Clustering based on the Density based clustering dissimilarity measure is robust to handle large amount of data in the dataset and able to estimate the number of clusters automatically by avoid overlap. The dissimilarity values are evaluated and clustering process is carried out with the density values.

Similarity measures that take into consideration on the context of the features have also been employed but refer to continuous data, e.g., Mahalanobis distance. Dino Ienco proposed context-based distance for categorical attributes. The motivation of this work is to measure the distance between two values of a categorical attribute *Ai* can be determined by which the values of the other attributes *Aj* are distributed in the dataset objects: if they are similarly of the attributes distributed in the groups of data objects in correspondence of the distinct values of *Ai* a low value of distance was obtained. Author also Propose also a solution to the critical point choice of the attributes *Aj* .The result was validated with various real world and synthetic datasets, by embedding our distance learning method in both partitional and a hierarchical clustering algorithm.

II. RELATED WORK

This work is based on hierarchical approach. So, the process is incremental clustering process.

A. Hierarchical Clustering

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). The agglomerative approach starts with each data point in a separate cluster or with a certain large number of clusters. Each step of this approach merges the two clusters that are the most similar. Thus after each step, the total number of clusters decreases. This is repeated until the desired number of clusters is obtained or only one cluster remains. By contrast, the divisive approach starts with all data objects in the same cluster. In each step, one cluster is split into smaller clusters, until a termination condition holds. Agglomerative algorithms are more widely used in practice. Thus the similarities between clusters are more researched.

Hierarchical *divisive* methods generate a classification in a top-down manner, by progressively sub-dividing the single cluster which represents an entire dataset. Monothetic (divisions based on just a single descriptor) hierarchical divisive methods are generally much faster in operation than the corresponding polythetic (divisions based on all descriptors) hierarchical divisive and hierarchical agglomerative methods, but tend to give poor results. Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

B. Non Hierarchical Clustering

A non-hierarchical method generates a classification by partitioning a dataset, giving a set of (generally) non-overlapping groups having no hierarchical relationships between them. A systematic evaluation of all possible partitions is quite infeasible, and many different heuristics have thus been described to allow the identification of good, but possibly sub-optimal, partitions. Non-hierarchical methods are generally much less demanding of computational resources than the hierarchic methods, only a single partition of the dataset has to be formed.

Three of the main categories of non-hierarchical method are single-pass, relocation and nearest neighbour:

- Single-pass methods (e.g. Leader) produce clusters that are dependent upon the order in which the compounds are processed, and so will not be considered further;
- Relocation methods, such as *k*-means, assigns compounds to a user-defined number of seed clusters and then iteratively reassign compounds to see if better clusters result. Such methods are prone to reaching local optima rather than a global optimum, and it is generally not possible to determine when or whether the global optimum solution has been reached;
- Nearest neighbor methods, such as the Jarvis-Patrick method, assign compounds to the same cluster as some number of their nearest neighbors. User-defined parameters determine how many nearest neighbors need to be considered, and the necessary level of similarity between nearest neighbor lists.

III. METHODOLOGY

In Methodology, we discussed about the existing system ODAC and Proposed System of IHCA and MDSC algorithms.

A. Online Divisive Agglomerative Clustering

The Online Divisive-Agglomerative Clustering (ODAC) is an incremental approach for clustering streaming time series using a hierarchical procedure over time. It constructs a tree-like hierarchy of clusters of streams, using a top-down strategy based on the correlation between streams. The system also possesses an agglomerative phase to enhance a dynamic behaviour capable of structural change detection. The ODAC (Online Divisive-Agglomerative Clustering) system is a variable clustering algorithm that constructs a hierarchical tree-shaped structure of clusters using a top-down strategy. The leaves are the resulting clusters, with a set of variables at each leaf. The union of the leaves is the complete set of variables. The intersection of leaves is the empty set. The system encloses an incremental distance measure and executes procedures for expansion and aggregation of the tree-based structure, based on the diameters of the clusters. The main setting of our system is the monitoring of existing cluster's diameters. In a divisive hierarchical structure of clusters, considering stationary data streams, the overall intra-cluster dissimilarity should decrease with each split. For each existing cluster, the system finds the two variables defining the diameter of that cluster. If a given heuristic condition is met on this diameter, the system splits the cluster and assigns each of the chosen variables to one of the new clusters, becoming this pivot variable for that cluster. Afterwards, all remaining variables on the old cluster are assigned to the new cluster which has the closest pivot. New leaves start new statistics, assuming that only forthcoming information will be useful to decide whether or not this cluster should be split. This feature increases the system's ability to cope with changing concepts as, later on, a test is performed such that if the diameters of the children leaves approach the parent's diameter, then the previously taken decision may no longer recent the structure of data, so the system re-aggregates the leaves on the parent node, restarting statistics. We propose our work to solve the density based subspace system comparing to the different densities at each of the subspace attributes system, each time series data streams.

B .Disadvantages of the existing system

- The split decision used in the algorithm focus only focus on measuring the distance between the two groups, which implies high risk to solve the density problems at different densities.
- The different density at sub attribute values is changes to both intra and inter cluster.

C.IHCA (Improved Hierarchical Clustering Algorithm)

The Improved Hierarchical Clustering algorithm [IHCA] is an algorithm for an incremental clustering of streaming time sequence. It constructs a hierarchical tree-shaped structure of clusters by using a top-down strategy. The leaves are the resulting clusters, with each leaf grouping a set of variables. The system includes an incremental distance measure and executes procedures for expansion and aggregation of the tree based structure. The system will be monitoring the flow of continuous time series data. Then time interval will be fixed. Within the specific time interval the data points will be partitioned. In a partition the diameter is calculated. Diameter is nothing but the maximum distance between the two points. Each and every data point of the partition will be compare with the diameter value. If the data point is greater than the diameter value then the split process will be execute otherwise the Aggregate (Merge) process will be performed. Based on the above criteria the hierarchical tree will be growing. Here we have to observe the splitting process, because the splitting will decide the growth of clusters. In the proposed technique the Hoeffding Bound is used for to observe the splitting process. In IHCA the technique unequality vapnik Chervonenkis is used for splitting process. Using this technique the observation of splitting process is improved. So, the cluster is grouping properly.

In the Hoeffding Bound,

$$\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}. \qquad (1)$$

Where, the observations starting that after n independent observations of the real valued random variable r with range R, with confidence $1-\delta,$

In the proposed algorithm, the range value will be increase from R^2 [1]to R^N .[2] So the observation process is not a fixed one. Depends on the number of nodes the system will generating the observation process.

D. Modified Density Subspace Clustering

Instead of finding clusters in the full feature space, subspace clustering is an emergent task which aims at detecting clusters embedded in subspaces. A cluster is based on a high-density region in a subspace. To identify the dense region is a major problem. And some of the data points are forming out of the cluster range is called "the density divergence problem". We propose a novel Modified subspace clustering algorithm is to discover the clusters based on the relative region densities in the subspaces attribute, where the clusters are regarded as regions whose densities are relatively high as compared to the region densities in a subspace. Based on this idea, different density thresholds are adaptively determined to discover the clusters in different subspace attribute. We also devise an innovative algorithm, referred to as MDSC (Modified Density Subspace clustering), to adopt a divide-and-conquer scheme to efficiently discover clusters satisfying different density thresholds in different subspace cardinalities.

Advantages of the proposed system

- The proposed system efficiently discovers clusters satisfying different density thresholds in different subspace attributes.
- To reduce the Density Divergence Problem.
- To reduce the outlier. (The data points which are out of the range).
- To improve the Intra cluster and Inter cluster performance.

E. Algorithm for Modified Density Subspace Clustering

1. Initialize the selected attribute set $A_d^s$and An –total attribute set

2. Select a set of attribute subset $A_d^s$from at dataset d=1…n, $A_d^s \in An$

```
While ($A_d^s$ ==! null )
{
Set $\mathcal{E}$(eps)
Set minPts
$d(p,q) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$
If (q$\epsilon D | d(p,q) \leq minpts)$
{
C = 0
best distance =0
i=1…n

For each unvisited point P in dataset D
mark P as visited
N = getNeighbours (P, eps)
if sizeof(N) < MinPts
mark P as NOISE
else
C = next cluster
expandCluster (P, N, C, eps, MinPts)
Calculate the average distance cluster ($f_i$ )
```

$$f_i = \frac{\sum_{i=1}^{N_c} \left\{ \frac{\sum_{j=1}^{x_i} d(c_i p_{ij})}{xi} \right\}}{N_c}$$

p_{ij} denotes the j^{th} datapoint which belongs to cluster i.

Nc stands for number of clusters.

$d(c_i p_{ij})$ is the distance between datapoint p_{ij} and the cluster centroid c_i.

x_i stands for datapoint which belongs to cluster centroid c_i

```
best distance =0
for each value as $f_i$
if (fi >bestdistance)
{
bestdistance=fi
selected attribute set = $A_d^s$ from best distance fi
Hierarchical _clustering ( )
}
d=d+1
}
End while
}
expandCluster (P, N, C, eps, MinPts)
{
add P to cluster C
for each point P' in N
if P' is not visited
mark P' as visited
N' = getNeighbours(P', eps)
if sizeof(N') >= MinPts
N = N joined with N'
if P' is not yet member of any cluster
add P' to cluster C
Return the cluster C from the datapoint P.
}
Hierarchical _clustering ( )
```

$X = \{x_1, x_2, x_3... x_n\}$ be the set of data points.

I. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.

II. Find the least distance pair of clusters in the current clustering, say pair (C_i), (C_j), according to $d[(C_i), (C_j),] = \min d[(i),(j)]$ =best distance where the minimum is over all pairs of clusters in the current clustering.

III. Increment the sequence number:

m = m +1.Merge clusters C_i and C_j into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(C_i), (C_j),]

IV. Update the distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The distance between the new cluster, denoted (C_i), (C_j) and old cluster (C_k) is defined in this way: d [(C_k), (C_i), (C_i)] = min [d(k,i) ,d(k,j)]

V. If all the data points are in one cluster then stop, else repeat from step 2.

3. End

IV. EXPERIMENTAL RESULTS

The main objective of this chapter to measure the proposed system result with the existing system. Measuring the performance of cluster results and cluster analysis was measured in terms of the Cluster Quality (Intra Cluster and Inter Cluster) and Computation Time. The proposed system is very much adapted with the dynamic performance of time series data stream. We must evaluate our proposed system with real data produced by applications that generate time series data streams.

A. Evaluation Criteria for Clustering Quality

Generally, the criteria used to evaluate clustering methods concentrate on the quality of the resulting clusters. Given the hierarchical characteristics of the system, the quality of the hierarchy is constructed by our algorithm. And another evaluation criterion is computation time of the system.

B. Cluster Quality

A good clustering algorithm will produce high quality based on intra cluster similarity and inter cluster similarity measures. The quality of the clustering result depends on the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. The criteria for measuring the cluster quality of intra clusters similarity will be high. And the inter cluster similarity will be low. For analysing cluster quality will be in two forms, First one is finding groups of objects will be related to one another. And second one is finding the group of objects that differ from the objects in other groups.

C. Computation Time

Another evaluation of this work is calculating the computation time of the process. The complexity of execution time will be decreased when using the proposed work.

D. Outlier

Outlier is nothing but, the data points which are out of the range of the cluster. The outlier is calculated for the existing method of ODAC(Online Divisive Agglomerative Clustering) and the proposed method MDSC (Modified Density Subspace Clustering).

Outlier Calculation

Step 1: Intra Cluster value is calculated for all Clusters.

Step 2: Mean of the Intra cluster is found out.

Step 3: All the data points of the clusters will be comparing with the mean value.

Step 4: After comparison, each data point will be decided whether the point will position within a cluster or out of the cluster.

V.SYSTEM EVALUATION ON TIME SERIES DATA SET

This proposed method is evaluated with different kinds of time series data sets. Three types of data sets are used to evaluate the proposed algorithm. The data sets are namely ECG Data, EEG Data and Network Sensor Data. ECG Data set is used to find out the anomaly Identification. This data set have three attributes namely time seconds, left peek and right peek. EEG Data set is used to find out abnormal personality. The name of the attributes is Trial number, Sensor value, Sensor position, Sample number. The third type of data set is Network sensor. The name of he attributes is Total bytes, in bytes, out bytes, Total Package, in package, out package, Events.

A. Record Set Specification

TABLE I DATA SET SPECIFICATION

Data Set	**Number of Instance**	**Number of Attributes**
ECG	1800	3
EEG	1644	4
Sensor Network	2500	7

Using the above three kinds of data sets we have to calculate Execution time of the system, Intra cluster , Inter cluster and outlier of the cluster.

B. Result of Outlier

TABLE2 OUTLIER SPECIFICATION

Technique	**Outlier Points**
Existing System(ODAC)	152
Existing System(IHCA)	123
Proposed System(MDSC)	63

C.Result of Execution Time

The following table shows that the difference between the two techniques of the system execution time.

TABLE 3
EXECUTION TIME BETWEEN EXISIING AND PROPOSED

No of Clusters	Existing System Time in seconds	Proposed System Time in seconds	
	ODAC	IHCA	MDSC
2	2.0343	1.9066	1.3782
4	2.0493	1.9216	1.3664
6	2.1043	1.9766	1.3641
8	2.1115	1.9838	1.3901
10	2.0536	1.9259	1.3251

FIGURE 1
EXECUTION TIME BETWEEN THE EXISTING AND PROPOSED SYSTEMS

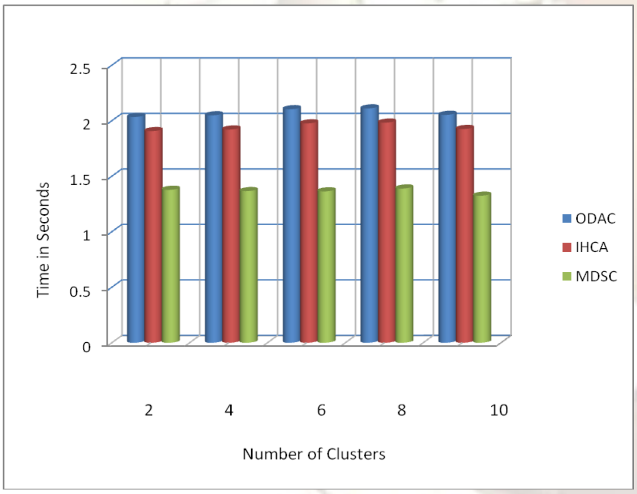


TABLE 4
INTRA CLUSTER BETWEEN EXISIING AND PROPOSED SYSTEM

No of Clusters	Existing System Intra Cluster	Proposed System Intra Cluster	
	ODAC	IHCA	MDSC
2	898.94	865.15	831.72
4	448.87	415.63	382.21
6	346.56	313.41	279.98
8	271.54	238.54	205.11
10	199.04	166.04	132.61

FIGURE 2
INTRA CLUSTER BETWEEN EXISIING AND PROPOSED SYSTEM

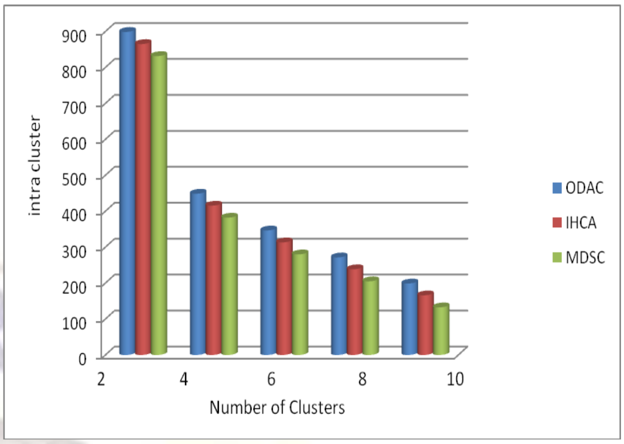
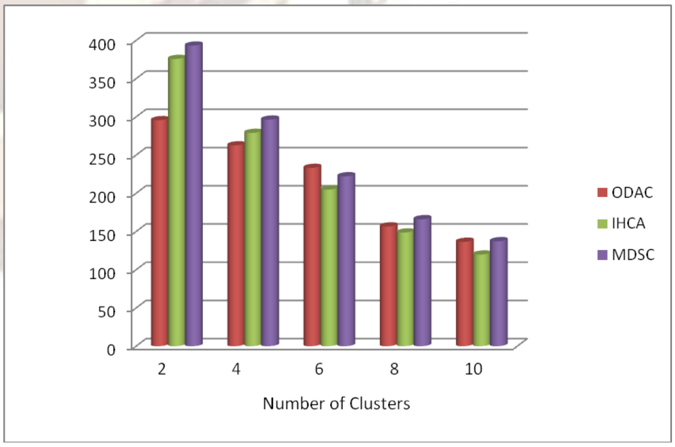


TABLE 5
INTER CLUSTER BETWEEN EXISIING AND PROPOSED SYSTEM

No of Clusters	Existing System Inter Cluster	Proposed System Inter Cluster	
	ODAC	IHCA	MDSC
2	295.64	375.84	393.26
4	262.72	279.07	296.42
6	233.27	204.84	222.26
8	156.65	148.74	166.16
10	136.54	119.89	137.31

FIGURE 3
INTER CLUSTER BETWEEN EXISIING AND PROPOSED SYSTEM



VI.CONCLUSION AND FUTURE WORK

Clustering of the time series data faced with curse of dimensionality, where real world data consist of many dimensions. Finding the clusters in feature space, subspace clustering is an growing task.

Density based approach to identify clusters in dimensional point sets. Density subspace clustering is a method to detect the density connected clusters in all subspaces of high dimensional data for clustering time series data streams Multidimensional data clustering evaluation can be done through a density-based approach. The Modified Density subspace clustering algorithm is used to find best cluster result from the dataset. Improve the Cluster Quality and evaluate the similarity between the data points in the clustering, The MDSC algorithm also concentrates the Density Divergence Problem (Outlier). Proposed system clustering results compared them with existing clustering algorithms in terms of the Execution time, Cluster Quality analysis. Experimental results show that proposed system improves clustering quality result, and less time than the existing clustering methods. The problem find out in the proposed work is, to optimize the centroid point using Multi view point approach. And to apply this technique in non time series data set also.

References

[1] Yi-Hong Chu, Jen-Wei Huang, Kun-Ya Chuang, DeNian Yang, "Density Conscious Subspace Clustering for High Dimensional Data" IEEE Transactions on Knowledge and Data Engineering. Vol 22, No 1, January 2010.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, E. Simoudis, J. Han, and U. Fayyad, Eds. AAAI Press, 1996, pp. 226–231.

[3] Y. Kim, W. Street, and F. Menczer, "Feature Selection in Unsupervised Learning via Evolutionary Search," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 365-369, 2000.

[4] M. Halkidi, Y. Batistakis, and M. Varzirgiannis, "On clustering validation techniques," Journal of Intelligent Information Systems, vol. 17, no. 2-3, pp. 107–145, 2001.

[5]. Pedro Pereira Rodriguess and Joao Pedro Pedroso, "Hierarchical Clustering of Time Series Data Streams," Sudipto Guha, Adam Meyerson, Nine Mishra and Rajeev Motiwani, "Clustering Data Streams: Theory and Practice", IEEE Transactions on Knowledge and Data Engineering. Vol. 15, no. 3, pp. 515-528, May/June 2003.

[6] Ashish Singhal, and Dale E Seborg, "Clustering Multivarriate Time Series Data," Journal of Chemometrics, vol. 19, pp. 427-438, Jan 2006.

[7] Sudipto Guha, Adam Meyerson, Nine Mishra and Rajeev Motiwani, "Clustering Data Streams: Theory and Practice", IEEE Transactions on Knowledge and Data Engineering. Vol. 15, no. 3, pp. 515-528, May/June 2003.

[8] Ashish Singhal, and Dale E Seborg, "Clustering Multivarriate Time Series Data," Journal of Chemometrics, vol. 19, pp. 427-438, Jan 2006.