INTERNATIONALJOURNALOFENGINEERING SCIENCES&MANAGEMENT

Optimal Grid Size Using Grid Clustering Algorithm

Nidhi Ashwin Shah* and Raj Kumar Paul

Computer Science & Engineering Department, Vedica Institute of Technology,
Bhopal, India
nidhishah285@gmail.com, rajkumar.rkp@gmail.com

ABSTRACT

The clustering plays a major role in every day-today application. Grid-based methods are highly popular compared to the other conventional models due to their computational efficiency but to find optimal grid size is a key feature in grid-based clustering algorithm. There exists some algorithm in that they achieve optimal grid size but in real life data can be dense or sparse.The grid-clustering algorithm is the most important type in the hierarchical clustering algorithm. The grid-based clustering approach considers cells rather than data points. In grid-based clustering all the clustering operations are performed on the segmented data space, rather than the original data objects. Grid-based methods are highly popular compared to the other conventional models due to their computational efficiency but to find optimal grid size is a key feature in grid-based clustering algorithm. There exists some algorithm in that they achieve optimal grid size but in real life data can be dense or sparse. So, in these research to develop an algorithm that can find optimal grid size in any type of dataset in dense or sparse with appropriate accuracy or maintaining the accuracy with less time.

Keyword: - *Data mining, KDD; Clustering, grid, GRPDBSCAN, GDILC,GGCA, OPT-GRID(S),STING,CLIQUE*

INTRODUCTION

There is a huge amount of data available in the Information Industry. This data is of no use until it is onverted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining also involves other rocesses such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

From large datasets, Data mining procedure is of taking out of unknown, analytical data patterns. Data mining definition can be described as a analyzing process and then rearranging the patterns of the data and finding co-relations in them in a manner that it goes in the benefit of the overall industries. In Today's world Mining of Data is useful specially, when there is enormous data quantity and identifying the useful portions of it container of a tiresome job in itself. Through Data mining we can be practical about situations rather than

Showing – that is now the future trends can be tried and predicted by us rather than detecting them before they taken place. Mining techniques said to be combination of three main factors: Data, Information and knowledge. Data are said to be mainly essential illustration of the entities, activities or events and/or transactions. Information is organized data which have some valuable meaning or some useful data. Knowledge defined as understanding information that is provided by known pattern or algorithms.

Data Mining

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. Data mining refers to "knowledge mining from data". Mining is a vivid term characterizing the

process that finds a small set of precious nuggets from a great deal of raw material. Through Mining of data technique, we able to find precious knowledge from the "large amount of dataset" [1][2].

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Some basic terminology related to Mining of data are defined as follows:

Data: Data are any facts, numbers, or text that can be processed by a computer.

Information: The processing among all the *data* can provide *information*.

Knowledge: Information can be converted into *knowledge* about historical patterns and future trends.

Knowledge Discovery From Data(KDD)

Data mining is also call knowledge discovery from the huge amount of data. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an Essential step in the process of knowledge discovery. The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases. It definite as “Knowledge Discovery Process(KDD)” from the vast data. This process consist of sequential steps may performed in iterative manner as follows:

Data Cleaning: This procedure consist of take out the noisy data, misplaced data, invalid data or unrelated data from the data set. Data integration may involve inconsistent data and therefore needs data cleaning. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data preprocessing step while preparing the data for a data warehouse.

Data integration: In this procedure multiple or diverse data are incorporated in to one. That merges the data from multiple heterogeneous data sources into a coherent data store.

Data Selection In this procedure specific task or analysis applicable data elected from the database. Sometimes data transformation and consolidation are performed before the data selection process.

Data Transformation: Data are transferred or converted into the suitable form of mining through “summary” or “aggregation”.

Data Mining: In this process through clever methods or techniques useful knowledge obtained.

Pattern Evolution: Through useful knowledge and some interesting parameters appealing pattern are obtained.

Knowledge presentation: “Visualizations” and “Knowledge Representation” methods are use to representing mine data to users.

Data Mining Applications

Data Mining Highly Useful In The Following Domains:

- Market Analysis And Management
- Corporate Analysis & Risk Management
- Fraud Detection
- Production Control
- Science Exploration
- Sports
- Internet Web Surf-Aid

Clustering Techniques

Partitioning Method In :The Partition method given dataset are partition into the no. of small part. Suppose n data object given n so partition methods generate k partition of that data objects.

- It means that it will classify the data into k groups, which satisfy the following requirements:
- Each group contains at least one object.
- Each object must belong to exactly one group.

Mainly two rules are there to follow partition method. (1) Each object must be belongs to one partition and (2) each partition at least contain one object. This method contains algorithms like k-mean and k-median.

Hierarchical
Method

This method hierarchical decomposition of the given set of data objects will be generate. This method creates a hierarchical decomposition of the given set of data objects.

It can be classified in either agglomerative or divisive, based on how the hierarchical decomposition is formed. The agglomerative method, also called the bottom-up approach, it starts with each object forming a separate group. It successively merges the objects or groups that are close to one another, until all of the groups are merged into one cluster, or until a termination condition holds. The divisive approach, also called the top-down approach, it starts with all of the objects in the same cluster.

Density-Based Method

Most partitioning methods of cluster objects work with the distance between objects. This method create problem while generating cluster in the arbitrary shapes. Methods which are developed based on the density of objects are call as a density based methods. Density is called as a no. of object or a data points within the given distance. Their general idea is to continue growing the given cluster as long as the density in the “neighborhood” exceeds

some threshold; that is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise (outliers) and discover clusters of arbitrary shape.

DBSCAN and its extension, **OPTICS**, are main types of density-based methods that grow clusters according to a density-based connectivity analysis.

Grid-Based Method

In this method data objects are divided into the finite number of the cell which is called a grid structure. The objects together form a grid. The object space is quantized into finite number of cells that form a grid structure. All the clustering methods are applied on that grid structure.

Main advantages of this type of method are that its processing time is less than any other method. It is independent of the number of object in the database and it is dependent
on the number of grids in the grid structure and its dimension in the cluster. STING is a example of the Grid based method.

Grid Based Clustering

The clustering methods discussed so far are data-driven—they partition the set of objects and adapt to the distribution of the objects in the embedding space. Alternatively, a **grid-based clustering** method takes a space-driven approach by partitioning the embedding space into *cells* independent of the distribution of the input objects.

The grid-clustering algorithm is the most important type in the hierarchical clustering algorithm. The grid-based clustering approach considers cells rather than data points. This is because of its nature grid-based clustering algorithms are generally more computationally efficient among all types of clustering algorithms. In fact, most of the grid-clustering algorithms achieve a time complexity of(n)where n is the number of data objects. It allows all clustering operations to perform in a gridded data space. Grid-based methods are highly popular compared to the other conventional models due to their computational efficiency.

The main variation between grid-based and other clustering methods is as follows. In grid-based clustering all the clustering operations are performed on the segmented data space, rather than the original data objects. Then any topological neighbor search is used to group the points of the closer grids. The grid-based clustering uses the multi resolution grid data structure. It is non-parametric means it does not require users to input parameter. Since cell density often needs to be calculated in order to sort cells and select cluster centers, most grid-based clustering algorithms may also be considered density-based. Some grid-based clustering algorithms also combine hierarchical clustering or subspace clustering in order to organize cells based on their density.

The computational complexity of most clustering algorithms is at least linearly proportional to the size of the data set. The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets. Gird-based method could be natural choice for data stream in which the infinite data streams map to finite grid cells. The synopsis information for data streams is contained in the grid cells. The example of grid-based clustering are **STING** (a STatistical INformation Grid approach), **CLIQUE** which is applied on high dimensional data and wave cluster.

The grid-based clustering method has the following advantages.

(1) Shapes are limited to union of grid-cells.

(2) It has fast Processing time in terms of it does not calculate distance and it is easy to determine which clusters are neighboring. Also, clustering is performed on summaries and not individual objects.

The grid-based clustering pproach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points. In general, a typical grid-based clustering algorithm consists of the following five basic steps (Grabusts and Borisov, 2002).

Creating the grid structure, i.e., partitioning the data space into a finite number of cells.

Calculating the cell density for each cell.

Sorting of the cells according to their densities.

Identifying cluster centers.

Traversal of neighbor cells.

LITERATURE SURVEY

Cluster Analysis :Imagine that you are given a set of data objects for analysis where, unlike in classification, the class label of each object is not known. This is quite common in large
databases, because assigning class labels to a large number of objects can be a very costly process. *Clustering* is the process of grouping the data into classes or *clusters*, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Often, distance measures are used. Clustering has its roots in many areas, including data mining, statistics, biology, and machine learning.

The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups of classes of objects, it requires the often costly collection and labeling of a large set of training tuples or

patterns, which the classifier uses to model each group. It is often more desirable to proceed in the reverse direction: First partition the set of data into groups based on data similarity (e.g., using clustering), and then assign labels to the relatively small number of groups. Additional advantages of such a clustering-based process are that it is adaptable to changes and helps single out useful features that distinguish different groups. Cluster analysis is an important human activity. Early in childhood, we learn how to distinguish between cats and dogs, or between animals and plants, by continuously improving subconscious clustering schemes. By automated clustering, we can identify dense and sparse regions in object space and, therefore, discover overall distribution patterns and interesting correlations among data attributes. Cluster analysis has been widely used in numerous applications, including market research, pattern recognition, data analysis, and image processing. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In biology, it can be used to derive plant and animal taxonomies, categorize genes with similar
functionality, and gain insight into structures inherent in populations. Clustering may also help in the identification of areas of similar land use in an earth observation database and in the identification of groups of houses in a city according to house type, value, and geographic location, as well as the identification of groups of automobile insurance policy holders with a high average claim cost. It can also be used to help classify documents on the Web for information discovery.

PROPOSED WORK

The proposed work is to find optimal grid size in grid-based clustering algorithm. There exists some algorithm in that they achieve optimal grid size but in real life data can be dense or sparse. So, proposed work is to develop an algorithm that can find optimal grid size in any type of dataset in dense or sparse with appropriate accuracy or maintaining the accuracy with less time. The proposed algorithm works in same manner except stopping criteria. With generation of clusters outlier detection is also required. If there exist outlier in the cluster it changes the characteristic of the cluster.

The proposed algorithm works in same manner except stopping criteria, To detect Outlier. The Algorithm calculate distance between all the data points for the same cluster. The distance is compared with the threshold value T. Any data point having distance greater than threshold value consider as outlier than and only than stopping condition is full filled, otherwise continue the partitioning process and cluster generation. Thus in our algorithm, we count total number of neighboring points within the same formed cluster are. Choose center point that share common vertices within the same cluster. Then all the neighboring points with respect to that points are counted. For any point to be neighboring it should be under maximum threshold value.

Proposed Algorithm

Step 1: Initialize grid structure G for given set of data points given data points with chosen dimension.

Step 2: Partition the initial grid into two equal volume of grids. Data point of G are distributed to this two grids which have non-empty and empty grids.

Step 3: After each round of partitioning of gird it is necessary to check the presence of the new cluster C.

Step 4 : Repeat step 2 TO 3 until the optimal grid structure is generated.

Step 5: Generate the clusters (say, C1,C2,…,Cl for some l) by grouping points in the grids which are connected by the common vertices.

Step 6: Find the number of boundary grids of all the clusters C1, C2,…, Cl

Step 7: If boundary grids of any cluster has outlier **then Go to** step 8 **Else Go to** Step 13 .

Step 8: Calculate distance between all data points for same cluster.(Say C1), Using Euclidean distance formula.

Step 9: Find the threshold value T by using the mean value of the calculated distance for same cluster.(Say C1).

Step 10: Select a Cluster Center data point randomly or by using mean value say C (imaginary cluster center) for same cluster.(Say C1).

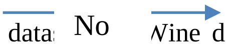
Step 11: Calculate the distance between C and any data point P in S, until no any single data point is remaining using Euclidean distance formula.

Step 12: If any single data point having a distance greater than T, Than it is outlier, go to step 8. **Else Go to** Step 13.

Step 13: Output the generated clusters C1, C2,…, Ck with respect to the grid size.

RESULT ANALYSIS

Dataset Description : In the proposed system, the Heart dataset, Iris data Wine dataset are used to perform grid clustering algorithm and it is taken from UCI Machine Learning Repository which is having the extension .txt [11]

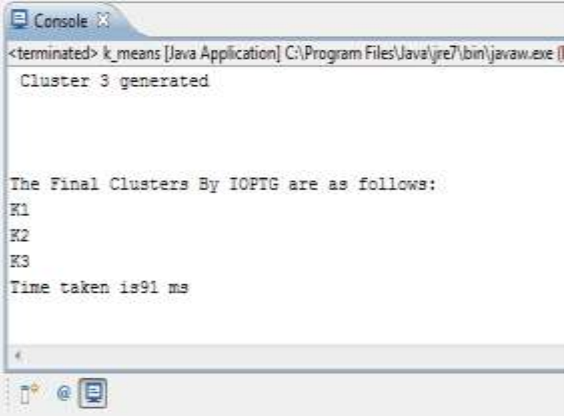


Data	No. of Attributes	Instances
Heart	13	270
Iris	4	150
Wine	13	178

Table 5.2. Dataset Description

Implementation Screenshot

Here, the presented result with complete execution as screen shot. The result of dataset with proposed algorithm is taken. Results show how many no of clusters are created and execution time. Based on that parameter, we will compare each result.

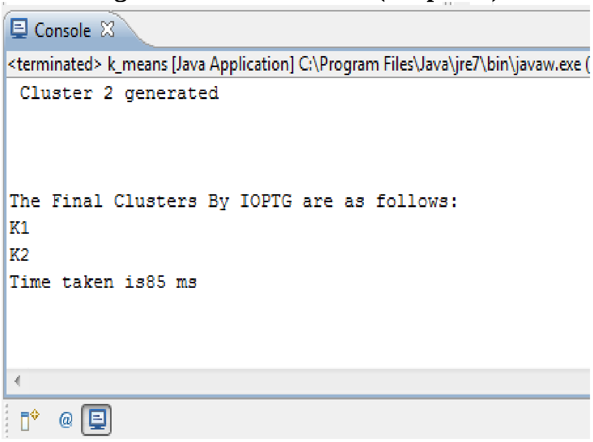


```
<terminated> k_means [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (
 Cluster 3 generated



The Final Clusters By IOPTG are as follows:
K1
K2
K3
Time taken is91 ms
```

Figure 5.1: Heart dataset(Proposed)

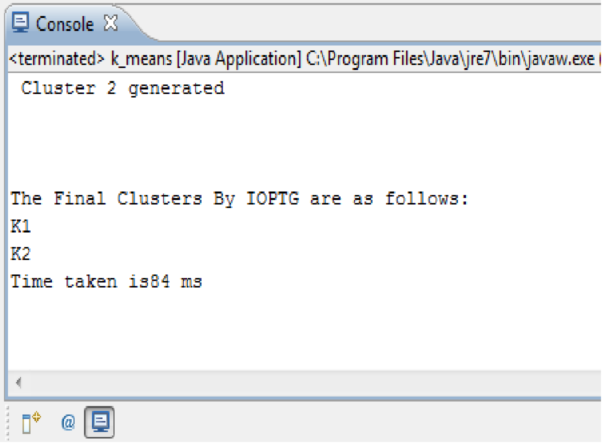


```
<terminated> k_means [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (
 Cluster 2 generated



The Final Clusters By IOPTG are as follows:
K1
K2
Time taken is85 ms
```

Figure 5.2:Iris dataset (Proposed)



```
<terminated> k_means [Java Application] C:\Program Files\Java\jre7\bin\javaw.exe (
 Cluster 2 generated



The Final Clusters By IOPTG are as follows:
K1
K2
Time taken is84 ms
```

Figure 5.3: Wine dataset (Proposed)

Results:

In this section, we have presented execution result of existing and proposed with specific dataset and parameter. We have taken execution time vs. algorithm with different dataset.

Dataset	No of Attributes	Instances	No of Clusters	Execution Time(ms)	
				Existing	Proposed
Iris	4	150	2	89	85
Wine	13	178	2	89	84
Heart	13	270	3	100	91

Table 5.3: Current VS Proposed (Time)

As shown in table, by comparing the modified algorithm with current algorithm with using time as parameter. Here the presented result for execution time vs algorithm. In this case two algorithms are used. First is current algorithm which is defined by researcher and second is modified algorithm.

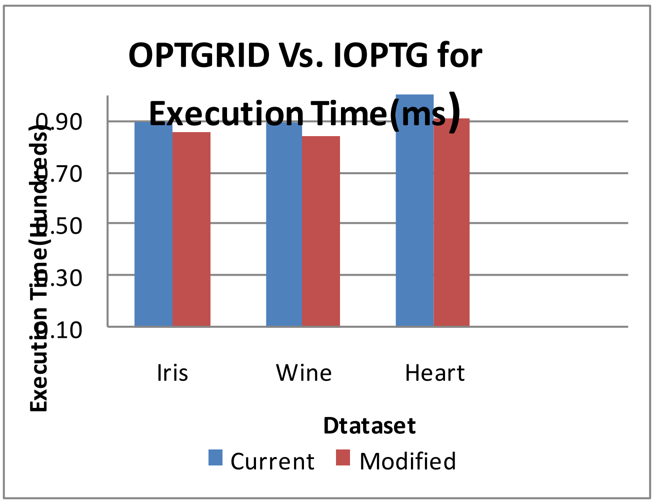


Figure 5.4 Current VS Proposed for Execution Time(ms)

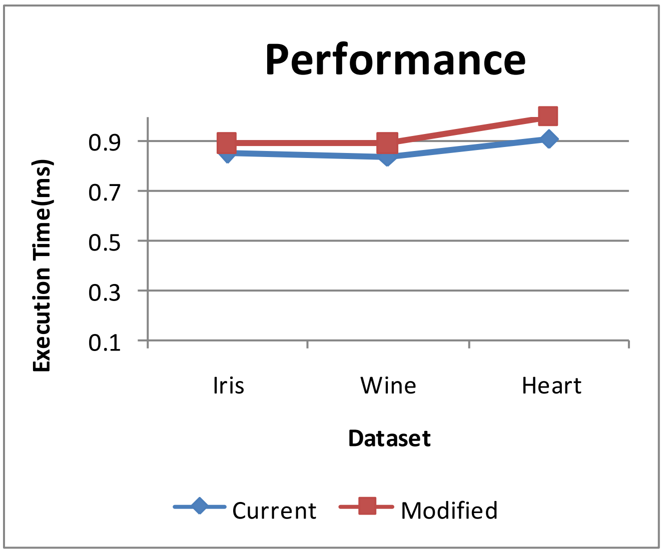


Figure 5.5 Performance Graph

Comparison Table

Table 5.4 Current VS Proposed (Time)

Data	**Execution Time(ms)**	
	OPTGRID	**IOPTG**
Heart	100	91
Iris	89	85
Wine	89	54

It can be observed that execution time is less while using modified algorithm. So using this result it can be clearly analyzed that modified algorithm is more efficient than current algorithm. Graphical represent of result is available above.

Conclusion and future work

The Experimental results show that the proposed approach consumes less execution time with same number of clusters as compared to existing approach. OPTGRID takes more time

to calculate the LOF whereas proposed approach uses less time because In the proposed method presented based on hybridization of density based and distance based outlier detection approach, where a point with maximum density value is used as an imaginary cluster center. The data points having distance greater than Threshold value are consider as outlier thus removed. In the proposed method we need to calculate the distance only once, while in the OPTGRID the calculation for the LOF increase the overhead. Even the k-

nn will also be able to remove the outliers same as the LOF but will optimize the grid partitioning. If the given data has clusters of variable densities, the notion of boundary grids may not results the optimal grid size.

REFERENCES

[1] Han, P.N., Kamber, M.: Data Mining: Concepts and Techniques,2nd (2006).

[2] Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining (2006).

[3]http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf

[4]https://lh4.ggpht.com/hqNrrhjbNEh1o1cYYrQh_YPCoGk8qQdDFyDKoY1px5Pui gMVUCO0FDN4p3yRbAzVkfHx7A=s170

[5] H. Darong and W. Peng, “Grid-based DBSCAN Algorithm with Referential Parameters,” Proc. International Conference on Applied Physics and Industrial Engineering (ICAPIE-2012), Physics Procedia, vol. 24(B), pp. 1166-1170, 2012

[6] N. Chen, A. Chen and L. Zhou, “An incremental grid density-based clustering algorithm,” Journal of Software, vol. 13, no. 1, pp. 1-7, 2002.

[7] E. W. M. Ma and T. W. S. Chow, “A new shifting grid clustering algorithm,” Pattern Recognition, vol. 37, pp. 503-514, 2004.

[8] Y. Zhao and J. Song, GDILC: A Grid-based Density-Isoline Clustering Algorithm,” Proc. International Conferences on Info-tech and Info-net (ICII-2001), vol. 3, pp. 140-145, October 29-November 1, 2001.

[9] Damodar Reddy Edla and Prasanta K. Jana ”A Grid Clustering Algorithm Using Cluster Boundaries” IEEE World Congress on Information and Communication Technologies 2012

[10]http://www.eclipse.org/home/newcomers.php

[11] UCI Machine Learning Repository,http://archive.ics.uci.edu/ml/datasets.html

[12] Monali Parikh,Tanvi Varma, “ IOG -An Improved Approach to Find Optimal Grid Size Using Grid Clustering Algorithm”, IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. IV (May-Jun. 2014).

[13]Jiawei Han, Micheline Kamber, “Data Mining:Concepts and Techniques Second Edition”,University of Illinois at Urbana- Champaign, 2006 by Elsevier Inc.209

[14] MR ILANGO,Dr V MOHAN, “A Survey of Grid Based Clustering, Algorithms”, Ilango et. al. /International Journal of Engineering Science and Technology Vol. 2(8), 2010, 3441-3446

[15] Kabiru Dalhat, Alex Tze Hiang Sim: AN IMPROVED DENSITY BASED k-MEANS ALGORITHM, ARPN Journal of Engineering and Applied Sciences, VOL. 10, NO. 23, DECEMBER 2015

[16] Breunig, M. M., Kriegel, H.-P., NG, R. T., and Sander, J. 2000. LOF: Identifying Density-Based Local Outliers. In Proceedings of 2000 ACM SIGMOD International Conference on Management of Data. ACM Press, 93-104.

[17] Madjid Khalilian, Norwati Mustapha, “Data Stream Clustering: Challenges and Issues”, IMECS 2010.

[18] Mahnoosh kholghi, Mohammadreza Keyvanpour, “An analytical framework of data stream mining techniques based on challenges and requirements”, IJEST, 2011.

[19] Pedro Pereira Rodrigues, João Gama, João Pedro Pedroso , ”Hierarchical clustering of Time series Data Streams”, IEEE Transactions on Knowledge and data engineering, May 2008 vol 20,no.5, pp. 615-627.

[20] T. Soni Madhulatha ,"overview of streaming-data algorithms, Department of Informatics, Alluri Institute of Management Sciences, Warangal, A.P. Advanced Computing: An International Journal ACIJ , Vol.2, No.6, November 2011.

[21]Sudipto Guha, Adam eyerson , Nine Mishra and Rajeev Motwani, "Clustering Data Streams: Theory and practice," IEEE Transactions on Knowledge and Data ngineering, vol. 15, no. 3, pp. 515-528, May/June.

[22] *Breunig, M. M.;* Kriegel, H.-P.*; Ng, R. T.; Sander, J. (2000)* .LOF: Identifying Density-based Local Outliers *(PDF). Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data.* SIGMOD. *pp. 93– 104.* doi*:*10.1145/335191.335388. ISBN 1-58113-217 4.

[23] Alexander Hinneburg and Daniel A. Keim. Optimal grid- clustering: Towards breaking the curse of dimensionality in high- dimensional clustering. In VLDB'99, pages 506–517.Morgan Kaufmann, 1999