Dynamic Clustering: Using Density Metrics

Megha S.Mane[1], Prof.N.R.Wankhade[2]

[1] PG student, Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India
[2] Head and Associate Professor, Department of Computer Engineering, Late G.N.Sapkal College of Engineering, Nashik, Maharashtra, India

ABSTRACT

One of the important tasks in data mining is data clustering. In clustering similar data is clubbed together for storage and analysis. The well known clustering algorithm is k-means algorithm. But this algorithm requires predefined cluster center count and hence unable to identify the ideal clusters from the dataset. The k means algorithm also unable to identify non-spherical clusters. RLCLu algorithm generates non- spherical clusters but required predefined parameters such as local density and minimum distance. To overcome the problems of existing system the proposed system defines cluster centers automatically using statistical testing. This is density based clustering system that generates arbitrary shape clusters. System also works on dynamic streaming data. System accommodates the incoming data points of a data stream in existing cluster using grid clustering and checks the need of re-clustering using statistical testing. The proposed work evaluates the system effectiveness and robustness over streaming data. System also compares the results with x-means cluster center identification process.

Keyword *clustering center identification, data streaming, outward statistical testing, re-clustering, X-means*

1. INTRODUCTION

In cluster generation process objects are clubbed together in a group. Each group contains objects with similar properties. Clustering has many application areas such as, machine learning, time series analysis, information retrieval, pattern identification etc. Previously there are several clustering algorithms have been proposed which basically generates the clusters from input data. The clustering based on Connectivity mechanism clubbed the objects close to each others in one cluster. It is basically organized in hierarchical format but unique partition is not present in it. Therefore, to generate approximate number of clusters still in this user required to pre-assign a distance threshold. Center based clustering approach depicts each cluster as a central vector then objects assigned to the nearest clustering center.

Most popular clustering algorithm such as, k-means, k-Medoids in which k-denotes the number of clusters. the value of k is predefined by user. Pre-assignment or define k in advanced is considered as one of the critical drawback of these previous algorithms. Distribution based clustering such as, Expectation Maximization (EM) algorithm. It utilizes fixed number of Gaussian distributions. Practically, to define object distribution in advanced as Gaussian distribution cannot possible. Therefore, such kind of algorithms still required pre-assigned parameter. To define areas with higher density based clustering is introduced in [5], it can detect clusters in non-spherical shape. DBSCAN is most popular clustering technique. This is a density based clustering technique. In this objects with high density than define threshold are linked together to formed clusters.

Comparing with existing clustering algorithms such as, k-Means and single-linkage algorithms are useful in n non-convex boundaries and execute on trial. Recently, a new clustering algorithm is RLClu is introduced which is the combination of all beneficial features of existing algorithms. It is similar to centroid based clustering. The density based clustering algorithm-RLCLU uses maximum local density function to define cluster centers. These cluster centers generates non-spherical clusters. But from literature survey analysis, it is determined that RLClu still required pre-assign threshold as an input parameter. Hence, in this research work, Statistical Test based Clustering (STClu) approach is proposed.

To calculate local density of each object a new metric is proposed. This technique uses statistical testing strategy to identify cluster centers. System comparesthe cluster count with x-means algorithm. System deals with streaming data to handle dynamic changes in data. System accommodates the incoming streaming points in existing clusters and check whether there is need of re-clustering or not using statistical testing.

2. REVIEW OF LITERATURE

k-means clustering is Lloyd's algorithm can also referred as filtering algorithm. It is based on kd-tree data structure. This structure stores the multidimensional data points. It is binary tree represents a hierarchical subdivision of the point. Every node of kd-tree is associated with closed box, called as cell. The proposed algorithm is simple and easy to understand and for implementation, it only required a kd-tree built once for the given data points. Better efficiency can be achieved with proposed algorithm due to the data points do not vary all over in computation. [1].

Image segmentation for segmentation of color image based on neural networks is discussed in[2]. It provides exhaustive solution for both the supervised and unsupervised segmentation of color images. It has ability to systematically address the problem of color image segmentation. It also contains uniformity in color representation, color reduction and clustering in unsupervised segmentation and color learning in supervised segmentation. Unsupervised segmentation is implemented by SOM-based color reduction, and SA-based color clustering. Using HPL learning and pixel classification supervised segmentation is achieved[2].

A time series clustering has been shown effective in providing useful information in various domains. There are three categories discussed in those are depending upon either time or frequency domain. Features are directly extracted from raw data with developed models from the raw data. Mainly, the fundamentals of time series clustering studies are highlighted. HAC clustering algorithm is introduced in [4]. It is more aggressive optimization for low level computations and it benefits from the most pair wise similarity. Higher dimensions problem is encounter in KNN classification in which dense centroids get avoided. The clustering approach for large vocabularies complete-link clustering can be more efficient than an unoptimized implementation of GAAC [3].

The problem of many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness in exploratory data analysis is addressed in [5][6][7]. Clustering is the combinatorial problem having and variations in assumptions and contexts in different communities have made the transfer of useful generic concepts and methodologies slow to occur. Pattern clustering methods from a statistical pattern recognition perspective also represented in terms of providing advice and references to fundamental concepts attainable to the broad community of clustering practitioners [5].

Clustering is a division of data into groups of similar objects. Data modeling brings clustering in historical perspective rooted in mathematics, statistics, and numerical analysis. Formally, clustering techniques do not categorized into noise nor abnormalities fit into clusters. There are several ways available to learn descriptive learning of managing outliers [6]. In connectivity based clustering works on distance function. The objects lies close to each other are present in one cluster. Objects those are far away from each other are present in different clusters. This type of clustering technique generates the hieratical structure of dataset. But the data is not completely partitioned in to distinct clusters. In such clustering technique distance threshold is user defined parameter.

A simple concave minimization model is introduced in [8]. k-Medoids Algorithm is very simple as well as efficient for quickly addressing useful stationary point. Working of proposed algorithm lies in its ability to manage huge databases and hence it would be useful tool for data mining. Mapping it with the k-Mean Algorithm, we have exhibited instances where the k-Medoids Algorithm is superior, and hence preferable. “k-means,"is Lloyd’s algorithm started working with k arbitrary "centers,"typically had chosen uniformly at random from the data points [9]. Each point is then assigned to the nearest center, and each center is recomputed as the center of mass of all points assigned to it.

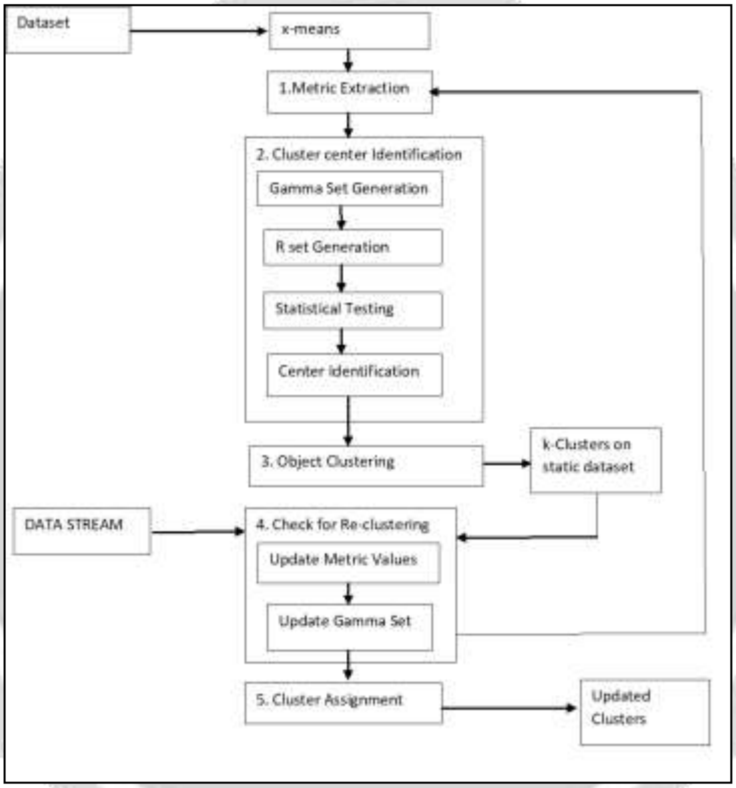
Methods to speed up the computationally procedure has been introduced in [10][11] . These methods are namely, EIGl and EIG1-IG. Post-processing for improvement in current results also introduced. EIGl and EIG1-IG should be applied to ratio cut partitioning for other CAD applications, especially test and the mapping of logic for hardware simulation. In final steps they represented the theoretical analysis for proposed algorithms or methods. The spectral clustering based algorithm uses eigenvalues for similarity matching. It uses dimensionality reduction technique to improve system performance and removes the unwanted dimensions from data set. This technique is used in multivariate statistics. [11].

3. PROBLEM DEFINITION

To design a system which can address limitations of previous clustering algorithms such as, k-Medoids, kmeans++ ,hierarchical algorithms, model based clustering algorithms which required preassigned parameter k from user. Along with the cluster count prediction dynamic data handling is important task. To assign cluster number to incoming streaming data points efficiently and find need of re-clustering based on statistical clustering.

4. SYSTE MARCHITECTURE

Fig. 1 represents the architecture of proposed system.



4.1 Xmeans:
The dataset is given to the X-means clustering algorithm and cluster centers are identified.

4.2 Density Grid Generation:
The input data is mapped onto the two-dimensional graph. The graph is divided in to l*m matrix form. The cell size is defined by partitioning each dimension of length one. The length is defined between 0.02 to 0.05 and it varies. For every data point the value of the grid is mapped. As we defined the moderate grid size, after clustering process records of a same grid cell belongs to one cluster. The information is mined from clustering process and grid to cluster mapping (GCM) data structure is created.

4.3 STClu Algorithm:
The Outward Statistical Testing based Clustering Algorithm STClu mainily contains 3 phases:

A] Metric extraction:
This phase is used to evaluate the two metrics named K-density metric ‘ρ’ and Minimum density-based distance metric ‘δ’ for all the ‘n’ objects in the density grid .

1. K-density: K-density (ρ_i) of object O_i is defined as,

$\rho_i = \frac{K}{\sum_{j=1}^{K} d_{i,j}}$………………..(1)

Where, $\{ d_{i,j} | 1 \leq j \leq K \}$ is the distance between objects O_i and its K nearest neighbors. Clustering centers generally located in the center of a dense area. So, the summation of the distances among a clustering center and its K-neighbors is generally smaller than the summation of the distances among other non-clustering centers and their K- nearest neighbors. According to equation 1, clustering centers generally have higher K-density. So it is rational to use K-density ρ to estimate the density of a particular object.

2. Minimum density-based distance metric: Minimum density-based distance (δ_i) is the smallest distance concerning object O_i and any other objects having higher ρ metric. \

$\delta_i = \min_{j \neq i \wedge \rho_i < \rho_j} (d_{i,j})$………(2)

The object O_i with the highest K-density ρ_i , its δ_i is stated as the maximum distance. That is, their δ_i is greater than their representative nearest-neighbor distance. Also, the clustering centers generally have maximum density. Hence, the cluster centers have anomalously big δ_i. Therefore minimum density-based distance pursues the long-tailed distribution.

But δ is still inadequate to detect the clustering centres as the outliers also have greater δ as per equation (2). But, the clustering centres generally have larger ρ and δ. Therefore, outliers can be detected by using both K-density ρ and minimum density-based distance δ.

B] Clustering center identification:

Gamma-set metric is used to compute the cluster centers. Gamma-set metric is also called as a centrality metric (γ). It is the product of ρ metric and δ metric. Gamma-set metric is evaluated as follows,

Gamma-Set (γ) = RhoSet * DeltaSet i.e. [γ= p. δ]… (3)

Evaluate the statistics $X_{1,n} \geq X_{2,n} \geq \cdots \geq X_{n,n}$ by using the centrality metric, where $X_{1,n}$ is the first maximal value in { γ_i , $1 \leq i \leq n$ }, $X_{2,n}$ is the second maximal, and so on. Construct a sequence of null hypothesis $H_{0,k}$. Null hypothesis $H_{0,k}$ considers that the k^{th} statistic $X_{k,n}$ follows the long-tailed distribution.

1. Ratio evaluation: Compute the ratio R_t by using the ordered statistics.

$R_t = \frac{X_{t,n}}{X_{t+1,n}}$…………….. (4)

where $(1 \leq t \leq n-1)$

2. Outward Statistical testing:

For the null hypotheses $H_{0,1}$, $H_{0,2}$,….,$H_{0,m}$ and the corresponding statistics R_1, R_2 , …, R_m, first analysing $H_{0,m}$ considering R_m if $H_{0,m}$ is not rejected, then check $H_{0,m-1}$ considering R_{m-1}. Follow this outward testing till a null hypothesis is rejected or all the m hypotheses are handled. If $H_{0,k}$ is the first hypothesis being rejected, the objects with reference to the statistics R_1, R_2 , …, R_m are recognized as the clustering centres. For the k^{th} hypothesis, critical value r_k can be computed as

$r_k = [1-(1-\alpha)^{1/m}]^{-1/(\lambda.k)}$………(5)

It gives the critical value evaluation for k^{th} hypothesis.

where α gives the level of significance and λ is the tail index. If $R_m > r_k$, reject the null hypothesis and obtain the clustering centres. So the object is a clustering centre if the null hypothesis with reference to its centrality metric is rejected.

In equation (5), λ is the tail index can be evaluated using Hill-type estimator [11].

$$\hat{\lambda}(\kappa) = \left[\frac{m}{\kappa - m + 1}\ln X_{m+1,n} - \frac{\kappa}{\kappa - m + 1}\ln X_{\kappa+1,n} + \frac{1}{\kappa - m + 1}\sum_{i=m+1}^{\kappa}\ln X_{i,n}\right]^{-1} \quad \text{… …..(8)}$$

where k is the leading index of object in X. Generally, to evaluate the clustering centers m= [0.1n] and k= [0.95n] is used.

C] Object clustering:

Once we get the clustering centers, all the objects apart from the clustering centers are clustered by allocating residual object to the cluster having their nearest neighbor of higher density ρ.

4.4 Data Stream:

Input stream is the data-steam containing two-dimensional m records arrives after every interval of time t. The records are read from stream and cluster centers are assigned to the streaming data points.

4.5 Check for Re-clustering:

After accommodating new streaming points in existing cluster the statistical test is again checked for cluster count. If previously generated cluster count is not same as the current statistical test result then re-clustering is performed by considering whole dataset.

4.6 Data visualization:

The graphical view is generated for clusters. It includes cluster centers, and points in a clusters. Every cluster points are distinguished using different colors.

5. ALGORITHMS

Algorithm 1: Outward Statistical Testing based Clustering Algorithm STClu

Input:

O= {O1, O2 …On} A set of ‘n’ objects

K= Number of nearest neighbor in k-density p.

Output:

CLU: Set of cluster

Processing:

```
Step1: Part 1: Matrix extraction
RhoSet    Ø, DeltaSet    Ø, NNSet    Ø, GamaSet    Ø;
Step2: Calculate distance
    distanceMatri    DistanceFunction(O)
Step3: Calculate p
    RhoSet    $F_p$(distanceMatrix, k);
Step4: Calculate δ and identify the nearest neighbor for each object
    [DeltaSet, NNSet] $F_\delta$ d(distanceMatrix, RhoSet);
Step5: GamaSet RhoSet . DeltaSet;
Ste6: Part 2: Clustering center identification
Step7: Sort GamaSet in descending order to get a set of ordered statistics X
Step8: R   {Ri    $X_{i.n}/X_{i+1}$;ng (1< i<n-1);
Step9: Start at the $m^{th}$ hypothesis
    m    [0:1n], k    0;
Step10: Identify the number of clusters k by outward statistical testing
Step11: while m > 2 do
Step12: Calculate the critical value $r_m$ according to Eq. (8);
Step13: if $R_m > r_m$ then
```

```
Step14: k    m;
Step15: break;
Step16:  end
Step17:  m   m-1;
Step18: end
Step19: Identify the objects corresponding to {R1; R2; ...; Rk} as the clustering centers {c1; c2; ...; ck}, and label ci as i;
Step20: Part 3: Object clustering
Step21: for i    1 to n do
Step22: if $O_i$ is unlabeled then
Step21: Mark $O_i$ the label of its nearest neighbor with higher p according to NNSet;
Step22: end
Step23: end
Step24: CLU    {$Clu_i$,1 < i< k}, where $Clu_i$ denotes the set of objects with label i;
Step25: return;
```

6. MATHEMATICAL MODEL

‘S ‘is the system of utility mining patterns such that
S = {I, F, O}
I is the input to the system
F is system functions
O is Systems output
Obj= {Obj-1, Obj-2, ...,Obj-n}
Set of Objects
K = Number of Clusters
F: {F1, F2, F3, F4, F5, F6, F7, F8, F9}
F1= Distance Evaluation
d{ji}$ denotes the distance between objects Obj-I and Obj-j
F2= K density evaluation
F3 = Minimum density-based Distance Matric
F4 = Gama set Evaluation
F5 = Sort Gama set in descending order
F6 = Calculate Long tail distribution ratio R
F7 = Statistical Testing
F8 = K center Identification
F9 = Object Grouping
O :{D, P,gamma, R, C, CLU}
D = matrix of d_{ij} represents distance between Obj-i and Obj-j
P=Matrix of P_{ij} represents Minimum density based distance metrics
R = Long tail distribution ratio for n Obj
C = {C1,C2,..Ck} set of Clustering Centers
CLU = {CLU-1, CLU-2,..CLU-k} Set of Clusters

7. IMPLEMENTATION

7.1 Experimental Setup

Desktop application using java development kit-1.7 is created for cluster creation. System is tested on core-i3 system with 4 gb RAM.

7.2 Dataset

Two dimensional dataset are considered for testing.

1. A-sets dataset: It is two dimensional dataset with number of clusters. Each cluster contains 150 objects
2. Shape sets datasets: It is 2D dataset such as, Aggregations, D31, flame and Spiral etc. represents complex clustering objects The number of objects in these four data sets is 788, 3100, 240, 312 for Aggregations, D31, flame and Spiral, respectively.

7.3 Performance Measures

1. **Efficiency:** -X- means algorithm is used for initial cluster creation. Efficiency of the system is compared with STClu initial cluster creation process and X-means.

2. **Visualization:** 2D cluster plot will be created to visualize the clusters.

3. Time: The time required for cluster center allotment for STClu and streaming data cluster allotment with grid generation is compared.

8. RESULTS

8.1 Compare Xmeans results with STClu:

Following figure 2 and 3 are the results for aggregation and flame dataset. Xmeans creates only 2 clusters for aggregation dataset whereas STCLu creates 7 clusters. For flame dataset both algorithms creates 2 clusters but the centroid and cluster shape is different.

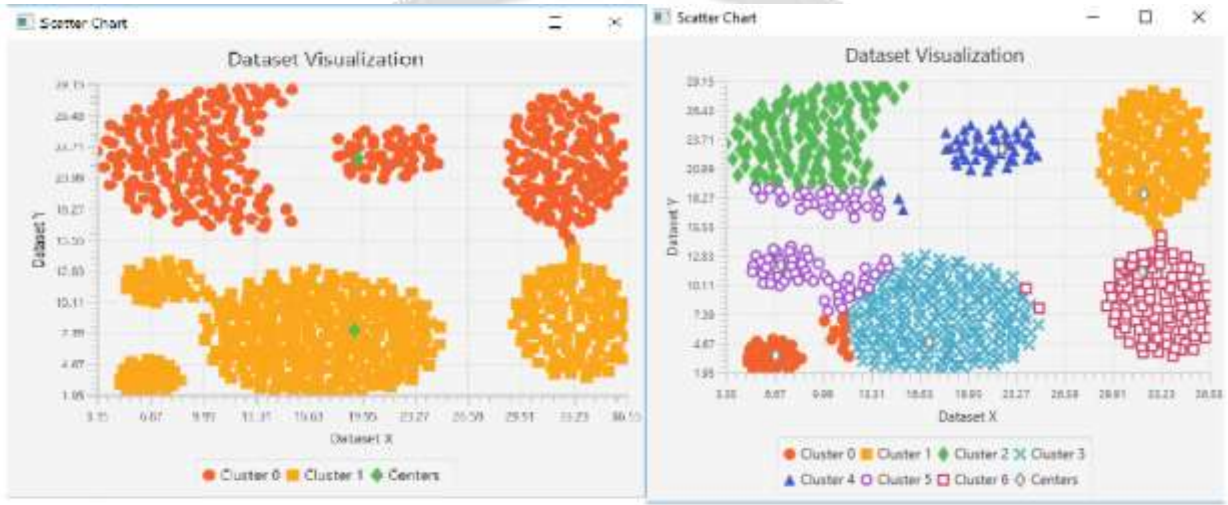


Fig -2: Comparison of XMeans and STClu for Aggregation dataset

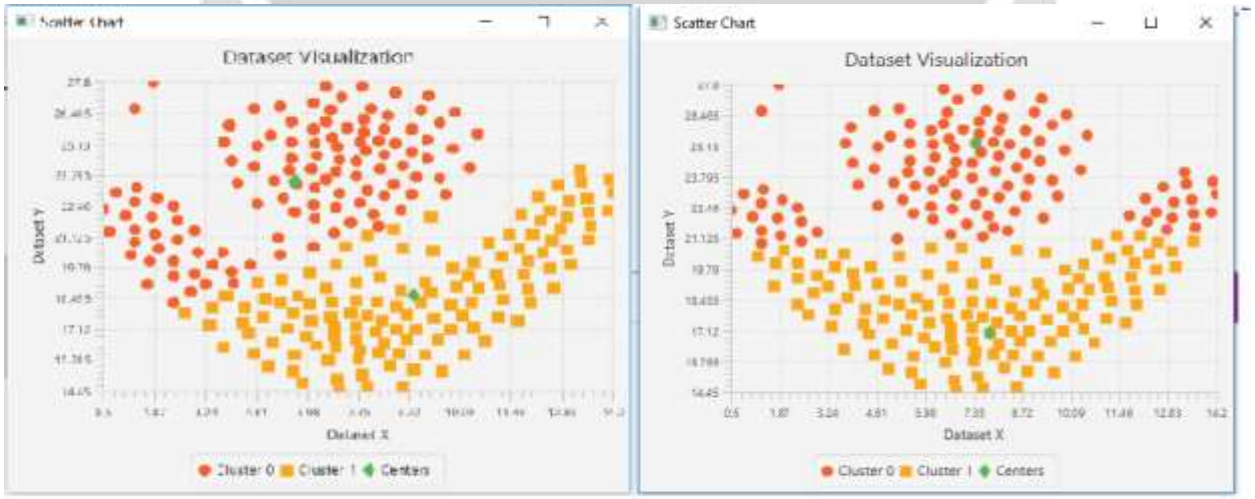


Fig -3: Comparison of XMeans and STClu for flame dataset

8.2. Streaming Data:

Following figure 4 and 5 represents the streaming data results for aggregation and flame datasets. In both of these figure 3 graphs are displayed. The first graph represents the incoming streaming points graph2 represents the cluster allotment. System allot cluster to the incoming streaming point and checks for the need of re-clustering. After re-clustering 8 clusters are created for aggregation dataset whereas 3 clusters are created for flame dataset.

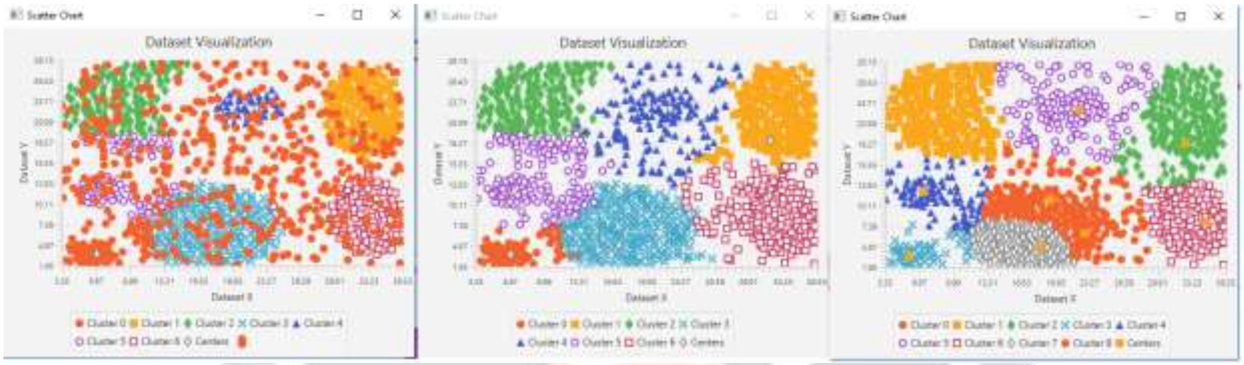


Fig -4: Streaming data results for aggregation dataset

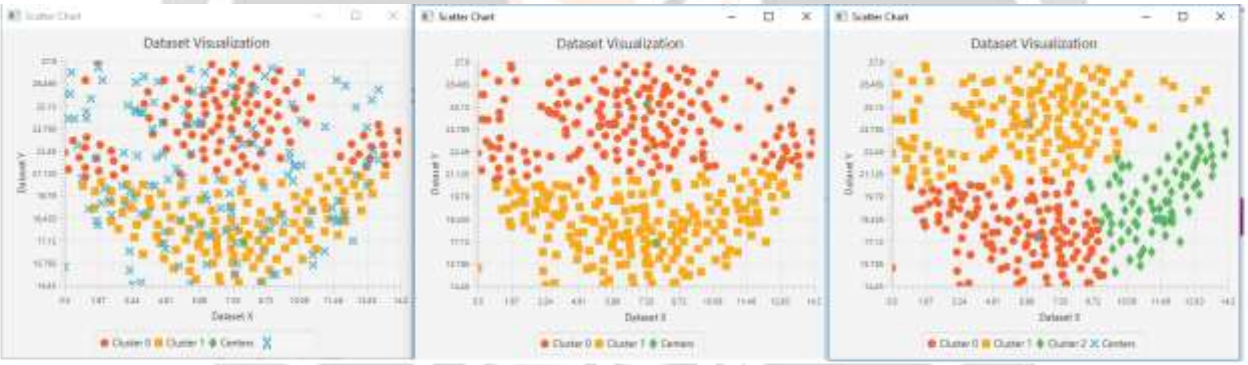


Fig -5: Streaming data results for flame dataset

8.3. Time Comparison:

The time required for object mapping is compared. In STCLu object mapping, the distance between object and cluster center is compared. As number of clusters increases the time required for individual object mapping also increases .The proposed system uses grid data clustering. The grid is assigned to a single cluster. As per the point co-ordinates cluster of object is identified. The proposed system requires less time than the existing system. Following graph shows the average time required for mapping individual object in cluster using STClu and grid clustering.

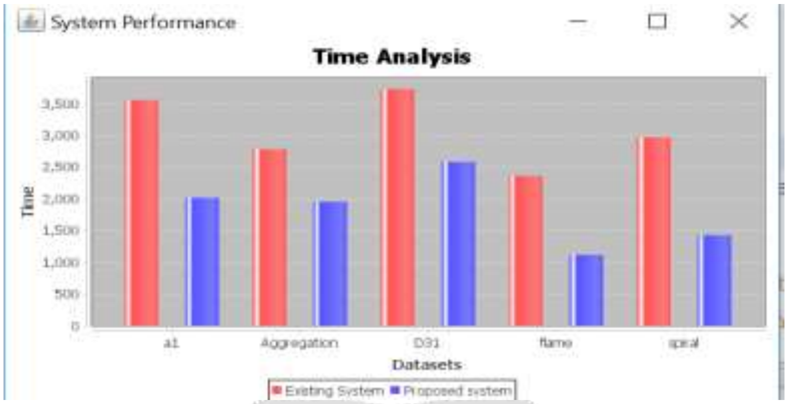


Fig -6: Time Analysis

9. CONCLUSIONS

Clustering is the process of data classification where no prior knowledge provided for classification. STClu algorithm generates the clusters from a given dataset without any prior knowledge. In this algorithm a new metric is defined to cluster the objects based on statistical testing. It uses local density function and minimum density based distance. System compares the estimation of number of cluster k with exiting x means algorithm. The proposed method also deals with dynamic streaming data. To accommodate the changes in cluster structure, re-clustering is performed. The re-clustering decision is taken based on the statistical test condition. In future System can be implemented for categorical data objects.

10. ACKNOWLEDGEMENT

I am deeply expressing my sincere gratitude to my project guide Prof. N.R.Wankhade who always supported and guided me with valuable comments. I express my immense pleasure and thankfulness to all faculty members of the Department of Computer Engineering of Late G.N.Sapkal College of Engineering, Nashik.

11. REFERENCES

[1]. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881-892, Jul. 2002.

[2]. G. Dong and M. Xie, "Color clustering and learning for image segmentation based on neural networks,"IEEE Trans. Neural Netw., vol. 16, no. 4, pp. 925-936, Jul. 2005.

[3]. T. W. Liao, "Clustering of time series data-a survey,"Pattern Recog., vol. 38, no. 11, pp. 1857-1874, Nov. 2005

[4]. N. Jardine and C. J. V. Rijsbergen, "The use of hierarchic clustering in information retrieval,"Inf. Storage Retrieval, vol. 7, pp. 217- 240, 1971.

[5]. A.K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review,"ACM Comput. Surveys, vol. 31, pp. 264-323, 1999.

[6]. P. Berkhin, "A survey of clustering data mining techniques,"in Grouping Multidimensional Data. New York, NY, USA: Springer, 2006, pp. 25-71.

[7]. A.K. Jain, "Data clustering: 50 years beyond k-means,"Pattern Recog. Lett., vol. 31, no. 8, pp. 651-666, 2010.

[8]. P. S. Bradley, O. L. Mangasarian, and W. N. Street, "Clustering via concave minimization,"in Proc. Adv. Neural Inf. Process. Syst., 1997, pp. 368-374.

[9]. D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding,"in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms, 2007, pp. 1027-1035.

[10]. L. Hagen and A. B. Kahng, "New spectral methods for ratio cut partitioning and clustering,"IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst., vol. 11, no. 9, pp. 1074-1085, Sep. 1992.

[11]. W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs,"IBM J. Res. Develop., vol. 17, no. 5, pp. 420-425, 1973.

[12]. http://cs.joensuu.fi/sipu/datasets/