A new approach to cluster high dimensional streaming data

rWFCM-HD

Diksha Upadhyay
Department of CSE
RITS, Bhopal, India
diksha.du31@gmail.com

Susheel Jain
Department of CSE
RITS, Bhopal, India
jain_susheel65@yahoo.co.in

Anurag Jain
Department of CSE
RITS, Bhopal, India
anurag.akjain@gmail.com

Abstract— In the given research paper we have proposed a method known as dimensionally reduced(optimized) weighted fuzzy clustering algorithm with the abbreviation (rWFCM-HD). The method can be used for multidimensional datasets having continuously arriving (streaming) behavior. This category of data we will observe (or obtain) in the area of wireless sensor networks, data obtained from web click stream and data generated by internet traffic flow etc. These categories of data have two special attributes which differentiate them from other category of datasets: a) Firstly they are of streaming behavior and b) Secondly they have multiple dimensions. Minimized fuzzy clustering algorithm has been already provided (proposed) for datasets having continuously arriving behavior or multi dimensions. But as per our survey yet, nobody has proposed any minimized fuzzy clustering method for incoming data sets with both the two functionalities, i.e., data sets with multiple dimensions and regularly arriving streaming behavior. Result outcomes shows that our provided method will enhance performance from the prospective of memory requirement as well as execution time and the abbreviation which we will use for our algorithm is (rWFCM-HD).

Index Terms— Data mining, Fuzzy C-Means, Dimension Reduction, Clustering, K-Means,. Weighted Fuzzy C-Means.

I. INTRODUCTION

In current years there are several sources, for producing data streams which is of continuous arriving behavior has Came in to existence, such kind of data will be obtained from sensor networks areas, data obtained by web click stream as per the action by web user and data stream obtained from internet traffic data transfer, now a days or simply for the recent researches data stream generated from various element become an important source of data. As a result, of which several researchers are paying their importance on it. Discovering efficient data stream mining method has become an interesting and popular research subject. Data stream from an element [1] is potentially of very large size (infinite), with the incoming arriving rate (speed) which is of uncertain nature and can be parsed in one pass. The overall processing of incoming data stream has to implement within a limited available space (memory) and with in a strict time constraint. Due to this, an efficient data stream mining algorithms must be of more capability.

The simple analysis in terms of different properties has been provided for various dimension reduction mechanisms and for various data clustering techniques (survey) in [20]. Cluster analysis is very important from the prospective of data mining. Different Clustering algorithm based on data stream composed model has gone to an extensive research [1], [2], [3], [4], [5]. Fuzzy C means (FCM) and its enhancements [6], [7] as one of the most important clustering techniques have been used in the large scale such as in the area of data mining, in the field of pattern recognition, in the area of machine learning and so on. In [8] the author has provided a weighted fuzzy c-means (sWFCM) clustering procedure for the datasets which have continuously arriving streaming behavior. The various issues and effects of high dimensionality property of sample data sets on clustering, and in solving the problem to find the nearest neighbor has been observed by various researchers in thorough detail. Due to multi dimensions the data becomes dense and sparse to handle in general; the traditional (previous) indexing and algorithmic procedures fail from the prospective of efficiency and effectiveness to do clustering. On multi dimensional data it has been seen that, the various parameters such as proximity measures, distance calculation or finding nearest neighbor may not be that much effective and meaningful as they could have to be. In the Recent research analysis outcomes shows the dimensionality attribute of sample data sets from the prospective of distance metrics which will be further used to find the similarity between various data objects [9]. Further, multi-dimensional data will create various typical issues for various traditional clustering algorithms which is very difficult to handle and require definite solutions. In high dimensional data, conventional similarity measures as used with in old clustering algorithmic approaches are usually not meaningful. Similar procedures to work with multi dimensional data are the concept subspace clustering, then after projected clustering, other one is pattern based clustering or correlation clustering proposed [10]. Due to the presence of various irrelevant attributes or of similarities among subsets of features will hugely effect the generation and visualization of obtained clusters in the full-dimensional space. The major issue the clustering algorithms will face is that the clusters will be generated on the basis of a factor which is subspace of features obtained of input data from the total provided feature space but there is the

possibility that the feature subspace for individual clusters may be of different nature.

The K-Means is one of the most famous clustering algorithm which is quite simple and hugely applicable partitioned clustering technique. The space complexity attribute of the K-Means algorithm is (O ((n + k) d) and the time complexity attribute is O (nKtd) where each letter have their individual meaning n shows the number of data, K shows the possible number of generated clusters, d will represents the dimension of the input data and t will represent the number of iterations. In [11] the authors have proposed a algorithm to convert multi dimensional input data into two dimensional data (two dimensional coordinate points) and then after simple K-Means procedure has been applied on the resulted dataset obtained in output. The major use of this new procedure is to reduce the dimension of the data such that the k-means algorithm will be get highly utilized.

In this study a dimension reduced weighted fuzzy c-means algorithm have been proposed having the abbreviation (rWFCM-HD). The given algorithm will work on the data sets those have higher dimensions and that having arriving streaming behavior. The practical example for such data sets is live high-definition videos available in the World Wide Web. These data's have two special attributes which differentiate them from other category of data sets: a) Firstly they have streaming behavior (continuously arriving) and b) secondly they have higher dimensions. As we discussed above that the optimized K-means procedures has already been provided for sample or incoming data sets which have streaming behavior or higher dimension. But as per our information, nobody has provided any optimized K-means procedure for data sets having both the attributes, i.e., data sets with multi dimensions and also continuously(regularly) arriving streaming behavior. So, our work will be a combination of the work done in [11] and [8].. Various fuzzy clustering algorithm proposed in the present will not be used directly with the data streams. The rest of the paper is organized as per following. In the next section we discussed related research works. Then after Section III will provide the background details required for this paper. Also the discussion of our proposed procedure is done in detail in section IV, and then after experimental generated comparisons and outcomes analysis is provided in section V. And at the last finally we conclude the paper in section VI.

II. RELATED WORKDONE

For performing clustering data stream algorithms have been studied. In the study [2], the STREAM procedure is provided to cluster data streams. STREAM algorithm first determines the size of sample data sets. If the condition arises where size of provided data chunk is larger in comparison to provided size of input data, then after a LOCALSEARCH procedure (algorithm) will be called for obtaining various clusters of the data chunks. And then after, the LOCALSEARCH algorithm is applied on all the cluster centers. generated previously

The k-means procedure is enhanced and the VFKM procedure is provided in [3]. It is guaranteed that the Generated model produced will not differ in comparison from the one that would be generated with infinite dataset. Another variant of the k-means algorithm, namely incremental k-means, is also provided to obtain high quality solutions. In [4] the authors have proposed a composite system (time series clustering technique) which is responsible to create the hierarchy of clusters on the incremental basis phenomena .The correlation between time series is used as similarity measure. Cluster formation and decomposition will be done at each iteration. In [5], another procedure namely CluStream is proposed to cluster newly evolving arriving data streams. CluStreams procedure aim is in partitioning the clustering method in the random working component which will afterwards periodically stores complete summary measures (statistics) and then after an offline component which has a work to use only this summary statistics. Micro-clustering approach with Pyramidal time frame parameters works in joint majority which is used to deal with the issues of generating efficient choice, providing storage, and use of the present statistical data parameters for a continuous fast data stream. For the objective of clustering image data which is of vey large size a algorithm has been provided based on sampling methodology in [12], where the working samples are chosen by the two parameters the hypothesis test or chisquare as per divergence parameter. In [13], speeding up is achieved by performing the quick and randomized sampling of the data and then after performing clustering on it. The centroids obtained are then after wards used for initializing the complete data set. Two well known mechanisms for dimensional reduction techniques are feature extraction and feature selection ; firstly prior to applying any data mining task, the most common way is to get rid of the issues of multi dimensional dataset in which various attributes are collaborated is to perform feature selection in the study [9]. For feature selection there may be unsupervised (PCA [14], LLE [15], ISOMAP [16]) learning mechanisms which Will quickly observe the low dimensional space that classify (represents) well the data without need to any specific task. Principal Component Analysis (PCA) procedure can be used to map directly the original provided data sets in various multiple dimensions to a limited or lesser dimensional data space where the points(centroids) may better cluster and the resulting clusters may be more meaningful to observe. For nonlinear approaches to perform data mining, Sammons mapping, multidimensional scaling and LTSA [17] are available. While another dimension reduction technique which is supervised in nature is Discriminative PLVM [18]) try to estimate a low dimensional representation which has sufficient information for predicting the task target values..

Different soft computing analyzing tools are also available for feature extraction of input data and also feature selection which can be studied in [19]. Next, the decision tree induction is another technique which can be used for attribute subset selection, a decision tree is constructed in this approach of the whole data and the attributes that did not appear in tree are assumed to be less dominant or of less use. After obtaining the tree where the attributes do appear are to be selected as

important attribute. Unfortunately, these dimensionality reduction methods cannot be work for finding the solution for clustering problems because these methods are global accepted and follow different way to work which is they generally compute only one joint subspace of the given original input data space in which the clustering is performed, considering complete set of points. In general, the issue of local feature relevance parameter and local feature correlation parameter justifies that several subspaces are required as each cluster may exist in a different subspace as per study [10]. In [11] dimension reduction mechanism has been proposed Which will first convert the multi dimensional data sets in to two dimensional data which can be easily handled and still rich in all the features and then for increasing the clustering efficiency K-Means clustering algorithm have been applied on the resultant data (two dimensional data).

The basic difference of the above studied procedures with the one proposed by us is that no one has composed a algorithm to work with the dataset having both the properties which is multiple or high dimensions as well as streaming behavior(continuously arriving behavior)

III. BACKGROUND

A. simple FCM algorithm

Consider a data set $X = \{x_1, x_2, x_3, \ldots, x_n\}$, the FCM algorithm firstly partitions X into c fuzzy clusters which will denote the number of clusters , and then find out center of each clusters such that the cost function (objective function) of dissimilarity measure should be minimized or below a certain threshold value. FCM algorithm works on the basis of analyzing membership value parameter of every dataset in every cluster, it is presented as follows:

Objective function:

$$J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m (d_{ik})^2 \qquad (1)$$

U and v can be calculated as:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{(m-1)}}} \qquad (2)$$

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \qquad (3)$$

Where u_{ik} is the membership value of the k^{th} data x_k in the i^{th} cluster. $d_{ik} = \| x_k - v_i \|$ is the Euclidean distance between data x_k and the cluster centroid v_i, $1 \le i \le c$, $1 \le k \le n$, exponent $m > 1$.

The FCM procedure finds the cluster centroid v_i and the after the membership matrix U through number of iterations using the
Following steps:

1. Initialize the membership matrix U , u_{ik} randomly comes from (0, 1) and satisfy:

$$\sum_{i=1}^{c}(u_{ik}) = 1, 1 \le k \le n$$

2. Calculate c fuzzy clusters v_i i=1,……. c using Equation 3.

3. According to Equation 1 Compute the objective function. Stop if objective function of dissimilarity measure is minimized or concentrated on a specific value or if its enhancement results over previous iteration outcomes is below a certain threshold or iterations reach a certain tolerance value.

4. Compute a new U using Equation 2. Go to step 2.

As FCM is clustering or simply operated on the total data set, and input data stream may contain a very large data set, so let FCM deal with data stream directly may utilize significant amounts of CPU time to converge, or result in an intolerable iteration quantity. Based on this situation, [8] proposed one alternative called weighted FCM (Fuzzy clustering algorithm) procedure (swFCM) for data stream as discussed in the next section

B. Weighted FCM (swFCM) Algorithm

Firstly, in this approach divide the arriving data stream into chunks X_1, X_2, …… X_s according to the arrival time of data, and the overall size of each data chunk is determined by main memory of the processing system, let n_1, $n_2, \ldots\ldots n_s$ be the data numbers of chunks X_1, X_2, …… X_s respectively. Due to its stream setting, a time weight w (t) is imposed on each data representing the datum influence extent on the clustering process, and

$$\int_{t_0}^{t_c} w(t)\,dt = 1$$

Where t_0 is the initial time of stream and t_c is the current time.

The main idea or approach of sWFCM is renewing the weighted clustering centers by multiple iterations until the cost function parameter reaches to a satisfying result or the number of iteration is to a tolerance value. Moreover, during the processing, we give the singleton a constant weight as 1. The procedure is presented as follow:

1) Import the chunk X_l $(1 \le l \le s)$.

2) Update the weight of cluster centroids.

- If l = 1: Apply FCM to gain cluster centroids v_i i =1,…… c ,and compute:

$$w_i^{'} = \sum_{j=1}^{n_1} (u_{ij}) w_j \qquad 1 \le i \le c$$

Where $w_j = 1, \forall 1 \le j \le n_1$

- If l 1 > 1:

$$w_i^{'} = \sum_{j=1}^{n_l + c} (u_{ij}) w_j \qquad 1 \le i \le c$$

Where $w_j = 1, \forall c+1 \le j \le n_l + c$

The centroid weight w_i then updates as $w_i = w_i^{'}$

3) Update cluster centroids:

$$v_i = \frac{\sum_{k=1}^{n_l + c} w_k (u_{ik})^m x_k}{\sum_{k=1}^{n_l + c} w_k (u_{ik})^m}$$

Where $x_k \in \{ v_i \mid 1 \le i \le c \} \cup X_l$

4) Compute objective function:

$$J_m(U, v) = \sum_{k=1} c + n_l \sum_{i=1}^{c} w_k (u_{ik})^m (d_{ik})^2$$

Stop if objective function parameter is minimization or concentrate on some specific value, or its enhancements over previous results obtained from iterations is below a certain threshold, or iterations reach a certain tolerance value.

5) Compute a new U using Equation 2. Go to step 2.

6) If $l = s$ then stop, else go to step 1.

C. Converting or transforming high (multi)dimensional dataset into two dimensional data set

The technique proposed in [11] is used for reducing dimension of multi dimensional datasets. In this technique each high dimensional data in the dataset is converted to a two dimensional co-ordinate point. So the clustering algorithm can take the converted two dimensional dataset as input instead of higher dimensional dataset. The working of the dimension reduction technique [11] is explained below: Let $O = o_i, o_2, \ldots, o_n$ be a *d*-dimensional dataset. Now to convert each *d*-dimensional data $o_i \in O$ two dimensional coordinate point (*Xi, Yi)* do the following:.
Calculate X_i and Y_i as

$$X_i = \frac{x_{i0} + x_{i1} + \ldots\ldots x_{id-1}}{d}$$

And

$$Y_i = \frac{y_{i0} + y_{i1} + \ldots\ldots y_{id-1}}{d}$$

For each j^{th} dimensional value of i^{th} data in *O* (i.e., o_{ij}), we can get a co-ordinate point (x_{ij}, y_{ij}). Where $x_{ij} = r_{ij} \cos\theta_j$ and. $y_{ij} = r_{ij} \sin\theta_j$ r_{ij} means the value of o_{ij} (value in j^{th} dimension of i^{th} data). $\theta_j = \theta_{j-1} + 360/d$, and $\theta_0 = 0^0$. In other words for each data $o_i \in O$, $1 \le i \le n$ there must be d numbers of coordinate points (x_{ij}, y_{ij}) , $1 \le i \le n$, $1 \le j \le d$ and with help of these coordinate point (x_{ij}, y_{ij}) we can get the mean value (X_i, Y_i). Plot all the n numbers of mean points on the two dimensional plane and then apply clustering algorithm on the plotted mean points.

IV. OUR PROPOSED TECHNIQUE (RWFCM-HD)

The major flaw of using multi dimensional datasets in clustering algorithms is already explained in section I. A dimension reduction algorithm is provided in [11] to overcome such difficulties. But if the dataset has streaming behavior then even after converting it into a smaller dimensional dataset but the problem still remains consistent[8], [11]. In section I, we have provided various disadvantages of applying FCM procedure on the input dataset of large size or having streaming behavior. We work by combining both dimension reduction algorithm[11] and sWFCM technique[8] together to propose a better fuzzy clustering algorithm for large size high dimensional stream datasets. We call our propose algorithm as rWFCM-HD as we used sWFCM(for our procedure we will write rWFCM instead of sWFCM) and a dimension reduction technique(HD) for higher dimensional

streaming dataset but instead of using letter "s" we will use the letter" r" in the beginning of our used abbreviation for our proposed algorithm. Our algorithm is discussed as follows:

***Algorithm**:* rWFCM-HD

***Provided Input**:* Large Dataset *O* which is High dimensional (*d*-dimensional) in nature with streaming behavior.

1) In this first we convert the *d*-dimensional dataset *O* into two dimensional dataset *X* using the dimension reduction technique discussed in section III-C. Procedure discussed in [11].

2) Then after we will apply sWFCM procedure on the transformed two dimensional dataset *X*. The sWFCM algorithm is discussed in section III-B also studied in[8].

As we know that, the input dataset *O* has streaming behavior it is not possible to reduce the dimension of the entire dataset at a time. But it will not generate any problem because sWFCM algorithm uses a chunk of data from dataset at a time. We can easily understand from section III-B that prior to applying sWFCM; we need to divide the dataset into number of data chunks. Main reason for this is because in real scenario these data are streaming in nature and will not be feed or kept into main (Physical) memory as a whole. So because of this reason firstly the data chunks will be formed and then after the dimension reduction technique will be applied on these chunks individually but not all together.

V. EXPERIMENTAL ANALYSIS

The result analysis is done by taking multi dimensional datasets as input and afterwards converting them into two dimensional data space as discussed in section III-C. Then after reducing the dimensions of the provided input dataset we run sWFCM procedure on it. Though sWFCM is already available we used it in our algorithm for clustering higher dimensional data after reducing their dimension. Experiment results shows that sWFCM algorithm performs better than FCM algorithm for multi dimensional dataset having streaming behavior. Our main focus here is to show that if we combine the procedures provided or simply proposed in [11] and [8] together for a clustering algorithm then performance will get improve much in comparison to the performance of any individual procedure. Note that, our proposed algorithm (rWFCM-HD) is a combination of the techniques proposed in [11] and [8] (see section IV). We use FCM procedure on the reduced (2D) dataset as baseline algorithm. For the experiments analysis of both the procedures we use three higher dimensional large size dataset: KDDCUP 1999, Nursery and Letter recognition having different attributes .All three datasets are available in http://archive.ics.uci.edu/ml/datasets.html. also the used KDDCUP 1999 is a very large dataset we used the first 5000 data from it.

A. Cluster Validity

We adopt validity functions [8] to compare cluster efficiency. The validity functions are based on partition coefficient and partition entropy of *U*.

Partition coefficient for FCM

$$V_{pc}(U) = \frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{2}\right)$$

Partition coefficient for sWFCM

$$V_{pc}(U) = \frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} w_i u_{ij}^{2}\right)$$

Partition entropy for FCM

$$V_{pe}(U) = -\frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\log u_{ij}\right)$$

Partition entropy for sWFCM

$$V_{pe}(U) = -\frac{1}{n}\left(\sum_{j=1}^{n}\sum_{i=1}^{c} w_i u_{ij}\log u_{ij}\right)$$

Where n is the total number of data in the dataset, w_i , u_{ij} ,U are weight of centroids and membership matrix respectively (see section III for details.)

Clusters	Partition Coefficient		Partition Entropy	
	Baseline	Proposed	Baseline	Proposed
4	0.7213	0.8836	44.5717	35.6725
6	0.6102	0.7629	72.8224	54.3179
8	0.5461	0.6992	89.2960	67.5357
10	0.5060	0.6434	103.9362	80.5267

Table I
CLUSTER VALIDITY BASED ON NURSERY DATASET

.

Clusters	Partition Coefficient		Partition Entropy	
	Baseline	Proposed	Baseline	Proposed
4	0.9070	1.1106	17.5581	10.3254
6	0.8176	1.0790	34.2595	15.8415
8	0.7527	1.0243	47.3129	23.4805
10	0.7518	1.0501	50.0947	21.2085

Table II
CLUSTER VALIDITY BASED ON KDDCUP 1999 DATASET

Table I, II and III shows cluster validity in terms of partition coefficient and partition entropy for the three datasets: nursery, KDDCUP 1999 and letter recognition respectively.

Clusters	Partition Coefficient		Partition Entropy	
	Baseline	Proposed	Baseline	Proposed
4	0.5578	0.6654	82.7003	68.1335
6	0.4878	0.5895	106.1772	86.6471
8	0.4491	0.5403	121.7372	101.3327
10	0.4177	0.4922	135.2946	115.0537

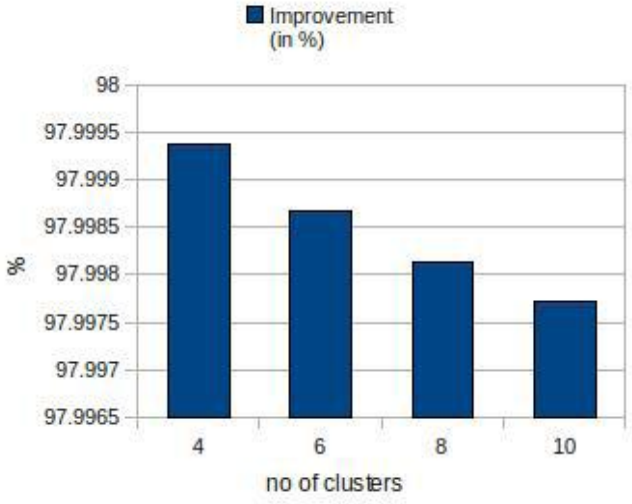
Table III
CLUSTER VALIDITY BASED ON LETTER RECOGNITION DATASET

B. Memory Used

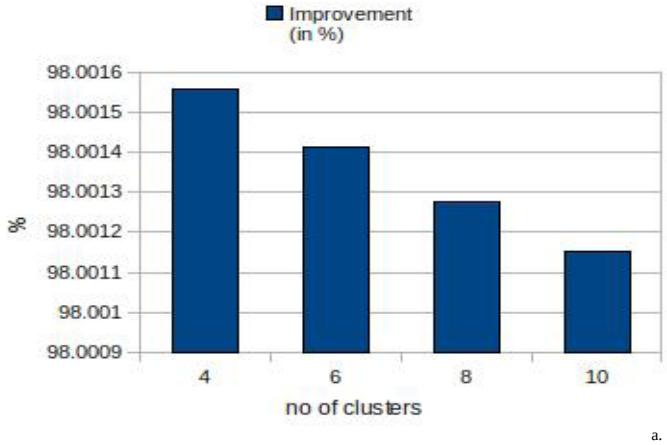
Since sWFCM procedure will process data as with the number of chunks we calculated the memory consumption of each data chunk individually and then after take the largest value as the final memory consumption for our proposed rWFCM-HD. Since the dataset is of continuously arriving streaming behavior, it is not required for rWFCM-HD to work with more than one chunk at a time. Figure 1 in the study shows the percentage of enhancements in terms of memory consumption by proposed (rWFCM-HD) as compared to the baseline algorithm. The enhancement is more than 97% for all three input sample datasets. The major requirement with Baseline Algorithm (FCM) is that it uses entire dataset at a time and hence it requires enough memory to hold the complete dataset. So because of this reason baseline need much higher memory than our proposed algorithm (rWFCM-HD).

C. Execution Time

Similar as memory consumption comparison we have also included calculation for execution time required for each chunk isolately and then after takes the highest parameter value as the final or major execution time for our proposed algorithm. Our main aim here is to calculate the execution time of proposed and baseline algorithm rWFCM-HD procedure will only process single chunk at a time and there is no time bound as when the next chunk will arrive. Figure 2 in the chart presents percentage of enhancement with rWFCM-HD as compared to baseline procedure in terms of execution time. The huge enhancement shown is possible because we compare the execution time of baseline with the largest execution time by a individual chunk in rWFCM-HD. The total execution time (adding the execution time of all the chunks individually) is also less than baseline but we have not shown it here in our study.

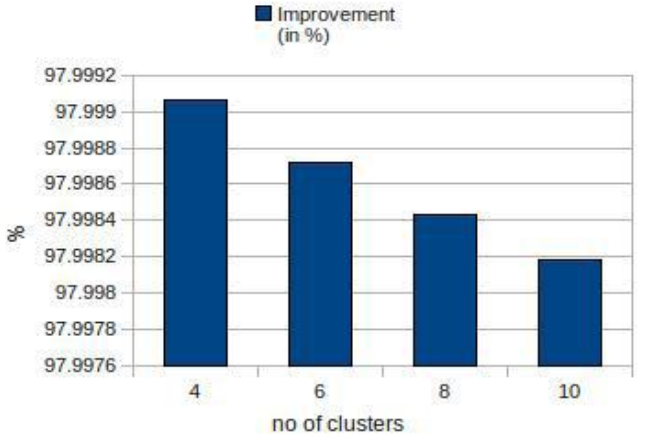


(a) Nursery Data Set



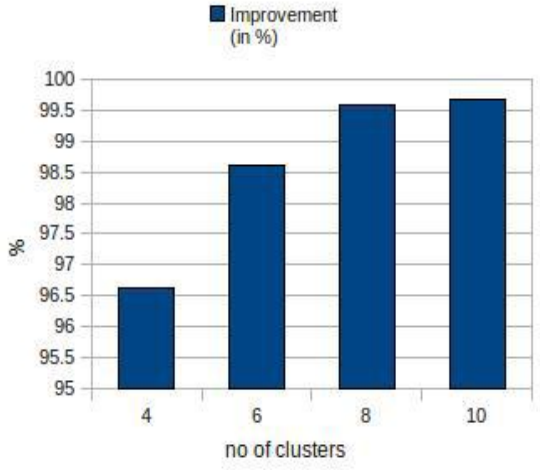
a.

(b) KDD Cup 1999 Data Set

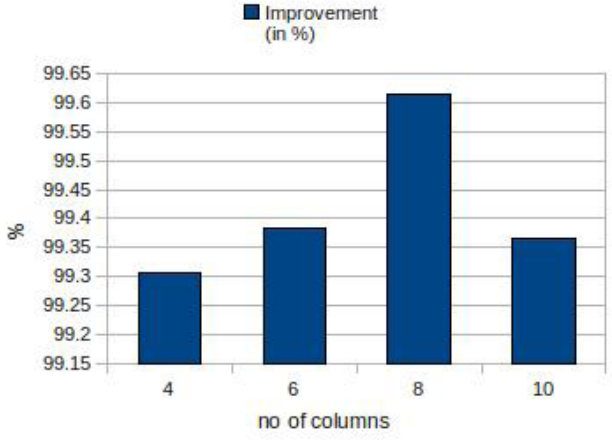


(c) Letter Recognition Data Set.

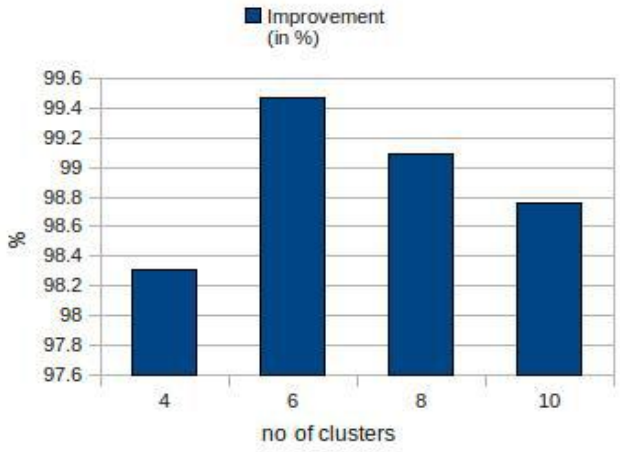
Figure 1. Percentage improvement for memory consumption in proposed rWFCM-HD over baseline (FCM)**.**



(A) NURSERY DATA SET



(B) KDD CUP 1999 DATA SET.



(C) Letter Recognition Data Set

Figure 2. Percentage improvement for execution time in Proposed rWFCM-HD over baseline (FCM).

D. Simple Description

There are two parameters namely partition coefficient and partition entropy which have been calculated on all the three data sets i.e. on Nursery dataset, KDDCUP 1999 dataset and letter recognition dataset. Each time we will calculate the results on these datasets:

A) By calculating the partition coefficient (PC) and partition entropy (PE), on the three data sets with BASELINE algorithm using 4, 6,8,10 no. of clusters.
(i)With Nursery dataset:
Total outcomes for PC with baseline algorithm are: 4
Total outcomes for PE with baseline algorithm are: 4
(ii)With KDDCUP 1999 dataset are:
Total outcomes for PC with baseline algorithm are: 4
Total outcomes for PE with baseline algorithm are: 4
(iii) With Letter recognition dataset 1999 dataset are:
Total outcomes for PC with baseline algorithm are: 4
Total outcomes for PE with baseline algorithm are: 4
Total number of outputs with Base algorithm in our Case is: 24

B) By calculating the partition coefficient (PC) and partition Entropy (PE) on the three data sets with PROPOSED algorithm using 4, 6,8,10 no. of clusters.
(i)With Nursery dataset:
Total outcomes for PC with proposed algorithm are: 4
Total outcomes for PE with proposed algorithm are: 4
(ii)With KDDCUP 1999 dataset are:
Total outcomes for PC with proposed algorithm are: 4
Total outcomes for PE with proposed algorithm are: 4
(iii)With Letter recognition dataset 1999 dataset are:
Total outcomes for PC with proposed algorithm are: 4
Total outcomes for PE with proposed algorithm are: 4
Total number of outputs with proposed algorithm in Our Case is: 24

We have done our result analysis with total 48 outputs

VI. CONCLUSION

To mine the data from the data streams is very difficult because of the limited amount of memory availability and real time query response requirement. The major task to perform mining on any input data is through clustering. On the other hand, high-dimensional data poses different problem (challenges) for clustering algorithms that require specialized solutions. In high dimensional data, for clustering traditional similarity measures as which used in conventional clustering algorithms are usually not meaningful. In this research we have provided a dimension reduced weighted fuzzy clustering procedure with the abbreviation (rWFCM-HD). The procedure

can be highly utilized for multi dimensional datasets having continuously arriving behavior. Such source of data are obtained from sensor networks, data generated by web click stream and output data stream from internet traffic data transfer rate etc, this category of data's have two special attributes which isolates them from other datasets: a) Firstly they have streaming behavior and b) Secondly they have multi dimensions. Minimized fuzzy clustering algorithm has been already provided for incoming bunch of data having streaming behavior or multi dimensions. But as per our survey yet, nobody has proposed any minimized fuzzy clustering for data sets having both the functionalities, i.e., data sets with multiple dimensions and also continuously arriving streaming behavior. Result analysis shows that our provided procedure (rWFCM-HD) improves efficiency from the prospective of memory requirement as well as in required execution time.

.

REFERENCES

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream Systems," in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, ser. PODS '02, 2002, pp. 1–16.

[2] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani, "Streaming-data algorithms for high-quality Clustering," in Data Engineering, 2002. Proceedings. 18thInternational Conference on, 2002, pp. 685–694.

[3] P. Domingos and G. Hulten, "A general method for scaling up machine learning algorithms and its Application to Clustering," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01, 2001, pp. 106–113.

[4] P. Rodrigues, J. Gama, and J. Pedroso, "Hierarchical clustering of time-series data streams," Knowledge and Data Engineering, IEEE Transactions on, vol. 20, no. 5, pp. 615– 627, 2008.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "A framework for clustering evolving data streams," in Proceedings of the 29th international conference on Very large data bases - Volume 29, ser. VLDB '03, 2003, pp. 81–92.

[6] S. Eschrich, J. Ke, L. Hall, and D. Goldgof, "Fast fuzzy clustering of infrared images," in IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, vol. 2, 2001, pp. 1145–1150 vol.2.

[7] M. B. Al-Zoubi, A. Hudaib, and B. Al-Shboul, "A fast fuzzy clustering algorithm," in Proceedings of the 6thConference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases - Volume 6, ser. AIKED'07, 2007, pp. 28–32.

[8] R. Wan, X. Yan, and X. Su, "A weighted fuzzy clustering algorithm for data stream," in Proceedings Of the 2008 ISECS International Colloquium on Computing, Communication, Control, and Management - Volume 01, ser. CCCM '08, 2008,pp. 360–364.

[9] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proceedings of the 8th International Conference on Database Theory, ser. ICDT '01, 2001, pp. 420– 434.

[10] H.-P. Kriegel, P. Kr¨oger, and A. Zimek, "Clustering high dimensional data: A survey on subspace Clustering, patternbased clustering, and correlation clustering,"ACMTransKnowl. Discov. Data, vol. 3, no. 1, pp. 1:1–1:58, Mar. 2009.

[11] P. Bishnu and V. Bhattacherjee, "A dimension reduction technique for k-means clustering algorithm," in Recent Advances in Information Technology (RAIT), 2012 1st International Conference on, 2012, pp. 531–535.

[12] N. R. Pal and J. C. Bezdek, "Complexity reduction for "large image" processing," Trans. Sys. Man Cyber. Part B, vol. 32, no. 5, Oct. 2002.

[13] D. Altman, "Efficient fuzzy clustering of multi-spectral images," in Geoscience and Remote Sensing Symposium,1999. IGARSS '99 Proceedings. IEEE 1999 International, vol. 3, 1999, pp. 1594–1596 vol.3.

[14] I. Fodor. (2002) A Survey of Dimension Reduction Techniques.[Online].Available:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5098

[15] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding,"SCIENCE, vol.290, pp. 2323–2326, 2000.

[16] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, no. 5500, pp. 2319–2323, 2000.

[17] L. Teng, H. Li, X. Fu, W. Chen, and I.-F. Shen, "Dimension reduction of microarray data based on Local tangent space alignment," in Proceedings of the Fourth IEEE International Conference on Cognitive Informatics, ser. ICCI '05, 2005, pp. 154–159.

[18] R. Urtasun and T. Darrell, "Discriminative gaussian process latent variable model for classification,"In Proceedings of the 24th international conference on Machine learning, ser. ICML '07, 2007, pp. 927–934.

[19] L. Tan and Y. Zhang, "A comparative study of dimension reduction based on data distribution," in Intelligent Systems (GCIS), 2010 Second WRI Global Congress on, vol. 3, 2010, pp. 309–312.

[20]Diksha Upadhyay, Susheel Jain, Anurag Jain "Comparative Analysis of Various Data Stream Mining Procedures and Various Dimension Reduction Techniques" International journal of Advanced Research in computer science "Volume 4, No.8, May-June 2013.