Particle of Swarm Based Automatic Variable Weighting Clustering Algorithm for Large Database

Abhijeet Dinkar Cholke[1] Vasimraj Siraj Tamboli[2] Ravindra Sundarlal Kakade[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Padmashri Dr.V.V.Patil Instt. of Tech. & Engg. (Polytechnic) Pravaranagar

Abstract— The clustering techniques play an important role in data mining process. For the mining of large data faced a lot of problem of noise and large number of iteration. The process of pattern generation used two type of technique such as supervised learning and unsupervised learning. In unsupervised learning clustering process is used. The varieties of clustering technique are used such as k-means, FCM and weighted clustering technique. The weighted clustering technique gives the two solution approach one is seed selection and another is mapping of seed in terms of weight of centre. In this dissertation modified the seed selection process using particle of swarm optimization technique. The particle of swarm optimization process select variable value one is seed value and another is weight of centre value. In weighted cluster techniques used some value of centre and generate new centre value of new cluster for the better generation of cluster. For more improvement of weighted clustering technique used two level weighted clustering techniques for better improvement of cluster technique. For the validation of clustering technique used various methods such as weighted clustering technique and multi-level weighted clustering technique. In the last decade, several approaches able to notice several clustering resolutions have been presented. According to the review, they can briefly be characterized into approaches effective on the novel data-space, methods performing space transformations, and methods analysing subspace projections. The main conception is to reflect every subspace as a state in a fuzzy inference, with rule allowed.

Key words: Clustering, K-means, FCM, Data-space, Fuzzy

I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters essentially loses certain fine elements, but attains simplification. It models data by its clusters. Data modelling places clustering in a historical perception engrained in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters resemble to hidden patterns, the exploration for clusters is unsupervised learning, and the consequential system signifies a data concept. From a practical perspective clustering plays an due role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology. These methods can be used to determine associations and structures within a data set that may not have been recognized. Cluster analysis has been broadly used in the biological and social sciences to help define classification schemes or taxonomies. It has also been used to propose new customs of relating a population in business and marketing applications. Cluster analysis techniques can be largely separated into two methods, hierarchical and non-hierarchical. The hierarchical approach forms clusters of consecutively larger size using some amount of resemblance or distance. Typical algorithms used in this technique include single linkage method (nearest neighbour), complete linkage method (furthest neighbour), and Ward's Method, which reduces the mean square distance among the centre of a cluster and each member. Non-hierarchical clustering approaches also occur, including the K-means technique.

A. The K-Means Algorithm:

One of the most popular clustering approaches is k-means clustering algorithm. It generates k points as initial centroids randomly, where k is a user itemized parameter. Each point is then allocated to the cluster with the neighbouring centroid. Then the centroid of each cluster is restructured by taking the mean of the data points of each cluster. Some data points may transfer from one cluster to other cluster. Again new centroids are computed and allocate the data points to the suitable clusters. The assignment is repeated and update the centroids, until convergence standards is met i.e., no point changes clusters, or equivalently, until the centroids continue the same. In this algorithm mostly Euclidean distance is used to discover distance between data points and centroids. The Euclidean distance between two multi-dimensional data points X = (x1, x2, x3... xm) and Y = (y1, y2, y3... ym) is described as follows:

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\ldots+(x_m-y_m)^2}$$

The superiority of the final clustering results of the k-means algorithm greatly depends on the arbitrary selection of the initial centroids. In the original k-means algorithm, the initial centroids are selected arbitrarily and hence different clusters are found for different runs for the same input data. It consists of two distinct phases: First phase is to decide k centers at random one for each cluster. Resulting phase is to determine distance between data points in Dataset and the cluster centers and assigning the data point to its nearest cluster. Euclidean distance is generally considered to determine the distance. When all the data points are included in some clusters an initial grouping is done. New centers are then designed by taking the average of points in the clusters. This is done because of insertion of new points may lead to modification in cluster centers. This process of center modifying is repeated till a condition where centers do not update any longer or criterion function becomes minimum. This signifies the convergence criteria. The fuzzy k-modes clustering algorithm begins with an initial set of cluster prototypes and uses the alternating minimization method to solve a non-convex optimization problem in finding cluster solutions. However, in the clustering process, the update formulas of the membership matrix and cluster prototypes are based on the within-cluster information only, i.e., the within-cluster compactness. The between-cluster information, i.e., the between-cluster separation, is not considered, which often results in the clustering results with weak between-cluster separation.

All rights reserved by *www.ijsrd.com*

II. CLUSTERING TECHNIQUE

Data Mining is defined as mining of knowledge from huge amount of data. Using Data mining we can guess the nature and performance of any kind of data. The past two decades has seen an strong increase in the amount of information being warehoused in the electronic format. This assembly of data has occupied place at an unstable rate. It was recognized that information is at the heart of the business developments and that decision makers could make the practice of data kept to gain the valued vision into the business. DBMS gave admittance to the data warehoused but this was only small part of what could be added from the data. Analysing data can further provide the knowledge about the business by going beyond the data explicitly stored to derive knowledge about the business. Learning valuable information from the data made clustering techniques widely applied to the areas of artificial intelligence, customer – relationship management, data compression, data mining, image processing, machine learning, pattern recognition, market analysis, and fraud – detection and so on. Cluster Analysis of a data is an important task in Knowledge Discovery and Data Mining. Clustering is the process to group the data on the basis of similarities and dissimilarities among the data elements. Clustering is the process of finding the group of objects such that object in one group will be similar to one another and different from the objects in the other group. A good clustering method will produce high quality clusters with high intra cluster distance similarity and low inter cluster distance similarity.

A. Data Mining Methods:

There are several major data mining techniques have been developed and used in data mining:

- Association
- Classification
- Clustering
- Prediction
- Sequential Patterns

Text mining can be broadly defined as a knowledge-intensive process in which a user interacts with a document collection over time by using a suite of analysis tools. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the identification and exploration of interesting patterns. In the case of text mining, however, the data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections. Clustering is the process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. It helps users to understand the natural grouping or structure in a dataset. A good clustering method will produce high quality clusters in which the intra-class (i.e., intra-clusters) similarity is high and the inter-class similarity is low. The quality of clustering result also depends on both 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 the entire hidden pattern.

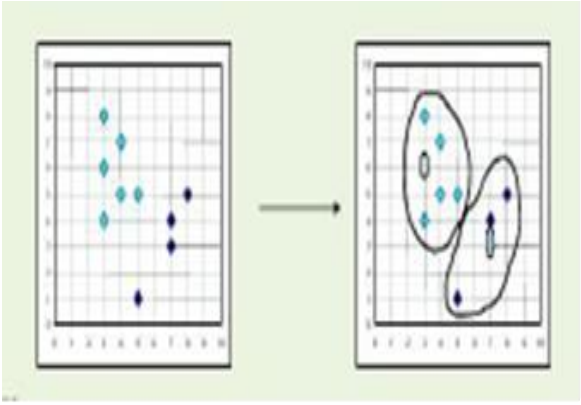

Fig. 1: The result of cluster analysis.

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.

B. Fuzzy C-Means Clustering:

A splitting method first generates an initial set of k partitions, where parameter k is the preferred number of clusters as output. It then uses an iterative repositioning technique that endeavours to recover the partitioning of the data points. K-Means is a numerical, unsupervised iterative method. It is Unique and very fast, so in numerous practical applications this technique is evidenced to be very effective technique that can produce decent clustering results. But the computational complexity of original k- Means is very high, particularly for large datasets. Furthermore, this algorithm results in changed type of clusters reliant upon the arbitrary selection of initial clusters. Fuzzy clustering lets every feature vector to fit in to more than one cluster with different association degrees (between 0 and 1) and imprecise or fuzzy boundaries between clusters. Fuzzy C-Means (FCM) is a method of clustering which permits one piece of data to belong to two or more clusters. This method is generally recycled in pattern recognition.

C. Pattern Based Clustering:

Pattern based clustering, also known as bi-clustering, is originally used in analysis of microarray gene expression data (Cheng and Church 2000). In a 2D microarray data-set, the genes are the objects and the samples are the attributes. Similar to subspace clustering, pattern based clustering mines clusters where a cluster is a set of objects that are homogeneous in a set of attributes, and overlapping of clusters is allowed. However, there are two subtle differences in these two types of clusters. First, the sub matrix defined by the objects (rows) and attributes (columns) of a pattern based cluster exhibits a pattern. Second, the objects and attributes of a pattern based cluster are treated equally, but this is not the case for a subspace cluster. This equal treatment of pattern based cluster encourages more flexibility in its homogeneity; its homogeneity can be on the attributes, on the objects, or on both attributes and objects. Homogeneity on the attributes means that the objects are homogeneous in each attribute, e.g. the objects have similar values in each attribute, as shown in the pattern based cluster of Figure 2.a. Homogeneity on the objects means that each object is homogeneous in the attributes, e.g. the attributes have similar values for each

All rights reserved by *www.ijsrd.com*

object, as shown in the pattern based cluster of Figure 2.b. On homogeneity on both objects and attributes, a simple example is where all values in the cluster have similar values. A more common type of this homogeneity is shifting or scaling of values across attributes and objects. Figure 2.c shows a pattern based cluster of shifting homogeneity on both attributes and objects.

	a_1	a_2	a_3
o_1	1	5	3
o_2	1	5	3
o_3	1	5	3

(a)

	a_1	a_2	a_3
o_1	1	1	1
o_2	5	5	5
o_3	3	3	3

(b)

	a_1	a_2	a_3
o_1	1	2	3
o_2	4	5	6
o_3	7	8	9

(c)

Fig. 2: 2.a a pattern based cluster with homogeneity on the attributes, 2.b a pattern based cluster with homogeneity on the objects, and 2.c a pattern based cluster with homogeneity on both attributes and objects.

D. Hierarchical Clustering:

Hierarchical clustering creates a cluster hierarchy or, in other words, a tree of clusters, also known as a dendro gram. Every cluster node comprises child clusters; sibling clusters partition the points enclosed by their mutual parent. Such a method permits discovering data on different levels of granularity. Hierarchical clustering methods are classified into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more maximum suitable clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster. The process lasts till a discontinuing standard (frequently, the requested number k of clusters) is achieved. Advantages of hierarchical clustering include.

– Embedded flexibility regarding the level of granularity.
– Ease of handling of any forms of similarity or distance.
– Consequently, applicability to any attributes types.

The classic approaches to hierarchical clustering are presented in the sub-section Linkage Metrics. Hierarchical clustering based on linkage metrics results in clusters of proper (convex) shapes. Active contemporary efforts to build cluster systems that incorporate our intuitive concept of clusters as connected components of arbitrary shape, including the algorithms CURE and CHAMELEON, are surveyed in the sub-section Hierarchical Clusters of Arbitrary Shapes. Divisive techniques based on binary taxonomies are presented in the sub-section Binary Divisive Partitioning. The sub-section Other Developments contains information related to incremental learning, model-based clustering, and cluster refinement.

III. LITERATURE SURVEY

A. Cams-Rs: Clustering Algorithm for Large-Scale Mass Spectrometry Data Using Restricted Search Space And Intelligent Random Sampling:

In this paper, author present an efficient algorithm, CAMS-RS (Clustering Algorithm for Mass Spectra using Restricted Space and Sampling) for clustering of raw mass spectrometry data. CAMS-RS utilizes a novel metric (called F-set) that exploits the temporal and spatial patterns to accurately assess similarity between two given spectra. The F-set similarity metric is independent of the retention time and allows clustering of mass spectrometry data from independent LC-MS/MS runs. A novel restricted search space strategy is devised to limit the comparisons of the number of spectra. An intelligent sampling method is executed on individual bins that allow merging of the results to make the final clusters. Our experiments, using experimentally generated data sets, show that the proposed algorithm is able to cluster spectra with high accuracy and is helpful in interpreting low S/N ratio spectra. The CAMS-RS algorithm is highly scalable with increasing number of spectra and our implementation allows clustering of up to a million spectra within minutes. We presented an efficient algorithm, called CAMS-RS, suitable for clustering of large-scale mass spectrometry data. A novel similarity metric (called F-set) is formulated and used in the algorithm based on the spatial locations and intensity of the peaks in a spectrum. The independence of the similarity metric from retention time allows multiple sets of independently run LC-MS/MS experimental data sets to be clustered together. A graph-theoretic framework is introduced that allows the use of the introduced F-set metric in an efficient manner. In order to make the algorithm scalable with increasing number of spectra we introduce two different kind of search space restrictions. Each of the search space restriction dramatically decreased the search space for the algorithm while giving accurate clustering results.

B. A Preliminary Survey On Optimized Multi-Objective Meta-Heuristic Methods For Data Clustering Using Evolutionary Approaches:

MOEAs have substantial success across a variety of MOP applications, from pedagogical multifunction optimization to real-world engineering design. The survey paper noticeably organizes the developments witnessed in the past three decades for EAs based meta heuristics to solve multi objective optimization problems (MOP) and to derive significant progression in ruling high quality elucidations in a single run. Data clustering is an exigent task, whose intricacy is caused by a lack of unique and precise definition of a cluster. The discrete optimization problem uses the cluster space to derive a solution for Multi objective data clustering. Discovery of a majority or all of the clusters (of illogical shapes) present in the data is a long-standing goal of unsupervised predictive learning problems or exploratory pattern analysis. An imperative surveillance was success of most MOEAs depends on the careful balance of two conflicting goals, exploration (searching new Pareto Optimal Solution) and exploitation (refining the obtained Pareto Solutions). EAs are easy to portray and execute, but rigid to analyze hypothetically. In spite of much experiential acquaintance and successful application, only little theoretical fallout pertaining to their effectiveness and competence are available.

C. Multi-Cluster Based Approach For Skewed Data In Data Mining:

The class imbalance problem define as the sample of one class may be much less number than another class in data set. There are many technology developed for handling class imbalance. Basically designed approaches are divided into two types. First is designed a new algorithm which improves the minority class prediction, second modify the number

All rights reserved by www.ijsrd.com

samples in existing class, it also known as data pre-processing. Under-sampling is a very popular data pre-processing approach to deal with class imbalance problem. Under-sampling approach is very efficient, it only use the subset of the majority class. The drawback of under-sampling is that it removes away many useful majority class samples. To solve this problem we propose multi cluster-based majority under-sampling and random minority oversampling approach. Compared to under-sampling, cluster-based random under-sampling can effectively avoid the important information loss of majority class. Performance of classifier increase when preprocess data by MCMUS algorithm before applying classifier. K-means clustering algorithm used for clustering majority class samples in to k clusters. In this project k=3 used. Once data is clustered two methods are used to select the data samples from each clustered. These selected samples then combine with minority class sample and new training dataset will generate. The size of new training samples is small but helpful to classify the imbalance dataset. Comparison between SVM and KNN classifier demonstrates that performance of KNN classifier with MCMUS algorithm is better than SVM classifier. Although SVM classifier has good theoretical foundation in classification, performance will degrades as the class imbalance ratio increased.

IV. METHODOLOGY AND ARCHITECTURE

A. Feature Extraction Process:

In general, a feature extraction procedure consists of four steps: smoothing, baseline removal, peak detection, and peak quantification. Note that some steps such as smoothing and baseline removal may switch their locations in the pipeline [4]. Furthermore, it is also possible that some steps are merged or some additional steps such as calibration are included. Feature extraction is the most important step and has a great impact on the accuracy of biomarker identification. This is because all subsequent analysis steps have to utilize the output of feature extraction as input. In other words, this is probably the only opportunity to recover real signals from noisy raw MS data in the whole analysis flow.

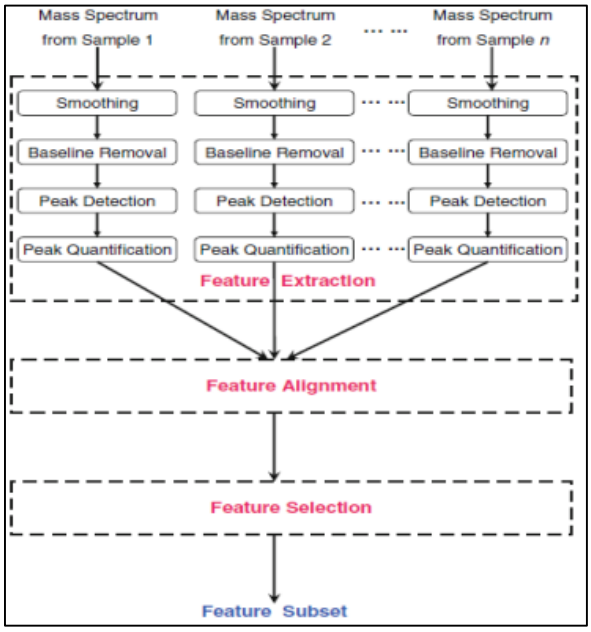

Fig. 3: Feature extraction and selection process.

The selection of seed and weight value in soft clustering technique play a major role for better generation of pattern and cluster. The soft clustering technique basically is a combination of K-means algorithm and fuzzy logic. The process of fuzzy logic gives a better selection of seed value of cluster. In consequence of cluster technique improvement for the large data used weighted seed centre cluster technique. In weighted cluster technique used some value of centre and generates new centre value of new cluster for the better generation of cluster. For more improvement of weighted clustering technique used two level weighted clustering techniques for better improvement of cluster technique. In this dissertation modified the weighted clustering technique for improvement. In the process of improvement used particle of swarm optimization technique. Particle of swarm optimization technique gives the better selection of seed for large database. In the continuity of chapter discuss the K-means algorithm, FCM algorithm, weighted clustering technique, particle of swarm optimization, modified algorithm and finally discuss proposed model.

B. K-Means Algorithm:

Partition based clustering technique is also called mean clustering technique. The mean based clustering process calculate the value of mean and put the number of cluster and done pattern analysis. Partitioning clustering attempts to decompose a set of N objects into k clusters such that the partitions optimize a certain criterion function[5]. Each cluster is represented by the centre of gravity (or centroid) of the cluster, e.g. k-means, or by the closest instance to the gravity centre (or medoid), e.g. k-medoids. Typically, k seeds are randomly selected and then a relocation scheme iteratively reassigns points between clusters to optimize the clustering criterion. The minimization of the square-error criterion - sum of squared Euclidean distances of points from their closest cluster centroid, is the most commonly used. A serious drawback of partitioning algorithms is that there are a number of possible solutions. In particular, the number of all possible partitions P (n, k) that can be derived by partitioning n patterns into k clusters is:

$$P(n,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-1}\binom{k}{i}(i)^{n}$$

An example is shows that having to partition n = 10 patterns into k = 4 clusters the total number of different partitions is P (10, 4) = 34105. However, for n = 19 and k = 4, P becomes huge, approximately 11,259,666,000.

Some representative examples of partitioning methods are:

K-Means

K-means is perhaps the most popular clustering method in metric spaces. Initially k cluster centroids are selected at random; k-means then reassigns all the points to their nearest centroids and recomputed centroids of the newly assembled groups [6]. The iterative relocation continues until the criterion function, e.g. square-error converges. Despite its wide popularity, k-means is very sensitive to noise and outliers since a small number of such data can substantially influence the centroids. Other weaknesses are sensitivity to initialization, entrapments into local optima, poor cluster descriptors, and inability to deal with clusters of arbitrary shape, size and density, reliance on user to specify the number of clusters.

All rights reserved by *www.ijsrd.com*

Simply put, k-Means Clustering is an algorithm among several that attempts to find groups in the data. In pseudo code, it is shown to follow this procedure:

- Initialize mi, i = 1,…,k, for example, to k random xt
- Repeat
- For all xt in X
- bit ← 1 if || xt - mi || = minj || xt - mj ||
- bit ← 0 otherwise
- For all mi, i = 1,…,kmi ← sum over t (bit xt) / sum over t (bit)Until mi converge

The vector m contains a reference to the sample mean of each cluster. x refers to each of our examples, and b contains.

1) Choose some manner in which to initialize the mi to be the mean of each group (or cluster), and do it.
2) For each example in your set, assign it to the closest group (represented by mi).
3) For each mi, recalculate it based on the examples that are currently assigned to it.
4) Repeat steps 2-3 until mi converge..

Now that we have some rudimentary understanding of what k-means is, what are some practical applications of it?

V. FCM ALGORITHM

FCM algorithm is hard on data sets too ,so the data sets must be quite regular, in order to solve problems, first of all we use information entropy to initialize the cluster centers to determine the number of cluster centers. It can be reduce some errors, and also can improve the algorithm introductions a weighting parameters .after that, combine with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering. Document similarity measures.

The algorithm as follows

Get the class prior prabablities {Pr }cC=1 Set the class growth rate nc=n X Prc Where c=1,…..C

If H(0) I not given then

Construct an initial clusters of N

For t initialize Cj (cluster centers) Initialize □(threshold value)

Repeat

For i=1 to n :update μj(Xi)

For k=1 to p ;

Sum=0

Count=0

For i=1 to n;

f μ (Xi) is maximum in Ck then

If μ (Xi)>= α Sum=sum+Xi Count= count+1

Ck=sum/count Until Cj estimate stabilize.

The clustering framing as follows Set value for cluster numbers algorithm stop threshold ε ≥ 0,

A set clusters C= {C1, C2, C3,…….Ck}

A. Tw-K-Means Clustering Algorithm:

The clustering process to partition X into k cluster with weights for both views and individual variables is modeled as minimization of the following objective function

$$p(U,Z,V,W)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in Gi}^{1} uiwtvjd(xij,zij) + n\sum_{j=1}^{m} vjlog(vj)+\lambda\sum_{t=1}^{T} wtlog(wt)\ldots\ldots(1)$$

Subject to

$$\begin{cases}\sum_{l=1}^{k} ui.l=1, ui,l\in(0,1), 1\le i\le n\\ \sum_{t=1}^{T} wt=1, 0\le wt\le 1, \quad \ldots..(2)\\ \sum_{j\in Gi} vj=1, 0\le vj\le 1, 1\le t\le T,\end{cases}$$

Where

U is a n×k portion matrix whose element ui,l are binary where ui,l=1 indicates that object I is allocated to cluster l;

Z={Z1,Z2,……..,Zk} is a set of k vectors representing the centres of the k clusters

W={W1,W2,……..Wt} are T weight for T view

V={v1,v2,……vm} are m weight form variable

d(xij,zlj) is a distance or dissimilarity measure on the jth variable between the ith object and the centre of the lth cluster. if the variable is numerical , then

d(xij ,zlj)= (xij-zlj)2…….(3)

if the variable is categorical, then

d(xij,zlj)={0 (xi.j=zl.j)……..…..(4)
1 (xi.j=/zl.j)

The first term in (1) is the sum of the within cluster dispersions, the second and third terms are two negative weight entropies. two positive parameter are control the strength of cluster.

VI. PARTICLE OF SWARM OPTIMIZATION

Particle of swarm optimization is dynamic population based searching technique. The process of working define in manner of particle of swarm optimization is birds fork. The fork of birds fly in sky and doesn't collide in path and maintain certain distance for the process of operation. In this algorithm find two parameter value one is local best and another is global best. The global best solution is better result in case of optimality. The process of optimization apply on seed parameter and weight parameter for the process of cluster operation. The process of particle of swarm optimization describe as. In Particle Swarm Optimization [10] optimizes an objective function by undertaking a population based search. The population comprise of possible solutions, named particles, which are metaphor of birds in flocks. These particles are at random initialized and freely fly across the multi-dimensional seek space. During flight, each particle updates its own velocity and position based on the best experience of its own and the entire population. The different steps involved in Particle Swarm Optimization Algorithm are as follows:

Step 1: All particles' velocity and position are randomly place to within pre-defined ranges.

Step 2: Velocity update – At every iteration, the velocities of all particles are updated based on below expression

$$v_i = v_i + c_1R_1\left(p_{i,best}-p_i\right)+c_2R_2\left(g_{i,best}-p_i\right) \quad ....(4.3)$$

All rights reserved by *www.ijsrd.com*

where pi is the position and vi are the velocity of particle i, pi,best and gi,best is the position with the 'best' objective value found so far by particle i and the entire population respectively; w is a parameter controlling the dynamics of flying; R1 and R2 are random variables in the range [0,1]; c1 and c2 are factors controlling the related weighting of equivalent terms. The random variables facilitate the PSO with the ability of stochastic searching.

Step 3: Position updating – The positions of all particles are updated according to,

$$p_i = p_i + v_i \qquad ...(4.4)$$

Following updating, pi should be verified and limited to the allowed range.

Step 4: Memory updating – Update pi,best and gi,best when condition is met,

$$p_{i,best} = p_i \quad if\ f(p_i) > f(p_{i,best})$$
$$g_{i,best} = g_i \quad if\ f(g_i) > f(g_{i,best}) \qquad ...(4.5)$$

Where f(x) is to be optimized and it is a objective function.

Step 5: Stopping condition – The algorithm repeats steps 2 to 4 until certain stopping circumstances are met, such as a pre-defined number of iterations. Once closed, the algorithm reports the values of gbest and f(gbest) as its solution[8].

PSO utilizes several searching points and the searching points gradually get close to the global optimal point using its pbest and gbest.

VII. PROPOSED METHODOLOGY

In this section discuss the improved algorithm of weighted k-means algorithm. The weighted k-means algorithm decide the weight value according to their sum of seed and merger the process of cluster selection. The process of seed selection used particle of swarm optimization. The process of particle of swarm optimization gives the better result in concern of weight value and seed value. The seed selection process recall with fitness function. The particle fitness function decide the selection process of seed parameter according to recall value. The fitness constraints parameter decide the selection criteria of weight and center of cluster.

1) Auto = (X,C) ←empty //initialize data and randomly center point
2) C_list ← TWK-means (Ci_list, K_{auto})
3) Input C_list X , the clustering number pn , population scale XN , probability auto P stop conditions cS ;
4) Code the data in real number and initialize population S(i),i = 0 at random;
5) Evaluate the fitness of all individual in the current instant D(s);
6) CR clustering requires optimization of cluster center, which way thrashing of data of waiting cluster. Hence the fitness function of algorithm is determined by f(x).
7) Umpire the termination conditions. $G(s) = \frac{N(s)}{D(s)} = \frac{\sum_{i=0}^{n-1} A_i s^i}{\sum_{i=0}^{n} a_i s^i}$ If the termination situation are satisfied then turn to step 9, if not, turn to step 10;

$$p(U,Z,V,W) = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{t=1}^{T}\sum_{j\in Gi}^{1} uiwtvjd(xij,zij) + n\sum_{j=1}^{m} vjlog(vj) + \lambda\sum_{t=1}^{T} wtlog(wt) \dots\dots(1)$$

8) Subject to

$$\begin{cases} \sum_{l=1}^{k} ui.l = 1, ui,l \in (0,1), 1 \le i \le n \\ \sum_{t=1}^{T} wt = 1, 0 \le wt \le 1, \\ \sum_{j\in Gi} vj = 1, 0 \le vj \le 1, 1 \le t \le T, \end{cases} \qquad \dots..(2)$$

Where

9) U is a n×k portion matrix whose element ui,l are binary where ui,l=1 indicates that object I is allocated to cluster l;
10) Z={Z1,Z2,…….,Zk} is a set of k vectors representing the centers of the k clusters
11) W={W1,W2,……..Wt} are T weight for T view
12) V={v1,v2,……vm} are m weight form variable
13) d(xij,zlj) is a distance or dissimilarity measure on the jth variable between the ith object and the center of the lth cluster. if the variable is numerical , then
14) d(xij ,zlj)=(xij-zlj)2…….(3)
15) if the variable is categorical, then
16) d(xij,zlj)={0 (xi.j=zl.j)……..(4)
17) 1 (xi.j=/zl.j)
18) The first term in (1) is the sum of the within cluster dispersions, the second and third terms are two negative weight entropies. Two positive parameters are control the strength of cluster.

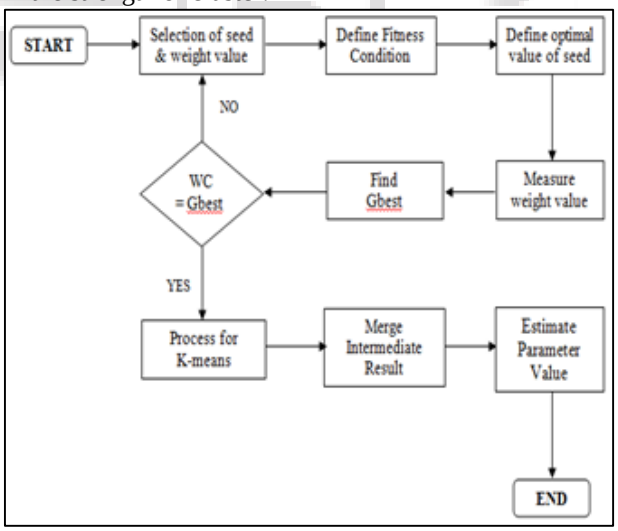

Fig. 4: Proposed model of TW-K-MEANS algorithm

VIII. EXPERIMENTAL RESULT

In this paper we perform experimental process of WKMeans with TW-KMeans. The proposed method implements in matlab 7.14.0 and tested with very reputed data set. I have measured classification accuracy, precision, recall, f-measure and execution time of ensemble method. To evaluate these performance parameters I have used three datasets namely segmentation, water treatment, and Yeast Cell Cycle data set. All the three dataset are large datasets.

All rights reserved by *www.ijsrd.com*

IX. EXPERIMENTAL ANALYSIS

Here we created a dataset which includes segmentation, water treatment, and Yeast Cell Cycle.

Input Value	Method Name	Precision (%)	Recall (%)
2	WK Means	87.000000	86.500000
	TW-K-Means	89.500000	88.170000
	PROPOSED METHOD	93.000000	92.810000
3	WK Means	84.000000	86.650000
	TW-K-Means	86.500000	87.620000
	PROPOSED METHOD	90.000000	92.800000
4	WK Means	83.000000	86.020000
	TW-K-Means	85.500000	87.160000
	PROPOSED METHOD	89.000000	92.610000
5	WK Means	82.000000	86.260000
	TW-K-Means	84.500000	87.160000
	PROPOSED METHOD	88.000000	92.960000

Table 1: Shows the comparative Precision and Recall of Dataset segmentation.

All the three dataset are large datasets. The result analysis of image clustering based on number of class of image based on three methods namely WK-Means, TW-KMeans & Proposed method. These three methods are applied on three different data sets.

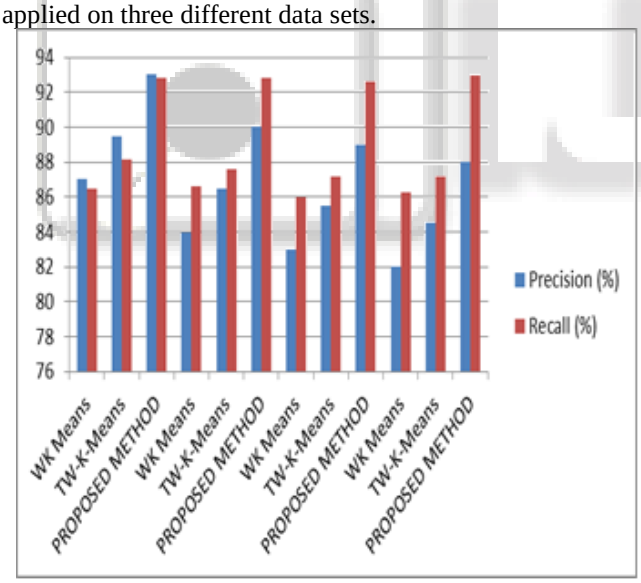

Fig. 5: Shows that performance of data set Segmentation counts of data and rate of precision and recall with input value 2,3,4,5.

X. CONCLUSION

In this dissertation modified the weighted clustering technique using particle of swarm optimization. The particle of swarm optimization used for the selection of seed and weight value. The optimal selection of seed and weight value increase the accuracy of cluster technique. The cluster technique imposed the two process for the selection of seed and weight parameter. Proposed weighting clustering algorithm for clustering of large data, Proposed can compute weights for views and individual variables simultaneously in the clustering process. With the two types of weights, compact views and important variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, Proposed can obtain better clustering results than individual variable weighting clustering algorithms from large data. We used two real-life data sets to investigate the properties of two types of weights in Proposed. We discussed the difference of the weights between Proposed and TW-means algorithms. The experiments also revealed the convergence property of the view weights in Proposed. We compared Proposed with three clustering algorithms on three real-life data sets and the results have shown that the proposed algorithm significantly outperformed the other three clustering algorithms in four evaluation indices. As such, it is a new variable weighting method for clustering of large data.

For the evaluation of performance of algorithm used MATLAB software and three real life data are used. The proposed algorithm work with particle of swarm optimization, so pos function of MATLAB is used. For the measuring the parameter used standard formula such as accuracy, precision, f-measure and recall.

Our empirical result shows that our proposed algorithm shows better result in comparison of W-k-means and TW-K-means algorithm. The exiting two algorithms not controlled the level weight of cluster and loss some data during the grouping of cluster. The proposed algorithm is very efficient for large data clustering technique.

XI. FUTURE WORK

The proposed algorithm is very efficient clustering technique for large data. The algorithm used particle of swarm optimization for controlling the weight variable of cluster level generation during formation of cluster. The POS algorithm takes more time for the selection of estimated value of weight. The values of weight influence the cluster quality during process of data. In future used optimization technique for self-selection of optimal cluster for large data.

REFERENCES

[1] Fahad Saeed, Jason D. Hoffert, Mark A. Knepper "CAMS-RS: Clustering Algorithm for Large-Scale Mass Spectrometry Data Using Restricted Search Space and Intelligent Random Sampling" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL-11, 2014. Pp 128-141.

[2] Ramachandra Rao Kurada, K Karteeka Pavan, AV Dattareya Rao "A preliminary survey on optimized multiobjective metaheuristic methods for data clustering using evolutionary approaches" International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, 2013. Pp 57-78.

[3] Rushi Longadge, Snehlata S. Dongre, Latesh Malik " Multi-Cluster Based Approach for skewed Data in Data Mining" IOSR Journal of Computer Engineering (IOSR-JCE) vol 12, 2013. Pp 66-73.

[4] Kelvin Sim, Vivekanand Gopalkrishnan, Arthur Zimek, Gao Cong "A survey on enhanced subspace clustering" 2012. Pp 34-41.

All rights reserved by *www.ijsrd.com*

[5] Adnan Alrabea, A. V. Senthilkumar, Hasan Al-Shalabi, Ahmad Bader "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA" Journal of Advances in Computer Networks, Vol-1, 2013. Pp 137-142.
[6] Wei Lu Yanyan Shen Su Chen Beng Chin Ooi "Efficient Processing of k Nearest Neighbor Joins using Map Reduce" 2012. Pp 1016-1027.
[7] E.N.Sathishkumar, K.Thangavel, T.Chandrasekhar "A Novel Approach for Single Gene Selection Using Clustering and Dimensionality Reduction" 2013. Pp 234-240.
[8] Liang Bai, JiyeLiang, ChuangyinDang, FuyuanCao "A novel fuzzy clustering algorithm with between-cluster information for categorical data" Elsevier ltd. 2012. Pp 1-19.
[9] Xiaoyan Wan "The Research of Fast Clustering Algorithm of High Dimension Data mining" International Journal of Digital Content Technology and its Applications (IJDCTA) Volume-7, Number2, January 2013. Pp 604-611.
[10] Suchithra Chandran, Bright Gee Varghese.R "A Survey On Clustering Techniques For Identification Of Extract Class Opportunities" International Journal of Research in Engineering and Technology, Vol-2, 2013. Pp 426-429.

All rights reserved by *www.ijsrd.com*