International Journal of Scientific Research & Engineering Trends
Volume 2, Issue 5, Sept.-2016, ISSN (Online): 2395-566X

An Genetic Based Fuzzy Approach for Density Based Clustering by Using K-Means

Karuna Kant Tiwari
Radha Raman Institute of Technology & Science, Bhopal, India
Email: Karunakant24@gmail.com

Anurag Jain
Radha Raman Institute of Technology & Science, Bhopal, India

***Abstract* – Density based clustering is an emerging field of data mining now a days. There is a need to enhance Research based on clustering approach of data mining. There are number of approaches has been proposed by various author. In this paper, the new algorithm density based clustering is proposed which efficiently overcome the major drawbacks viz. right number of cluster and initial seed (center point) problem. Proposed Genetic Based Fuzzy k-mean clustering algorithm is based on two specific factors, threshold factor which initial decide the number of cluster and specific factor which merge the clusters according the similarity. The careful selection of threshold value and specific factor which control merging of clusters yields efficient algorithmic results. In the process of generation of cluster, the seed generation is select randomly. The randomly select seed encoded in the form of binary format.**

***Keywords* – Data Mining, Clustering, DBSCAN, Genetic Algorithm, Fuzzy Set.**

I. INTRODUCTION

Today, data is automatically received from various types of equipment. Satellites, X-rays and traffic cameras are some of them. For the information / data understandable to us, should be treated. When working with large data sets is useful in many scenarios to separate the information by dividing the data into smaller categories, and finally the identification of class. No less important is it important in the treatment of large spatial databases. A satellite, for example, collects an image that moves around our land. You want to classify portions of images of houses, cars, roads, lakes, forests, etc. From the database of the image is large, a good classification algorithm is necessary. Classification can, for example, is made by means of clustering algorithms which similar data is grouped in different groups. However, the use of clustering algorithms involves some problems: It is often difficult to know which are the input parameters to be used for a specific database, if the user does not have sufficient domain knowledge. In addition, spatial data sets can contain large amounts of data, and try to find patterns of the various cluster sizes is very computationally expensive. Short calculation time is always favorable. Finally, the shapes of the groups may be arbitrary, and in severe cases very complex. Find these forms can be very heavy.

II. CLUSTERING ALGORITHM

Clustering and classification are two fundamental tasks of data mining. The classification is primarily used as a supervised learning method, the combination of unsupervised learning (clustering models are both). The goal of the group is descriptive classification is predictive. Since the objective of the group is to find a new set of categories, new groups are of interest in themselves, and evaluation is intrinsic. In classification tasks, however, a large part of the assessment is extrinsic, because the groups must reflect a certain set of reference classes. "World Understandingour requires the conceptualization of similarities and differences between the entities that compose.

III. CLASSIFICATION OF CLUSTERING ALGORITHM

There are some good clustering algorithms used there; one of them is the famous CLARANS. Other methods include K-means, K-medoid, hierarchical clustering and self-organizing maps. However, none of these algorithms can handle all three problems mentioned in the right direction. This report does not deal with these methods, but focus on DBSCAN (based spatial clustering applications with noise density) [1] algorithm, which offers solutions to these problems.

Partitioning Algorithm:

Construct various partitions then evaluate them by some criterion (CLARANS, O(n) calls). This type of algorithm constructs a partition of a database D of n objects into a set of k groups. k is an input parameter for these algorithms is that the domain knowledge is unfortunately not available for many applications is needed.

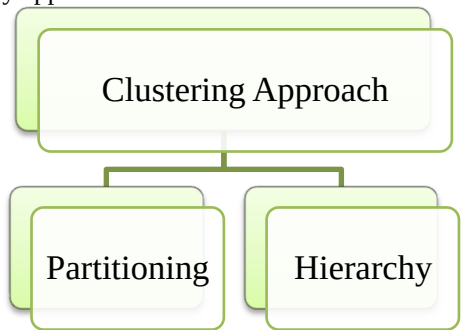

Fig.1. Basic Classification of Clustering Approach

The partitioning algorithm usually starts with initial partition D and then uses an iterative control to optimize an objective function. Each group is represented by the center of gravity of the cluster or a group of objects located near its center. Accordingly, the separation algorithms using a procedure in two stages. First, determine the k representatives minimizing the objective function. Second, assign each object in the class with his "closest" representative of the object in question. The second step involves a partition is equivalent to a Voronoi diagram and each group is contained in one of the Voronoi cells. Therefore, the form found in all groups by a partitioning algorithm is very restrictive convex.

Hierarchy Algorithm:

Create a hierarchical decomposition of the set of data (or objects) using some criterion (merge & divisive, difficult to find termination condition).In the hierarchical decomposition of D. The hierarchical decomposition is represented by a dendrogram, a tree that is iteratively divided into smaller subsets until each subset D of a single object is made. In such a hierarchy, each node of the tree represents a group of D. The dendrogram can be created from the leaves to the root (agglomeration approach) or from the root to the leaves (approach of division) merger or division of groups in each step. Unlike separation algorithms, algorithms need not hierarchical k as input. However, a condition of termination must be set indicating that the process of merger or division must be completed. An example of a termination approach agglomeration state Dmin is the critical distance between all groups Q. Until now, the main problem with hierarchical clustering algorithms has been the difficulty of deriving the appropriate settings for the termination condition for example, a value of Dmin is small enough to remove all the "natural" groups, while large enough so that no group is divided into two parts. Recently, in the field of signal processing Ejcluster hierarchical algorithm was presented automatically derive a termination condition. Its main idea is that two points belong to the same group if you walk in the first point of the second stage of a "sufficiently small". Ejcluster follows the approach of the division. It requires no intervention by domain knowledge. In addition, experiments show that is very effective in the discovery of non-convex groups. However, the computational cost of Ejcluster is O (n2) due to the calculation of the distance for each pair of points. This is acceptable for applications such as character recognition with moderate values of n, but is prohibitive for applications in large databases.

IV. DBSCAN

DBSCAN (Density based spatial clustering of application with noise) [14] is density based method which can identify arbitrary shaped clusters where clusters are defined as dense regions separated by low dense regions. DBSCAN starts with an arbitrary object in the dataset and checks neighbor objects within a given radius (Eps). If the neighbours with in that Eps are more than the minimum number of objects required for a cluster, it is marked as core object and if the objects in it surrounding within given Eps are less than the minimum number of objects required, then this object is marked as noise. The search continues for all the objects in the dataset. Later on if the minimum numbers of objects within a given radius are met subsequently previously marked objects as noise are renamed, in this way the DBSCAN differentiate between the border points of a cluster and noisy objects.

V. THE DBSCAN ALGORITHM

The DBSCAN algorithm can identify clusters of large spatial data sets watching the local density of blocks of data using a single input parameter. In addition, the user gets a suggestion that the parameter value which would be appropriate. Therefore, a minimum area of knowledge is required. The DBSCAN can also determine what information should be classified as noise or outliers. Despite this, it is the work process is fast and scales well with the size of the database almost linearly. By using the density distribution of nodes in the database, those nodes DBSCAN be classified into distinct groups defining different classes. DBSCAN can find clusters of arbitrary shape, as shown in Figure 1 [1]. However, groups that are close together tend to belong to the same class.

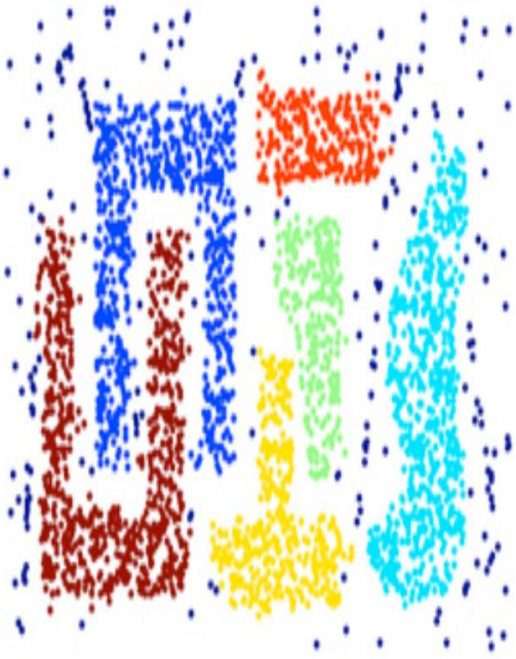

Fig.2. Example of Density based Clustering

VI. APPLICATIONS OF DBSCAN

An example of software program that has the DBSCAN algorithm implemented is WEKA. The following of this section gives some examples of practical application of the DBSCAN algorithm.

Satellites images

A large number of satellite data is received worldwide and these data must be translated into understandable information, eg, classification of satellite images taken in accordance areas with forests, water and mountains. Before the DBSCAN algorithm can classify these three elements in the database, a work must be done with image processing. Once the image processing is given, the data appears as spatial data where DBSCAN can sort the groups if desired.

X-ray crystallography

X-ray crystallography is another practical application that locates all the atoms in a crystal, which causes a large amount of data. The DBSCAN algorithm can be used to find and classify the atoms in the data.

Anomaly Detection in Temperature

This type of application data focuses on the anomalies in the data model, which is important in many cases, for example, credit fraud, health, etc. This application measures the temperature anomalies [3], which is important because of environmental changes (global warming). You can also find computer errors and so on. These unusual patterns must be identified and examined to take control of the situation. The DBSCAN algorithm is able to discover these patterns in the data

VII. LITERATURE REVIEW

Classification method of research based on the density is an important task of data mining. To improve methods based on the density of the space attribute (such as DBSCAN, Camarilla, optical, etc.) which do not take into account the relationship between objects and methods based on the density of the network (eg , SCAN, DCSBRD etc.) to ignore the attribute information of the object, a clustering algorithm based on density weighted network attribute information (WN DCA) proposed in the document. After setting the network-based distance weighted attribute, the algorithm updates the definition of the object nearest neighbor and the object, and provides the appropriate policy group. To take into account both the information of attributes and relations, the algorithm increases the accuracy of clustering improves the results of clustering, and distinguish the nature and aberrant objects effectively.

The data stream clustering attracted many researchers and applications that generate data streams have become more popular. Several clustering algorithms have been introduced for data streams based on the distance they are incompetent to find clusters of arbitrary shapes and can not handle outliers. Classification algorithms based on density are remarkable not only for finding clusters arbitrarily, but also to deal with noise in the data. In classification algorithms based on the density of the dense zones of the objects in the data space are considered groups are separated by areas of low density. Another group of methods of classification of data streams is based on the combination of the network wherein the data space is quantized in finished cells which form the network structure and perform the grouping number of networks. On the basis of the cluster network assigns infinite number of data records in the data flow to a finite number of networks. In this article, the clustering algorithms using algorithms based on a grid based on the concept of the density or density are considered for clustering. We call grouping algorithms network density. We explore algorithms in detail and the advantages and limitations of them. The algorithms are also summarized in a table on the basis of important features. In addition, it describes how well the algorithms deal with difficult issues in clustering data streams.

In this article, based on the density outlier mining similarity-neighbor technique for data preprocessing for data mining algorithm is proposed. First, the notion of a k-density object is presented, and the number of similar density (SDS) of the object to the evolution of the density of objects and densities is established on the neighboring base. Second, the series of average cost (AUC) of the object on the basis of the weighted sum of the distance between adjacent objects to the object is obtained SDS. Finally, the density of outliers similarity factor based on neighbor (DSNOF) of the object is calculated using both the AUC and the object of the k-ASC neighbor distance object, and the degree of object is an outlier is indicated by DSNOF. The experiments were performed on sets of synthetic data and real data to evaluate the effectiveness and performance of the proposed algorithm. The results of experiments to demonstrate that the proposed algorithm has an outlier extraction better and do not increase the complexity of the algorithm.

Social networking has been considered a timely and cost effective source of spatio-temporal information for many fields of application. However, while some research groups have successfully developed methods for detecting current issue of the text for a while, and even some popular microblogging services such as Twitter to provide information of the main trend themes for the selection remains incapable of fully support users to collect all the items on the events in real time with a point of complete spatio-temporal view to meet their information needs. This work aims to study how micro-blogging social network (eg Twitter) can be used as a reliable source of emerging events to extract the spatio-temporal characteristics of messages to increase awareness event information. In this paper, a method for online classification based on density flow micro blogging text mining, in order to obtain spatial and temporal characteristics of real-world events are applied. By analyzing events detected by our system, temporal and spatial impacts of emerging events can be estimated for the attainment of situational awareness and risk management.

The advent of modern for scientific data collection techniques has led to the massive accumulation of data from various fields. Cluster analysis is one of the main methods of data analysis. It is the art of all similar items in large data sets without the need to detect the specified

groups by explicit functions. The problem of detection is difficult when the groups are of different size, density and shape. This paper provides a new approach to clustering based on the approach of the density. DBSCAN is considered one of the pioneers of density on the technical basis of clustering; this paper makes a step towards the detection of groups within a cluster. On the basis of various parameters necessary for proper clustering algorithm is estimated that the number of groups formed, the noise in the change of distance, time to form a group where non-cluster and incorrectly.

VIII. RESEARCH SCOPE

This section primarily reflects the comparison and contrast of the above reviewed literature regarding the different DBSCAN variations and modifications. It identifies the similarities and differences among the various research works on the DBSCAN algorithm enhancements. This will help for the future research in the DBSCAN modification and enhancements.

Liu et al. [11] have modified the DBSCAN to deal with the datasets that are varied in densities. Their algorithm is called VDBSCAN. VDBSCAN is able to calculate the density threshold parameters automatically based on the K-distance plotting. Its computational complexity is same as that of DBSCAN. The same work is explored in GRIDBSCAN [12] to deal with the dataset that have cluster with different densities. The research work proposed in [11, 12] are identical in that they do not require any user supplied input parameters. The study carried out by [12] can cluster the dataset efficiently as that of [11] but [12] is expensive as compare to that of [11]. Fahim et al. [13] carried out the research in the same dimension as that of [11] in the sense that it does not require any user supplied density threshold parameters.

Uncu et al. [12] have introduced an extension of DBSCAN such that it can cluster the datasets having different densities. The author in [12] has used the concept of grid while performing clustering. Its clustering results are more efficient than results produced by DBSCAN. Similar grid based technique is also used by Mahran et al. [14] to generate efficient clustering output from the underlying dataset and it has proved more faster than DBSCAN . The method in [12] was more costly than that of [14] when applied on the large volume of datasets.

YU et al. [15] also used the local density in its clustering technique for large datasets. EDBSCAN [16] also focused on the local density variation and provided an enhancement to DBSCAN.

The clustering techniques described in [14, 15] have achieved the efficient clustering result by using the local density in their clustering technique. The density-based techniques discussed in [11, 12] does not need density threshold to be input by the end users. The technique described in [16] requires the user input density threshold manually.

IX. GENETIC ALGORITHM

Genetic algorithms [27] are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved. Algorithm is started with a set of solutions (represented by chromosomes or also called string) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one.

A. Coding to Strings

In GA, each individual in a population is usually coded as coded as a fixed-length binary string. The length of the string depends on the domain of the parameters and the required precision.

B. Initial Population

The initial process is quite simple. We create a population of individuals, where individual in a population is a binary string with a fixed-length, and every bit of the binary string is initialized randomly.

C. Evaluation

In each generation for which the GA is run, each individual in the population is evaluated against the unknown environment. The fitness values are associated with the values of objective function.

D. Genetic Operators

Genetic operators drive the evolutionary process of a population in GA, after the Darwinian principle of survival of the fittest and naturally occurring genetic operations. The most widely used genetic operators are reproduction, crossover and mutation.

To perform genetic operators, one must select individuals in the population to be operated on The selection strategy is chiefly based on the fitness level of the individuals actually presented in the population. There are many different selection strategies based on fitness. The most popular is the fitness proportionate selection.

After a new population is formed by selection process, some members of the new populations undergo transformations by means of genetic operators to form new solutions (a recombination step). Because of intuitive similarities, we only employ during the recombination phase of the GA three basic operators: reproduction, crossover and mutation, which are controlled by the parameter *Pr*, *Pc* and P*m* (reproduction probability, crossover probability and mutation probability), respectively. Let us illustrate these three genetic operators. As an individual is selected, reproduction operator only copy it form the current population into the new population (i.e., the new generation) without alternation.
The crossover operator starts with two selected individuals and then the crossover point (an integer between 1 and *L*-1, where *L* is the length of strings) is selected randomly. The third genetic operator, mutation, introduces random changes in structures in the population, and it may occasionally have beneficial results: escaping from a local optimum. In our GA, mutation is just to negate every bit of

the strings, i.e., changes a 1 to 0 and vice versa, with probability P*m*.

X. PROPOSED METHODOLOGY

Proposed improved k-means algorithm based on modified genetic algorithm The proposed algorithm consists of SIXES phases. All six phases describe here with block diagram.

Phase I

The general idea about selection of initial cluster centers using genetic algorithm In this algorithm, we first use random function to select K data objects as initial cluster centers to form a Chromosome, a total of M chromosomes selected, then have K-means operation on each group of cluster center in the initial population to compute fitness, select individuals according to the fitness of each chromosome, select high-fitness chromosomes for the crossover and mutation operation eliminating low fitness chromosomes, format next generation group finally. In this way, within each new generation of groups, the average fitness are rising, each cluster center is closer to the optimal cluster center, and finally select chromosome that have the highest fitness as the initial cluster center. Basically in this phase of algorithm we select the population of data for the processing of seed generation of cluster. Initially here set the population size in this process population size of data is automatically select the size of population, by default the size of population is 50. The proposed algorithm block diagram shown in figure 3

Phase II

Chromosome Coding In this algorithm, we use real-coded, the value of chromosome gene corresponds to cluster center number, length of the chromosome is the number of cluster, and the

Specific code form is:

X= (XI, X2 . . . . X k)…… (1)

K is the number of cluster center of a chromosome.

After the completion of first phase selection of population then perform the encoding of data in the form of binary (1010101010110). The binary encoding technique depends on the process of data selection method. Basically these are a part of survival of fitness. Some set encoded in the form of 0string and some form in 1 strings.

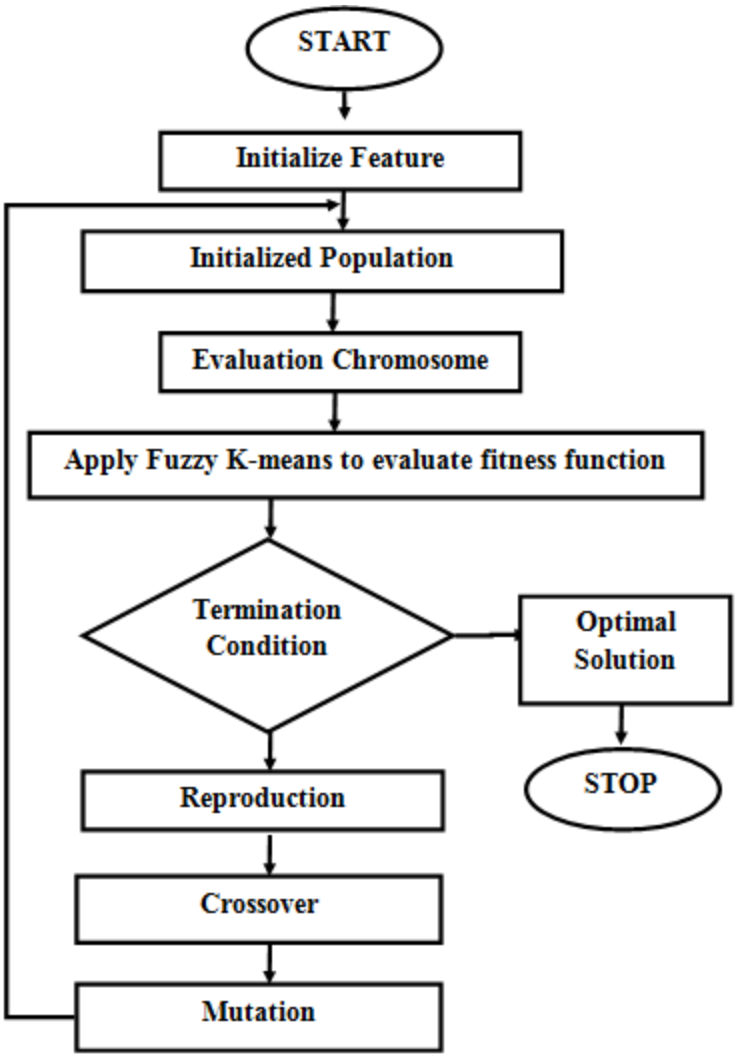

Fig.3. Proposed Model

Phase III

The range of M is 50-178. Specific operation is as follows: select K cluster centers randomly to form a chromosome Ran, if the center randomly selected has already exist in the same chromosome, then remove the center and reselect until it reaches K centers, until the population size to M .This algorithm use the inverse of objective function J as the fitness function, that is

$$F=1/J \quad (2)$$

The smaller J is, the greater fitness function will become, so the better clustering effect is.

E Genetic Operation This algorithm use proportional selection operator, single-point crossover operator and uniform mutation operator. To avoid premature or slow convergence phenomenon using a fixed probability, this algorithm use self-adaptive genetic operator that is dynamically adjust the crossover rate and mutation rate. Among them, f_{ave} means average fitness value of each generation group; f_{max} means the largest individual fitness value in the group; f_1 means the larger fitness value of the two crossing individuals; f indicates the fitness value of mutating individual. The formula makes individuals with high fitness have lower crossover rate and mutation rate; individuals with small fitness have a higher crossover rate and mutation rate. This helps protect the best individual, but also can make individuals with lower fitness cross and mutate at higher rate, producing excellent model [19]. Finally in this phase we produced offspring list of data set, now in this phase we select the best set of data for the generation of cluster.

Phase IV

In the process of list of selection : if the highest fitness in current group is larger than the best individual's fitness so far, then use the best individual in the current group as the new best individual so far., otherwise, replace the worst individual in current generation with the best individual so far .

Phase V

This algorithm , we use termination algebra μ as running end condition of genetic algorithm, which indicate that the genetic algorithm stop running after it runs to the specified evolution algebra, and output the best individual in current group as optimal solution of the problem. Generally range from 50 to 178.

Phase VI

1) Set the parameters: population size M, the maximum number of iteration μ, the number of clusters K, etc.

2) Generate m chromosomes randomly, a chromosome represents a set of initial cluster centers, to form the initial population.

3) According to the initial cluster centers showed by every chromosome, carry out K-means clustering, each chromosome corresponds to once K-means clustering, then calculate chromosome fitness in line with clustering result, and implement the optimal preservation strategy.

4) For the group, to carry out selection, crossover and mutation operator to produce a new generation of group.

5) To determine whether the conditions meet the genetic termination conditions, if meet then withdrawal genetic operation and tum 6, otherwise tum 3.

6) Calculate fitness of the new generation of group; compare the fitness of the best individual in current group with the best individual's fitness so far to find the individual with the highest fitness.

7) Carry out K-means clustering according to the initial cluster center represented by the chromosome with the highest fitness, and then output clustering result.

XI. Result Analysis

Here we used two different data set for result analysis one is abalone data set and another one data set is wine data set both data set provided by UCI machine Respiratory. In our analysis we put the threshold value for the generation of seed and cluster. Threshold is nothing basically this is intermediate value between data set. Here used different - different threshold value for analyses of seed and cluster. Here change the value of threshold also automatically changes the iteration rate of data.

Result set of abalone data set

Table 1: shows that comparison of cluster generation of threshold value.

Threshold value	Fuzzy GA-K-means based DBSCAN	Distance based DBSCAN
0.24	4	6
0.52	5	7
0.45	3	5
0.69	3	3

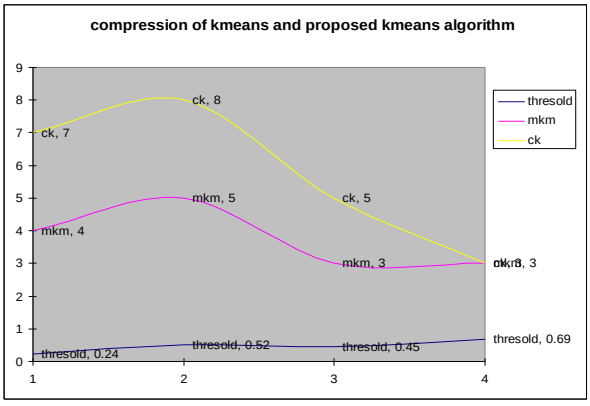

Fig.4. Comparison between conventional k means algorithm and modified k means

Table 2: Comparison of cluster generation error of threshold value

Threshold value	Fuzzy GA-K-means based DBSCAN	Distance based DBSCAN
0.24	2.79	4.52
0.52	3.70	4.69
0.45	3.72	4.65
0.69	3.721	4.699

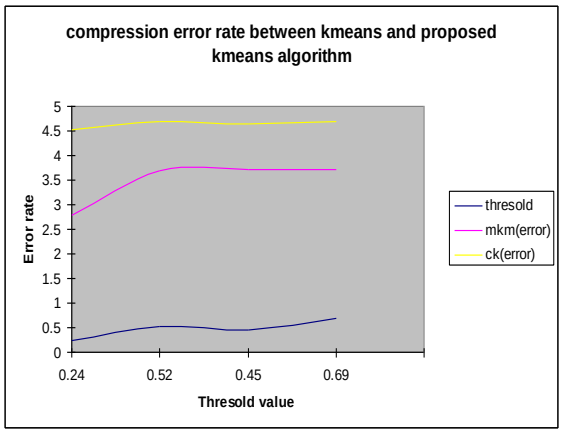

Fig.5. Comparision between conventional k means algorithm and modified k means algorithm error rate.

XII. CONCLUSION

In this Paper, the new algorithm for DBSCAN clustering is proposed which efficiently overcome the major drawbacks viz. right number of cluster and initial seed (center point) problem. Proposed k-mean clustering algorithm is based on two specific factors, threshold factor which initial decide the number of cluster and specific factor which merge the clusters according the similarity. The careful selection of threshold value and specific factor which control merging of clusters yields efficient algorithmic results. In the process of generation of cluster, the seed generation is select randomly. The randomly select seed encoded in the form of binary format. Merging of a cluster is a very difficult task for the optimizations of cluster generation .here we merge cluster on the basis of most similarity property using the concept of nearest neighbor. In the process of result generate the optimized no of cluster. In this paper we test the algorithm used two different data set one is abalone data set another one data set is wine data set both data set provided by UCI Machine respiratory.

REFERENCES

[1] H. Sun, J. Huang, J. Han, H. Deng, P. Zhao, and B. Feng, "Gskeletonclu: Density-Based Network Clustering via Structure-Connected Tree Division or Agglomeration," IEEE 2010, pp. 481-490, 2010.

[2] M. Girvan and M.E.J. Newman, "Community Structure in Social and Biological Networks," IEEE 2002, vol. 99, no. 12, pp. 7821-7826.

[3] A. Clauset, C. Moore, and M.E.J. Newman, "Hierarchical Structure and the Prediction of Missing Links in Networks," IEEE 2008, vol. 453, pp. 98-101

[4] J.M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," IEEE 1998, pp. 668-677

[5] P. Domingos and M. Richardson, "Mining the Network Value of Customers," ACM 2001, pp 57-66

[6] Wu Lingyu, Gao Xuedong, "A Density-based Clustering Algorithm for Weighted Network with Attribute Information", 3rd International Conference on Advanced Computer Control, IEEE 2011, pp 629-633.

[7] Amineh Amini, Teh Ying Wah, Mahmoud Reza Saybani, Saeed Reza Aghabozorgi Sahaf Yazdi, "A Study of Density-Grid based Clustering Algorithms on Data Streams", Eighth International Conference on Fuzzy Systems and Knowledge Discovery, IEEE 2011, pp 1652-1656.

[8] Hui Cao, Gangquan Si, Yanbin Zhang and Lixin Jia, "Enhancing effectiveness of density-based outlier mining scheme with density-similarity-neighbor-based outlier factor", Expert Systems with Applications, elsevier 2010, pp 8090-8101

[9] Chung-Hong Lee, "Mining spatio-temporal information on microblogging streams using a density-based online clustering method", Expert Systems with Applications, elsevier 2012, pp 9623–9641

[10] Glory H.Shah, "An Improved DBSCAN, A Density Based Clustering Algorithm with Parameter Selection for High Dimensional Data Sets", IEEE 2012,pp 1-6.

[11] P. Liu, D. Zhou, and N. J. Wu,"VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise," in proceedings of IEEE International Conference on Service Systems and Service Management, Chengdu, China, pp 1-4, 2007.

[12] O. Uncu, W. A. Gruver, D. B. Kotak, D. Sabaz, Z. Alibhai, and C. Ng, "GRIDBSCAN: GRId Density-Based Spatial Clustering of Applications with Noise," 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan.

[13] A. M. Fahim, A. M. Salem, F. A. Torkey, and M.A. Ramadan, "Density Clustering Based on Radius of Data (DCBRD)," World Academy of Science, Engineering and Technology 2006.

[14] S. Mahran and K. Mahar, "Using Grid for Accelerating Density Based Clustering," Computer and Information Technology, CIT2008, 8th IEEE International Conference on. 08/08/2008, ISBN: 978-1-4244-2357-6, Sydney, NSW.

[15] X. P. Yu, D. Zhou, and Y. Zhou, "A New Clustering Algorithm Based on Distance and Density," presented in proceedings of International Conference on Services Systems and Services Management (ICSSSM-2005), Vol. 2.

[16] A. Ram, A. Sharma, A. S. Jalall, R. Singh, and A. Agrawal, "An Enhanced Density Based Spatial Clustering of Applications with Noise," 2009 IEEE International Advance Computing Conference (IACC2009) Patiala, India, 6-7 March 2009.