Methodology of the Heuristic Based Hybrid Clustering Technique for Pattern Classification and Recognition

Sajal Kanta Das,
Department of Computer Science & Technology,
Women's Polytechnic,
Hapania,Agartala,Tripura,India.
Email:sajalmtechcse@gmail.com

Tanmay De,
Department of Computer Science & Engg,
NIT Durgapur(WB), India.
Email:tanmayd12@gmail.com

Abstract

In this paper we investigate the problem in different data sets to form similar objects into identical groups. Our technique is an unsupervised based algorithm. Unsupervised portion so high that the no input are given by user. Automatically judge the threshold applying threshold which is selected heuristic manner. It can also be resolve Singleton sets which can be identified in some special condition. Clustering is the clubbing of similar objects into identical groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common feature - often proximity according to some defined distance measure. Clustering is the clubbing of similar objects into identical groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common feature - often proximity according to some defined distance measure. The capability of recognizing and classifying patterns is one of the most fundamental characteristics of human intelligence. The primary goal of pattern recognition is supervised or unsupervised classification.

Keywords: unsupervised, singleton, k-means, partitional clustering, hierarchical clustering, optimal clustering, divisive and agglomerative clustering.

I. INTRODUCTION

Clustering determines which elements in a dataset are similar. It works to group records together according to an algorithm or mathematical formula that attempts to find centroids, or centers, around which similar records gravitate. It is the process of automatically dividing a dataset into mutually exclusive subgroups, without relying on predefined classes.It plays a key role in perception as well as at the various levels of cognition. As a field of study, pattern recognition has been evolving since the 1950s, in close connection with the emergence and evolution of computer technology.

From a general point of view, pattern recognition may be defined as a process by which we search for structures in data and classify these structures into categories such that the degree of association is high among structures of the same category and low among structures of different categories. Relevant categories are usually characterized by prototypical structures. Among the various frameworks in which pattern recognition has been traditionally formulated, the statistical approach has been most intensively studied and used in practice. More recently, neural network techniques and methods imported from statistical learning theory have been receiving increasing attention. The design of a recognition system requires careful attention to the following issues: definition of pattern classes, sensing environment, pattern representation, feature extraction and selection, cluster analysis, classifier design and learning, selection of training and test samples, and performance evaluation. New and emerging applications, such as data mining, web searching, retrieval of multimedia data, face recognition, and cursive handwriting recognition, require robust and efficient pattern recognition techniques. The objective of this present work is to summarize and compare some of the well-known methods and finally designed and developed one Heuristic Based Hybrid Clustering(HBHC).

II. CLUSTERING PROCESS

The first step of any clustering process is pattern representation. Starting with the initial data, we must choose what data to actually process. This may involve removing outlying data or choosing the particular features that we're actually interested in. Next, we must choose a distance measurement for the clustering process. If patterns can be represented as vectors, it may be easiest to simply use the Euclidean distance. This works particularly well for compact and isolated clusters. This may not work well if features have very different scales, or if some features are not continuous (i.e. color classified as red, green, or blue). Potential solutions include normalization of data, the use of other distance measurements such as the Mahalanobis distance, or representing the relationships between patterns in a tree structure [1], [4], [5], [7].

978-0-7695-4941-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ACCT.2013.17

There are many clustering techniques which produce many different types of clusters. After forming clusters, it may be desirable to perform data abstraction. This might be as simple as providing measurements for each cluster such as the size, center, and density. The specific abstraction scheme is dependent upon the actual data being analyzed.

Finally, we may want to do cluster validity analysis.
If the results just have a few very large clusters or many really small clusters, it may be that the data was simply evenly distributed. Second, assuming that the data does in actuality contain clusters, did the algorithm find the correct clusters? Answering these questions is difficult since the concept of clusters is somewhat subjective. However, there are various statistical techniques that attempt to determine whether or not the results could have occurred by chance or are best explained by an actual property of the system being measured. Each cluster can be represented by its average, or centroids, here we see "x" in the approximate center of each cluster for the location of its centroids.

Detailed study of algorithms like k-means [11], [12], Optimal Clustering Algorithm [2], Hierarchical Clustering (Agglomerative Clustering and Divisive Clustering) [1], [3], [13] is done. Comparative study between the various clustering algorithms is done, along with their time complexity and error rate.

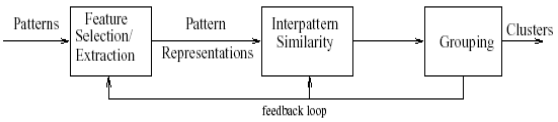


Figure 1: Cluster formation from Pattern

Clustering methods can be divided into two basic types: hierarchical and partitional clustering.

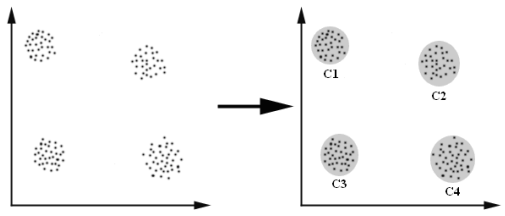


Figure 2: Basic Clustering

Cluster analysis is, essentially, the process of taking data and grouping it according to similarities. It might also be called pattern recognition. In the area of supervised classification, we start with pre-determined categories and then label the data accordingly. For example, a common problem in the biomedical field is image segmentation. This involves taking an image and identifying particular features, perhaps various tissue types. Cluster analysis, if properly implemented, can automatically divide such an image into similar regions. Another broad area for which cluster analysis is desirable is data mining. Increasingly, researchers are collecting vast quantities of data--physical, chemical, biological, social, or on nearly anything which can't possibly be analyzed by humans. So, we use cluster analysis to break the data down into its essential components, and then analyze each cluster together.

- There are several requirements that we would like our clustering algorithm to have:
- Ability to use various dissimilarity measures.
- Ability to deal with many observations.
- Ability to deal with high dimensional observations.
- Flexibility to investigate sub-optimal solutions.
- Ability to visualize resulting clusters.

Our goal here is to make efficient algorithms with less time complexity. We have first studied the existing algorithms on Partitional Clustering Techniques and Hierarchical Clustering Algorithms like different k-means [11], [12], Optimal Clustering Algorithm [9] followed by divisive and agglomerative approach of Hierarchical clustering technique [1],[3],[13].

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- Interpretability and usability.

III. DISTANCE MEASURE

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance[equ. (d)] metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling.

For pixels *p, q* and *z* with co-ordinates *(x, y), (s, t)* and *(v, w)* respectively, *D* is a distance function or metric if

(a) $D(p, q) \geq 0$ ($D(p, q) = 0$ iff $p = q$)
(b) $D(p, q) = D(q, p)$, and
(c) $D(p, z) \leq D(p, q) + D(p, z)$.

The *Euclidean distance* between *p* and *q* is defined as

$$D_e(p, q) = [\,(x-s)^2 + (y-t)^2\,]^{1/2} \quad \ldots. (d)$$

For this distance measures, the pixels having a distance less than or equal to some value r from *(x, y)* are the points contained in a disk of radius r centered at *(x, y).*

IV. TYPES OF CLUSTERING ALGORITHMS

Cluster analysis methods divide the set of processed patterns into subsets (clusters) based on the mutual similarity of subset elements. Most clustering algorithms are based on two popular techniques known as *Hierarchical* and *Partitioned* clustering[1], [3]. We have presented an elaborate discussion of hierarchical and partitional clustering algorithms. A hierarchical clustering algorithm is based on the union between the two nearest clusters.

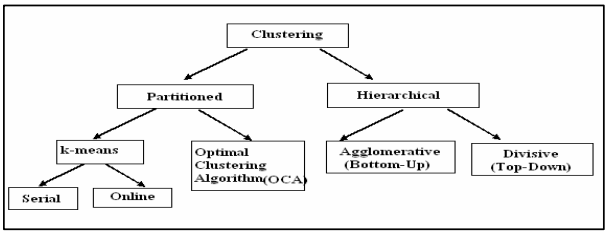


Figure 3: Types of Clustering

Each of these algorithms belongs to one of the clustering types listed above. *K*-means [11][12], Optimal Clustering Algorithm [2] etc. are the examples of Partitioned Clustering Algorithm. Divisive and Agglomerative algorithms are Hierarchical Clustering Algorithm [1], [3]. Clustering models identify relationships in a dataset that we might not logically derive through casual observation. Algorithms divide input data distribution into a number of clusters such that the sum of distances over the items to their cluster centers is minimal.

THE BASIC CLUSTERING PROCESS

The first step of any clustering process is pattern representation. Starting with the initial data, we must choose what data to actually process. This may involve removing outlying data or choosing the particular features that we're actually interested in. Next, we must choose a distance measurement for the clustering process. If patterns can be represented as vectors, it may be easiest to simply use the Euclidean distance. This works particularly well for compact and isolated clusters. This may not work well if features have very different scales, or if some features are not continuous (i.e. color classified as red, green, or blue). Potential solutions include normalization of data, the use of other distance measurements such as the Mahalanobis distance, or representing the relationships between patterns in a tree structure [1], [4], [5], [7].

The specific abstraction scheme is dependent upon the actual data being analyzed. Finally, we may want to do cluster validity analysis.

If the results just have a few very large clusters or many really small clusters, it may be that the data was simply evenly distributed. Second, assuming that the data does in actuality contain clusters, did the algorithm find the correct clusters? Answering these questions is difficult since the concept of clusters is somewhat subjective. However, there are various statistical techniques that attempt to determine whether or not the results could have occurred by chance or are best explained by an actual property of the system being measured.The clustering algorithm trains the model strictly from the relationships that exist in the data and from the clusters that the algorithm identifies.

Partitioned Clustering Algorithms

The aim of the partitioned clustering algorithms is to decompose directly the data set into a set of disjoint clusters, obtaining a partition which should optimize a certain criterion. Moreover, partitioned algorithms are generally unable to handle isolated points and to discover clusters with non-convex shapes.

An important issue in clustering related to clustering validity is the problem of choosing the right number of clusters, and given this number, selecting the partition that better fits a data set. Addressing this problem may not be an easy task if no a priori information exists, as to the expected number of clusters in the data. Even when we know the right number of clusters, due to an inappropriate choice of algorithm parameters or wrong choice of the clustering algorithm itself, the generated partitions may not reflect the desired clustering of the data.

V. METHODOLOGY OF THE PRESENT HEURISTIC BASED HYBRID CLUSTERING TECHNIQUE

Heuristic Based Hybrid Clustering (HBHC) is an efficient heuristic based clustering algorithm, suitable for

overlapping, singletons are proposed for effective clustering and prototype selection for pattern classification [6], [8], [10], [14]. It is another simple and efficient technique which uses incremental clustering principles to generate a hierarchical structure for finding the sub groups or sub clusters within each cluster. Classification accuracy is obtained using the representatives generated by hybrid structure is found to be better than that of using others partitioned or hierarchical algorithm. Even if more number of prototypes is generated, classification time is less as only a part of the hierarchical structure is searched. In this chapter at first describe *Heuristic Based Clustering(*HBC*) Algorithm* which one used to implement one part of the *HBHC* algorithm.

HEURISTIC BASED CLUSTERING ALGORITHM

This is an unsupervised clustering algorithm. In this type of clustering method the algorithm specifies the threshold which is a measure of the degree of similarity amongst the members of the same cluster and that of dissimilarity amongst members of different clusters.

Algorithm
Input: A database containing n elements.
Output:A set of clusters that minimizes the intra-cluster dissimilarity
Heuristic_Based_Clustering *(all objects in the data set)*
begin
Steps:
n: total number of data-set.
C_i :Cluster ,where i=1 to m.
for each of the remaining object in the data-set
do
begin
- *randomly select an object O that has not been visited been visited before, from n;*
- ***call function*** *{FIND_THRESHOLD(O)};*
- *increment i;*
- *T_i =T;*
- *mark objects O1 and O2 ,so that they are visited for the last time;*
- *Create a cluster C_i with threshold T_i and include objects O1 and O2 obtained in last step, into cluster C_i;*
- *Calculate the mean of cluster C_i ;*

for each of the remaining unvisited objects in the data-set
do
begin
d = minimum distance between the existing objects in cluster C_i and new object;
if *(d <= T_i)*
then,
- *mark the object;*
- *include the object within C_i ;*
- *calculate the new mean;*

end if
end for
end for
end
FIND_THRESHOLD(O)
Begin
- *select nearest two objects to O say B and C;*
- *find distance between all the respective points to O and let them be d1,d2 and d3.*
- *find the maximum distance between d1,d2 and d3.*
- *Return highest value among d1,d2 and d3 as threshold(T) with the corresponding object O1 and O2.*
- ***end.***

VI. EXPLANATION

In Figure 4, we have taken an image data-set, to be clustered. In the data set two type data distributions are present, one denser and another rarer. But the distributions of data are uniformed.

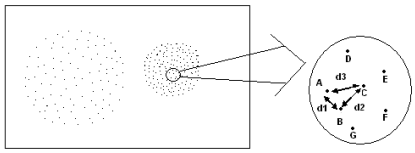


Figure 4: Initial Image with magnifying part of one dataset.

Considering the *zoom* version of the data points or objects we choose to apply the *Heuristic Based Clustering Algorithm*.

- We randomly choose an object *A* and take the two nearest objects to *A* that is, *B* and *C*.
- Let, the distance between *A-B*, *B-C* and *A-C* be *d1, d2* and *d3* respectively.
- Assume, the highest distance be *d2* (distance between objects *B* and *C*).
- We form a cluster C_1 constituting of objects *B* and *C*.
- Hence, *threshold* distance of cluster C_1 becomes T_1. where $T_1 = d2$
- Objects *B* and *C* now become marked objects, so that they are not visited for the second time.
- Now, *mean* of cluster C_1 is calculated.
- Next, we consider either object *B* or *C*.
- Suppose, we take object *C*. *D* is a point nearest to object *C* as we see from the above given figure.
- Let the distance between objects *C* and *D* be *d'*. If this distance is smaller than the threshold distance T_1, then object *D* is included within the obtained cluster C_1.
- Calculate the *mean* of the cluster C_1.

- This procedure continues until we get a greater distance than T_1 (*the existing threshold distance*).
- Henceforth, cluster C_1 is formed and we calculate the *final mean* of this cluster- ($C_{1.}$), which is X_1 and Y_2 with threshold T_1. The output is shown in the following Figure-

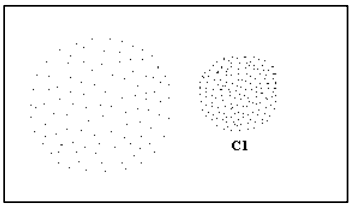


Figure 5: First cluster C_1 is formed

We get a cluster $C_{1.}$ As we encounter a threshold distance greater than that of T_1, we again randomly select another object point. In Figure 6, which is depicted in the next page. In the following figure zooming the remains portion of the object in the data set. Also cluster C_1 is defined which one already visited by the algorithms therefore are not accessed in next time. So, randomly select one object *I* from the remaining unvisited objects of the data-set. Similarly, we choose two of its nearest points *H* and *K*. Then follow the same algorithms.

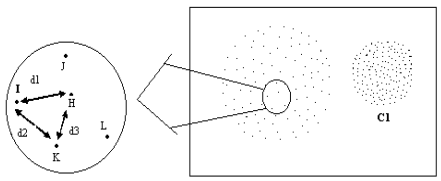


Figure 6: Magnify remaining object in the data set.

- As previously done, here again distances between *I, H* and *K* are considered, and the distances between *I-H* , *I-K* and *K-H* are taken to be *d1, d2* and *d3* respectively.
- Assume, the highest distance be *d2* (distance between objects *I* and *K*).
- We form a cluster C_2 constituting of objects *I* and *K*.
- Hence, *threshold* distance of cluster C_2 becomes $T_2 = d2$.
- Objects *I* and *K* now become marked objects, so that they are not visited for the second time.
- Now, *mean* of cluster C_2 is calculated.
- Now, we consider either *K* or *H* object.
- Suppose, we take object *K*. *L* is an object nearest to object *K* as we see from the above given figure.
- Let the distance between objects *K* and *D* be *d'*. This distance is smaller than the threshold distance T_2, and object *L* is included within the obtained cluster $C_{2.}$
- Calculate the mean of the cluster $C_{2.}$
- This procedure continues until all the remaining objects in the data-set are included in cluster C_2.
- Henceforth, cluster C_2 is formed and we calculate the *final mean* of this cluster is X_2 and Y_2 with *threshold* $T_{2.}$

On applying Heuristic Based Clustering Algorithm to the data set given in Figure 7, we get the final output as shown in the following figure.

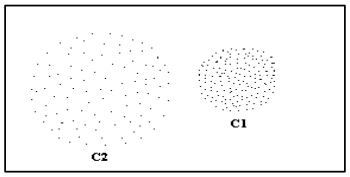


Figure 7: Clusters C_1 andC_2 are formed.

Here, we find two clusters C_1 and C_2 are formed out of the given data-set on considering the *threshold* distances between (T_1 *and* T_2) the objects. With there corresponding *mean* position are X_1, Y_1 and X_2, Y_2 respectively. Threshold values of the two clusters are not equal. Obviously T_2 is grater than the $T_{1.}$ But, the threshold remains same for that cluster.

A heuristic is a particular unsupervised technique of directing one's attention in learning, discovery, or problem-solving. Heuristics are simple, efficient rules of thumb which have been proposed to explain how people make decisions, come to judgments and solve problems, typically when facing complex problems or incomplete information. These rules work well under most circumstances, but in certain cases lead to systematic cognitive biases.

Similarly, in our algorithm, we do not supervise the threshold distance unlike optimal clustering algorithm, where we supervise the threshold distance. The algorithm automatically generates the threshold distance and forms clusters based on that.

VII. REPRESENTING DIGITAL IMAGES

All the Digital Image Frame contains some kind of representation which is very much essential at the time of Clustering Algorithm implementation [9]. The Image Frame has N columns and M rows. Each object in the Image Frame is called pixel and the object is represented in *f(x, y)* format. The values of the coordinates *(x, y)* now become discrete quantities. We shall use integer values for these discrete coordinates. Thus the values of the coordinates at the origin are *(x, y) = (0,0)*. The next coordinate values along the first row of the image are represented as *(x, y) = (0, 1)*.

The notation *(0, 1)* is used to signify the second sample along the first row.

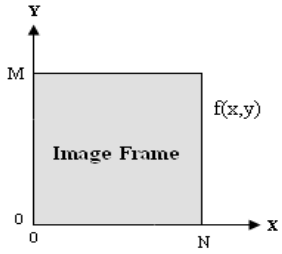


Figure 4.1: Coordinate convention used to represent digital Image.

So N×M Image Frame in the following compact matrix form:

$$f(x,y) = \begin{bmatrix} f(0,0) & f(0,1) & \dots & f(0,N-1) \\ f(1,0) & f(1,1) & \dots & f(1,N-1) \\ . & & & . \\ . & & & . \\ . & & & . \\ f(M-1,0) & f(M-1,1) & \dots & f(M-1,N-1) \end{bmatrix}$$

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	30	25	34	29
2	55	150	50	169
3	55	100	61	96
4	78	50	74	42
5	145	150	143	166
6	140	95	149	100
7	155	40	160	47

(T_1)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	54	189	56	168
2	165	176	146	165
3	28	104	37	93
4	181	112	152	104
5	28	32	53	36
6	86	43	75	94
7	165	35	148	40

(T_2)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	54	189	42	168
2	165	176	146	165
3	28	104	33	89
4	181	112	152	104
5	28	32	50	34
6	86	43	76	82
7	165	35	148	40
8	95	142	72	167
9	101	59	66	111

(T_3)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	54	189	56	168
2	165	176	146	165
3	28	104	38	94
4	181	112	151	105
5	28	32	53	36
6	86	43	79	94
7	165	35	148	40

(T_4)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	54	189	42	168
2	165	176	147	165
3	28	104	34	92
4	181	112	150	104
5	28	32	42	31
6	86	43	71	70
7	165	35	148	40
8	96	130	72	167
9	101	56	66	102

(T_5)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	66	169	103	110
2	93	133	103	110

(T_6)

Cluster Number	Initial Mean		Final Mean	
	X	Y	X	Y
1	66	169	103	110
2	93	133	103	110
3	96	110	105	74

(T_7)

Table T_1: Initial and Final means of 1st TIF in k-means with k = 7,**Table T_2:** Initial and Final means of 2nd TIF in k-means with k = 7,**Table T_3:** Initial and Final means of 3rd TIF in k-means with k = 9,**Table T_4:** Initial and Final means of 4th TIF in k-means with k = 7,**Table T_5:** Initial and Final means of 5th TIF in k-means with k = 9,**Table T_6:** Initial and Final means of 6th TIF in k-means with k = 2 , **Table T_7:** Initial and Final means of 7th TIFin k-means with k = 3

All the Image Frame is in *200×200* format. Therefore *M = 200* and *N = 200*.

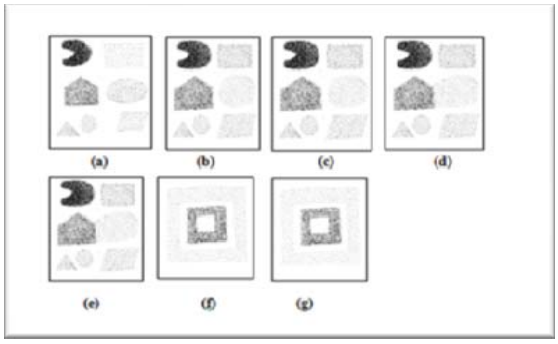
Decisions about values for *M, N* and for the number *L* of discrete gray levels allowed for each pixel. The number of gray levels typically is an integer power of 2:

$$L = 2^k$$

The discrete levels are equally spaced and that they are integers in the interval *[0, L-1]*. We shall take *k = 8*, 8 bit, where 0 (2^0) and 255 (2^8) indicate back and white respectively.

We shall consider mainly the density and the distance between the conjugate pixels not the gray levels of the image, only taken binary image where pixels are either black or white.

VII. TYPE OF IMAGE FRAME



In the Image Frame (a), inter cluster dissimilarity is highly present and the distance can be acknowledged by human intelligence also. Intra cluster distance is so large that simple clustering techniques as well as natural intelligence can easily distinguish each cluster. There is neither overlapping of clusters nor the presence of singleton. *The distances between each data point in each cluster is uniformly distributed and will remain unique throughout our work. But the inter cluster data point density is non uniform.*

In the Image Frame (b),inter cluster similarity is highly present with respect to density and intra cluster distance is not so high. Image Frame is not properly judged by the simple clustering techniques as well as natural intelligence. There does not exist overlapping of clusters or singletons.

In the Image Frame (c), Image Frame along with inter cluster dissimilarity singleton are present.

Clusters are easily identified by the natural intelligent but singleton, which is not judged properly.

In the Image Frame (d), Image Frame along with inter cluster dissimilarity which is high, overlapping clusters are present, which is easily identified by natural intelligence but simple clustering techniques cannot.

In the Image Frame (e) This Image Frame is similar to the Fourth Type Image Frame. Only singletons are added to the image frame. As in the preceeding image frame, overlapping clusters are present here, inter cluster dissimilarity is high and intra cluster dissimilarity is low.

In the Image Frame (f) This is a special type of cluster combination where one large cluster includes one small cluster. There is neither overlapping of clusters nor presence of singleton.

In the Image Frame (g),This image frame same,only difference is the presence of a singleton almost at the center. There is no overlapping of clusters the inter cluster dissimilarity is high and intra cluster dissimilarity is low.

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	34	29	3
2	50	169	2
3	61	96	3
4	74	42	4
5	143	166	7
6	149	100	4
7	160	47	5

Table :1

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	34	30	3
2	57	94	2
3	56	168	2
4	72	42	5
5	148	40	7
6	146	165	6
7	151	104	9

Table :2

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	34	30	3
2	57	94	2
3	56	168	2
4	72	42	5
5	96	141	singleton
6	102	58	singleton
7	148	40	7
8	146	165	6
9	151	104	9

Table :3

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	34	30	3
2	52	94	2
3	56	168	2
4	72	42	5
5	151	103	9
6	148	40	7
7	146	165	6

Table :4

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	34	30	3
2	72	42	5
3	56	168	2
4	86	89	9
5	52	95	2
6	146	165	6
7	148	40	7
8	103	55	Singleton
9	98	139	Singleton

Table :5

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	103	110	8
2	102	103	3

Table :6

Cluster Number	Mean Coordinate		Threshold
	X	Y	
1	96	110	Singleton
2	103	110	8
3	102	103	3

Table :7

Table 1 to Table 7:

Unlike *OCA*, where we had only one threshold value, in this case different threshold values are generated for each data distribution. This concept applies to both figure a,b.in c image frame, though inter cluster distances are less, *HBHC* can detect all different clusters. We have analyzed that both *OCA* and Hierarchical Clustering algorithms can also successful detect but it is not possible for *k*-means to give logical clusters.Singletons are properly judge which was not judged properly in *HBC*. This is the advantage of *HBHC* over *HBC* and *k*-means.

This algorithm is able to judge overlapped cluster logically, due to hybrid approach. At a certain level the overlapped clusters are detected as a single cluster and at the very next level, the two data distribution of different densities are broken down into two separate logical clusters. As we have seen, this is the only approach which can classify overlapped data distributions, which was not possible in k-means, *OCA* or Hierarchical algorithms. Cluster *C2* and *C5* are properly classified.*HBHC* algorithm, as we have seen can logically detect overlapped clusters and along with that it can also detect singleton clusters.This is a special image frame we have considered. Like all the other algorithms *HBHC* also works successfully in this kind of image frame.

This type of data frame can be logically clustered by HBHC algorithm. HBC may fail at certain times to give proper results and k-means can never successfully detect this sort o data distribution. While OCA can give logical results if proper threshold value is supplied.

VIII. CONCLUSION

Our technique is an unsupervised based algorithm. Unsupervised portion so high that the no input are given by user. Automatically judge the threshold applying threshold which is selected heuristic manner. It can also be resolve Singleton sets which can be identified in some special condition.

If overlapping clusters exist and the dense portion is selected at random then, clusters can be identified separately, but in case if the less dense portion is selected at random then, this algorithm cannot properly judge two different clusters.

If only one singleton exists and this is selected by the algorithm at the end then, it can be properly identified. Otherwise, if the singleton is selected at the midst of execution, then it cannot be judged as a different cluster. If more than one singleton exists then this algorithm cannot at all identify them as different clusters.

References

[1] KING, B. 1967. Step-wise clustering procedures. *J. Am. Stat. Assoc. 69*, 86–101.

[2] K. Shah, N. Chatterjee, B.Saha, A New Approach to a single – Pass Optimal Clustering Algorithm in Data Mining International Seminar on Communication Device and Intelligent System (CODIS-2004), pp 698 – 701, held at Kolkata.

[3] SNEATH, P. H. A. AND SOKAL, R. R. 1973. *Numerical Taxonomy*. Freeman, London, UK.

[4] Fixed Input Output 1-WTA Learning System Shell for Intelligent Pattern Recognition—G. Sarker, : Opportunities and Challenges in the Millennium on 10th & 11th March, 2000, W.B State Center pp. 64-67.

[5] SELECTWIN- A New K-WTA Optimal Learning Technique for Pattern *Classification and Recognition* – G. Sarker, *Journal of the Institution of Engineers(I)*,Vol. 83, pp 16-21, May 2002.

[6] A New Winners Take All Algorithm for Pattern Classification and Recognition – G. Sarker,CODIS 2004, 8-10 January 2004; Kolkata, pp.721-724.
[7] A Survey on Machine Intelligence and Machine Learning – G. Sarker, Institution of Engineers(India), December 15-18, 2005, Kolkata, paper no.47.
[8] A Heuristic Based Hybrid Clustering for Natural Classification in Data Mining—G. Sarker, 20th Indian Engineering Congress, December 15-18,INDIA, paper no. 4.
[9] "Digital Image Processing"- Second Edition by Rafael C. Gonzalez, Richard E. Woods.
[10] "An efficient hierarchical clustering algorithm for large data sets"-P. A. Vijaya, M. Narasimha Murty and D. K. Subramanian , Indian Institute of Science, Bangalore 560 012,.Available online 20 January 2004.
[11] DAY, W. H. E. 1992. Complexity theory: An introduction for practitioners of classification. In *Clustering and Classification*, P. Arabie and L. Hubert, Eds. World Scientific Publishing Co., Inc., River Edge, NJ.
[12] NAGY, G. 1968. State of the art in pattern recognition. *Proc. IEEE 56*, 836–862.
[13] A.K. Jain, M.N Murty, P.J Flynn, Data Clustering: A Review, ACM Computing Surveys,Vol.31,No. 3,Sept.1999.
[14] Overlap pattern synthesis with an efficient nearest neighbor classifier P.Viswanath,M. Narasimha Murty, Shalabh Bhatnagar. ,Institute of Science, Bangalore 560 012,8 October 2004