

## Data Mining Clustering Analysis is a Source of Finding Similar Function of Genetic Factor, Genome and Protein

Gul Rukh Khan<sup>1</sup>, Hamoodur Rehman Durrani<sup>2</sup>, Israr Iqbal Awan<sup>3</sup>, Mansoor Qadir<sup>4</sup>, Zahid Khan<sup>5</sup>

<sup>1</sup>Ph.D Scholar, I.T Division, Higher Education Commission, Regional Center Peshawar, Pakistan

<sup>2</sup>Assistant Professor, Department of computer Science, Iqra National University, Peshawar, Pakistan

<sup>3</sup>PhD Scholar, Department of computer Science, Iqra National University, Peshawar, Pakistan

<sup>4</sup>Lecturer, Department of computer Science, Iqra National University, Peshawar, Pakistan

<sup>5</sup>M.Phil Scholar, Department of computer Science, Iqra National University, Peshawar, Pakistan

Received: April 25, 2014

Accepted: June 2, 2014

### ABSTRACT

K-Means algorithm looks for k different clusters and divides the data into k-1 different clusters for summarization and enhanced accepting. For instance, cluster analysis has been used to extract related data (streams), to find a similar function of genetic factor, genome and proteins, and to squeeze high-dimensional data. In this segment, we provide a brief introduction to “Curse of dimensionality” cluster analysis. We propose a new technique, a simple view, which uses a clustering method based on the concept of regression and autocorrelation.

**KEYWORDS:** *E.D:* Euclidean Distance, *BD:* Biological Data, *SVD:* singular value decomposition, *PCA:* Principal Component Analysis, *HCT:* Hierarchical Clustering Theory, *GBC:* Grid Based Clustering, *JP-Clustering:* Jarvis–Patrick Clustering).

### INTRODUCTION

The purpose of the clustering is to organize data by finding some ‘reasonable’ group of data items. As clustering is unsupervised learning but more commonly it is just classification that seeks to search rules for classifying objects given a set of pre-classified objects, clustering does not use pre-defined label types related to data items. Clustering algorithm is designed to find structure in the current data, and not for the classification of future data.

Cluster is the process of grouping observation of similar kinds into smaller groups within the larger population. Clustering is used to get an idea of the nature or structure of the data: i.e. to help with hypothesis generation or anomaly detection. For the sake to get meaningful cluster, then the result should catch phrase structure “nature” of the data. Structure analysis is just a good starting point for other purposes, such as data compression, or effectively finds the point neighbors. To understand the usefulness of cluster analysis has been widely used in various fields just like machine learning, information retrieval, pattern recognition and data mining. In this segment we place a short introduction to cluster analysis. A brief look regarding recent technique of concept bases is presented here. In this very case, the approach to clustering high-dimensional data is to deal with “lack of direction”.

#### Material and Methods/ Concept:

Cluster analysis has been used to extract useful related data (streams), to find a similar function of genetic factor. Clustering based on similarity matrix and distance found between the objects (streams) in the high-dimensional data. The target is the object of a similar (or related) to each other and different (or unrelated) from other groups object group. Grouping based on information in the description of the data objects and their relationships are found in the analysis of the object/ data stream (observable events). The main objective is to find the similarity and dissimilarity in-between the groups in high-dimensional data. Clustering will be better if the data are showing more homogeneity in the group or more *Heteroscedasticity* (type of heterogeneity restricted to in-quality of variances). In statistics, a collection of random variables is heteroscedastic if there are sub-populations that have different variances from others. Cluster analysis is a classification of objects from the data, but by “sorting” we mean, a trademark of the object class (group) labels. Clustering finds only the points most tightly connected to each other by discarding background and noise points aspects. True Cluster may further sub-divided into sub-clusters. The important requirement in above case is sub-clusters that are precise and mostly all the points are inter-related and inter-connected. It is also sometimes acceptable to produce a set of clusters where a group is actually divided into many sub-clusters (often combine later with other techniques).



\*Corresponding Author: Gul Rukh Khan, Ph.D Scholar, I.T Division, Higher Education Commission, Regional Center Peshawar, Pakistan. Email: grkhan@hec.gov.pk

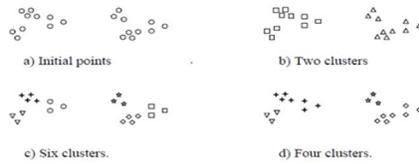


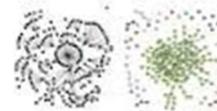
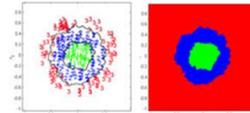
Figure 1: example of different clusters

<p>a) points</p>	Points	Xx	Yy				
	P1	0	2				
	P2	2	0				
	P3	3	1				
<p>d) proximity graph</p>	P1	P2	P3	P4			
	P1	0.00	2.83	3.20	5.10		
	P2	2.83	0.00	1.41	3.20		
	P3	3.13	1.41	0.00	2.00		
P4	5.10	3.20	2.00	0.00			
Graph		Proximity Matrix					

**Table 1: Data (4-Points) with their nearest Graph**

A cluster is a set of data/streams that close to each other (proximity & similarity) than the data outside the cluster. Second important point is that cluster is a set of data/streams that close to centroid of the same cluster as compared to the centroid of other clusters not belonging to that. Therefore, many clustering algorithms use the following criteria.

1. Proximity or Nearest Neighbor.
2. Cluster composed of a set of data streams that close to each other (proximity & similarity) on the basis of distance and other similarities with their neighbors than the data outside the cluster.
3. Cluster definition based on Density: cluster comprises of compressed area of dots showing that this is high density area along-with low-density area showing that the cluster contains lots of irregularities and noise effect in the cluster.
4. Some clusters are similar, on the basis of their objects/streams similarities and vice versa. The dispersion is only due to the high and low density areas and their irregular shapes in the cluster. In this very case to check the proximity, the type of object/data stream. Typical three types of attribute values are in Binary, discrete and continuous. Some SPSS terms regarding data is as follows:
  - a. Categorical values: is information gathered from a study i.e. descriptive and not based on numbers. This type data is then not measurable. Qualitative Nominal values are the values of Names, NIC, Address, zip-code etc. That the data entered have no superiority on one another (may be in any order).
  - b. Ordinal data: the data that has to be entered has the superiority with one another i.e. the data is in order such as:
    - i. Effected, Least-Effected, More-Effected,
    - ii. Daily-Basis, Weekly-Basis, Monthly-Basis,
    - iii. Good, Better, Best.
  - c. Quantitative: the values in number are meaningful. In quantitative-Interval, the difference between the values is significance. E.g. salaries, temperature in Degree Centigrade (Celsius) or Fahrenheit or Kelvin.



**Euclidean Distance:** The very famous proximity measure is the Euclidean Distance equation to measure scales with an absolute zero (Minkowski Space). The formula is given by:-

$$P_{ij} = \left( \sum_{k=1}^d |x_{ik} - x_{jk}|^r \right)^{1/r}$$

Accordingly, the proximity distance between any two points in XY-coordinates is as given below:

$$\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

By putting some values such as (2, -1) and (-2, 2), in the above formula, as given below:

$$\begin{aligned} \text{Dist}((2, -1), (-2,2)) &= \sqrt{(2-(-2))^2+((-1) - 2)^2} \\ &= \sqrt{(2+2)^2+((-1-2)^2)} \\ &= \sqrt{(4)^2+((-3)^2)} \\ &= \sqrt{16+9} \\ &= \sqrt{(4)^2+((-3)^2)} \\ &= \sqrt{16+9} \rightarrow \sqrt{25} \rightarrow 5 \end{aligned}$$

**Euclidean Distance (E.D):**As we know that Euclidean proximity distance matrix is  $n \times n$ . Consider the following Euclidean distance equation with high dimensional space denoted by A,

$$A = (a_{ij});$$

$$a_{ij} = ||x_i - x_j||_2^2$$

Let's consider the following different dimensions in calculating Euclidean Plane:-

Two dimensions: if  $p = (p_1, p_2)$  and  $q = (q_1, q_2)$  then the distance is as under:-

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

If the polar coordinates of the point P, are  $(r_1, \theta_1)$  and q are  $(r_2, \theta_2)$ , then the distance between the points is as follows:

$$\sqrt{r_1^2 + r_2^2 - 2r_1r_2 \cos(\theta_1 - \theta_2)}.$$

Three dimensions: Euclidean space for three dimensions, the equation is as follows:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

Similarly, for N dimensional space Euclidean distance is as follows:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_i - q_i)^2 + \dots + (p_n - q_n)^2}.$$

For standard Euclidean distance, the equation may be given as under:-

$$d^2(p, q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_i - q_i)^2 + \dots + (p_n - q_n)^2.$$

Standard Euclidean distance is used in optimization problem where distance only has to be compared.

**Euclidean Distance with some variations:**

The E.D is most commonly used in proximity measure, while Minkowski Metric is used for ratio scale with an absolute zero. Which is the generalization of distance between points in Euclidean Space.

$$P_{ij} = \left( \sum_{k=1}^d |x_{ik} - x_{jk}|^r \right)^{1/r}$$

Dimensionality in the above equation is represented by the letter d, r is a parameter. The objects  $I^{th}$  and  $J^{th}$  have the  $K^{th}$  components denoted by  $|X_{ik}|$  and  $|X_{jk}|$  respectively. There are L1 Norm or city block distance where  $r = 1$ , and L2 Norm where  $r=2$ , which occurs most commonly (Euclidean Distance). Sometimes periodically when r tends to infinity denoted by  $L_{max}$  Norm. By using Table 1 data the matrix for proximity for L1, L2 and  $L_{\infty}$  is shown below:-

Points	Xx	Yy
P1	0	2
P2	2	0
P3	3	1
P4	5	1

L2	P1	P2	P3	P4
P1	0.00	2.83	3.20	5.10
P2	2.83	0.00	1.41	3.20
P3	3.13	1.41	0.00	2.00
P4	5.10	3.20	2.00	0.00

L1	P1	P2	P3	P4
P1	0.00	4.00	4.00	6.00
P2	4.00	0.00	2.00	4.00
P3	4.00	2.00	0.00	2.00
P4	6.00	4.00	2.00	0.00

$L_{\infty}$	P1	P2	P3	P4
P1	0.00	2.0	3.0	5.0
P2	2.0	0.00	1.0	3.0
P3	3.0	2.0	0.00	2.0
P4	5.0	3.0	2.0	0.00

Table 2 data matrix, L1, L2 & Proximity matrix

Thus the calculated Minkowski distance is exactly the metric distance as it satisfies the mathematical functions reflex: Symmetry:  $\text{Dist}(Xx, Xx)=0$ ,  $(\text{dist}(Xx, Yy) = \text{dist}(Yy, Xx))$  and the data in triangle quality  $(\text{dist}(Xx, Zz) \leq \text{dist}(Xx, Yy) + \text{dist}(Yy, Xx))$ .

According to Jaccard Matrix, it deals with some difficulties in the matrices of proximity in clusters in the sense that  $Xx$  near to  $Yy$  and  $Yy$  near to  $Zz$ , which is a proximity problem. The main difference in clustering approaches is between hierarchical and Partitional approaches. Hierarchical and partitional clustering have some features on the basis of their different behavior in running time which are as follows:-

1. Partitional clustering is more efficient and faster as compared to hierarchical but the result of Hierarchical clusters are more meaningful (more suitable for qualitative data).
2. Hierarchical clustering forms nested partitions, while Partitional clusters forms un-nested partitions of the high dimensional data, but it is worth to mention that in case of partitional approach some clusters are to be needed for nesting smooth running
3. Hierarchical one deals only with the proximity matrix, while partitional one deals with the centroids of the cluster data.
4. In case of Hierarchical clustering there is no need of parameters to start with.

**Specific Partitional Clustering Techniques K-Means:**

K-Means clustering algorithm Looks for k (non-overlapping) different clusters, and break data into K-1 different partitions by placing some points as centroids. If  $k = 3$ , the data will be break into 3 clusters as shown in Fig 1(A). Three seeds are randomly placed by K-Means algorithm to decide which cluster is associated with its nearest centriod. Centriods are only for the sake to minimize the ERROR term. For the same we draw stright line between two seeds as shown in graph:-

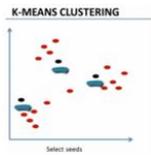


Fig. 1(A)



Fig.1(a)



Fig. 1(b)(Mid-Point)

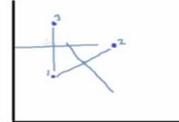


Fig. 1(c)

Calculate the Mid-Piont of the line and draw perpendicular line over it as as in Fig 1(b):-

So according to the High School Geometry any point to the left of this point will be near to seed No.1 and any point to the right of this line is near to seed No.2. Similarly if we draw line between seed No.1 and seed No.3 and then draw a perpendicular bisector of this line as shown in Fig 1(c).

Now, any observation below this perpendicular line belongs to seed No.1, and any observation above the line belongs to seed No.3. The combination of both perpendicular bisectors, any point below line belongs to seed No.1, as given in Fig. 1(d).

Similarly for seed No.2 and seed No.3 are as given in Fig. 1(e) respectively.

Here we assigned all the records into three different groups; assigned them to one of the three seeds, the below shown graph is the first set of cluster that are formed.

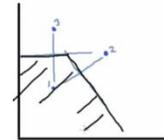


Fig.1(d)

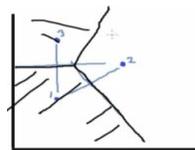


Fig.1(e)

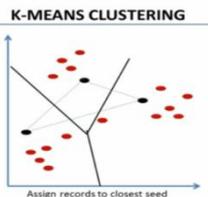


Fig.1(f)

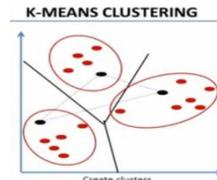


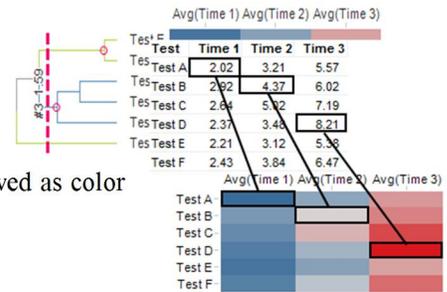
Fig.1(g)

How K-Means Algorithm finds Random Clusters:

1. Selection of random clusters with the identification of their centroids.
  2. Differentiate Points with respect to its nearest points.
  3. On the basis of above point 2 centroids are again to be taken.
  4. Above Points 2 & 3 will be executed again and again until centroids remain the same.
- K-Means algorithm (vector property) itself select some random points for the separation of different clusters as centroids. This is the point where process of similarity and proximity matrices are to be calculated.

**Overview of Hierarchical Clustering Theory (HCT):**

Hierarchical Clustering arrange items with a treelike structure based on the distance/similarity between them – called dendrogram. Dendrograms are strongly connected to Heat Map visualization (Heat map is to think of a cross table or spreadsheet which contains colors instead of numbers. Default color is set to lowest number i.e. dark blue and the highest color is Bright Red and the mid- range to Light Gray. Head maps are well-suited for visualization large amounts of multi-dimensional data and may be used to identify clusters of rows with similar values as given below: e.g. the below example display how the values are showed as color gradients in the Head map cells.



Data in short/wide format: Like in other visualizations, highlighting and marking in the heat map are applied to one or more rows in the underlying data table. In the example below, the data is in short/wide format, and each row in the data table corresponds to a row in the heat map.



Algorithm: the algorithm uses the agglomerative method of Hierarchical clustering.

**Agglomerative hierarchical clustering:** -

One of the clustering approaches is Agglomerative Hierarchical Clustering which is a bottom-up clustering method. The concept of sub-clusters found here e.g. species taxonomy in biology, a species is one of the basic units of biological classification and taxonomic rank i.e. kingdom, phylum, genus, species. In genetics a species is often defined as a group of organisms capable of mating and producing fertile children.

Gene Expression high dimensional data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Gene contains high dimensional data (multiple functions and each function have high dimensional multiple characters). In each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster. Agglomerative hierarchical clustering methods occupy a prominent position in the science of classification. In summarizing generic characteristics of the clustering methods we wish to investigate, the agglomerative methods start with set of n objects to be clustered, they group these objects into successively fewer than n sets, arriving eventually at a single set containing all n objects. They are non-overlapping methods that specify a sequence  $P_0, \dots, P_n$  of partitions of the objects in which  $P_0$  is the disjoint partition and  $P_n$  is the conjoint partition. The algorithm is used to generate iterative  $P_{i+1}$  from  $P_i$  for all  $0 \leq i < n$ .

$$L_p(x,y) = \left\{ \sum_{i=1}^k |x_i - y_i|^p \right\}^{\frac{1}{p}}$$

This family includes the L1 or Manhattan metric, the L2 or Euclidean metric or Chebychev metric for which  $L_\infty(x,y) = \max\{|x_i - y_i| : 1 \leq i \leq k\}$ .

There are some other concepts as well, as when there is minimum gap between the high dimensional data the proximity matrix will be efficient and there will be more similarity due to nearest matrix (Single Link Approach). On the other hand when there is gap between the high dimensional data is maximum the proximity matrix will not be efficient and the data will not show similarity, and the clusters will show some problems in their data during their breakage into small clusters (Complete link Approach). Group Average approach lie between single link approach and complete link approach, and is given by:-

$$proximity (cluster1, cluster2) = \frac{\sum_{\substack{p_1 \in cluster1 \\ p_2 \in cluster2}} proximity(p_1, p_2)}{size(cluster1) * size(cluster2)}$$

The “Curse of Dimensionality” (By Richard Bellman, in a book of “Control Theory”). Data Dimensionality refers to various phenomenon that arise when analyzing and organizing data in high dimensional spaces (often when hundreds and thousands of dimensions) e.g. analysis of genes and genome, that are more complex and high dimensional data approach to extract data streams comes from different sources.

- ✓ In bioinformatics each tissue sample contains tens of thousands of genes.
- ✓ In social activities networks, each person may be linked with thousands of events (Likes, tweets, Add friend etc).
- ✓ Weather and climate prediction contains multivalued attributes i.e. (temperature, rainfall, cyclones, precipitation and sunshine etc.) at each movement across Europe (<http://eca.knmi.nl/>).
- ✓ World is multidimensional (vertebrates, bees, fishes, ants, neuronsetc).

**Discretization:**

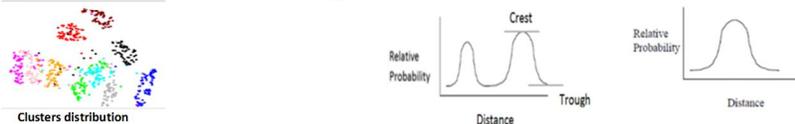
It is the process of converting real gene expression data into a typical number of finite values. One of the related problems to the “Curse of Dimensionality” is that analyzing of these data are more and more difficult and require new and most sophisticated techniques than that use in past.

Consider some points say 100 uniformly while randomly distributed with the interval (1, 0). Now, break the interval into ten cells i.e. all cells will have some extra values/points. By keeping the number same while disturbing their distribution over the unit square (2-Dim) with the discretization unit of 0.1 for each Dim. This will result in 100 of 2-Dim cells, in which some of the cells will be empty. But in three Dim 1000 cells will be empty and there is lots of empty cells will be. As this will create lots of blank spaces where we will lose much of our data due to high dimensionality. Computer language i.c. Array in C-language may handle up to 7-Dim data, which is still not implemented in real life and of course studios job and still a question mark how to implement it?

Curse of Dimensionality mainly concerns with the similarity or proximity measure between the objects/ voices of the same clusters rather than neighboring clusters.

To analyze the data regarding plotting their histogram (density function) of all its pairwise points involves lots of computations as given in MS Excel sheet below as compared to the clusters that is easier to interpret

If we plot a graph of the data containing clusters, it will produce trough and crest, crest represent the average distance between points in the cluster. If there is only one peak present it means that this will be without clusters and honestly proximity measure will be very difficult in such cases.



By analyzing the voices from different sources in high dimensional data on the basis of genes/ genome, is a complex task, but clustering make a sense to do the same on the basis of proximity measure. In such a situation the gap between the selected data with relation and with respect to their performance is to be measured, and the difference of their adjacent points (may be near or farthest) will reach to zero in high dimensional data. Such facts occur if the attributes within the cluster are identically and independently distributed.

$$\lim_{d \rightarrow \infty} \underline{MaxDist} - \underline{MinDist} = 0$$

Observation from real-world problems are often high dimensional vectors i.e. containing many variables. Relatively in Neural Networks a small number of high dimensional data is very difficult to handle. In high-dimensionality data we are interested to separate meaningful information (voices) but at the same time the result reflects the problem of “Curse of dimensionality”. In this case absolute difference instead of relative difference is preferred as: -

$$\underline{MaxDistance} - \underline{MinDistance}$$

For  $L = 1$  metric *MaxDistance–MinDistance* is directly proportional to dimensionality, while for  $L=2$  metric *MaxDistance–MinDistance* remains relatively constant but for  $L>=3$  metric *MaxDistance–MinDistance* tends to zero as dimensionality increases (i.e. meaningless), this problem may overcome if we decrease data dimensionality without losing required information (apriority) i.e. only most frequent items has to be extracted that are of interest and rest are discarded (i.e. with little variations or with high correlation).As a result dimensionality reduces in this way.

It is also worth to mention that high dimensional data is also a cause of noise in data, which is another problem. And such problem may be resolved by keeping relatively small number of dimensions from bulk of dimensions. In such case some “true” information may lose and will enhance data analysis but is an efficient way to remove noise from data. Statistics and linear algebra techniques such as Principal Component Analysis (PCA) or Singular Value Decomposition (SVD) has a wide contribution to make high dimensional data more concrete and brief.

### The Collecting and Analyzing Complex Biological Data Such As Genetic Codes - Bio-Informatics.

The analysis and extraction of gene expression in Bio-Informatics effectively and efficiently is a great challenge. A gene contains functions and each function contains different characters which contains itself a high-dimensional data streams. Tools such as singular value decomposition (SVD) and Principal Component Analysis (PCA) are used to extract character matrix from multiple function of genetic factor. PCA is map the data to lower dimensional. In order for PCA to do that it calculate and rank the importance of features/dimensions. Consider the following formula:-

$$\mathbf{X} = \mathbf{USV}^T$$

Where  $U$ ,  $S$  &  $V^T$  are  $(M \times N)$ ,  $(N \times N)$  and  $(N \times N)$  matrix respectively. Values Columns  $U$  is called left singular vector,  $V^T$  contains right singular vector  $\{v_k\}$ , an orthonormal structure of genetic factor is formed that is responsible for Transcriptional responses.

Diagonally the vector data will be less than or greater than zero, that is a problem i.e.  $S_k > 0$  for  $1 \leq k \leq r$ , and  $S_i = 0$  for  $(r+1) \leq k \leq n$ .

$$(x)^i = \sum_{k=0}^i (v_k^T) \mathbf{U}_k S_k$$

By using SVD the magnitudes of values decreasing in a continuous manner.

### Some introduction to recent work:

One of the recent works is Grid-Based Clustering (GBC) statistically where unbounded high-dimensional data (massive data/voices) coming from different sources in streams in a continuous and successive fashion. In these complex situations the Heuristic Search is beneficial as algorithms fail to give precise output here due to the processing time which is adversely affected to the maintenance of massive streams of dat.

Similarly, lots of issues are there by analyzing data by performing regression in time series high dimensional data as Covariance deal with the measures how much the movement of variable in high dimensional data predicts the movement in a corresponding variables. But sometimes in some cases two dependent variables in high dimensionality depend on each other (auto-collinearity issue); in that specific situation we cannot ignore that variable(s) as they create some other problems by ignoring the same variable(s). While in some cases the error term is measured. If the error terms in time series data are related then correlation assumptions are violated making any results invalid. Autocorrelation occurs in many ways but specifically in time series data. Commonly the error is ignored and run the test. For identification the issue Durbin Watson test is significantly used to find out the 1N-form. If the error terms follow the pattern after one year called (1N), after 2 year (2N) and so on.

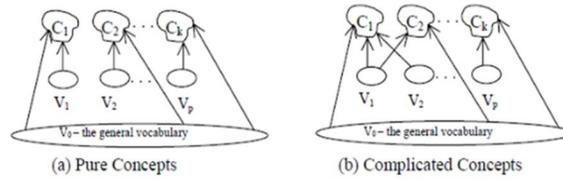
In-spite these, other different types of real time issues also there in high dimensionality clustering analysis as a neighbor which is most nearest and most similar to an object is not highly related as compared to the object which is not similar. In these cases we will look into behavior of distances to overcome such type of issues.

### Concept Oriented Approach:

Thus a concept oriented approach to cluster of objects and then sub-cluster of such clusters shows that an object is a collection of data (streams) and each stream/ word has different types of DNA (genes) with Dominant or Recessive characteristics that come from one or more concepts/living organism. This will further be discussed in my next paper with the Grace of Al-Mighty God as data streams may be in Radio-waves, supersonic-waves (frequency above our hearing) or Ultrasonic-waves(frequency below our hearing) etc.

Please examine the following conceptual models:-

First model (a) is referred to as the pure concept model or simplest model.



In this conceptual model words/ stream ( $c_1 \dots c_i$ ) comes from different sources (general vocabulary as  $V_0$  or from one specialized vocabularies as  $V_1, V_2, \dots, V_p$ ), vocabularies in such a case are just combination of words (stream) and no other form which may overlap.

Second model (b) slightly more completed than previous one which is realistic modal called multiple concepts model, where a data stream/ word may come from different vocabularies probabilistically. Consider the following equation:

$$P(w | C_i) = \sum P(w | V_j) \times P(V_j | C_i)$$

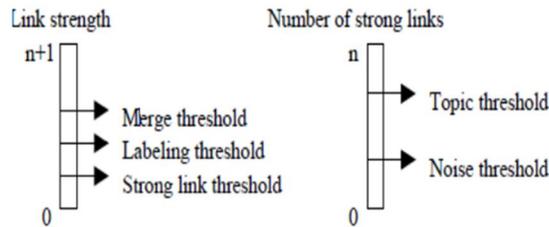
Where  $w$  represents a word/stream in a document/object from a cluster  $C_i$  containing with a sources.

To identify the data (stream) we begin with clustering approach by calculating the similarity matrix for their neighbors close to it. Main thing and which is the issue is that we should recognize the type of streaming means in which type the streams exist in i.e. frequency and bandwidth of the streams i.e. they may be in Radio waves, supersonic waves (frequency above our feeling) or Ultrasonic waves (frequency below our feeling).

**Threshold:** The level or point at which one starts to experience something, or at which something starts to happen or change.

If we plot a graph for the nearest neighbor there will be links between two objects say  $m, n$ ; in each shared cluster there is a link between objects/streams in their neighbor list.

Threshold could be applied at this stage by keeping in view the concerned neighbors as least cluster. Note that this will not provide appreciated outcome as arrays (matrix) in such a case is not noteworthy. That's why high threshold will apply in this case and as a result cluster will be split into sub-clusters because of its variation in the similarity within the same cluster. Then we look for the strength of the links between neighbors and will also looks towards the strong links within the cluster in comparison to the threshold. If a link with a greater strength than the threshold then it will be considered/ labeled as strong link.



**Shared Nearest Neighbor Clustering Algorithm:** details are as given below:-

1. Detection of connectivity i.e. to calculate the level of strength for each link.
2. Consider a point  $k$  in the dataset:-
  - a. Consider if  $con[k] < Noise(Threshold)$ , this means such value (dot) is not part of that cluster by not possessing the similarities with most of its neighbors.
  - b. On the other hand if the  $con[ k ] > Topic-Threshold$  then that point will be considered as part of the clustering as it possess the similarities with most of the neighbors.
3. If a pair  $k$  and  $L$  closely linked with their valuable neighbors, and if the relation between  $K$  and  $L$  is  $> Merge(Threshold)$  then the pair will be considered as a part of the same cluster. Similarly, consider two different objects that are not related directly with each other but they are put in the same cluster due to the presence of other objects that are closely connected to each other with strong links and may be used as agent for that specific neighbors.
4. Labeling-Threshold: consider scanning of streams with respect to their neighbors belonging to a cluster. If they have links to some other points that are not part of any other cluster, and their link strength  $> labeling(Threshold)$ , then some points will be withdrawn that are not in a position to survive the Merge-Threshold even they have defined the agent points along-with the points that are strongly related to them.

**Jarvis – Patrick (JP) clustering encounters Nearest-Neighbor-Approach** to clusters points. Distance between two points/ objects are measured. Let's  $J$  be the size of the closest neighbors and  $P$  be the number of closest neighbors. The following steps must be encountered.

1. Determine the size of closest neighbor in terms of J for each object in the cluster.
2. The objects A & B will be considered in the same cluster if B contains in A neighbors list and A contains in B neighbors list.
3. Both A & B at least P nearest neighbors in common.

By applying the Jarvis – Patrick (JP) clustering approach a singleton clusters formed here is different than that of formed during the Shared-Nearest-Neighbor approach.

To introduce Topic-Threshold, if an object(stream) in the cluster having same streams around remains singleton but labeled with Topic does not mean that this will be noise, as this will have links with their neighbors but not seems to be very strong due to their singleton-ness. Some singletons with label as Background are left out and there is no other algorithm to produce valuable singleton cluster.

The main theme of this paper is to find streams of data with their neighbors. The probability of the variation in the stream of a class is relatively high as compared to its closest class. And this probability decreases as its neighbors increases in number. Such type of shared nearest neighbors approach may be applied on different sort of data as to extract the data streams/voices related to past/ ancient times, which will be the most tedious task having the highest importance. In this specific case high dimensional data regarding to genes related to data streams/voices are to be clustered by using the above concept. In future our approach will be towards the extraction of data streams/voices that may be in Radio waves, supersonic waves (frequency above our hearing) or Ultrasonic waves(frequency below our hearing) that include high-Dimensional data.

### **Conclusions:**

Discussion regarding clustering analysis of streams composed of massive data; K-Mean’s algorithm looks for k different clusters and divide the data into k-1 different clusters with the issue of “Curse of Dimensionality”. The solution of the issue regarding “Curse of Dimensionality”. Cluster analysis has been used to extract related streams, to find a similar function of genetic factor, genome and proteins.

The above said approaches are applied there in lots of different areas but we expect and stress that the single type clustering approach may be applied everywhere and on every data, also in high-dimensional data. But there is some problems still exist in High-dimensional data i.e. scalability to large streams of data, estimation based on High-dimensional data and their interpretation etc. Our future focus will be towards the said versatile topics insha Allah. We will also focused on clustering of High-Dimensional massive data through which we may detect some streams/voices from the ancient times as genetic factor that are in Radio waves, supersonic waves (frequency above our hearing) or Ultrasonic waves (frequency below our hearing) that need some enhanced study with the comparisons of different techniques, advantages and their drawbacks and limitation as well.

### **Acknowledgment:**

With the Grace of Al-Mighty Godmy this paper is made possible through my parents, my family members, friends, teachers, and all’s help and support. But frankly speaking especially my parents have the great contribution in it to praying for me from their core of hearts.

It is worth to mention to please allow me to dedicate my acknowledgment to Mr. Nasir Shah, Deputy Director, Higher Education Commission (HEC), Pakistan. I would like to thank my two professors namely Prof. Dr. Himayatullah Khan and Prof. Dr. Zahoor Jan who provided me the valuable advices and one of my friends Mr. Ghulam Nabi who assisted me financially. My Research Paper would not possible without of the above personalities. God bless all of them.

### **REFERENCES**

1. Difference between Hierarchical and Partitional Clustering by Indika .Posted on May 29<sup>th</sup>, 2011.
2. CharuAggarwal, Cecilia Procopiuc, Joel Wolf, Phillip Yu & Jong Park “Fast Algorithm in Projected Clustering” (Conference 1998).
3. Anil K.Jain and Richard C.Dubes, Algorithms for Clustering Data, Prentice Hall (1988).
4. RA Jarvis, EA Patrick -, IEEE Transactions on Computers (1973). RA Jervis, EA Patrick, “Clustering Using Similarity Measure Based on Shared Nearest Neighbor”.
5. William H.E.Day, Herbert Esdlsbrunner, “Efficient Algorithms for Agglomerative Hierarchical Clustering Methods”.
6. Alberto Fernandez, Sergio Gomez, “Solving non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms”.
7. Parallel algorithms for hierarchical clustering, Volume 21, Issue 8, August 1995.
8. Ying Zhao, George Karypis, “Evaluation of hierarchical clustering algorithms for document datasets” article 2002.