International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Clustering Algorithms: Brief Review in Bioinformatics

Jayant Mishra[1], Vivek Agarwal[2], Megha Sharma[3], J.K Srivastava[4]

[1]Dept. of Biotechnology, PhD Research Scholar, Himalayan University, Arunachal Pradesh

[2]Dept. of Computer Science & Engineering, SRMGPC, Dr APJ Abdul Kalam Technical University, Lucknow

[3]Dept. of Computer Science & Engineering, SRMGPC, Dr APJ Abdul Kalam Technical University, Lucknow

[4]HOD, Dept. of Biotechnology, Amity University, Lucknow

Abstract: ***Pattern recognition is a task of categorising some objects to a correct class based on certain measurement or features of the object. Clustering is a technique for finding similar groups in data called clusters. It groups data instances that are similar to each other in one cluster and data instances that are very different from each other are grouped into different cluster. This paper gives a brief review of the various aspects of clustering approaches that are currently in vogue. Moreover the author summarize types and utilities of clustering approaches prevalent in practical field like Bioinformatics, comparative analysis such as that K-mean clustering is truly advantageous of all approaches and is therefore used in computational recognition studies.***

Keywords: Clustering, Patterns, Recognition, Bioinformatics, K-mean Clustering

1. Introduction

Machine learning is the field of research for the study of learning system. Machine learning basically refers to the task performed using Artificial intelligence. It is generally taken to encompass the automatic computing procedure based on logical and binary operations. Machine learning aims to generate classifying expressions simply enough to be comprehensible easily by the humans. In machine learning, pattern recognition is assignment of label to given input value [1]. A pattern is taken as an entity for the further process through which tasks can be performed. The pattern can be a fingerprint, DNA sample which is used for further learning purposes. Recognition is an act of associating a classification with a label, in recognition given objects are assigned to prescribed classes. Pattern recognition is the scientific discipline whose goal is the classification of data, objects pattern and categories.

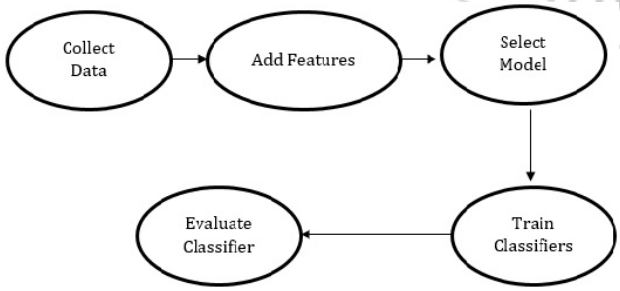


Figure 1: Flowchart for Pattern Recognition [1]

2. Pattern Recognition

Pattern recognition is a task of categorising some objects to a correct class based on certain measurement or features of the object. Object categorisation is an automation process. It is the discipline of building machines to perform perceptual task which humans perform easily by recognizing faces, voice, identifying different species etc. Pattern recognition is mainly categorised according to the type of learning technique which is present through which output can be generated .In pattern recognition identification of a pattern can be done basically on following two ways-

1) **Classification** –To define the object in pre-define classes.
2) **Clustering**- To group the object having same feature.

Pattern recognition algorithms are based on probabilistic and non-probabilistic approach. In probabilistic approach statistical method is used to generate single output to find the best case for a particular pattern which provides the best method as a solution [2]. Non-probabilistic approach has many methods as their solutions.

This approach provides N-best case output with associated probabilities for a given value of N instead of providing single best case for a particular problem. It is more advantageous to use probabilistic algorithm as compared to non-probabilistic algorithm due to following reason firstly probabilistic algorithm provide a mathematical value which is grounded in probabilistic theory whereas non-probabilistic algorithm does not provide refined and specific output. Secondly probabilistic algorithm are more effectively incorporated into large machine task as compare to non – probabilistic methods. Pattern Recognition is an applied as a medium in following machine learning techniques.

A. Unsupervised Learning

Unsupervised learning algorithm aims to create group or subset of data where data point belonging to a cluster are similar to each other as possible, while making the difference between the clusters as high as possible[3][4].

A classic example is the clustering of the customers by their demographics. The learning algorithm may help you to discover distinct groups of customer by region, age, ranges, gender and other attributes in such way that marketing

program can be targeted. The principle of working of unsupervised learning is outlined in Fig 2.

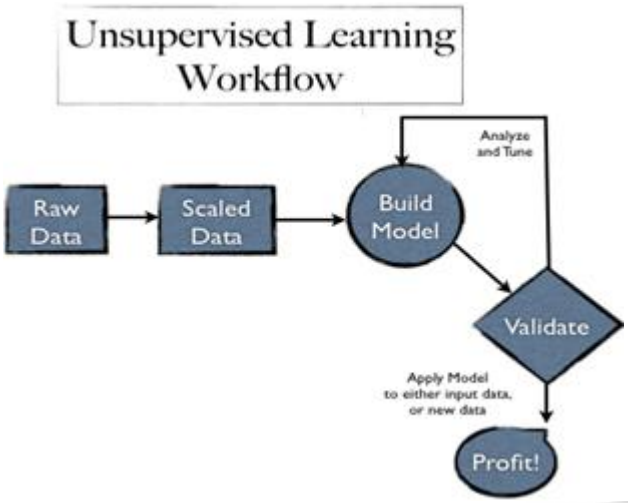


Figure 2: Unsupervised Learning

As is evident there are 4 main steps of Unsupervised learning, Firstly to scale and prepare raw data, Secondly to build model, thirdly to validate and satisfy the with cluster which is created, Fourthly once satisfied with cluster created there is no need to run the model the raw data.Unsupervised learning is of two types.They are discussed as below:

Clustering – Clustering is a techniques for finding similar groups in data called clusters. It groups data instances that are similar to each other in one cluster and data instances that are very different from each other
are grouped into different cluster. This process of grouping a set of pattern into classes of similar objects is called Clustering.

Blind Signal Separation- It is a process of separating a signal or a single data from a mixed signal without having information about the source of data [5].

B. Supervised Learning

In Supervised learning algorithm as an author consider set of data points or observation for which we know the desired output, class, target variable or outcome. The outcome may take one of many values called classes or labels. It consist of predetermined classes, these classes are the finite set of data .A certain segment of data will be labelled as b classification .The main task is to search for patterns and construct the mathematical models. The training chiefly consist of unlabelled data [3][4]. The principle of working of supervised learning methods is outlined in Fig 3.

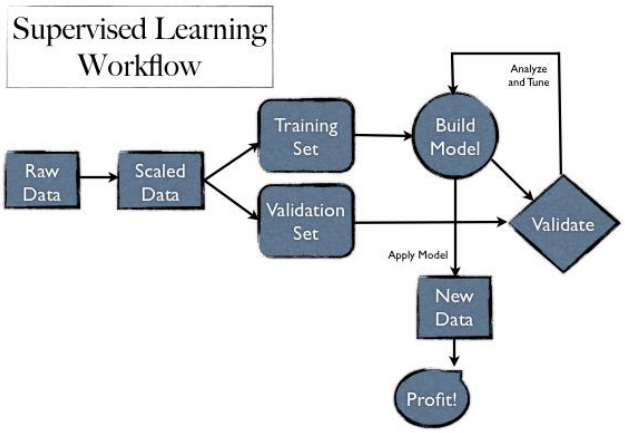


Figure 3: Supervised Learning

As is evident there are 4 main steps of Supervised learning, Firstly to scale and prepare Training Data, Secondly creation of training set and then validation of set by randomly splitting the universe of data, thirdly train the model, Fourthly validation and timing and Lastly, validate the model"s performance. Supervised Learning can be classified into following two types-

Classification –A classification problem is considered when the output variable is a category, such as "present" or "absent" .This type of learning is used when we have to classify an entity or pattern.

Regression – A regression problem is when the output variable is a real value, such as "dollars" or "weight" [6].

C. Semi Supervised Learning

This type of learning can be consider as a learning with features of both supervised and unsupervised learning .In this type of problem the input data is in huge amount and only some of the data is labeled. Many real time problem falls under semi-supervised category as it is not efficient and its time consuming to label each part of machine and it becomes easy to work with unsupervised learning and it became simple to use structure of supervised learning method due to this reason real world problems are categorized under semi-supervised learning. A classic example is a photo archive only some photos are labelled and rest are unlabelled

3. Clustering

Clustering is defined as an "act of differentiating the unlabeled dataset with the labeled dataset ".It can also be define as a phenomena in which we group all dataset which share some common characteristic or in some way they are similar to each other .It is one of the most important data processing technique which is used in many different biological and computer field .Some popular areas where it is used are as follows – artificial intelligence, bio-informatics ,computer - vision, city planning ,data-mining ,data-compression and many others.

There are some properties which each formed cluster should posess.These properties are define as follows-
1) Low Inter –Class Similarity
2) High Intra -Class Similarity
Clustering occurs mostly in unsupervised learning.it also help in gaining overall distribution of pattern and correlation among data object [2].

A. Data Clustering

It is also known as cluster analysis. Cluster analysis is a group of multivariate technique whose main goal is to group abstract object under a specified class. Data clustering simply mean grouping object based upon attributes that makes object similar in such a manner that if graph is plotted geometrically then object within cluster will be close together while distance between clusters will be far apart. The main use of cluster analysis is for taxonomy description, data simplification, relationship identification etc. [2].The data-clustering can be explain through following figure-

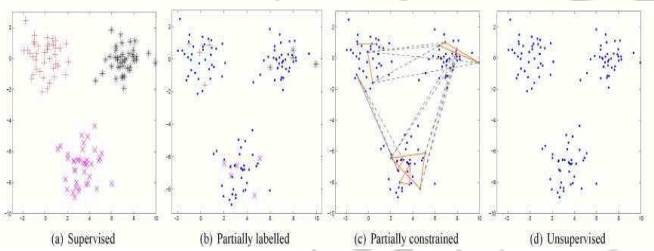


Figure 4: Fig. 1: Learning problems: dots correspond to points without any labels. Points with labels are denoted by plus signs, asterisks, and crosses. In (c), the must-link and cannot-link constraints are denoted by solid and dashed lines[9].

B. Definition of Cluster

The primary objective of cluster analysis is to define the structure of the data by placing the resembling object into groups .This groups are known as cluster. The simple rule for forming cluster is to identify the two most similar observation not already in the same cluster and combine them. In order to explain cluster in brief consider following example. Consider a representation of „n‟ object ,find K groups based on a measure of similarity .These object share the similarities between object through the figure it is clear that in one group similarity among object are high while in another similarity is low .Fig 5 depicts that cluster can differ in term of their shape, size and density[2].

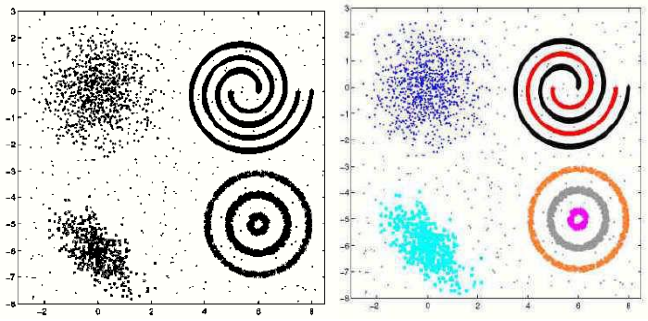


Figure 5: (a) Input Data (b) Desired Clustering [2]

C. Types of Cluster

Cluster formation are based on many feature such as shape, size, distance-proximity, pattern variation etc. Cluster which are formed are categorized in following types –
1) Connectivity Based Clustering
2) Centroid Based Clustering
3) Distribution Based Clustering
4) Density Based Clustering
5) Grid Based Clustering
6) Spectral Based Clustering

1) Connectivity Based Clustering: This type of clustering is also known as Hierarchical Based Clustering. In this type of clustering objects are clustered according to the distance between objects. A cluster is described by the maximum distance needed to connect part of cluster. In this after a certain distance different cluster are form which are represented through dendogram. In dendogram, y-axis represent the distance at which cluster merge while along x-axis objects are placed. In connectivity based cluster not only distance but linkage criteria is also consider. Based on linkage clustering that is single-linkage, complete-linkage clustering, discrimination is done. When the minimum distance object is consider it is called single-linkage cluster and when the maximum of object distance is consider is called complete linkage [5].

This technique works as a great tool for decomposition of metabolic network into functional modules based on global connectivity structure of reaction graph. In this modular network are arranged in modular hierarchical manner and then rational decomposition of metabolic network into relatively subset are analyzed and decompose. This technique is also

used for analyzing and validating the post-genomic data and also decomposition of metabolic network into functional module based on global connectivity. In order to depict the working of hierarchical clustering algorithm Figure-6 has been used

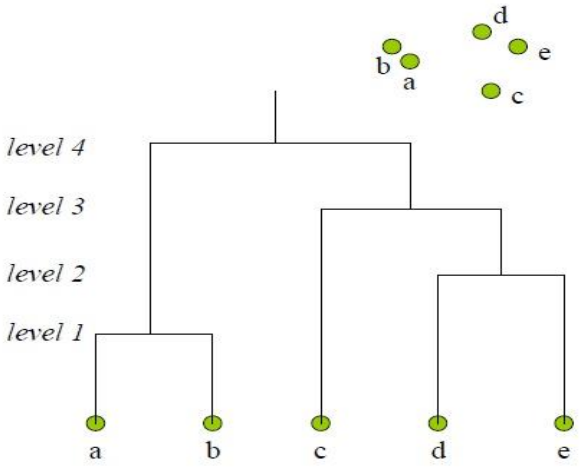
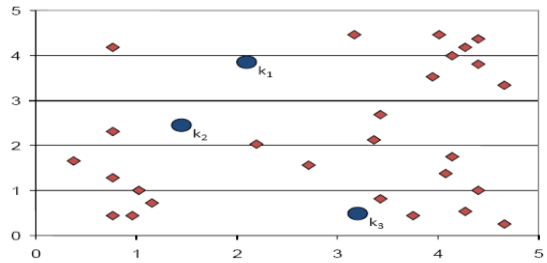


Figure 6: Hierarchical Clustering

2) Centroid Based Clustering- - It is most commonly known as K-Mean Clustering. In this type of clustering clusters are represented by a central vector which may not necessarily be a member of the data –set .This type of clustering can be define "as an optimization problems" find the k cluster centres and assign the object to nearest cluster centre such that squared distance from cluster are minimized .Lloyd"s Algorithm comes under this type of clustering .This approach is not acknowledge because according to this approach we have to predefine number of cluster which is not practically problem in every real world problem. In order to depict the working of k-mean clustering algorithm consider Figure-7.With the help of centroid based clustering technique sequencing of cluster is done. In this sequencing method alignment comparison are done inorder to challenge known sequence cluster. It is used to analyse and compare the different student performance [6].



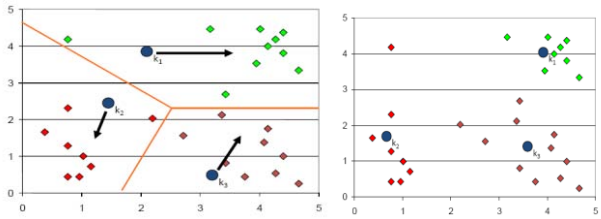


Figure 7: K- means Clustering

3) Distribution Based Clustering- This type of clustering is related to statistics. It is based on distribution model. This is the easiest approach for formation of cluster, in distribution based clustering object which closely resembles are arranged. Then resemblance of data is done by sampling random object from a distribution. Gaussian mixture model is an example of this clustering approach. This clustering technique is used to define, analyse and display the whole genome and also used to represent and analyse the DNA pattern and also the different protein pattern [5][6].

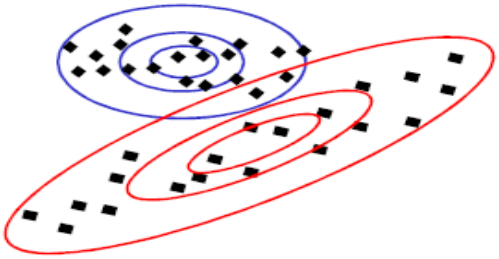


Figure 8: Gaussian mixture model [7]

4) Density Based Clustering- In this type of clustering approach clusters are defined as areas of higher density than the remainder of data-set .It plays an important role for cluster formation where inputs are in non-linear shapes. This approach form cluster based on the density of the input structure. DBSCAN is the most widely used density based algorithm. It uses concept of density reachability and density connectivity. This clustering technique is used to understand the principle of cellular organization and also the function of cellular organization can be enhanced .It is also used to detect and predict still undiscovered protein complexes within the cell"s protein-protein interaction network. This provide an accurate and scalable approach to protein complex identification [8].

a) Density Reachability - A point "p" is said to be density reachable from a point "q" if point "p" is

within ε distance from point "q" and "q" has sufficient number of points in its neighbors which are within distance ε.In order to depict the complete working of density reachability consider Figure-8(a) .

b) Density Connectivity - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p". In order to depict the complete working of density reachability consider Figure-8(b)

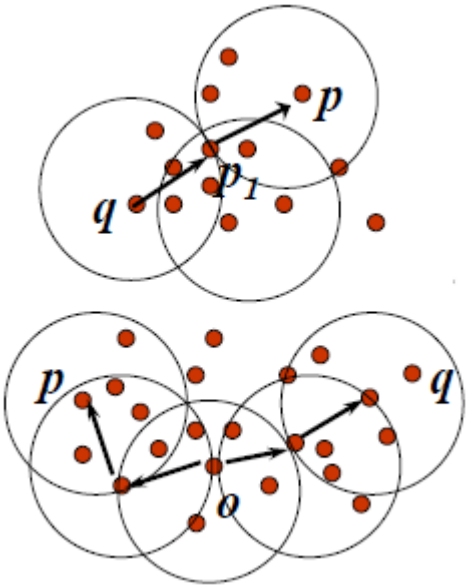


Figure 9: a) Density Reachability b) Density Connectivity

5) Grid Based Clustering: This differs from other conventional based algorithm. Conventional based algorithm are based on distance calculation while grid based algorithm does not organize the pattern instead it depend on the values space which surround the pattern .To organize the value space a variation of the multi-dimensional data structure of the grid –file is used which is known as grid structure. Sting/Clique Algorithm is based on this approach. It is a fully objective method for defining most relevant interaction areas in complex deriving pharmacophore model from 3 Dimensional molecular structure information .It is based on logical and clustering operation with 3 dimensional map computed by the grid program on structurally known molecular complex [8].

6) Spectral Clustering: This technique make use of spectrum of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimension. Similarity matrix is consider as an input and it consist of a quantitative assessment of the relative similarity of each pair of points in the dataset .They are also known as segmentation based categorization. Eigen Value Algorithm is based on this approach [7][8].

4. Clustering Algorithms [7][8]

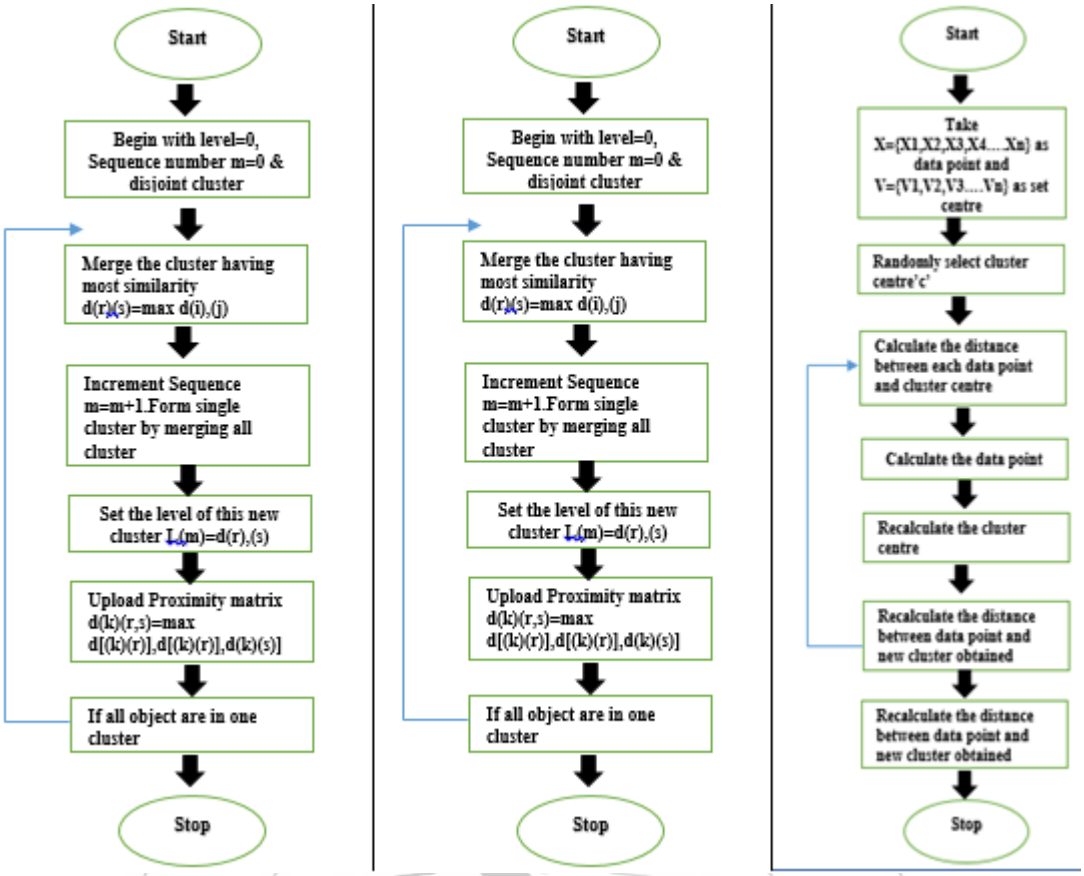


Figure 10: a) Single Linkage Algorithm b) Complete Linkage Algorithm c) K-means Clustering

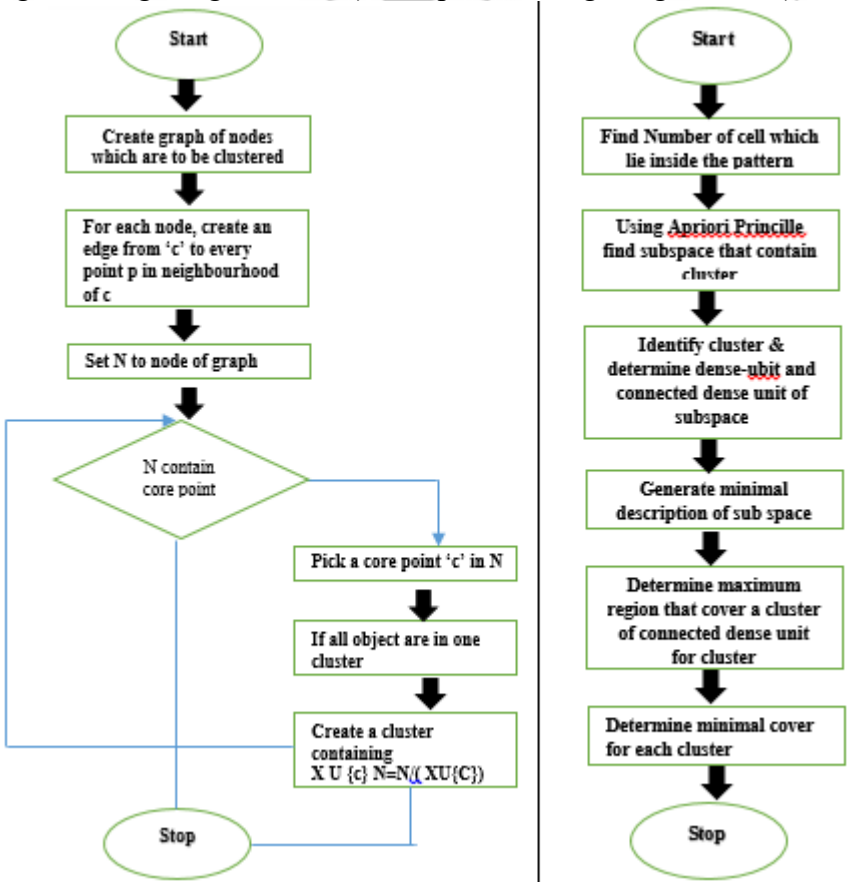


Figure 10: a) DBSCN b) STING

Table 1: Summarization of Important Clustering Algorithm

Name of Algorithm	*Type of Clustering*	*Working*	*Advantage of Algorithm*	*Disadvantage of Algorithm*	*Complexity of Algorithm*	*Application In Bioinformatics [17]*
Single Linkage	Connectivity Based	It is a bottom up approach. It is used to form cluster when input pattern are non-global in shape	1) They can handle non-global shapes input.	They are sensitive to noise and outliers. Problem Of chaining occurs.	O(N^2)	It is used to determine the different trait from genes in the different trait from genes
Complete linkage	Connectivity Based	This technique is used to form cluster when there is a huge distance between the input patterns.	Less susceptible to noise and outliers. Robust in nature.	They tends to break big cluster.	O(N^2)	It is consider as a best technique in order to analyse lung and breast cancer using join latent variable model.
K-Mean clustering	Centroid Based	It is an unsupervised clustering algorithm which is used in formation of cluster.	1)If input are great in number then it require less time for cluster formation 2) Strong cluster are formed in comparison to connectivity based cluster.	1) Difficult to depict value of k. 2) Does not work well for global cluster. 3) Does not give appropriate cluster if input is of different density.	O(KN)where k is number of cluster and n is number of points.	It is used as a best technique to improve protein structure prediction by merging large number alternative model.
Gaussian Mixture Model	Distribution Based	It is an iterative approach for forming of cluster. Through huge input with different density can be clustered.	1)They are flexible in nature 2) They can handle input of different data type.	Its time consuming	O(nlogn)	It is used to display and analyse whole genome DNA microarray expression data .It is also used for protein analysis through DNA
DBSCAN	Density Based	It is a density based approach .In this cluster is form depending upon the density of input pattern.	1.It is robust, not effected by noise 2.Does not require to specify number of cluster	1. It is non-deterministic. 2.Cnnot be used when inputs are with large difference density	O(logn)	It is used to analyse DNA structure.
STING/CLIQUE	Grid Based	It works by randomly arranging input pattern into a grid and then forming cluster.	1.Insensitive to the order of input 2. Scales linearly with size of inputs.	Sometime cluster form is inaccurate	O(n)	It is used for representation of genes through directed graph for protein analysis and other component of DNA
EIGEN VALUE	Spectral Based	It uses spectral range in order to form cluster.	1. They are easy to understand.	Time consuming and accuracy of cluster form is not certain.		It is used to determine the vibration, shape, function and structure of different nodes represented through graph.

5. Conclusion

An overview of different types of clustering techniques and also the algorithm, it"s been clear that each and every designed algorithms are designed for a special purpose and are unique in their own way. By studying these following points are clear: Firstly If the input pattern are small in size and dissimilar in shape so we can get best cluster in less time is from single linkage clustering as it is used to form cluster when the input patter are not same. If the distance between the input patterns is large then we use complete linkage clustering .If the input pattern varies through density then the density based approach should be follow in order to get an efficient cluster .DBSCAN cluster algorithm approach is used to form cluster. One can also use Gaussian Mixture Model Algorithm but this algorithm takes little time consuming than DBSCAN and it work efficient for the case where input pattern are huge in number and also vary in density.

Once one has prior information about the number of cluster to be follow and the input pattern are in regular fashion then centroid based approach is used to form a cluster. Also K-mean clustering algorithm inorder to form a cluster. If the input pattern are arrange randomly then we can use STING/CLIQUE algorithm to form cluster which is based on grid based approach.

6. Acknowledgment

The author thank the Dept. of Biotechnology, Himalayan University and Dept. of Computer Science & Engineering, SRMGPC, Dr APJ Abdul Kalam Technical University. No funds and grants were used for this work.

References

[1] A Comprehensive Overview of Clustering Algorithms in Pattern Recognition: Namratha M,Prajwala T R(Dept. of information science, PESIT/visvesvaraya technological university, India)
[2] Pattern Recognition Algorithms for Cluster Identification Problem: Depa Pratima CMR Institute of Technology, Hyderabad, India & Nivedita Nimmakanti CMR College of Engineering & Technology, Hyderabad, India E-mail : pratima.depa@gmail.com, nimmakantinivedita@gmail.com
[3] http://www.springer.com/computer/image+processing/book/978-0-387-31073-2
[4] http://www.scribd.com/document/66685419/Machine-learning-and-PR-Approaches
[5] Jean-Francois Cardoso "Blind Signal Separation: statistical Principles"
[6] Clustering 15-381 Artificial Intelligence Henry Lin Modified from excellent slides of Eamonn Keogh, Ziv Bar-Joseph, and Andrew Moore
[7] http://documentslide.com/documents/cluster-analysis-56a4a2f57990f.html
[8] http://www.cs.bu.edu/fac/gkollios/ada05/LectNotes/lect27-05.ppt
[9] Black hole: A new heuristic optimization approach for data clustering Abdolreza Hatamlou Islamic Azad University, Khoy Branch, Iran Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology, University Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia