International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 11, Issue 10 (October 2015), PP.55-60

Study on Clustering of Data

Monika Singh[1]
Department of Computer Science

Abstract:- Clustering can be defined as the unsupervised classification of patterns (observations, data, or feature vectors) into groups (clusters).The main objective of clustering is to find similarities between any given data and use these similarities to assist in understanding the relationship between the sample data. In this paper, a brief reference to machine learning and description of supervised and unsupervised learning is given. Various approaches to clustering data and cluster analysis are discussed. Each approach has its own pros and cons. The taxonomy of clustering, different methods of clustering and validity indices that determines how compact and well separated the clusters are is also discussed. The application of clustering techniques is also discussed in brief.

Keywords:- unsupervised learning, hierarchical clustering, K-means, K-mediods, density based clustering, fuzzy clustering, validity indices

I. INTRODUCTION

Clustering is unsupervised learning. A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. This means that if any program or system starts to learn from a given set of data and apply it on a new data, and its performance increases, then that can be defines as learning. There are two types of learning- Supervised and Unsupervised learning.

In supervised learning, we assign labels with each example in a training set of data. We have to generalize the properties of the training set to be able to make predictions about data that is not yet explicitly given. There are various conditions that we have to keep in mind in supervised learning. For example, how to divide training data to minimize training error, number of training data to make it possible to come to a generalization, the accuracy of labels, issue of feature extraction or similarity index, the past data being the correct representative of the future data and so on.

In unsupervised learning, we have training data but we do not have labels. Here, we have to group unlabelled data or patterns into meaningful groups. The labels are associated with the groups or clusters but these labels depend only on the given data which is not the case in supervised learning. The labels are data driven.

This paper focuses on unsupervised learning/ Clustering. It provides an overview of various types of clustering, their pros and cons, validity indices used to determine the compactness of clusters and application of clustering algorithms.

The rest of the paper is organized as follows: Section II describes the various stages in clustering. Section III the taxonomy of clustering and various approaches for clustering .Section IV describes various clustering methods and their algorithms. Section V describes the use of validity indices and their importance in comparing the clustering techniques. Section VI concludes the paper.

II. STAGES IN CLUSTERING

[1]There are four stages in clustering: Pattern representation, feature extraction/selection, interpattern similarity and grouping. Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. Feature selection is the process of identifying or selecting the most effective subset of the original features to be used in clustering. Feature extraction is the using one or more transformations of the input features to produce new identifying features. One or both of the techniques described above can be used to obtain an appropriate set of features to use in clustering. Pattern proximity is measured by a distance function defined on pairs of patterns.

There are various similarity measures such as Minkowski metric and variance.
Minkowski metric is defined as
$dist(X1,X2,p)= {}^{len}\Sigma_{k=1} (abs(X1k-X2k)^p)^{1/p}$ which is the distance between two points X1 and X2.

p = 1: Manhattan Distance
p = 2: Euclidean Distance
Variance is defined as: Variance(C) =$\sum X \in C$ (Mean(C)-X)^2

We have to mention the constraints for the optimum number of clusters because if there are no constraints and we want to know the maximum number of clusters then each element is considered as a cluster as the variance is zero.

III. TAXONOMY OF CLUSTERING

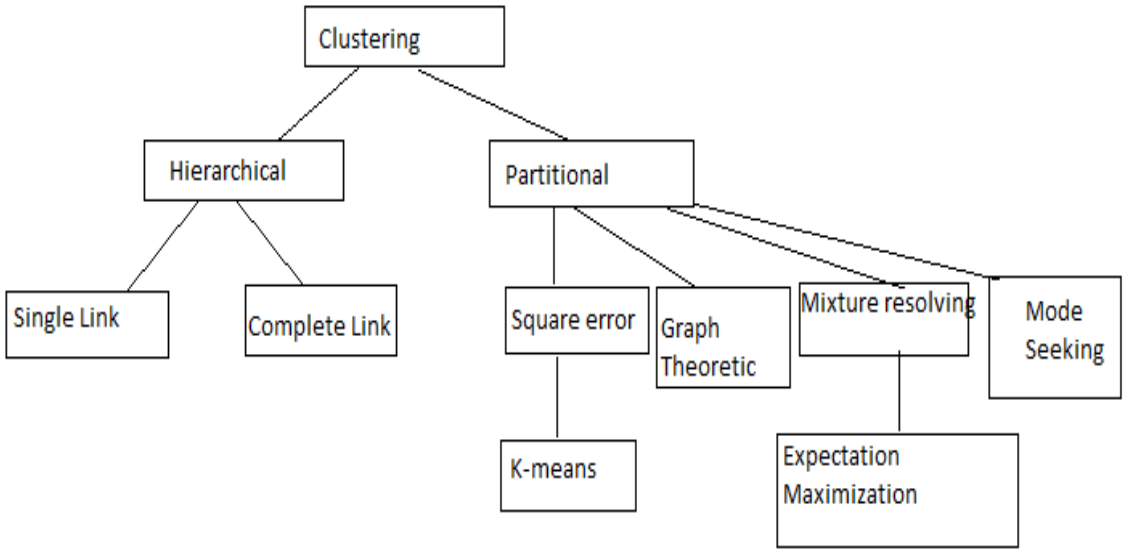

Fig.1 Taxonomy of Clustering [1]

Types of clustering discussed in the paper are:

- Agglomerative clustering and divisive clustering- An agglomerative approach begins with each pattern in a separate cluster, and then successively merging clusters together until a terminating criterion is satisfied. It follows a bottom up approach. Divisive clustering starts with all patterns/data items in a single cluster and performs splitting successively until a stopping criterion is satisfied. It follows a top down approach.
- Hard and Fuzzy clustering –A hard clustering algorithm allocates each pattern to a single cluster. The fuzzy clustering approach assigns degree of membership in several clusters for each pattern.

IV. TYPES OF CLUSTERING

A. Hierarchical clustering[2]

In data mining, hierarchical clustering aims at building hierarchy of structures. There are two type of hierarchical clustering: Agglomerative and Divisive clustering. Agglomerative clustering is a bottom up approach in which pattern starts with its own cluster and pairs of clusters are merged as one moves up the cluster. The divisive approach is a top down approach. All the patterns are in one cluster and move recursively as one moves down the cluster. The complexity of agglomerative clustering is O (n^3) whereas complexity for divisive clustering is O (2^n). Hierarchical clustering id represented using a dendrogram.

Agglomerative clustering can be further divided into single linkage and complete linkage agglomerative clustering. The operation of a hierarchical clustering algorithm is illustrated. There are eight patterns labelled A, B, C, D, E, F, G and H in three clusters. The dendrogram represents the iterative grouping of patterns and similarity levels at which the groups change.

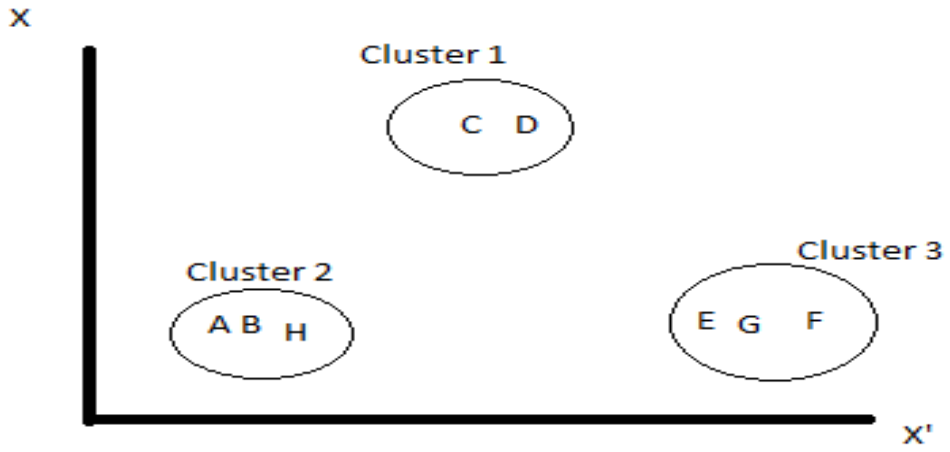

Fig 2. Clustering of data

This means that the hierarchy of clustering follows as:
Level 0: A, B, C, D, E, F, G, H are separate clusters in the beginning.
Level 1: A, (B, H), C, D, E, F, G. Now B and C are in one cluster.
Level 2: A, (B, H), C, D, E, F, G. Now C and D join to form a single cluster.
Level 3: A, (B, H), (C, D), E, F, G. Now based on similarity, F and G combine to form one cluster
Level 4: A, (B, H), (C, D), E, (F, G). Now A and (B, H) combine.
Level 5: (A, B, H), (C, D), E, (F, G)
This goes on until only one single cluster is left i.e. (A, B, C, D, E, F, G, H). This is how Agglomerative clustering works. Each cluster combines with another cluster based on some similarity (feature that we initially selected to be the measure of similarity).

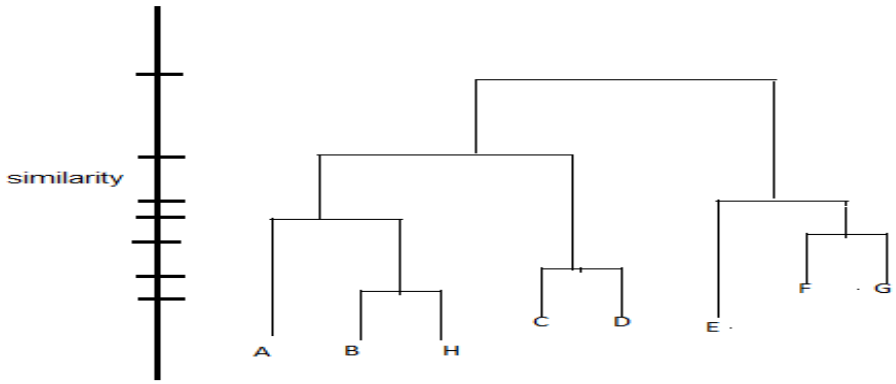

Fig 3. A dendrogram representing agglomerative hierarchical clustering

Divisive Clustering is just the opposite of agglomerative clustering. All the objects are initially in one cluster. Then we start based on a selected feature, we determine the similarity levels and split the clusters into groups. This process continues until all the objects form separate cluster. It is a top down approach.

B. K-means Clustering[3]

K-means clustering is another method of cluster analysis in which we partition n observations into k clusters and in this each observation belongs to the cluster with the nearest mean. It is the simplest method of unsupervised learning that classifies the data set in a given number of clusters which fixed beforehand. The main idea is to define k centres, one for each cluster. These centres are random and should be chosen such that they are farther from each other. Then the next step is to take each point belonging to the dataset and associate it to the nearest centre. When all points are associated with one of the centres, we recalculate k new centroids of the new clusters formed. After we have these k new centroids, recalculate the distance between the points and the new centre. A loop has been generated. The k centres change their location after each step until no more changes are done or centres do not move anymore i.e. no more data points are reassigned. This algorithm aims at minimizing an objective function knows as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\|x_i - v_j\| \right)^2$$

Where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centres.

We calculate the new centre for each cluster using the formula below:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, 'ci' represents the number of data points in i^{th} cluster.

The advantage of K means clustering is that it is fast, robust and easier to understand. But we need to determine the number of clusters beforehand. It is applicable only when mean of data is defined. It is not applicable for categorical data.

C. K-mediods Clustering[2]

K-mediods algorithm is a clustering algorithm which is related to k-means clustering. Similar to k-means, we determine the number of clusters beforehand. It also minimizes the distance between points belonging to the same cluster. A mediod can be defined as a point inside the cluster whose average dissimilarity to all the other objects in the clusters is minimal. The most common algorithm for k –mediod is Partitioning around mediods (PAM). In this instead of randomly selecting points in the beginning we select k points from the n data objects to be the mediods (k is the number of required clusters and n is the number of data objects). For each mediod selected, we swap the mediod with a data point and calculate the cost of configuration. Then we select the configuration with the lowest cost. This process is repeated and a loop is generated which stops when there is no change in the mediods.

Cost is calculated in terms of Euclidean, Manhattan distance or Minkowski distance.

$$\text{cost}(x, c) = \sum_{i=1}^{d} |x_i - c_i|$$

D. Density Based Clustering[6]

Density based clustering is most importantly used in finding non linear structure based on density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most commonly used density based clustering algorithm. There are two terms that need to be defined while understanding density based clustering i.e. Density reachability and density connectivity.

Density Reachability: Any point "a" is said to be density reachable from a point "b" if point "a" is within ε distance (minimum distance mentioned beforehand) from point "b" and "b" has appropriate number of points as its neighbours which are within ε distance.

Density Connectivity : A point "a" and "b" are said to be density connected if there exists a point "c" which has sufficient number of points in its neighbours and both the points "a" and "b" are within the ε distance. So, if "b" is neighbour of "c", "c" is neighbour of "d", "d" is neighbour of "c" which in turn is neighbour of "a" implies that "a" is neighbor of "b".

Some terms used in Density based clustering are:

Core points: A core point is a point which has a specified number of points within a given distance. These points make the interior of the cluster.

Border point: A border point is a point that lies in neighbourhood of a core point but itself is not a core point.

Noise: Any point that is neither a core nor a border point is termed as noise.

Density based clustering works on the basic principle of connecting points that satisfy the density criteria i.e. the minimum number of points within a given radius. Also another noteworthy point is the low complexity of density based clustering.

In density based clustering, we start with an unvisited point, if the point is found to be a core point then it is marked as visited else it is marked as noise. If the point selected is a part of a cluster then its ∈ neighbourhood points are also the part of the cluster and the process is reiterated until all the points in the cluster are determined. Then again a new unvisited point is taken and the same process is repeated until all the points

are marked visited. The advantage of this approach is that we do not require determining the number of clusters in the beginning, we able to separate noise from the clusters and it also is able to find non linear shaped clusters. The disadvantage is it does not work with high dimensional data as well as when the dataset is neck type.

E. Fuzzy Clustering[4][5]

All the clustering algorithms before this point were assigning data points into one cluster i.e. each data point is a member of one single cluster. This is known as hard clustering. Fuzzy Clustering is also known as soft clustering because in this data points belong to a number of clusters simultaneously. In this type of clustering, data can belong to more than one cluster. The data object has a degree of membership with clusters i.e. the strength of association of the data object with the cluster. In fuzzy clustering, we assign membership to each data point corresponding to each cluster center on the basis of distance between the data point and the centre of the cluster. The membership increases with an increase in the closeness to the centre of the cluster. Clearly, summation of membership of each data point should be equal to one. After each repetition, membership and cluster centers are updated according to the formula:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

Where,
n is the number of data points.
'vj' represents j^{th} cluster center.
'm' is the fuzziness index m $\in [1,\infty]$.
'c' represents the number of clusters.
'µij' represents membership of i^{th} data to j^{th} cluster.
'dij' represents the Euclidean distance between i^{th} data point and j^{th} cluster centre.
Main objective of fuzzy c-means algorithm is to minimize:

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \left\| x_i - v_j \right\|^2$$

Where '||xi – vj||' is the Euclidean distance between i^{th} data and j^{th} cluster center.

The algorithm randomly selects 'c' clusters and calculates fuzzy membership using

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

Then compute the fuzzy centres using the given formula

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, ...c$$

:

Repeat the process until the termination criterion is met i.e. 'J' value is achieved or ||U (k+1) - U (k) || < β.
Where
k is the iteration step.
'β' is the stopping criterion between [0, 1].
'U'= (µij) n*c' is the fuzzy membership matrix.
'J' is the objective function
The use of fuzzy clustering is better in case of overlapped data.

V. VALIDITY MEASURES FOR CLUSTERING

[7]There are two important criteria for evaluating an optimality of a clustering algorithm. They are as follows:-
1. Compactness: The members of each cluster should be as close to each other as possible. Variance which is a common measure for compactness should be minimized. The clustering algorithm should group data points/pattern such that the intracluster similarity should be maximized.
2. Separation, the clusters themselves should be widely spaced. The clusters should be so formed that the inter cluster similarity is minimized. There are three most widely used ways to measure distance between two clusters:
• Single linkage: It measures the distance between the closest members of the clusters.
• Complete linkage: It measures the distance between the farthest members.

• Centroid comparison: It measures the distance between the centres of the clusters.

There are various indices defined to evaluate the optimality of clusters. Some of them are as follows:

1) Dunn Index
2) Davies Bouldin Index
3) Calinski Harabasz index
4) Silhoutte index
5) PBM Index

VI. CONCLUSION

This paper focuses on providing an overview on Clustering of data and various clustering algorithm developed till now. Various approaches to cluster data have been discussed and the validity measures for clustering have also been discussed in brief.

ACKNOWLEGEMENT

It is my pleasure to get this opportunity to thank my beloved and respected teachers who imparted valuable knowledge specifically related to unsupervised learning and clustering. We are grateful to my friends for providing me moral support.

REFERENCES

[1]. A.K. Jain, M.N. Murty, P.J. Flynn, "Data clustering: A review", ACM Computing Surveys, Vol. 31, No. 3, September 1999.
[2]. Glenn Fung, "A Comprehensive Overview of Basic Clustering Algorithms"
[3]. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, et. Al, "An Efficient k-means Clustering Algorithm: Analysis and Implementation".
[4]. Balaji K and Juby N Zacharias, "Fuzzy c-means"
[5]. Weiling Cai, Songcan Chen and Daoqiang Zhang, "Fast and Robust Fuzzy C-Means Clustering Algorithms Incorporating Local Information for Image Segmentation".
[6]. https://sites.google.com/site/dataclusteringalgorithms
[7]. web.itu.edu.tr/sgunduz/courses/verimaden/paper/validity_survey.pdf
[8]. en.wikipedia.org/wiki