Chapter Publication

Domain: Data Mining and Warehousing

Subdomain: Data Preprocessing

Author By:

Mrs. Nilam S. Patil

Assistant Professor,

D. Y. Patil College of Engineering Akurdi Pune,

Teaching Experience: 17+ Years

PhD Pursuing From Vel Tech University, Chennai

All data related to this chapter will copyright@aspirepublishers.com
Publication Date: December 2018
Web: www.aspirepublishers.com
Email: info@aspirepublishers.com

INDEX

List of Figures

CHAPTER 1

CLUSTERS ANALYSIS

1.1 WHAT IS CLUSTER ANALYSIS

Cluster analysis groups data objects based only on information found in the data that describes objects & their relationship. The goal is that objects within a group be similar to one another and different from objects in another group.

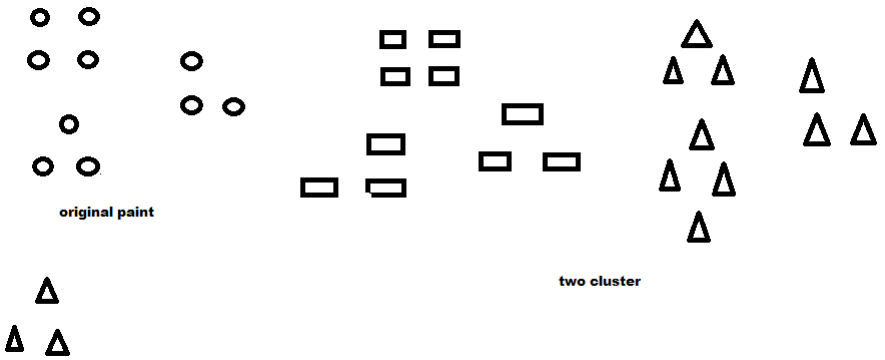

Figure 1.1: Group of Cluster

-Clustering can be regulated as form of classification in that it creates a labeling of objects with class (cluster) labels.

- Clustering is unsupervised classification.

- It is Called as segmentation and partitioning.

Ex:- Image can be split into segments based on pixel intensity and color.

1.2 TYPES OF CLUSTERING:

1.2.1 Hierarchical vs Partitional Clustering :

1.2.1.1 Partitional Clustering :

Division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset.

Ex:- Above fig. shows clusters

1.2.1.2 Hierarchical:

- Clusters have sub clusters which is set of nested clusters organized as tree.

- Each node (except children) is union of its children (sub clusters) and root of tree is a cluster with all objects.

- In above example (a-c) in cluster is hierarchical clustering with respectively 1,2,4 clusters at each level.

1.2.1.3 Exclusive vs overlapping vs Fuzzy

- Clusters shown above fig. are all exclusive as they assign each object to a single cluster.

If a point is placed in more than one cluster then its nonexclusive or overlapping

Ex:- person at university both enrolled as student and employee.

1.2.1.4 Fuzzy:

Every object belong to every cluster with a membership weight that is between 0 and 1 clusters are treated as fuzzy sets.

1.2.1.5 Probabilistic

It compute possibility with which each point belong to each other and probability must also sum to 1. Probabilistic cluster are converted to exclusive.

1.2.1.6 Complete vs Partial

A complete clustering assigns every object to a cluster

Partial: Does not design every object to other many time objects in the data sets represent noise, outliers or un-intresting Background.

Ex:- some newspapers share common themes while other stories are more generic . Thus to find the important topics in the last months stories , we may want to search cluster of documents that are tightly to common themes.

1.2.2 Different types of clusters

1) well-separated :

- Each point is cluster to all other points in its cluster than to any other point in another cluster.

- The distance between any two point in different group is larger than the distance between any two points in the same clusters.

- They can have any shape.

2) Prototype based:

A cluster is set of objects in which each objects is closes to the prototype that defines the cluster than to the prototype of any other cluster.

For continuous data , prototype is centroid i.e; the average (mean) of all the points in the cluster For categorical attribute. Most central point is used as prototype, there for also called center based cluster **3) Graph based :**

Figure 1.2: Well seprated and prototype cluster

Data is represented as graph where all the node are objects if links shows the connection on many objects. Cluster is called connected component i.e; group of objects connected to one another.

Ex:- contiguity based cluster shown below : Each part is closes to at least one point in its cluster than any other in another cluster.

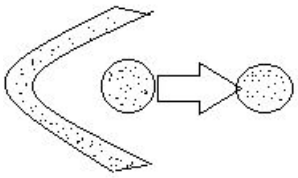

Figure 1.3: graph based cluster

Clique: set of nodes in a graph that are completely connected to each other.

4) Density based :

A cluster is dense region of objects that is ssurrounded by a region of low density. Region of high density are separated by region of low density

5) shared -property (conceptual clusters)

A cluster as set of objects that share same property.

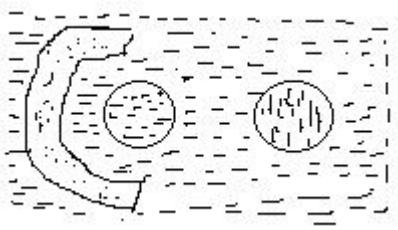

Figure 1.4: Density based cluster

Ex. In centroid based clustering property is centering finding such clusters is called conceptual clustering used in pattern recognition.

Figure 1.5: Conceptual based cluster

1.2.3 Techniques:

1) K-means: prototype based , partitional clustering technique that attempts to find uses speech number if cluster (k) which are represented by their centroid.
2) Agglomerative : Hierarchical clustering : storing each object as a singleton cluster then repeatedly merging the two clusters until a single, all encomposing cluster remaining.
3) DBSCAN : Density based clustering algorithm that produces partitional clustering in which number of clustering is automatically determinable by the algorithm. Points in low density regions are classified as noise and omitted. Thus it does not produce complete clustering.

1.2.3.1 K-means:

0. Start with initial guesses for cluster centers (centroids)
1. For each data point, find closest cluster center (partitioning step)
2. Replace each centroid by average of data points in its partition

3. Iterate 1+2 until convergence Write $x_i = (x_{i1}, ...x_{ip})$:

   If centroids are $m_1, m_2, ...m_k$, and partitions are

   $c_1, c_2, ...c_k$, then one can show that K-means converges to a *local* minimum of

$$\sum_{k=1}^{K} \sum_{i \in c_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

   (within cluster sum of squares)

   In practice:

   - Try many random starting centroids (observations) and choose solution with smallest of squares

   How to choose K?

   - Difficult – details later

- All clustering algorithms start with a dissimilarity measure for j^{th} feature

$$d_j(x_{ij}, x_{i'j}) \text{ and define}$$

$$D(x_i, x_{i'}) = \sum_{j=1}^{P} d_j(x_{ij}, x_{i'j})$$

  Usually $d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$

Other possibilities:

- Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

  $\bar{x}_i$ = mean of observation i

- If observations are standardized:

$$x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_i}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2}}$$

then $2(1 - \rho(x_i, x_{i'})) = \sum_j (x_{ij} - x_{i,j'})^2$

So clustering via correlation $\equiv$ clustering via Euclidean distance with standardized features

Partitioning (Clustering) Algorithms

- Group assignment function ("encoder") $C(i)$

$C : 1, 2, ...N \rightarrow (1, 2, ...K)$

- **Criterion**: choose C to minimize

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

(within cluster scatter)

Fact:

- K-means minimizes $W(C)$ when $D = ||x_i - x_{i'}||^2$

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} ||x_i - x_{i'}||^2 \\ &= \sum_{k=1}^{K} N_k \sum_{C(i)=k} ||x_i - \bar{x}_k||^2 \end{aligned}$$

- K-means solves *enlarged* problem:

$$\min_{C, m_1 ... m_k} \sum_k \sum_{C(i)=k} ||x_i - m_k||^2$$

to find assignment function C

Strength And Weakness:

Strength-

1)Efficient, variables are more effective

2)Less susspecteble to initialized problem.

Weekness:1)Not sutable for all types of dats.

2)Can't handle clusterof diffrent size of densites

3)Problems if obtien present it.

Bisecting -means: To obtain k cluster , split the set of all points into two cluster ,select one of these clusters to split and so on, until k cluster have been produced.

Agglomerative Hierarchical clustering:

2 types of Hierarchical

1) Agglomerative:Start as a path as individual cluster and at each step mearge the closest pair of cluster

2)Divisive clustering:Stets with all inclusive cluster and at each step split a cluster until only singleton cluster of individual points remain. In this case ,we need to decide which cluster to split at each step & how to do the spliting

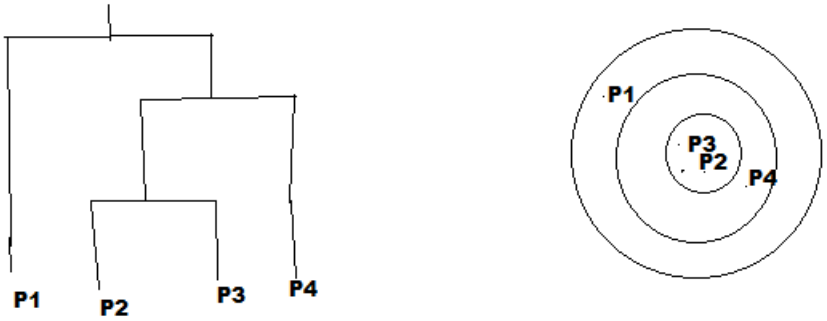

Figure 1.6: dendogram and nested cluster

Algorithm: Agglomerative 1:complete proximity matrix if necessary

2:repeat

merge the closest two clusters

update the proximity matrix to reflect

the proximity between new cluster and the original cluster

3:until only one cluster remain

Defining proximity between clusters. *min:proximity between closest, two points that are different cluster

max:proximity between farthest, that are in different cluster

average:cluster proximity to be average pairwise proximity of all pairs of points from different cluster

Time and space complexity:

space:O(m2) M:number of clusters

time:O(m2log m)

1.2.3.2 Wards method:

- Here proximity between 2 cluster is define as the increased in squared error that results when two cluster are merged
- Similar to group average

1.2.3.3 Centroid Method:

Calculate proximity between 2 clusters by calculating distance between centroids of clusters.

Issues in Hierarchical Clustering-

Objective function is not globally optimizing various criteria for merging decided locally. **Strength and Weekness:**

Strength: More general Technique. Typically used of underling application ex. creation of taxonomy requires hierarchy.

2 Better Quality clusters.

Weakness: Expensive in terms of storage requirement. Merging may trouble for noisy, high dimensional data such as documents.

1.3 DBSCAN:

Density based Clustering. It locates regions of high density that are separated from one another by regions of low density. It is Simple ands Effective.

1.4 INTRODUCTION

DBSCAN is an unsupervised clustering algorithm. As the name suggests, the key idea of this algorithm is based on how dense the data points are located. More details about the steps involved are discussed in the following sections.

Some other famous Algorithms are k-means clustering, Fuzzy c-mean, etc..

But the advantage of DBCAN being that it is more immune to noise, and the number of clusters are not fixed before the algorithm is run as in case of k-means.

1.5 ALGORITHM

Labelling convention:

- 0: Unlabelled
- -1: Noise
- 1: Cluster number 1

  ...
- k: Cluster number k

Procedure:

1. Index all points,and label all as '0'
2. foreach point:
   (a) If it is labelled already, goto next point.
   (b) Get points (neighbours) within 'ε' distance from the chosen point.
   (c) If # of neighbours $<$ 'minPts', label as NOISE(-1) and goto next point.
   (d) Else, label it as a new cluster (c_{new}).

(e) select the neighbouring points as a new set 'S', for each point in 'S':

i. If labelled as -1, relabel to new cluster number (c_{new}), and goto next point in the set.

ii. If point is already labelled, skip it.

iii. If it is unlabelled, then label it (c_{new}),
Get newighbours in ε boundary and if count $>$ minPts add to the set 'S'

iv. continue to next point in the set, if set is empty, the go back to step (a)

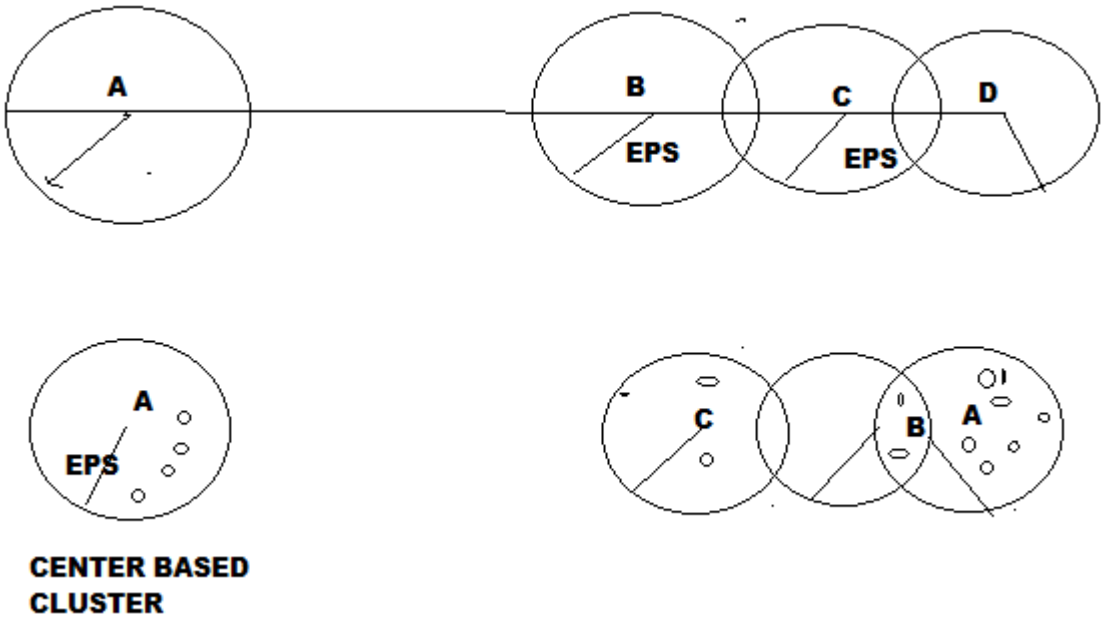

Figure 1.7: DBSCAN example

Algorithm Steps: Label all points as core border or noise points.

Eliminate noise points .

Put an edge between all core points that are within Eps of each other.

Make each group of connected core points into a separate cluster.

Addign each border point to one of the clusters of associated core points.Eps= radius

Classification of points: Core path: These are in the interior of a density based cluster. A point is a core point if the number of points within a given neighborhood around the pair as determine d by the distance function and a user specified distance function and a user specified parameter .

Border points: Its not a core point but it falls within neighborhood of core point. Point B is border point.

Noise point: Its niether a core core point nor border point.

Time and Space complexity: Time: O(mlogm) where m is number of points Space : O(m)

Strength and weakness: Strength: It is resistant to noise and can handle clusters of shapes and sizes. It can found many cluster that could not be found using k-means.

Weakness: It has trouble when cluster have widely varying densities.

In high dimensional data density os more difficult to define

It is expensive when computation of nearest neighbor requires all pairwise proximities.

1.6 CLUSTER EVALUTION

Since each clustering algorithm defines its own type if clusters, cluster evaluation is required. Cluster evaluation need not be part of cluster analysis .

Issue for cluster validation:

Determine cluster tendency of a set of data.

Distinguish whether non-random structure exist

Determine correct number of clusters

Evaluating how well the results of clusters analysis fits the data without references to external info. Comparing results of a cluster analysis of externally know results, such as externally provided class label.

Comparing two sets of cluster to determine which is better.

1.6.1 Evaluation of measures or indices applied to judge

1. Unsupervised: measure goodness of a clustering structure without respect to extend info. Cluster cohesion: compactness and tightness which determine how closely related the object into a cluster.

Cluster separation: Measure determine how distant or well separated clusters from each other. they are intend indices as they use into present in data set.

Supervised: Measures extent to which the clustering structure discovered by clustering algorithm matches same external structure / Ex. Entropy They are external indices as they do not use into present in the dataset.

Relative: Compares different clustering/ clusters both supervised and unsupervised. Ex. K means can be compared using SSE/Entropy

REFERENCES

Wikipedia, "https://en.wikipedia.org/wiki/DBSCAN"

https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/

https://home.deib.polimi.it/matteucc/Clustering/tutorial$_html$/