|| Jai Sri Gurudev |

Sri AdichunchanagiriShikshana Trust (R)

SJB INSTITUTE OF TECHNOLOGY

SJBIT

Study Material

Subject Name: Machine Learning

Subject Code: BCS602

By

Faculty Name: Mrs. Pavithra

Designation: Assistant Professor

Semester: VI

Department of Information Science & Engineering

Aca. Year: Even Sem /2024-25

MODULE 5

ClusterAnalysis:BasicConceptsandMethods-ClusterAnalysis,partitioningmethods, Hierarchical Methods and evaluation of Clustering

ClusterAnalysis:BasicConceptsand Methods:

- WhatIsClusterAnalysis?
- RequirementsforClusterAnalysis
- OverviewofBasicClusteringMethods

Whatisa Cluster?

Thegivendataisdividedintodifferentgroupsbycombiningsimilarobjectsintoagroup.Thisgroupis nothing but a cluster.

Aclusterisnothingbutacollectionofsimilardatawhichisgroupedtogether.
Clusteringisknownas unsupervisedlearningbecausetheclasslabelinformationisnotpresent.For this reason, clustering is a form of learning by observation, rather than learning by examples.

Agoodclusteringalgorithmaimstoobtainclusterswhose:

- The**intra-clustersimilarities** arehigh,Itimpliesthatthedatapresentinsidetheclusteris similar to one another.
- The**inter-clustersimilarity**islow,anditmeanseachclusterholds datathatisnotsimilarto other data.

WhatIsCluster Analysis?

Clusteranalysisorsimplyclustering is theprocess ofpartitioningasetofdataobjects (orobservations)into subsets.
Eachsubsetisacluster,suchthatobjectsinaclusteraresimilartooneanother,yetdissimilartoobjectsi nother clusters.
Thesetofclustersresultingfromaclusteranalysiscanbereferredtoasa **clustering**.

For example,

consideradatasetofvehicles giveninwhichitcontainsinformationaboutdifferentvehicles likecars,buses, bicycles, etc.
Asitisunsupervisedlearningtherearenoclasslabelslike Cars,Bikes,etcforallthevehicles,allthedat ais combined and is not in a structured manner.
Clustering is also called data segmentation in some applications because clustering partitions large data setsintogroupsaccordingtotheirsimilarity.Clusteringcanalsobeusedforoutlierdetection,where outliers(values that are “far away” from any cluster) may be more interesting than common

cases.
Applications
ofoutlierdetectionincludethedetectionofcreditcardfraudandthemonitoringofcriminal activitiesinelectroniccommerce.Forexample,exceptionalcasesincreditcardtransactions, suchasvery expensive and infrequent purchases, may be of interest as possible fraudulent activities.

RequirementsforClusterAnalysis:

The following are requirements of clustering in data mining:

Scalability: Many clustering algorithms work well on small data sets containing fewer than several hundred data objects; however, a large database may contain millions or even billions of objects, particularly in Web search scenarios. Clustering on only a sample of a given large data set may lead to biased results. Therefore, highly scalable clustering algorithms are needed.

Ability to deal with different types of attributes: Many algorithms are designed to cluster numeric (interval- based) data. However, applications may require clustering other data types, such as binary, nominal (categorical), and ordinal data, or mixtures of these data types. Recently, more and more applications need clustering techniques for complex data types such as graphs, sequences, images, and documents.

Discovery of clusters with arbitrary shape: Many clustering algorithms determine clusters based on Euclidean or Manhattan distance measures .Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape. Consider sensors, for example, which are often deployed for environment surveillance. Cluster analysis on sensor readings can detect interesting phenomena. We may want to use clustering to find the frontier of a running forest fire, which is often not spherical. It is important to develop algorithms that can detect clusters of arbitrary shape.

Requirements for domain knowledge to determine input parameters: Many clustering algorithms requireusers to provide domain knowledge in the form of input parameters such as the desired number of clusters. Consequently, the clustering results may be sensitive to such parameters. Parameters are often hard to determine, especially for high-dimensionality data sets and where users have yet to grasp a deep understanding of their data. Requiring the specification of domain knowledge not only burdens users, but also makes the quality of clustering difficult to control.

Ability to deal with noisy data: Most real-world data sets contain outliers and/or missing, unknown, or erroneous data. Sensor readings, for example, are often noisy—some readings may be inaccurate due to the sensing mechanisms, and some readings may be erroneous due to interferences from surrounding transient objects. Clustering algorithms can be sensitive

to such noise and may produce poor-quality clusters. Therefore, we need clustering methods that are robust to noise.

Incrementalclusteringandinsensitivitytoinputorder:Inmanyapplications,incrementalupdates (representing newer data) may arrive at any time. Some clustering algorithms cannot incorporate incremental updatesintoexistingclusteringstructuresand,instead,havetorecomputeanewclusteringfromscratch. Clusteringalgorithms mayalsobe sensitive to the input dataorder.Thatis,given aset of data objects,clustering algorithmsmayreturndramaticallydifferentclusteringsdependingontheorderinwhichtheobjects are presented. Incremental clustering algorithms and algorithms that are insensitive to the input order are needed.

Capability of clusteringhigh-dimensionality data: A data set cancontainnumerous dimensions orattributes. Whenclusteringdocuments,forexample,eachkeywordcanberegardedasadimension,andthereare often thousands of keywords. Most clustering algorithms are good at handling low-dimensional data such as data sets involvingonlytwoorthreedimensions.Findingclustersofdataobjectsinahighdimensionalspaceis challenging, especially considering that such data can be very sparse and highly skewed.

Constraint-based clustering: Real-world applications may need to perform clustering under various kinds of constraints. Suppose that your job is to choose the locations for a given number of new automatic teller machines (ATMs) in a city. To decide upon this, you may cluster households while considering constraints such as thecity's riversandhighwaynetworks andthetypesandnumberofcustomers percluster.Achallengingtask is to find data groups with good clustering behavior that satisfy specified constraints.

Interpretability and usability: Users want clustering results to be interpretable, comprehensible, and usable.It is important to study how an application goal may influence the selection of clustering features and clustering methods.

Thefollowingareorthogonalaspectswithwhichclusteringmethodscanbecompared:

The partitioning criteria: In some methods, all the objects are partitioned so that no hierarchy exists among the clusters. That is, all the clusters are at the same level conceptually. Such a method is useful, for example, for partitioning customers into groups so that each group has its own manager.

Separation of clusters: Some methods partition data objects into mutually exclusive clusters. When clustering customers into groups so that each group is taken care of by one manager, each customer may belong to onlyone group.

Similarity measure: Some methods determine the similarity between two objects by the distance between them. Such a distance can be defined on Euclidean space. Similarity measures play a fundamental role in the design of clustering methods.

Clustering space: Many clustering methods search for clusters within the entire given data space. These methods are useful for low-dimensionality data sets. With highdimensional

data, however, there can be many irrelevant attributes, which can make similarity measurements unreliable. Consequently, clusters found in the full space are often meaningless. It's often better to instead search for clusters within different subspaces of the same data set. Subspace clustering discovers clusters and subspaces (often of low dimensionality) that manifest object similarity.

OverviewofBasicClusteringMethods:
Theclusteringmethodscanbeclassifiedintothefollowingcategories:

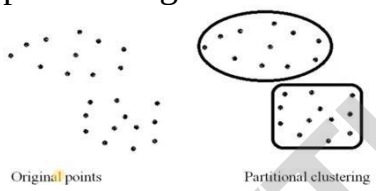
- PartitioningMethod
- HierarchicalMethod
- Density-basedMethod
- Grid-BasedMethod

PartitioningMethod

Itisusedtomakepartitionsonthedatainordertoform clusters.
If"n" partitions are done on"p" objects of the database then each partition is represented by a cluster and n < p.

ThetwoconditionswhichneedtobesatisfiedwiththisPartitioningClusteringMethodare:

- Oneobjectiveshouldonlybelongtoonlyonegroup.
- Thereshouldbenogroupwithoutevenasinglepurpose.
- In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning



Mostwell-knownandcommonlyusedpartitioningmethods:

1) The k-means algorithm: where each cluster is represented by the mean value of the objects in the cluster.(Each cluster is represented by the center of the cluster)
2) Thek-medoidsalgorithm:where eachclusteris representedbyone ofthe objects locatednear the center of the cluster.

Hierarchicalclustering

Hierarchical clustering is a method of data mining that groups similar data points into clusters by creating a hierarchical structure of the clusters.
Identify the 2 clusters which can be closest together, and Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.
In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called**Dendrogram**(A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy.
Ahierarchicalmethodcanbeclassifiedasbeingeitheragglomerativeordivisive

- **The agglomerative approach**, also called the bottom-up approach, starts with each object forming a separategroup.Itsuccessivelymergestheobjectsorgroupsclosetooneanother,untilalltheg roupsare merged into one (the topmost level of the hierarchy), or a termination condition holds.
- **The divisive approach**, also called the top-down approach, starts with all the objects in the same cluster. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.

Hierarchicalmethodssufferfromthefactthatonceastep(mergeorsplit)isdone,itcanneverbe undone.

A density-based method clusters objects based on the notion of density. It grows clusters either according to the density of neighborhood objects (e.g., in DBSCAN) or according to a density function (e.g., in DENCLUE). OPTICS is a density-based method that generates an augmented ordering of the data's clustering structure.

A grid-based method first quantizes the object space into a finite number of cells that form a grid structure,and then performs clustering on the grid structure. STING is a typical example of a grid-based method based on statisticalinformationstoredingridcells.CLIQUEisagrid-basedandsubspaceclusteringalgorithm.

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Figure:Overviewofclusteringmethods

PartitioningMethods:

- k-Means:ACentroid-BasedTechnique
- k-Medoids:ARepresentativeObject-BasedTechnique

k-Means:ACentroid-BasedTechnique:

Supposeadataset,D,containsnobjects inEuclideanspace.Partitioningmethodsdistributetheobjectsin Dinto k clusters, C1,...,Ck , that is, $C_i \subset D$ and $C_i \cap C_j = \emptyset$ for $(1 \leq i,j \leq k)$.

Anobjectivefunctionis usedtoassess thepartitioningqualitysothatobjectswithinaclusteraresimilartoone another but dissimilar to objects in other clusters. This is, the objective function aims for high intracluster similarity and low intercluster similarity.

K-MeansClusteringisanUnsupervisedLearningalgorithm,whichgroupstheunlabeleddatasetinto different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
Itisaniterativealgorithmthatdividestheunlabeleddatasetintokdifferentclustersinsuchawaythate ach dataset belongs only one group that has similar properties.

Tomeasurethedistancebetweendatapointsandcent roid, we can use any method such as

- Euclideandistanceor
- Manhattandistance.

Thek-meansclusteringalgorithmmainlyperformstwotasks:

- DeterminesthebestvalueforKcenterpointsorcentroidsbyaniterativeprocess.
- Assignseachdatapointtoitsclosestk-center.Thosedatapointswhichareneartotheparticulark- center, create a cluster.

A centroid-based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster.Conceptually, the centroid of a cluster is its center point.
The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assignedto the cluster. The difference between an object p ∈ Ci and ci , the representative of the cluster, is measured by dist(p,ci), where dist(x,y) is the Euclidean distance between two points x and y.
The quality of cluster Ci can be measured by the withincluster variation, which is the sum of squared error between all objects in Ci and the centroid ci , defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(\boldsymbol{p}, \boldsymbol{c_i})^2,$$

whereEis thesum ofthesquarederrorforallobjectsinthedataset;pis thepointinspacerepresentingagiven object; and ci is the centroid of cluster Ci (both p and ci are multidimensional).

“Howdoesthek-meansalgorithmwork?”

The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster. It proceeds as follows.
First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean.
The k-means algorithm then iteratively improves the within-cluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration.

All the objects are then reassigned using the updated means as the new cluster centers. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round.
The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as iterative relocation

Thek-meansprocedure:

The time complexity of the k-means algorithm is O(nkt), where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, k<<n and t<< n. Therefore, the method is relatively scalable and efficient in processing large data sets.
There are several variants of the k-means method. These can differ in the selection of the initial k-means, the calculation of dissimilarity, and the strategies for calculating cluster means.

Thek-meansmethodcanbeappliedonlywhenthemeanofasetofobjectsis defined.This maynotbe thecase in some applications such as when data with nominal attributes are involved.

The k-modes method is a variant of k-means, which extends the k-means paradigm to cluster nominal data by replacingthemeans of clusters withmodes.It usesnew dissimilaritymeasures todeal with nominal objects and afrequency-basedmethodtoupdatemodesofclusters.Thek-meansandthek-modesmethodscanbe integrated to cluster data with mixed numeric and nominal values.

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

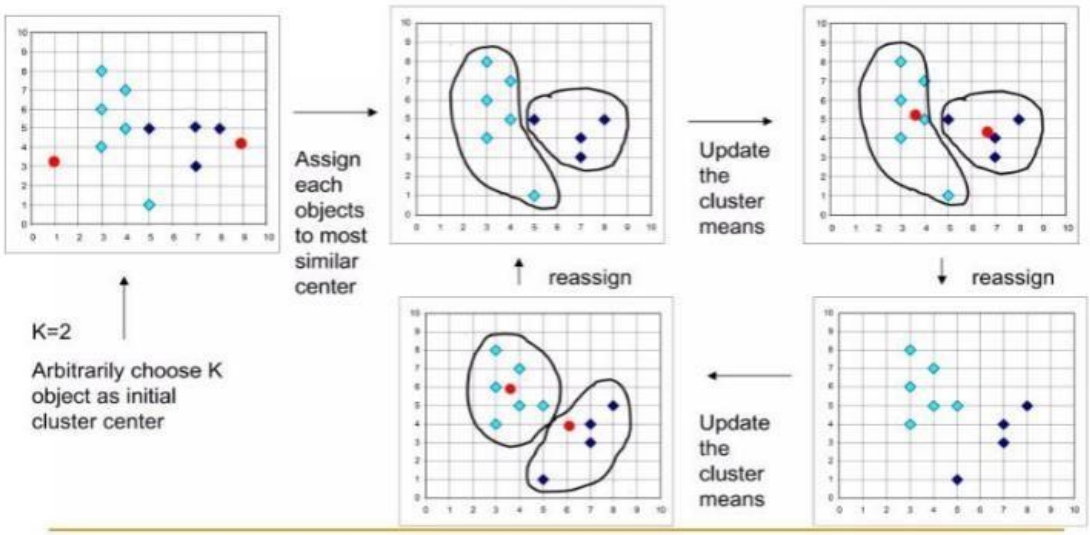
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Example:Clusteringby k-means partitioning. Consider a set ofobjectslocated in 2-D space,asdepictedin the following Figure Let k = 2, that is, the user would like the objects to be partitioned into two clusters.



Drawbacksofk-means:

- The k-means method is not suitable for discovering clusters with nonconvex shapes or clusters of very different size.
- Moreover,itissensitivetonoiseandoutlierdatapointsbecauseasmallnumberofsuchdatacan substantially influence the mean value.

"Howcanwemakethek-meansalgorithmmorescalable?"

- One approach to making the k-means method more efficient on large data sets is to use a good-sized set of samples in clustering.
- Another is to employ a filtering approach that uses a spatial hierarchical data index to save costs when computing means.
- Athirdapproachexploresthemicroclusteringidea,whichfirstgroupsnearbyobjectsinto "microclusters" and then performs k-means clustering on the microclusters.

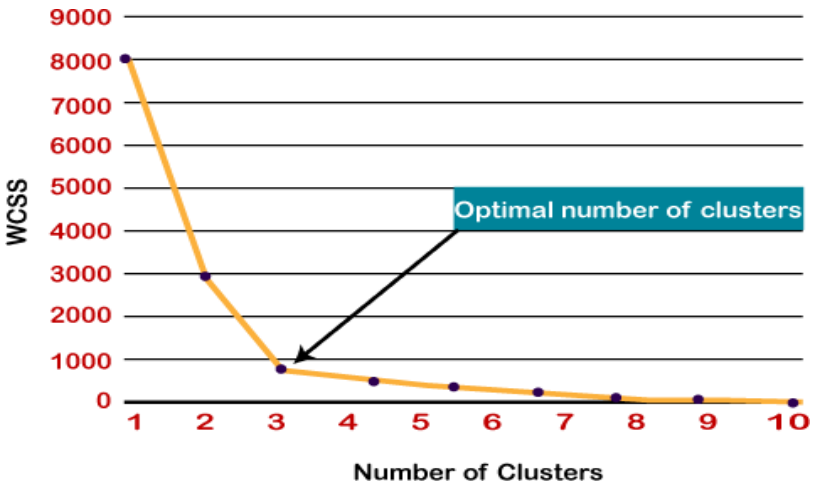
Howtochoosethevalueof"Knumberofclusters"inK-meansClustering? Elbow Method

TheElbowmethodisoneofthemostpopularwaystofindtheoptimalnumberofclusters.
This method uses the concept of WCSS value. WCSS stands forWithin Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$WCSS= \sum_{P_i \text{ in Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_i \text{ in Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_i \text{ in CLuster3}} distance(P_i\ C_3)^2$$

Tofindtheoptimalvalueofclusters,theelbowmethodfollowsthebelowsteps:

- ItexecutestheK-meansclusteringonagivendatasetfordifferentKvalues(rangesfrom1-10). For each value of K, calculates the WCSS value.
- PlotsacurvebetweencalculatedWCSSvaluesandthenumberofclustersK.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.
- Sincethegraphshowsthesharpbend,whichlookslikeanelbow,henceitisknownasthe elbow method. The graph for the elbow method looks like the below image:



k-Medoids: A Representative Object-Based Technique:

"How can we modify the k-means algorithm to diminish such sensitivity to outliers?"

Instead of taking the mean value of the objects in a cluster as a reference point, we can pick actual objects to represent the clusters, using one representative object per cluster. Each remaining object is assigned to the cluster of which the representative object is the most similar. The partitioning method is then performed based on the principle of minimizing the sum of the dissimilarities between each object p and its corresponding representative object.

That is, an absolute-error criterion is used, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i),$$

where E is the sum of the absolute error for all objects p in the data set, and oi is the representative object of Ci . This is the basis for the k-medoids method, which groups n objects into k clusters by minimizing the absolute error.

The Partitioning Around Medoids (PAM) algorithm is a popular realization of k-medoids clustering. Ittackles the problem in an iterative, greedy way. Like the k-means algorithm, the initial representative objects (called seeds) are chosen arbitrarily. We consider whether replacing a representative object by a non- representative object would improve the clustering quality. All the possible replacements are tried out.

The iterative process of replacing representative objects by other objects continues until the quality of the resulting clustering cannot be improved by any replacement.

This quality is measured by a cost function of the average dissimilarity between an object and therepresentative object of its cluster.

PAM, a k-medoids partitioning algorithm:

Specifically, let o1,...,ok be the current set of representative objects (i.e., medoids).

To determine whether a non-representative object, denoted by o random, is a good replacement for a currentmedoid oj ($1 \leq j \leq k$),
we calculate the distance from every object p to the closest object in the set {o1,...,oj−1,orandom,oj+1,...,ok}, and use the distance to update the cost function.
Thereassignmentsofobjectsto{o1,...,oj−1,orandom,oj+1,...,ok}aresimple.Supposeobjectpiscurrentlyassigned to a cluster represented by medoid oj .
Doweneedtoreassignptoadifferentclusterifojisbeingreplacedbyorandom?
Object p needs to be reassigned to either orandom or some other cluster represented by oi (i 6= j), whichever is the closest.

Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

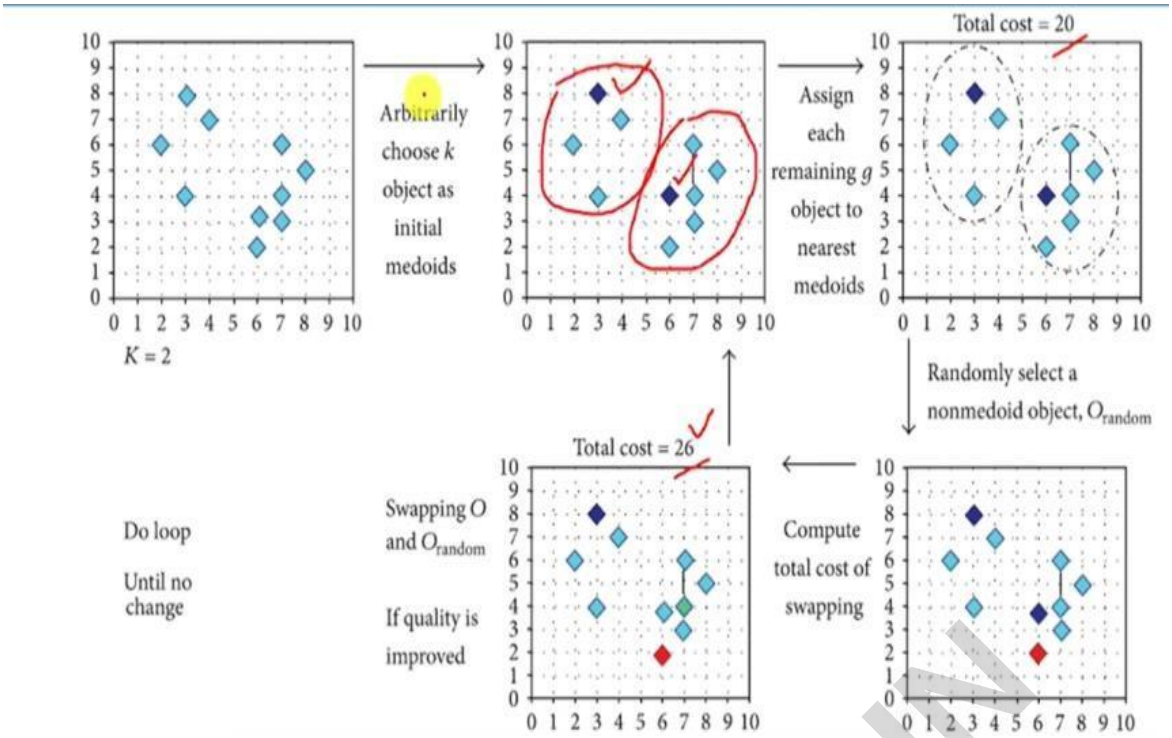
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

Each time a reassignmentoccurs, a difference in absolute error, E, is contributed to the cost function. Therefore, the cost function calculates the difference in absolute-error value if a current representative object is replacedbyanorepresentativeobject.Thetotalcostofswappingisthesumofcostsincurredbyallnorepresentative objects. If the total cost is negative, then oj is replaced or swapped with o random because the actual absolute- error E is reduced. If the total cost is positive, the current representative object, oj , is considered acceptable,and nothing is changed in the iteration.



Whichmethodismorerobust—k-meansork-medoids?"

The k-medoids method is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean. However, the complexity of each iteration inthe k-medoids algorithm is O(k(n − k) 2). For large values of n and k, such computation becomes very costly,and much more costly than the k-means method. Both methods require the user to specify k, the number of clusters.

"Howcanwescaleupthek-medoidsmethod?"

k-medoids partitioningalgorithm like PAM works effectively for smalldatasets, but does not scale wellfor large data sets.

To deal with larger data sets, a sampling-based method called CLARA (Clustering LARge Applications) can be used. Instead of taking the whole data set into consideration, CLARA uses a random sample of the data set.

CLARA builds clusterings from multiple random samples and returns the best clustering as the output. A k- medoids partitioning algorithm like PAM works effectively for small data sets, but does not scale well for large data sets.

"HowmightweimprovethequalityandscalabilityofCLARA?"

A randomized algorithm called CLARANS (Clustering Large Applications based upon RANdomized Search) presents a trade-off between the cost and the effectiveness of using samples to obtain clustering.

Recall that when searching for better medoids, PAM examines every object in the data set against every current medoid,

whereasCLARAconfinesthecandidatemedoidstoonlyarandomsampleofthedataset.

First, it randomly selects k objects in the data set as the current medoids. It then randomly selects a current medoid x and anobject ythat is not oneof the current medoids. Can replacing x by y improve the absolute-error criterion? If yes, the replacement is made. CLARANS conducts such a randomized search l times. The set of the current medoids after the l steps is considered a local optimum. CLARANS repeats this randomized process m times and returns the best local optimal as the final result.

HierarchicalMethods:

- AgglomerativeversusDivisiveHierarchicalClustering
- DistanceMeasuresinAlgorithmicMethods
- BIRCH:MultiphaseHierarchicalClusteringUsingClusteringFeatureTrees
- Chameleon:MultiphaseHierarchicalClusteringUsingDynamicModeling
- ProbabilisticHierarchicalClustering

HierarchicalMethods:

A hierarchical clustering method works by grouping data objects into a hierarchy or "tree" of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization.

A hierarchical method creates a hierarchical decomposition of the given set of data objects. The method canbe classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (e.g., in Chameleon), or by first performing microclustering (that is, grouping objects into "microclusters") and then operating on the microclusters with other clustering techniques such as iterative relocation (as in BIRCH).

Forexample,asthemanagerofhumanresourcesatAllElectronics,youmayorganizeyouremployeesinto major groups such as executives, managers, and staff. You can further partition these groups into smaller subgroups. For instance, the general group of staff can be further divided into subgroups of senior officers, officers, and trainees. All these groups form a hierarchy. We can easily summarize or characterize the data that are organized into a hierarchy, which can be used to find, say, the average salary of managers and of officers.

A hierarchical clustering method can be either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or topdown (splitting) fashion. Let's have a closer look at these strategies.

AgglomerativeversusDivisiveHierarchicalClustering:

AgglomerativeHierarchicalClustering

- **Anagglomerativehierarchicalclusteringmethod**usesabottom-upstrategy.
- It starts by letting each object form its own cluster and iteratively merges clusters into larger and larger clusters, until all the objects are in a single cluster or certain termination conditions are satisfied.
- Thesingleclusterbecomesthehierarchy's root.
- Forthemergingstep,itfindsthetwoclustersthatareclosesttoeachother(accordingtosomesimilarity measure), and combines the two to form one cluster. Because two clusters are merged per iteration, where each cluster contains at least one object, an agglomerative method requires at most n iterations.

DivisiveHierarchicalClustering:

- **Adivisivehierarchicalclusteringmethod**employsatop-downstrategy.
- It starts byplacingall objectsin one cluster, which is the hierarchy's root. It thendivides the root cluster into several smaller subclusters, and recursively partitions those clusters into smaller ones.
- The partitioning process continues until each cluster at the lowest level is coherent enough—either containing only one object, or the objects within a cluster are sufficiently similar to each other.

In either agglomerative or divisive hierarchical clustering, a user can specify the desired number of clusters as a termination condition.

Example:

Agglomerative versus divisive hierarchical clustering. Figure shows the application of AGNES (AGglomerativeNESting), an agglomerative hierarchical clustering method, and DIANA (DIvisiveANAlysis), a divisive hierarchical clustering method, on a data set of five objects, {a,b,c,d, e}.

Initially,AGNES,theagglomerativemethod,placeseachobjectintoaclusterofitso
wn. The clusters are then merged step-by-step according to some criterion.
Forexample,clustersC1andC2maybemergedifanobjectinC1andanobjectinC2formtheminimum Euclidean distance between any two objects from different clusters.

This is a single-linkage approach in that each cluster is represented by all the objects in the cluster, and the similarity between two clusters is measured by the similarity of the closest pair of data points belonging to different clusters.Thecluster-mergingprocessrepeatsuntilalltheobjectsareeventuallymergedtoformonecluster.

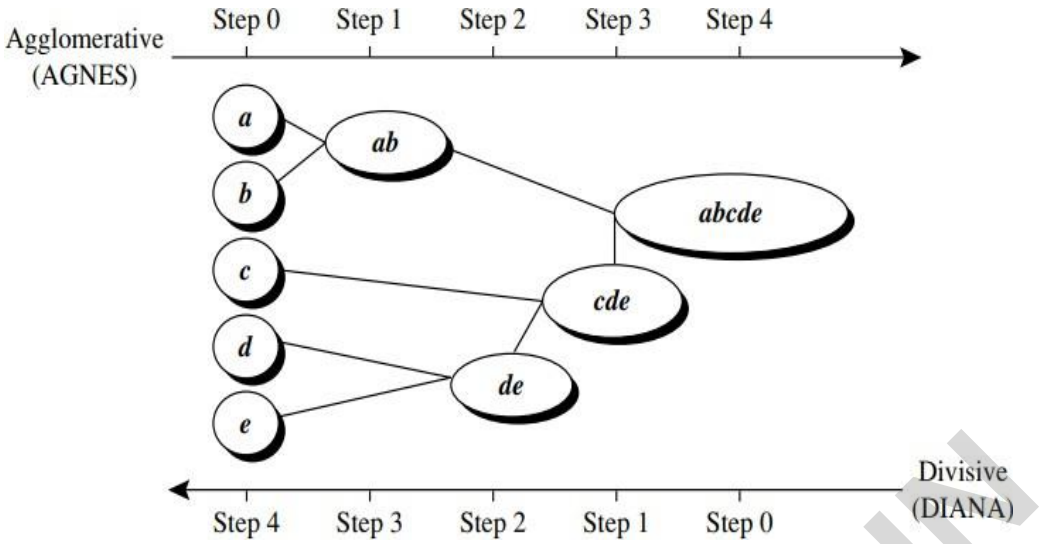


Figure-1:Agglomerativeanddivisivehierarchicalclusteringondataobjects{a,b,c,d,e}

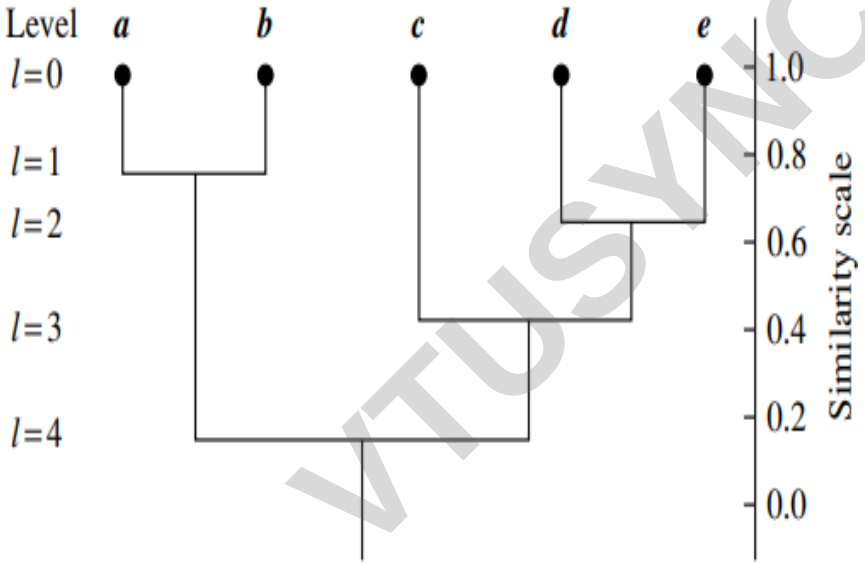


Figure-2:Dendrogramrepresentationforhierarchicalclusteringofdataobjects{a,b,c,d,e}.

DIANA, the divisive method, proceeds in the contrasting way. All the objects are used to form one initial cluster.
Theclusterissplitaccordingtosomeprinciplesuchasthemaximum Euclideandistancebetweentheclosest neighboringobjectsinthecluster.Thecluster-splittingprocessrepeatsuntil,eventually,eachnewcluster contains only a single object.

Dendrogram:

A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering. It shows how objects are grouped together (in an agglomerative method) or partitioned (in a divisive method) step-by-step. Figure-2shows a dendrogram for the five objects presentedin Figure-1, where l = 0shows the five objects as singletonclusters at level 0. At l = 1, objects a and bare grouped together to form the first cluster, and they stay together at all subsequent levels. We

can also use a vertical axis to show the similarity scale between clusters. For example, when the similarity of two groups of objects, {a,b} and {c,d, e}, is roughly 0.16, they are merged together to form a single cluster.

A challenge with divisive methods is how to partition a large cluster into several smaller ones. For example, thereare2n−1−1possiblewaystopartitionasetofnobjectsintotwo exclusivesubsets,wherenis thenumber of objects.

AdvantagesofHierarchicalclustering:

- Itissimpletoimplementandgivesthebestoutputinsomecases.
- Itiseasyandresultsinahierarchy,astructurethatcontainsmore information.
- Itdoesnotneedustopre-specifythenumberof clusters.

Disadvantagesofhierarchicalclustering

- Itbreaksthelarge clusters.
- ItisDifficulttohandledifferentsizedclustersandconvexshapes.
- Itissensitivetonoiseandoutliers.
- Thealgorithmcanneverbechangedordeletedonceitwasdone previously.

DistanceMeasuresinAlgorithmicMethods:

Whether using an agglomerative method or a divisive method, a core need is to measure the distance between two clusters, where each cluster is generally a set of objects.

Four widely usedmeasures for distance betweenclusters are as follows, where |p − p‘ | is the distance between two objects or points, p and p ‘ ; mi is the mean for cluster, Ci ; and ni is the number of objects in Ci . They arealso known as linkage measures.

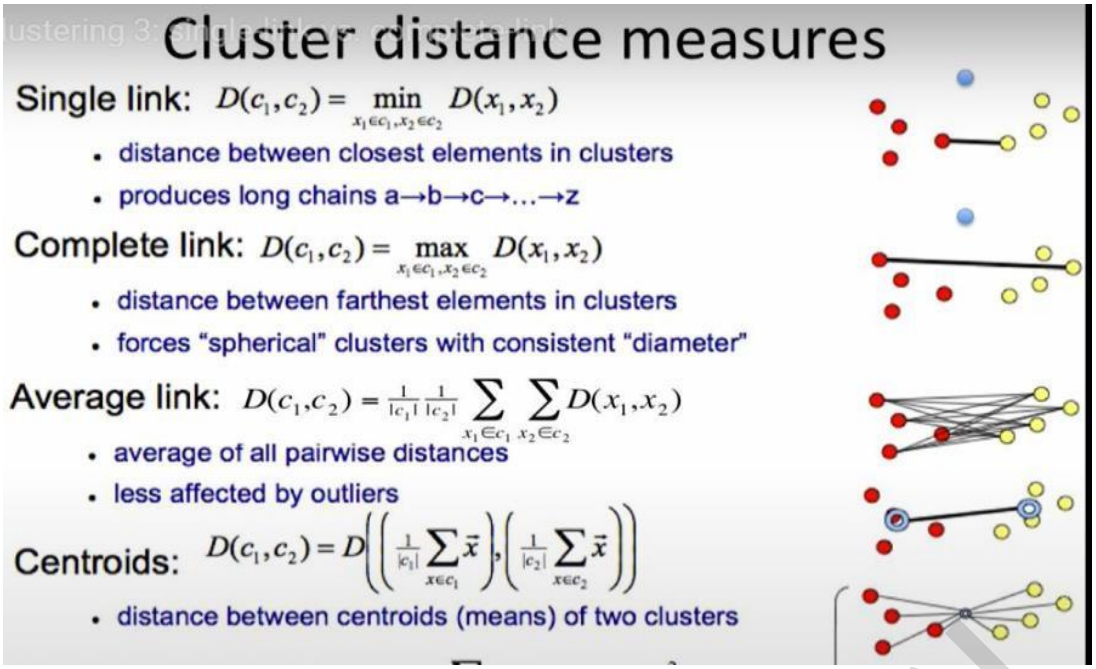
$$\text{Minimum distance: } dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$$

$$\text{Maximum distance: } dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$$

$$\text{Mean distance: } dist_{mean}(C_i, C_j) = |m_i - m_j|$$

$$\text{Average distance: } dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$$

When an algorithm uses the minimum distance, dmin(Ci ,Cj), to measure the distance between clusters, it is sometimes called a **nearest-neighbor clustering algorithm**.


Cluster distance measures
Single link: $D(c_1,c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
distance between closest elements in clusters
produces long chains a→b→c→…→z
Complete link: $D(c_1,c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1,x_2)$
distance between farthest elements in clusters
forces "spherical" clusters with consistent "diameter"
Average link: $D(c_1,c_2) = \frac{1}{|c_1|}\frac{1}{|c_2|}\sum_{x_1 \in c_1}\sum_{x_2 \in c_2} D(x_1,x_2)$
average of all pairwise distances
less affected by outliers
Centroids: $D(c_1,c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1}\vec{x}\right),\left(\frac{1}{|c_2|}\sum_{x \in c_2}\vec{x}\right)\right)$
distance between centroids (means) of two clusters