Data Mining: Concepts and Techniques

— Slides for Textbook —
— Chapter 10 —

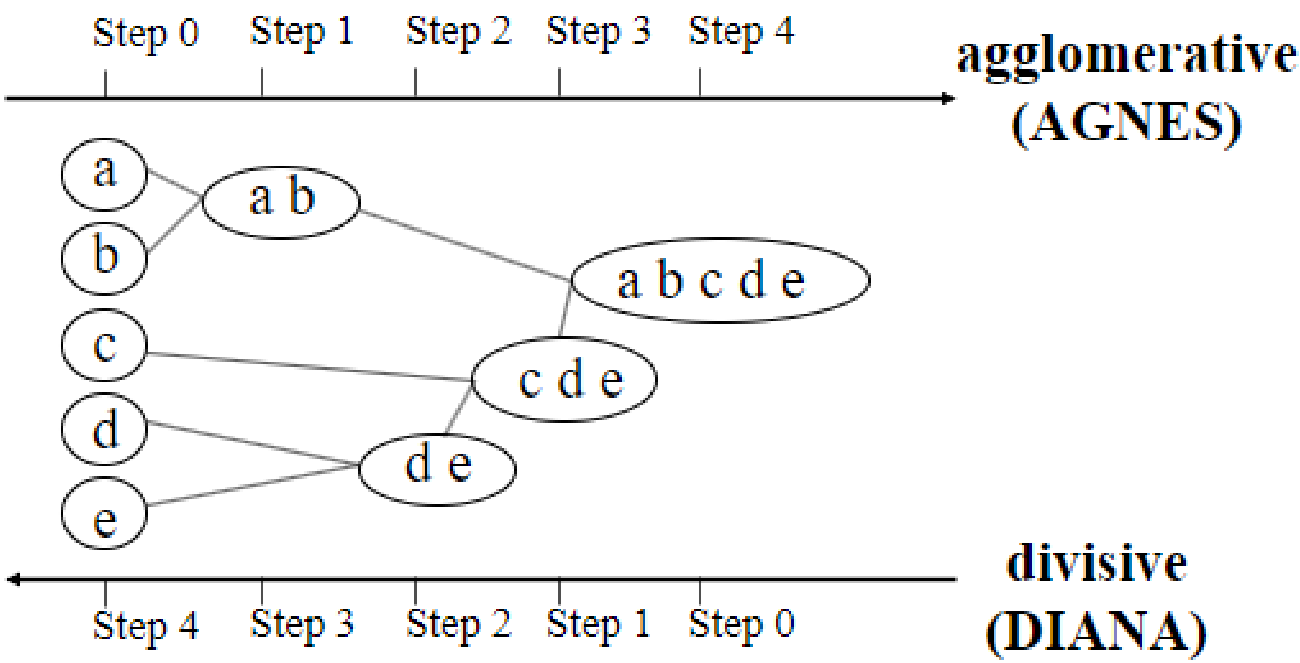
©Jiawei Han and Micheline Kamber
Intelligent Database Systems Research Lab
School of Computing Science
Simon Fraser University, Canada
http://www.cs.sfu.ca

Chapter 7. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- Major Clustering Approaches
  - Partitioning approach
  - Hierarchical approach
  - Density-based approach
- Evaluation of Clustering
- Summary

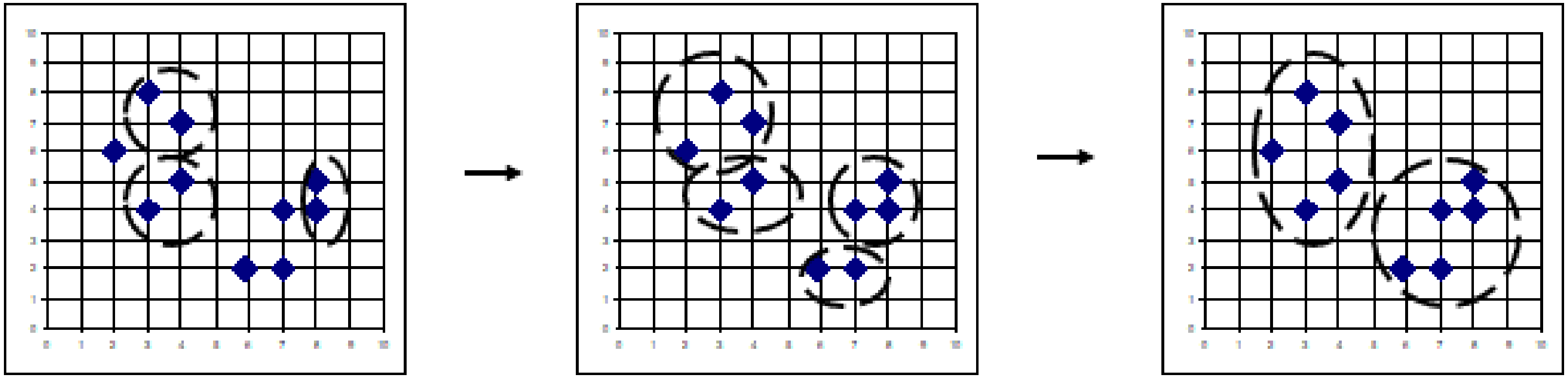
Hierarchical Clustering

- Use distance matrix as clustering criteria.
  - This method does not require the number of clusters *k* as an input, but needs a termination condition



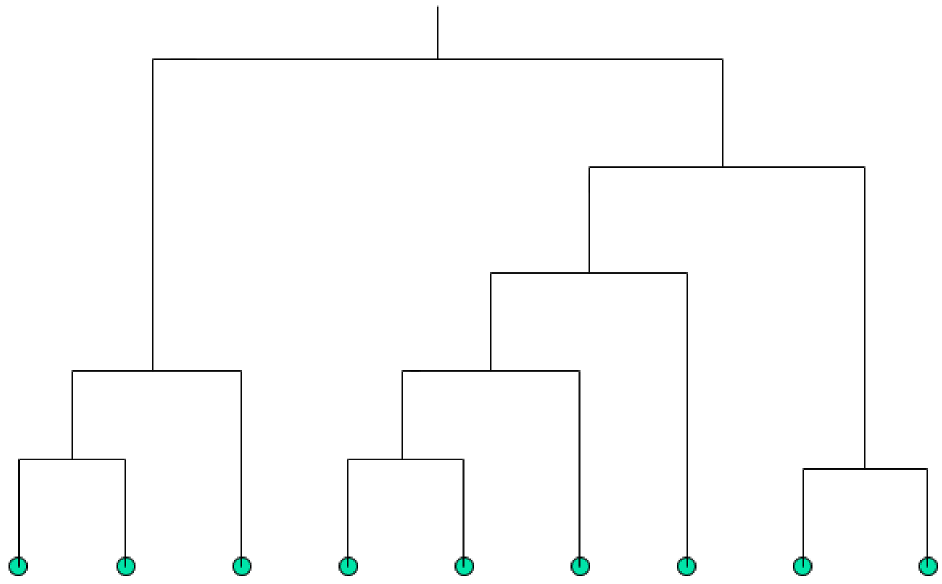
AGNES (Agglomerative Nesting)

- Use the **single-link** method and the dissimilarity matrix
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



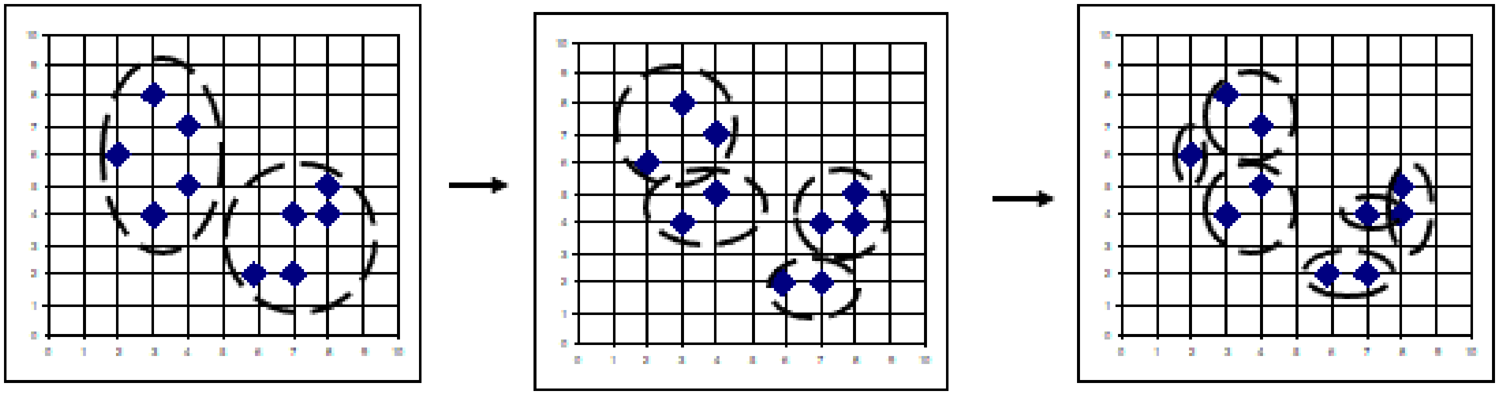
Dendrogram Visualization

- Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram
- A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster

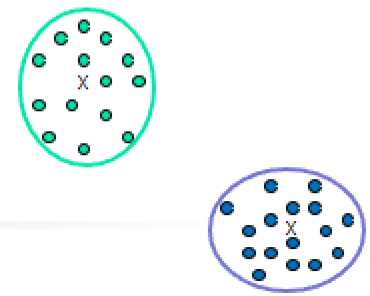


DIANA (Divisive Analysis)

- Inverse order of AGNES
- Eventually each node forms a cluster on its own



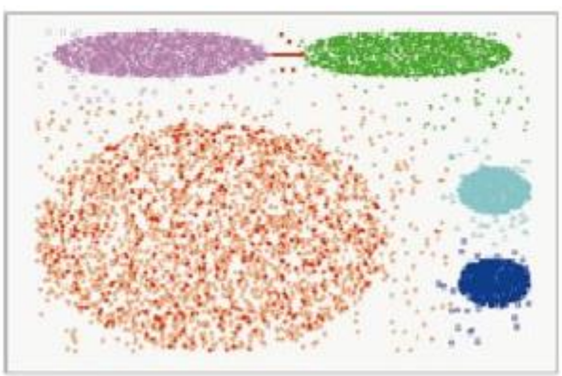
Distance between Clusters

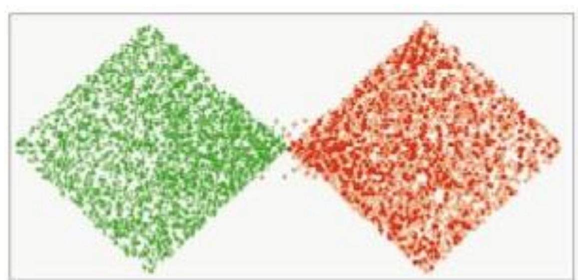


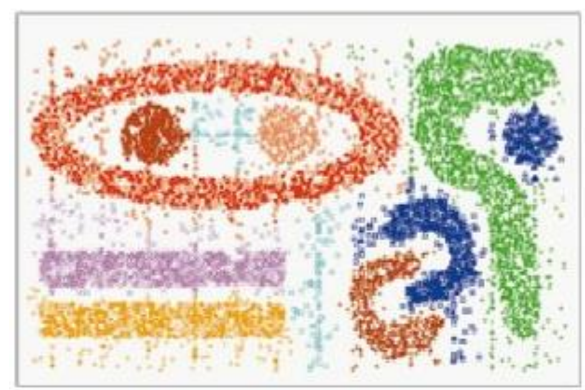
- Single link: smallest distance between an element in one cluster and an element in the other, i.e., dist(Ki, Kj) = min(tip, tjq)
- Complete link: largest distance between an element in one cluster and an element in the other, i.e., dist(Ki, Kj) = max(tip, tjq)
- Average: avg distance between an element in one cluster and an element in the other, i.e., dist(Ki, Kj) = avg(tip, tjq)
- Centroid: distance between the centroids of two clusters, i.e., dist(Ki, Kj) = dist(Ci, Cj)
- Medoid: distance between the medoids of two clusters, i.e., dist(Ki, Kj) = dist(Mi, Mj)
  - Medoid: a chosen, centrally located object in the cluster

Extensions to Hierarchical Clustering

- Major weakness of agglomerative clustering methods
  - Can never undo what was done previously
  - Do not scale well: time complexity of at least O(n2), where n is the number of total objects
- Integration of hierarchical & distance-based clustering
  - **BIRCH**: uses CF-tree and incrementally adjusts the quality of sub-clusters
  - **CHAMELEON**: hierarchical clustering using dynamic modeling







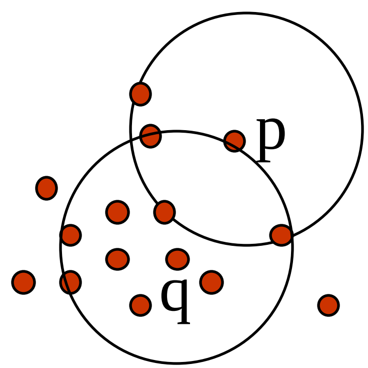
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - **DBSCAN**: Density-based spatial clustering of applications with noise
  - OPTICS: Ordering Points To Identify Cluster Structure
  - DENCLUE: DENsity-based CLUstEring
  - CLIQUE: CLustering in QUEst (more grid-based)

Density-Based Clustering: Basic Concepts

- Two parameters:
  - ***Eps***: Maximum radius of the neighbourhood
  - ***MinPts***: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: {q belongs to D | dist(p,q) ≤ Eps}
- Directly density-reachable: A point *p* is directly density-reachable from a point *q* w.r.t. *Eps*, *MinPts* if
  - *p* belongs to $N_{Eps}(q)$
  - core point condition:

$$|N_{Eps}(q)| \geq MinPts$$

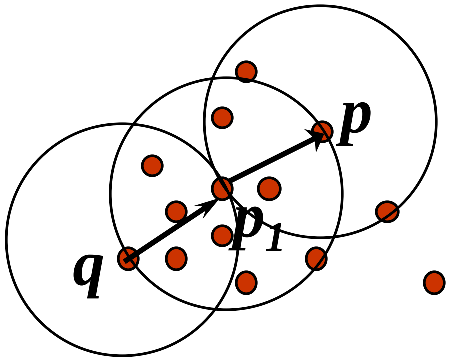


MinPts = 5

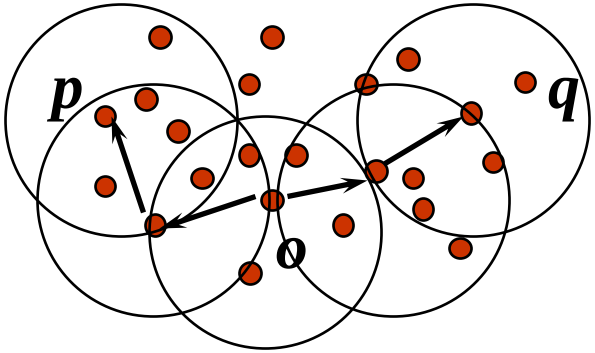
Eps = 1 cm

Density-Reachable and Density-Connected

- Density-reachable:
  - A point *p* is density-reachable from a point *q* w.r.t. *Eps, MinPts* if there is a chain of core points *p1,…,pn*

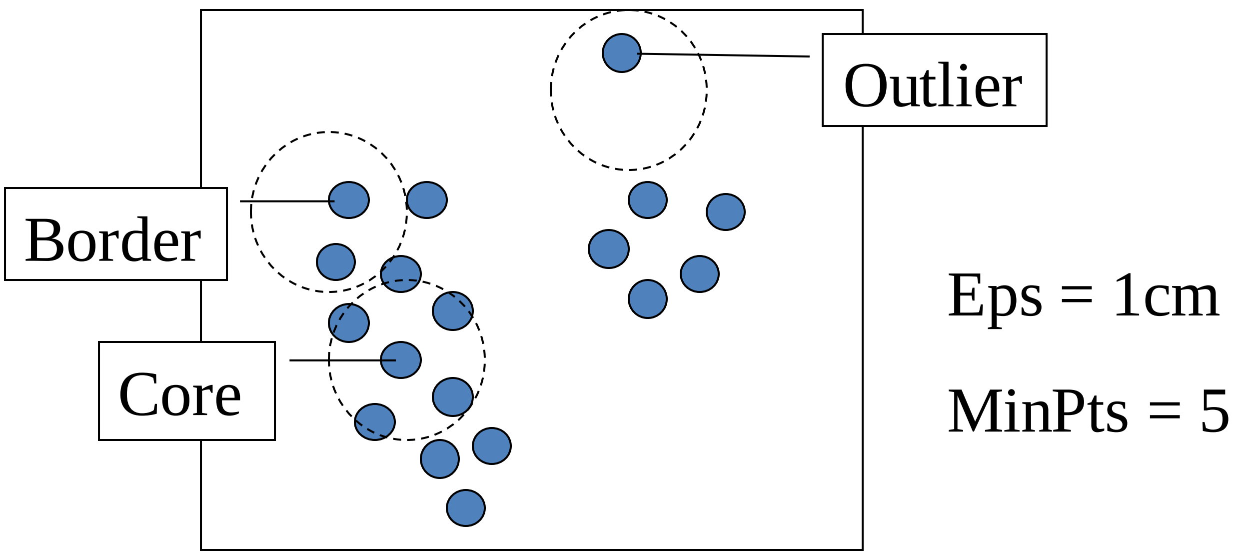


- Density-connected
  - A point *p* is density-connected to a point *q* w.r.t. *Eps, MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable



DBSCAN

- Density-Based Spatial Clustering of Applications with Noise
- Relies on a density-based notion of cluster: A **cluster** is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise

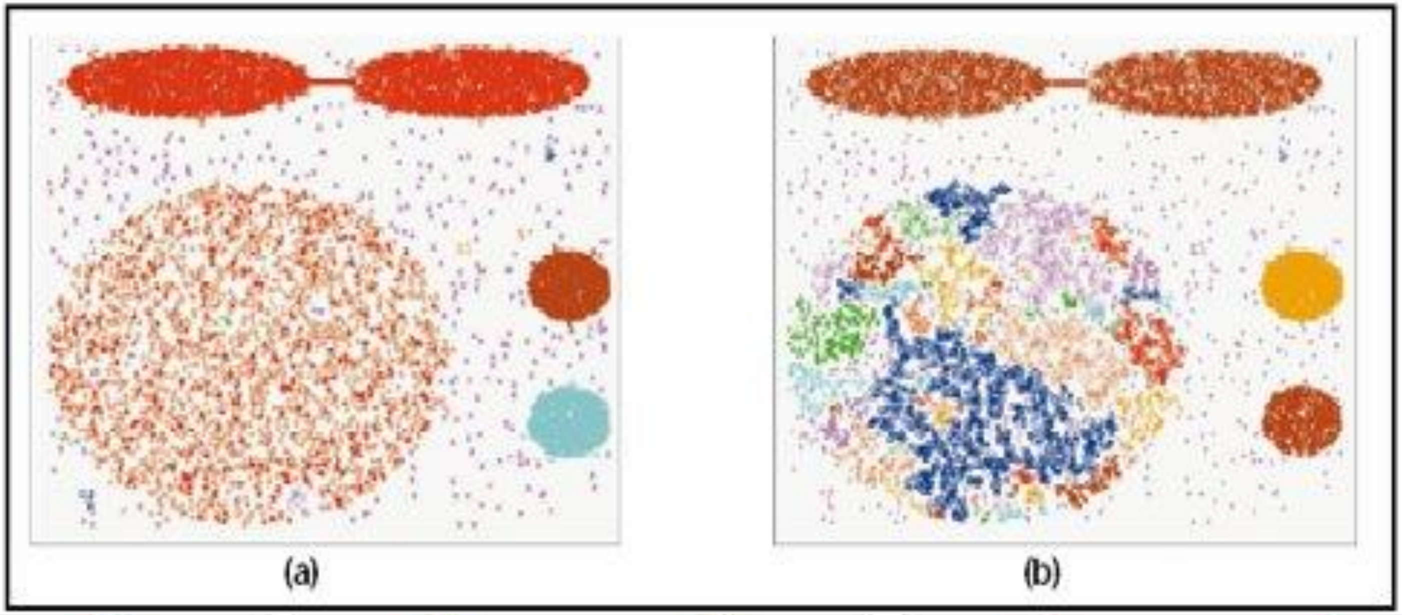


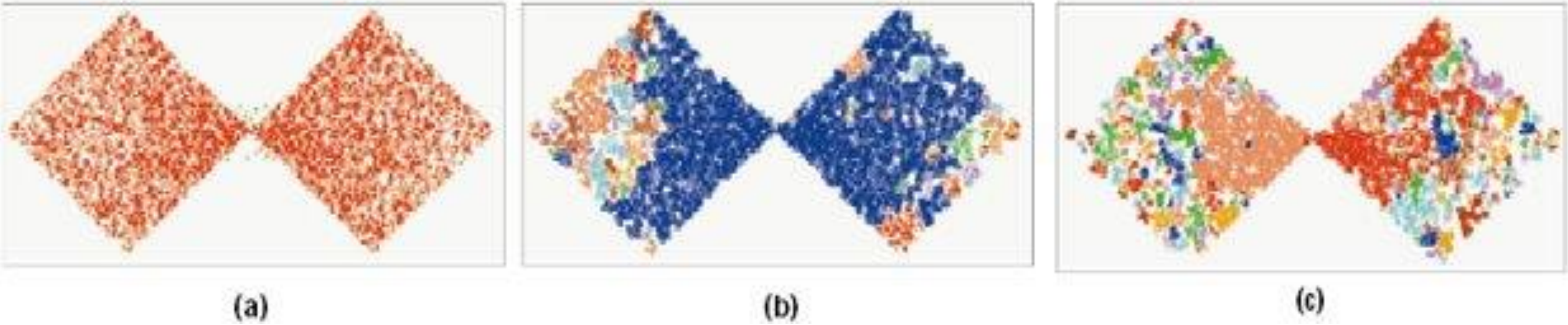
DBSCAN: The Algorithm

- Arbitrary select a point p
- Retrieve all points density-reachable from *p* w.r.t. *Eps* and *MinPts*
- If *p* is a core point, a cluster is formed
- If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database
- Continue the process until all of the points have been processed

DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.



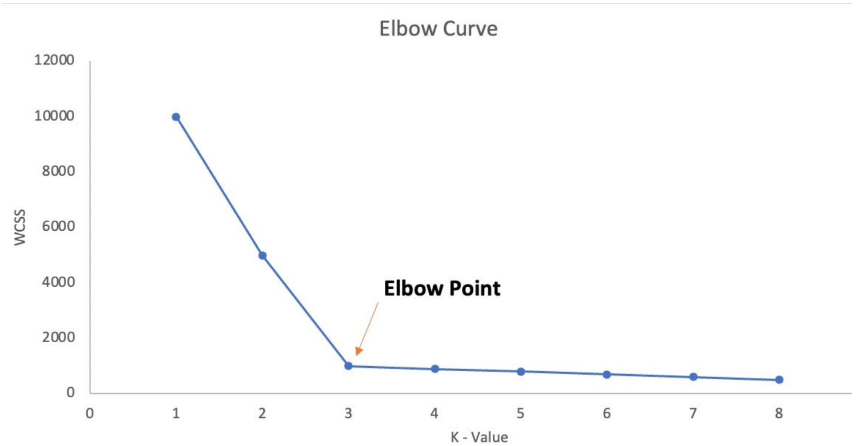


Chapter 7. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- Major Clustering Approaches
- Evaluation of Clustering
- Summary

Determine the Number of Clusters

- Empirical method
  - # of clusters ≈$\sqrt{n/2}$ for a dataset of n points
- Elbow method
  - Use the turning point in the curve of <u>sum of within cluster variance w.r.t the # of clusters</u> (sum of squared distance of members with centroid)
  - Choose number of clusters so that adding another cluster does not give much better modeling of the data

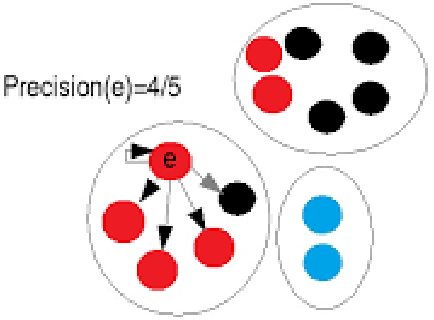


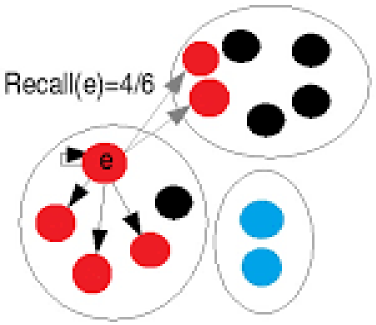
Determine the Number of Clusters

- Cross validation method
  - Divide a given data set into *m* parts
  - Use *m – 1* parts to obtain a clustering model
  - Use the remaining part to test the quality of the clustering
    - Example:
      - Given test set data points $\boldsymbol{x}=\{x_1,\ldots,x_n\}$
      - Assign each point in the test set, x_n to cluster based on closest centroid, C_i
      - Calculate the sum of squared distance between $\boldsymbol{x}$ and the closest respective centroids C_i
      - Distance measure how well the model fits the test set
  - For any $k > 0$, repeat it *m* times, compare the overall quality measure w.r.t. different *k*'s, and find # of clusters that fits the data the best

Measuring Clustering Quality

- Two methods: extrinsic vs. intrinsic
- Extrinsic: supervised, i.e., the ground truth is available
  - Compare a clustering against the ground truth using certain clustering quality measure
  - Ex. BCubed precision and recall metrics
- Intrinsic: unsupervised, i.e., the ground truth is unavailable
  - Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are
  - Ex. Silhouette coefficient
    - Mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample.
    - The Silhouette Coefficient for a sample is (b - a) / max(a, b)





Extrinsic Methods

- Clustering quality measure: $Q(C, C_g)$, for a clustering C given the ground truth C_g.
- Q is good if it satisfies the following **4** essential criteria
  - Cluster homogeneity: the purer, the better
  - Cluster completeness: should assign objects belong to the same category in the ground truth to the same cluster
  - Rag bag: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag (i.e., "miscellaneous" or "other" category)
  - Small cluster preservation: splitting a small category into pieces is more harmful than splitting a large category into pieces

Chapter 7. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- Major Clustering Approaches
- Evaluation of Clustering
- Summary

Summary

- Cluster analysis groups objects based on their similarity and has wide applications
- Measure of similarity can be computed for various types of data
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods
- K-means algorithm is a popular partitioning-based clustering algorithms
- DBSCAN is an interesting density-based algorithms
- Quality of clustering results can be evaluated in various ways

References (1)

- R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98
- M. R. Anderberg. Cluster Analysis for Applications. Academic Press, 1973.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure, SIGMOD'99.
- P. Arabie, L. J. Hubert, and G. De Soete. Clustering and Classification. World Scietific, 1996
- M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. KDD'96.
- M. Ester, H.-P. Kriegel, and X. Xu. Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification. SSD'95.
- D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2:139-172, 1987.
- D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. In Proc. VLDB'98.
- S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. SIGMOD'98.
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Printice Hall, 1988.

References (2)

- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98.
- G. J. McLachlan and K.E. Bkasford. Mixture Models: Inference and Applications to Clustering. John Wiley and Sons, 1988.
- P. Michaud. Clustering techniques. Future Generation Computer systems, 13, 1997.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
- E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition, 101-105.
- G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB'98.
- W. Wang, Yang, R. Muntz, STING: A Statistical Information grid Approach to Spatial Data Mining, VLDB'97.
- T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH : an efficient data clustering method for very large databases. SIGMOD'96.

http://www.cs.sfu.ca/~han

Thank you !!!