Clustering

Huiping Cao

Outline

- What is cluster analysis?
- Clustering approaches

What is cluster analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Unsupervised learning: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

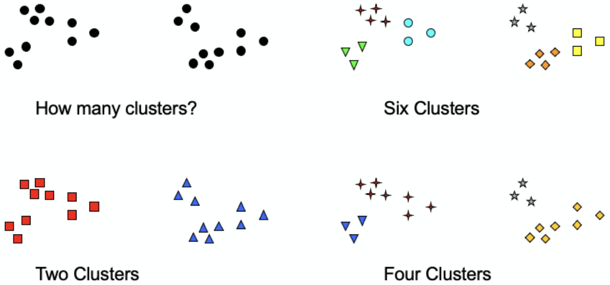
Applications of Cluster Analysis

- Understanding
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
  - Segment customers into a small number of groups for marketing activities
- Summarization
  - Reduce the size of large data sets

What is not Cluster Analysis?

- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
  - Clustering is a grouping of objects based on the data
- Supervised classification
  - Have class label information

Notion of a Cluster can be Ambiguous



Quality: what is good clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Requirements of clustering in data mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Major clustering approaches (1)

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: AGNES, DIANA, BIRCH, ROCK, CAMELEON
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue

Major clustering approaches (2)

- Grid-based approach:
  - Based on a multiple-level granularity structure
  - Typical methods: CLIQUE, STING, WaveCluster
- Model-based:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM, COBWEB

Major clustering approaches (3)

- Clustering high-dimensional data:
  - Consider large number of features and dimensions
  - Typical methods: CLIQUE, PROCLUS
- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: pCluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering

Centroid, Radius and Diameter of a cluster

- Centroid: the "center" of a cluster K_i

$$\bar{C}_i = \frac{\sum_{p=1}^{n} t_{ip}}{n}$$

Here, t_{ip} is a point in cluster K_i and n is the number of points in cluster K_i

- Radius: square root of average distance from any point of the cluster to its centroid

$$R_i = \sqrt{\frac{\sum_{p=1}^{n} dist(t_{ip}, \bar{C}_i)^2}{n}}$$

- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$D_{i,j} = \sqrt{\frac{\sum_{p=1}^{n} \sum_{q=1}^{n} dist(t_{ip}, t_{jq})^2}{n \cdot (n-1)}}$$

Typical alternatives to calculate the distance between clusters (1)

- Single link: smallest distance between an element in one cluster and an element in the other, i.e.,

$$dist(K_i, K_j) = min_{p,q} dist(t_{ip}, t_{jq})$$

- Complete link: largest distance between an element in one cluster and an element in the other, i.e.,

$$dist(K_i, K_j) = max_{p,q} dist(t_{ip}, t_{jq})$$

- Average: avg distance between an element in one cluster and an element in the other, i.e.,

$$dist(K_i, K_j) = avg_{p,q} dist(t_{ip}, t_{jq})$$

Typical alternatives to calculate the distance between clusters (2)

- Centroid: distance between the centroids of two clusters,

$$dist(K_i, K_j) = dist(\bar{C}_i, \bar{C}_j)$$

- Medoid: distance between the medoids of two clusters,

$$dist(K_i, K_j) = dist(M_i, M_j)$$

Medoid: one chosen, centrally located object in the cluster

Basic concept

- Partitioning method criterion: Construct a partition of a database D of N objects into a set of k clusters, s.t., min sum of squared error, which is also called **within-cluster variation**.

$$E = \sum_{i=1}^{k} \sum_{p \in K_i} (\bar{C}_i - p)^2$$

Partitioning algorithms

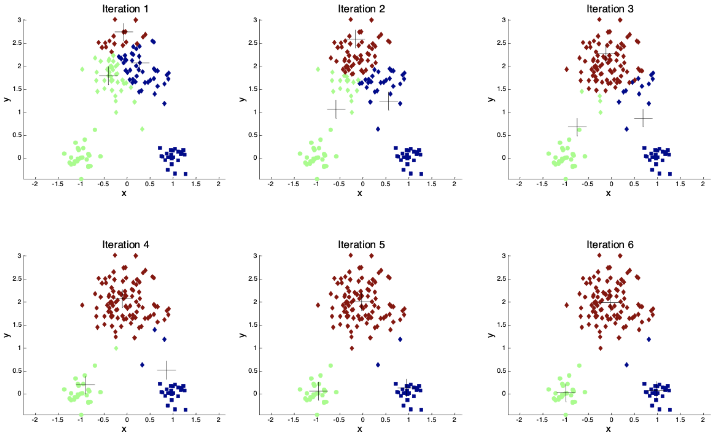
- Given a data set D of N objects and k, the number of clusters
- A partitioning algorithm is to find k partitions that optimize a given objective partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: k-means and k-medoids algorithms
    - k-means (MacQueen' 67): Each cluster is represented by the center of the cluster
    - k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw' 87): Each cluster is represented by one of the objects in the cluster

The K-Means clustering method

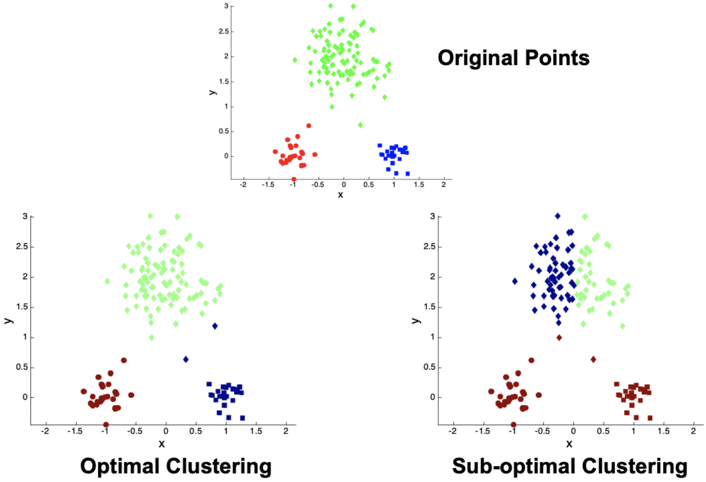
- Given k, the k-means algorithm is implemented in four steps:

  1. Partition objects into k non-empty subsets based on randomly chosen centroids
  2. Compute the centroids of the clusters represented by the current partition (the centroid is the center, i.e., mean point, of the cluster)
  3. Assign each object to the cluster with the nearest centroid
  4. Go back to Step 2, stop when no more new assignment

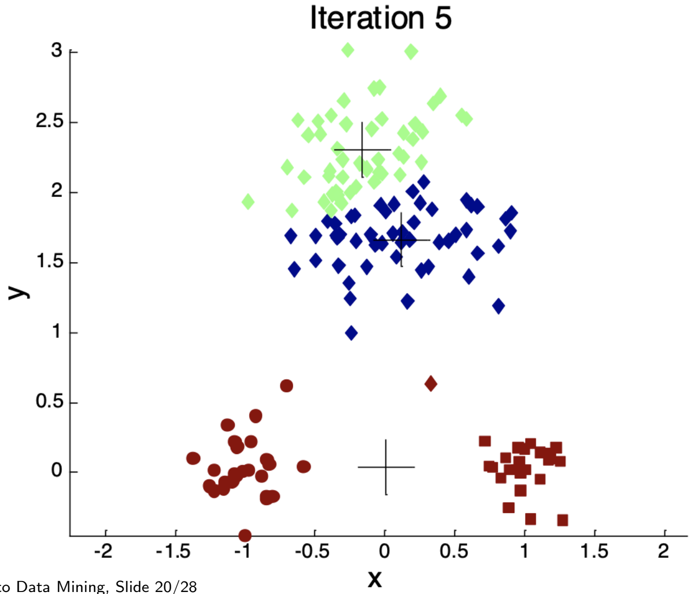
Example of K-means Clustering



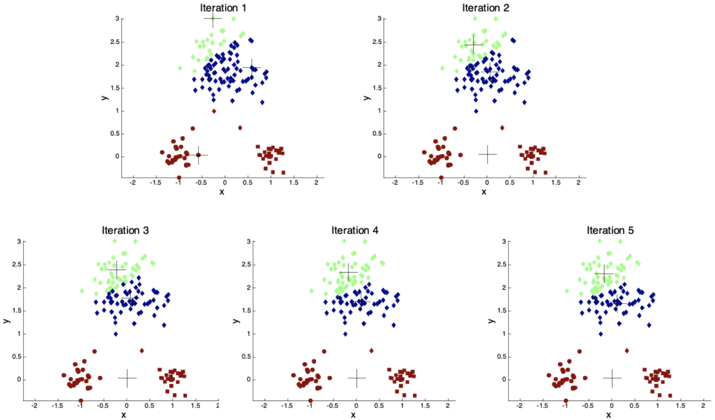
Two different K-means Clusterings



Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids (cont.)



Solutions to Initial Centroids Problem

- Multiple runs: Helps, but probability is not on your side
- Use some strategy to select the k initial centroids and then select among these initial centroids
  - Select most widely separated: K-means++ is a robust way of doing this selection
  - Use hierarchical clustering to determine initial centroids
- Bisecting K-means: Not as susceptible to initialization issues

K-means++

To select a set of initial centroids, C, perform the following

1. Select an initial point at random to be the first centroid
2. For k – 1 steps
3. For each of the N points, x_i, $1 \leq i \leq N$, find the minimum squared distance to the currently selected centroids, $C_1, \ldots, C_j$, $1 \leq j < k$, i.e., $\min_j d^2(C_j, x_i)$
4. Randomly select a new centroid by choosing a point with probability proportional to $\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$ is
5. End For

Comments on the K-Means method

- Strength: Relatively efficient: $O(IkN)$, where N is the number of objects, k is the number of clusters, and I is the number of iterations. Normally, $k, I \ll N$.
- Weakness
  - Applicable only when mean is defined, then what about categorical data?
  - Need to specify k, the number of clusters, in advance
  - Unable to handle noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes

Variations of the K-Means method

- A few variants of the k-means which differ in
  - Selection of the initial k means
  - Dissimilarity calculations
  - Strategies to calculate cluster means

What is the problem of the K-Means?

- The k-means algorithm is sensitive to outliers! An object with an extremely large value may substantially distort the distribution of the data.

  - Given seven points in 1D space: 1,2,3,8,9,10,25 and $k = 2$
    - Intuitively, {1,2,3},{8,9,10,25}

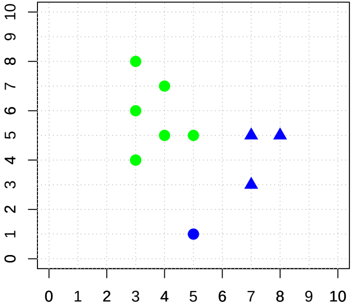
$$\begin{aligned} SSE =&(1-2)^2+(2-2)^2+(3-2)^2+(8-13)^2+(9-13)^2\\ &+(10-13)^2+(25-13)^2\\ =&196 \end{aligned}$$

    - Another partitioning (from K-means) {1,2,3,8}, {9,10,25}

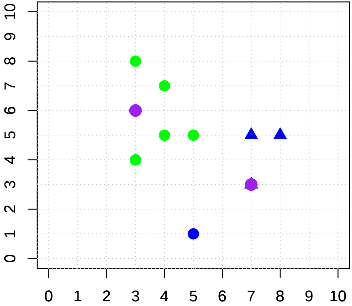
$$\begin{aligned} SSE =&(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(8-3.5)^2\\ &+(9-14.67)^2+(10-14.67)^2+(25-14.67)^2\\ =&189.67 \end{aligned}$$

The K-Medoids clustering method

K-Medoids: Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.



Clusters



Medoids

References

- Chapter 7: Introduction to Data Mining (2nd Edition) by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar
- Python K-means: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Python K-Medoids: https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html
- Python CLARAS: https://pyclustering.github.io/docs/0.10.1/html/d6/d42/classpyclustering_1_1cluster_1_1clarans_1_1clarans.html