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
- Evaluation of Clustering
- Summary

Cluster Analysis: Basic Concepts

- Cluster: A collection of data objects
  - **similar** (or related) to one another within the same group
  - **dissimilar** (or unrelated) to the objects in other groups
- Cluster analysis (or clustering, data segmentation, …)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- **Unsupervised learning**: no predefined classes (i.e., learning by observations vs. learning by examples: supervised)
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

Clustering as Preprocessing

- Summarization:
  - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
  - Image processing: vector quantization
- Finding K-nearest Neighbors:
  - Localizing search to one or a small number of clusters
- Outlier detection:
  - Outliers are often viewed as those "far away" from any cluster

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- 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.

Considerations for Cluster Analysis

- Partitioning criteria
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Clustering space
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Requirements of Clustering in Data Mining

- Scalability
  - Many clustering algorithms work well on small data sets containing fewer than several hundred data objects
  - A large database may contain millions or even billions of objects
  - Clustering on only a sample of a given large data set may lead to biased results
- Ability to deal with different types of attributes
  - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Discovery of clusters with arbitrary shape
  - Algorithms based Euclidean and Manhattan distance measures tend to find spherical clusters with similar size and density
  - A cluster could be of any shape that is not spherical

Requirements of Clustering in Data Mining

- Minimal requirements for domain knowledge to determine input parameters
  - Many algorithms require users to provide domain knowledge in the form of input parameters (results may be sensitive)
    - e.g. desired number of clusters
  - Requiring the specification of domain knowledge not only burdens users, but also makes the quality of clustering difficult to control
- Able to deal with noise and outliers
  - Most real-world data sets contain outliers and/or missing, unknown, or erroneous data
  - Algorithms sensitive to such noise and may produce poor-quality clusters.

Requirements of Clustering in Data Mining

- Insensitive to order of input records
  - Given a set of data objects, algorithms return similar clusterings
  - Some algorithms return significantly different clusters depending on the order in which the objects are presented
- High dimensionality
  - Most clustering algorithms are good at handling low-dimensional data (only two or three dimensions)
  - High-dimensional data can be very sparse and highly skewed

Requirements of Clustering in Data Mining

- Incorporation of user-specified constraints
  - Real-world applications may need to perform clustering under various kinds of constraints
  - Challenging task to find data groups with good clustering behavior that satisfy specified constraints
- Interpretability and usability
  - Tie clustering with specific semantic interpretations and applications
  - Important to study how an application goal may influence the selection of clustering features and clustering methods

Chapter 7. Cluster Analysis

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

Measure the Quality of Clustering

- Dissimilarity/Similarity metric:
  - Similarity is expressed in terms of a distance function, which is typically metric: ***d(i,j)***
  - The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
  - Weights should be associated with different variables based on applications and data semantics.
- Quality of clustering:
  - There is a separate "quality" function that measures the "goodness" of a cluster.
  - It is hard to define "similar enough" or "good enough"
    - the answer is typically highly subjective.

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*

$$d(i,j) = \sqrt[q]{(|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + ... + |x_{i_p} - x_{j_p}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and q is a positive integer

  - Hence the distance is defined as L-q *norm*

- Properties:
  - $d(i,j) \geq 0$ if $i \neq j$ and $d(i,i) = 0$ (Positive definiteness)
  - $d(i,j) = d(j,i)$ (Symmetry)
  - $d(i,j) \leq d(i,k) + d(k,j)$ (Triangle Inequality)
  - A distance that satisfies these properties is a metric

Similarity and Dissimilarity Between Objects (Cont.)

- Special cases of Minkowski distance are often used
- *If q = 1, d is Manhattan distance*

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

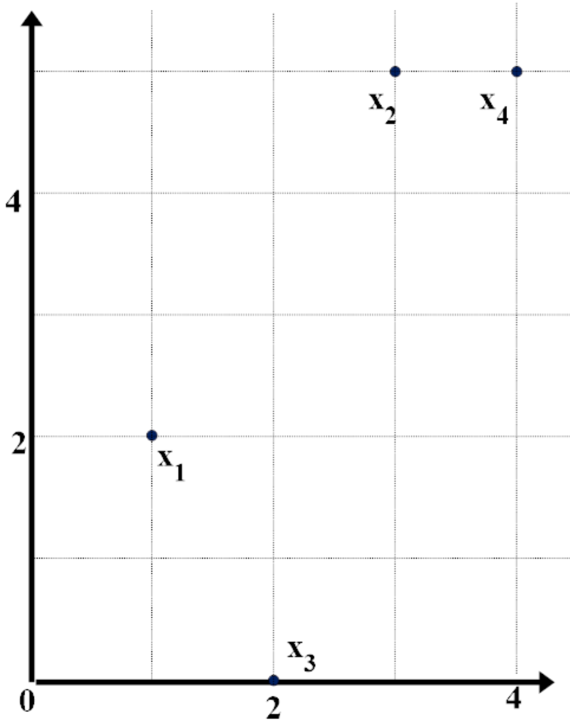
- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + ... + |x_{i_p} - x_{j_p}|^2)}$$

- One can also use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures.

Similarity and Dissimilarity Between Objects (Cont.)

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
  - creating a new binary attribute for each of the M nominal states

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object j		
		1	0	*sum*
	1	a	b	$a+b$
Object *i*	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Simple matching coefficient (invariant, if the binary variable is *symmetric*): $d(i, j) = \frac{b + c}{a + b + c + d}$
- Jaccard coefficient (noninvariant if the binary variable is *asymmetric*): $d(i, j) = \frac{b + c}{a + b + c}$

Dissimilarity between Binary Variables

- Example: Medical tests data

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

  - gender is a symmetric attribute
  - the remaining attributes are asymmetric binary
  - let the values Y and P be set to 1, and the value N be set to 0

		Object j		
		1	0	*sum*
Object i	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$d(i,j) = \frac{b+c}{a+b+c+d} \qquad d(i,j) = \frac{b+c}{a+b+c}$$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Attributes of Mixed Type

- A database may contain all attribute types
  - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

  - *f* is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
  - f is numeric: use the normalized distance

Chapter 7. Cluster Analysis

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

Major Clustering Approaches (I)

- **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: Diana, Agnes, BIRCH, CAMELEON
- **Density-based approach**:
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue
- **Grid-based approach**:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches (II)

- **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
- **Frequent pattern-based**:
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- **User-guided or constraint-based**:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- **Link-based clustering**:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database ***D*** of ***n*** objects into a set of ***k*** clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i and p is a data point)

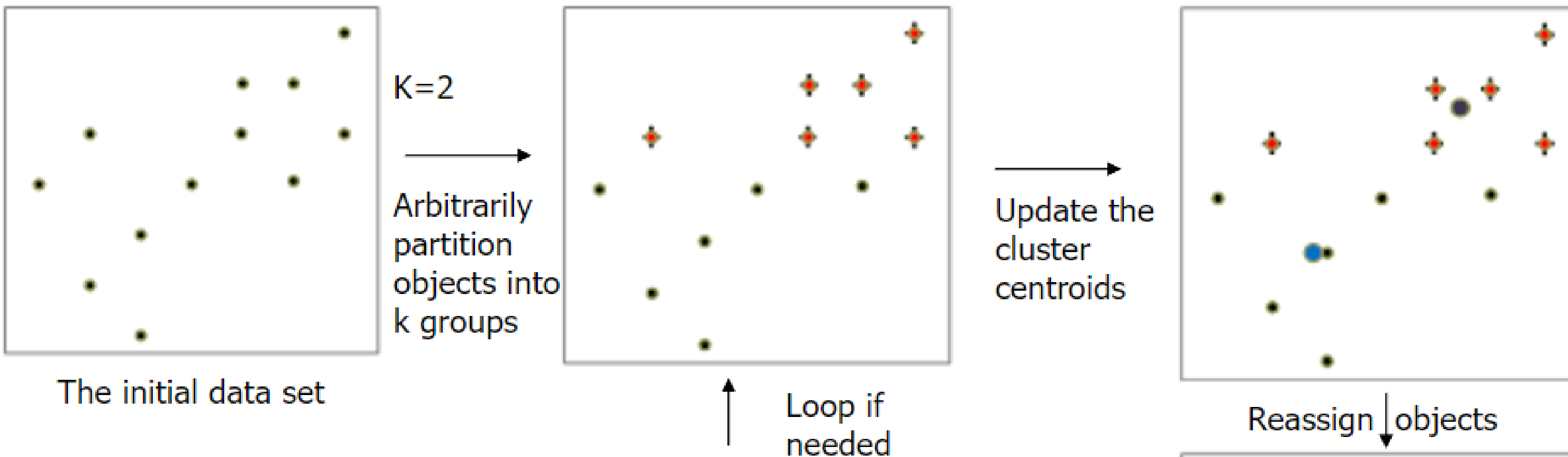
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

- Given *k*, find a partition of *k* clusters that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods:
    - k-means algorithm: Each cluster is represented by the center of the cluster
    - k-medoids or PAM (Partition around medoids) algorithm: Each cluster is represented by one of the objects in the cluster

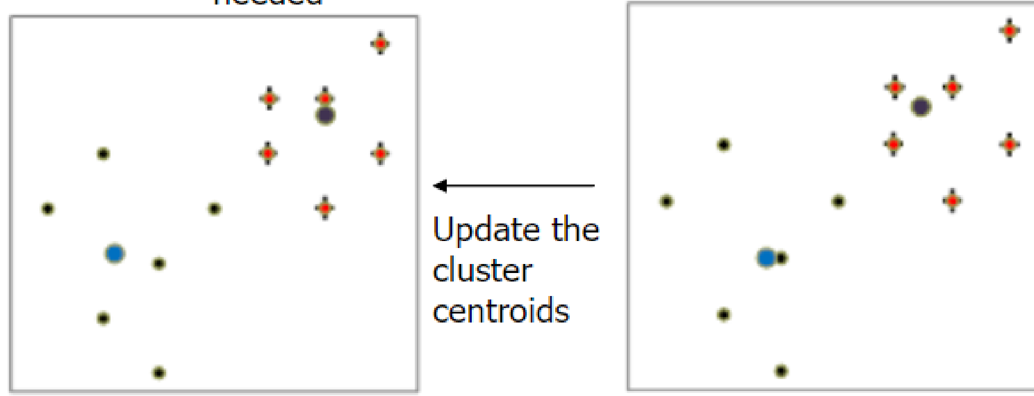
The *K*-Means Clustering Method

- Given *k*, the *k*-means algorithm is implemented in four steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster)
  3. Assign each object to the cluster with the nearest seed point
  4. Repeat Step 2, stop when the assignment does not change

The *K*-Means Clustering Method

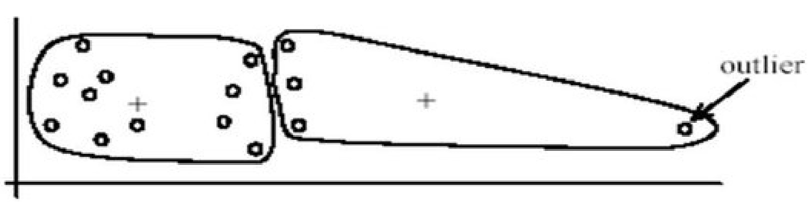


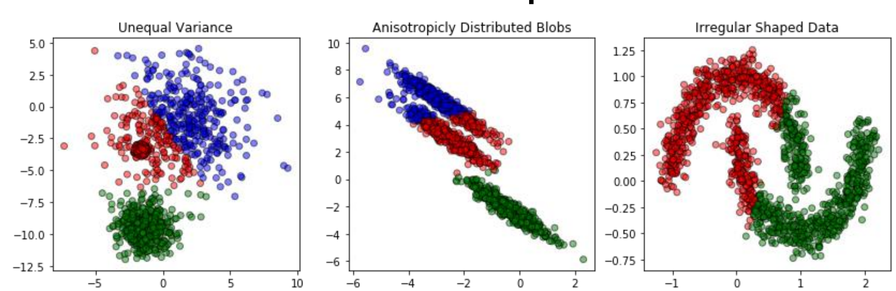
- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change



Comments on the K-Means Method

- Strength: Efficient compared to other methods of clustering
- Weakness
  - Often terminates at a *local optimal*
  - Applicable only to objects in a continuous n-dimensional space
    - Using the k-modes method for categorical data
    - In comparison, k-medoids can be applied to a wide range of data
  - Need to specify k, the number of clusters, in advance (there are ways to automatically determine the best k)
  - Sensitive to noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes





Variations of the K-Means Method

- Most of the variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes*
  - Replacing means of clusters with **modes**
  - Using new dissimilarity measures to deal with categorical objects
  - Using a **frequency**-based method to update modes of clusters
  - A mixture of categorical and numerical data: k-prototype method

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

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 !!!