Clustering

What is cluster analysis?

- What is a cluster?
  - A cluster is a collection of data objects which are
    - **Similar** (or related) to one another within the same group (i.e., cluster)
    - **Dissimilar** (or unrelated) to the objects in other groups (i.e., clusters)
- Cluster analysis (or clustering, data segmentation, …)
  - Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible
- Cluster analysis is unsupervised learning (i.e., no predefined classes)
  - This contrasts with classification (i.e., supervised learning)

What is cluster analysis?

- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms

What Is Good Clustering?

- A good clustering method will produce high quality clusters which should have
  - **High intra-class similarity: Cohesive within clusters**
  - **Low inter-class similarity: Distinctive between clusters**
- Quality function
  - There is usually a separate "quality" function that measures the "goodness" of a cluster
  - It is hard to define "similar enough" or "good enough" and is typically subjective

There exist many similarity and dissimilarity measures and different functions for different applications

Common Distance measures:

- Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters.
- They include:
- 1. The Euclidean distance (also called 2-norm distance) is given by:

$$d(x,y)=\sum_{i=1}^{p}|x_i-y_i|$$

- 2. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

$$d(x,y)=\sqrt[2]{\sum_{i=1}^{p}|x_i-y_i|^2}$$

-

Common Distance measures:

3. **The maximum norm is given by:**

$$d(x,y)=\max_{1\le i\le p}|x_i-y_i|$$

4. The Mahalanobis distance corrects data for different scales and correlations in the variables.

5. Inner product space: The angle between two vectors can be used as a distance measure when clustering high dimensional data

6. Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

Cluster Analysis: Applications

- **A key intermediate step for other data mining tasks**
  - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
  - Outlier detection: Outliers—those “far away” from any cluster
- **Data summarization, compression, and reduction**
  - Ex. Image processing: Vector quantization
- **Collaborative filtering, recommendation systems, or customer segmentation**
  - Find like-minded users or similar products

Cluster Analysis: Applications

- **Dynamic trend detection**
  - Clustering stream data and detecting trends and patterns

- **Multimedia data analysis, biological data analysis and social network analysis**
  - Ex. Clustering images or video/audio clips, gene/protein sequences, etc.

Considerations for Cluster Analysis

Partitioning criteria (Single level vs. hierarchical partitioning)

- **Single level:** All clusters are conceptually at the same level no hierarchy exists among clusters.
  - Eg: partitioning customers into groups so that each group has its manager.
- **Hierarchical level:** Clusters at different semantic levels.
  - Eg: general topics: “sports”, “politics” and subtopics in text mining.

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 cluster)

Considerations for Cluster Analysis

Similarity measure

- Distance-based (e.g., **Euclidean, road network, vector**) vs. similarity measure defined as connectivity-based **(e.g., density or contiguity)** not on absolute distance between two objects.

Clustering space

- Clustering methods search for clusters within entire given space (often when low dimensional) vs. subspaces (often in high-dimensional clustering due to presence of irrelevant attributes)

Requirements and Challenges

- **Scalability**
  - Highly scalable clustering algorithms are needed to work on large database to produce unbiased results.

Quality

- Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, mixture of multiple types and complex data types such as graphs,sequences images and documents.

Discovery of clusters with arbitrary shape

- Distance based clustering algorithm produces spherical clusters with similar size and density
- Important to develop clusters of any shape.
- Develop algorithms that can detect clusters of arbitrary shape

Requirements and Challenges

- **Ability to deal with noisy data**
  - Most data contains outliers, missing , unknown or erroneous data
  - Clustering algorithms are sensitive produce poor quality clusters
  - Need methods that are robust to noise
- **Incremental clustering and insensitivity to input order:**
  - Incremental updates requires recomputing from scratch and return different clusters depending of the order of the data given.
  - Algorithms may be sensitive to the input data order,different clusters depending on the order.
  - Incremental clustering algorithms and algorithms insensitive to the input order are needed.
- **High dimensionality:** Need to handle high dimension data

Requirements and Challenges

- **Constraint-based clustering**
  - User-given preferences or constraints; domain knowledge; user queries
- **Interpretability and usability**
  - Clustering results should be interpretable, comprehensible and usable
  - Can able to tie with specific semantic interpretations and applications

Type of data in clustering analysis

- Interval-scaled variables:
- Binary variables:
- Nominal, ordinal, and ratio variables:
- Variables of mixed types:

Interval-valued variables

- Standardize data
  - Calculate the mean absolute deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + ... + |x_{nf} - m_f|)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - Calculate the standardized measurement (z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

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:

$$d(i,j)=\sqrt[q]{(|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\ldots+|x_{ip}-x_{jp}|^q)}$$

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

- If q = 1, d is Manhattan distance

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

Similarity and Dissimilarity Between Objects (Cont.)

- If q = 2, d is Euclidean distance:

$$d(i,j)=\sqrt{\left(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\ldots+|x_{ip}-x_{jp}|^2\right)}$$

- Also, one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarit y measures

Binary Variables

- A contingency table for binary data

	1	0	*sum*
1	a	b	$a+b$
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):

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

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., red, yellow, blue, green
- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

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

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - replace x_{if} by their rank
  - map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - compute the dissimilarity using methods for interval-scaled variables

Ratio-Scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale, approximately at exponential scale, such as Ae^{Bt} or Ae^{-Bt}
- Methods:
  - treat them like interval-scaled variables—*not a good choice!* (why?—the scale can be distorted)
  - apply logarithmic transformation

$$y_{if} = log(x_{if})$$

  - treat them as continuous ordinal data treat their rank as interval-scaled

Variables of Mixed Types

- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- One may use a weighted formula to combine their effects

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

  - *f* is binary or nominal:

    $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ o.w.

  - *f* is interval-based: use the normalized distance
  - *f* is ordinal or ratio-scaled
    - compute ranks r_{if} and
    - and treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Basic Clustering Methods

- Partitioning approach:
  - Construct various k partitions and then evaluate them by some criterion
  - Adopts exclusive separation, distance methods, uses iterative relocation technique to improve partition.
  - Uses heuristic methods, finds spherical-shaped clusters,local optimum
- **Typical methods: k-means, k-medoids, CLARANS**

Basic Clustering Methods

- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Classified using agglomerative (bottom-up) or divisive approach(top-down)
  - adopts distance measures or density measures
  - Considers clusters in subspaces, cannot possible to correct erroneous decision since once done cannot be undone.
  - Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEO

Basic Clustering Methods

- Density Based Methods:
  - Based on connectivity and density functions
  - Methods used to filter out noise and outliers.
  - Divides set of objects in mutually exclusive or hierarchy of clusters.
  - Extended to full space or subspace clustering
- **Typical methods: DBSACN, OPTICS, DenClue**

Basic Clustering Methods

- Grid Based Methods:
  - based on a multiple,level granularity structure
  - Quantize the object space into finite number of cells that form a grid structure.
  - Fast processing and dependent on number of cells

Typical methods: STING, WaveCluster, CLIQUE