Principles of Knowledge Discovery in Databases

Fall 1999

Chapter 8: Data Clustering

Dr. Osmar R. Zaïane

University of Alberta

Source:
Dr. Jiawei Han

Summary of Last Chapter

- What is classification of data and prediction?
- How do we classify data by decision tree induction?
- What are neural networks and how can they classify?
- What is Bayesian classification?
- Are there other classification techniques?
- How do we predict continuous values?

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- Data mining operations
- Data summarization
- Association analysis
- Classification and prediction
- **Clustering**
- Web Mining
- Similarity Search
- *Other topics if time permits*

Chapter 8 Objectives

Learn basic techniques for data clustering.

Understand the issues and the major challenges in clustering large data sets in multi-dimensional spaces.

Data Clustering Outline

- What is cluster analysis?
- What do we use clustering for?
- Are there different approaches to data clustering?
- What are the major clustering techniques?
- What are the challenges to data clustering?

What is a Cluster?

According to the Webster dictionary:

- a number of similar things growing together or of things or persons collected or grouped closely together: BUNCH.
- two or more consecutive consonants or vowels in a segment of speech.
- a group of buildings and esp. houses built close together on a sizable tract in order to preserve open spaces larger than the individual yard for common recreation.
- an aggregation of stars, galaxies, or super galaxies that appear close together in the sky and seem to have common properties (as distance).

➔ **A cluster is a closely-packed group (of people or things).**

What is Clustering in Data Mining?

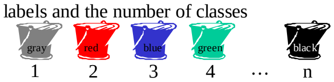
Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes, called **clusters**.

– Helps users understand the natural grouping or structure in a data set.

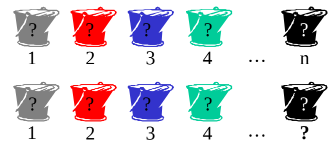
- Cluster: a collection of data objects that are “similar” to one another and thus can be treated collectively as one group.
- Clustering: unsupervised classification: no predefined classes.

Supervised and Unsupervised

Supervised Classification = Classification
➔ We know the class labels and the number of classes



Unsupervised Classification = Clustering
➔ We do not know the class labels and may not know the number of classes



What Is Good Clustering?

- A good clustering method will produce high quality clusters in which:
  – the **intra-class** (that is, intra-cluster) similarity is high.
  – the **inter-class** similarity is low.
- The **quality** of a clustering result also 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.
- The quality of a clustering result also depends on the definition and representation of cluster chosen.

Requirements of Clustering in Data Mining

- Scalability
- Dealing 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
- Interpretability and usability.

Data Clustering Outline

- What is cluster analysis?
- What do we use clustering for?
- Are there different approaches to data clustering?
- What are the major clustering techniques?
- What are the challenges to data clustering?

Applications of Clustering

- Clustering has wide applications in
  – Pattern Recognition
  – Spatial Data Analysis:
    - create thematic maps in GIS by clustering feature spaces
    - detect spatial clusters and explain them in spatial data mining.
  – Image Processing
  – Economic Science (especially market research)
  – WWW:
    - Document classification
    - Cluster Weblog data to discover groups of similar access patterns

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.
- **Earthquake studies:** Observed earthquake epicenters should be clustered along continent faults.

Data Clustering Outline

- What is cluster analysis?
- What do we use clustering for?
- Are there different approaches to data clustering?
- What are the major clustering techniques?
- What are the challenges to data clustering?

Major Clustering Techniques

- Clustering techniques have been studied extensively in:
  - Statistics, machine learning, and data mining

  with many methods proposed and studied.
- Clustering methods can be classified into 5 approaches:
  - **partitioning algorithms**
  - **hierarchical algorithms**
  - **density-based method**
  - **grid-based method**
  - **model-based method**

Five Categories of Clustering Methods

- **Partitioning algorithms**: Construct various partitions and then evaluate them by some criterion.
- **Hierarchy algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion. There is an agglomerative approach and a divisive approach.
- **Density-based**: based on connectivity and density functions.
- **Grid-based**: based on a multiple-level granularity structure.
- **Model-based**: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.

Data Clustering Outline

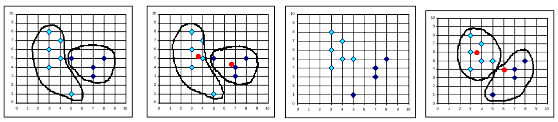
- What is cluster analysis?
- What do we use clustering for?
- Are there different approaches to data clustering?
- What are the major clustering techniques?
- What are the challenges to data clustering?

Partitioning Algorithms: Basic Concept

- **Partitioning method:** Construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen 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 4 steps:
  1. Partition objects into *k* nonempty subsets
  2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  3. Assign each object to the cluster with the nearest seed point.
  4. Go back to Step 2, stop when no more new assignment.



Comments on the *K-Means* Method

- Strength of the *k-means*:
  - *Relatively efficient*: *O*(*tkn*), where *n* is # of objects, *k* is # of clusters, and *t* is # of iterations. Normally, *k*, *t* << *n*.
  - Often terminates at a *local optimum*.
- Weakness of the *k-means*:
  - 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.
- Handling categorical data: *k-modes* (Huang'98):
  - 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.

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
  - To achieve this goal, only the definition of distance from any two objects is needed.
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.
  - *PAM* works effectively for small data sets, but does not scale well for large data sets.
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling.
- Focusing + spatial data structure (Ester et al., 1995).

PAM (Partitioning Around Medoids) (1987)

- PAM (Kaufman and Rousseeuw, 1987), built in S+.
- Use real object to represent the cluster.
  1. Select ***k*** representative objects arbitrarily
  2. For each pair of non-selected object ***h*** and selected object ***i***, calculate the total swapping cost TC_{ih}.
     - If $TC_{ih} < 0$, ***i*** is replaced by ***h***.
  3. Then assign each non-selected object to the most similar representative object
  4. Repeat steps 2-3 until there is no change

$O(k(n-k)^2)$

CLARA (Clustering Large Applications) (1990)

- *CLARA* (Kaufmann and Rousseeuw in 1990)
- Built in statistical analysis packages, such as S+.
- It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output.
- Strength of *CLARA*: $O(kS^2+k(n-k))$
  - deal with larger data sets than *PAM*.
- Weakness of *CLARA*:
  - Efficiency depends on the sample size.
  - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased.

CLARANS ("Randomized" CLARA) (1994)

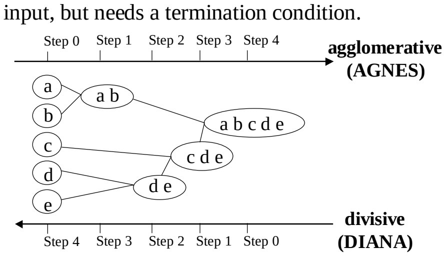
- *CLARANS* (A Clustering Algorithm based on Randomized Search) by Ng and Han in 1994.
- CLARANS draws sample of neighbours dynamically.
- The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of *k* medoids.
- If the local optimum is found, *CLARANS* starts with new randomly selected node in search for a new local optimum.
- It is more efficient and scalable than both *PAM* and *CLARA*.
- Focusing techniques and spatial access structures may further improve its performance (Ester et al.'95).

Two Types of Hierarchical Clustering Algorithms

- **Agglomerative** (bottom-up): merge clusters iteratively.
  - start by placing each object in its own cluster.
  - merge these atomic clusters into larger and larger clusters.
  - until all objects are in a single cluster.
  - Most hierarchical methods belong to this category. They differ only in their definition of *between-cluster similarity*.
- **Divisive** (top-down): split a cluster iteratively.
  - It does the reverse by starting with all objects in one cluster and subdividing them into smaller pieces.
  - Divisive methods are not generally available, and rarely have been applied.

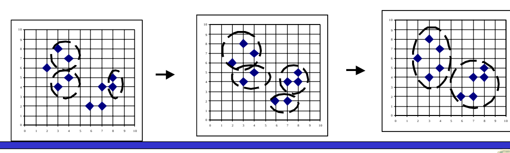
Hierarchical Clustering

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



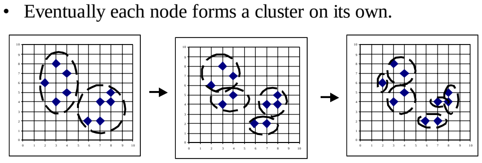
AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, such as S+.
- 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



DIANA (Divisive Analysis)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, such as S+.
- Inverse order of AGNES.
- Eventually each node forms a cluster on its own.



More on Hierarchical Clustering

- Major weakness of agglomerative clustering methods:
  - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo what was done previously.
- Integration of hierarchical clustering with distance-based method:
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters.
  - CURE (1998): selects well-scattered points from the cluster and then shrinks them towards the center of the cluster by a specified fraction.
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling.

BIRCH (1996)

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies, by Zhang, Ramakrishnan, Livny (SIGMOD'96).
- Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering:
  - Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)
  - Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree.
- *Scales linearly*: finds a good clustering with a single scan and improves the quality with a few additional scans.
- *Weakness:* handles only numeric data, and sensitive to the order of the data record.

Clustering Feature Vector

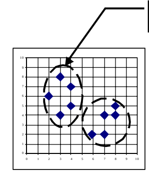
Clustering Feature: $CF = (N, \vec{LS}, SS)$

N: **Number of data points**

LS: $\sum_{i=1}^{N} \vec{X_i}$

SS: $\sum_{i=1}^{N} \vec{X_i}^2$

CF = (5, (16,30),(54,190))



(3,4)
(2,6)
(4,5)
(4,7)
(3,8)

See class presentation for algorithm details

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: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - DENCLUE: Hinneburg & D. Keim (KDD'98)
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

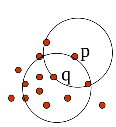
DBSCAN: A Density-Based Clustering

- DBSCAN: Density Based Spatial Clustering of Applications with Noise.
  - Proposed by Ester, Kriegel, Sander, and Xu (KDD'96)
  - 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

Density-Based Clustering: Background

- 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 wrt. ***Eps***, ***MinPts*** if
  - 1) p belongs to $N_{Eps}(q)$
  - 2) core point condition:

    $|N_{Eps}(q)| >= MinPts$

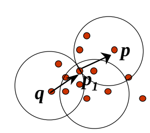


MinPts = 5

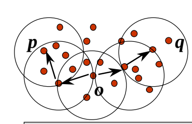
Eps = 1 cm

Density-Based Clustering: Background

- Density-reachable:
  - A point p is density-reachable from a point q wrt. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



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



See class presentation for algorithm details (on-line)

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99).
  - Extensions to DBSCAN.
  - Produces a special order of the database with regard to its density-based clustering structure.
  - This cluster-ordering contains information equivalent to the density-based clusterings corresponding to a broad range of parameter settings.
  - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure.
  - Can be represented graphically or using visualization techniques.

CLIQUE (1998)

- CLIQUE (Clustering In QUEst) by Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98).
- Automatic subspace clustering of high dimensional data
- CLIQUE can be considered as both density-based and grid-based
- Input parameters:
  - size of the grid and a global density threshold
- It *partitions* an *m*-dimensional data space into non-overlapping rectangular units.
- A unit is *dense* if the fraction of total data points contained in the unit exceeds the input *model parameter*.
- A *cluster* is a maximal set of connected dense units.

CLIQUE: The Major Steps

- Partition the data space and find the number of points that lie inside each cell of the partition.
- Identify the subspaces that contain clusters, using the DNF expression
- Identify clusters:
  - Determine dense units in all subspaces of interests.
  - Determine connected dense units in all subspaces of interests.
- Generate minimal description for the clusters
  - Determine maximal regions that cover a cluster of connected dense units for each cluster.
  - Determination of minimal cover for each cluster.

See class presentation for algorithm details (on-line)

Grid-Based Clustering Method

- Grid-based clustering: using multi-resolution grid data structure.
- Several interesting studies:
  - **STING** (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - BANG-clustering/**GRIDCLUS** (Grid-Clustering) by Schikuta (1997)
  - **WaveCluster** (a multi-resolution clustering approach using wavelet method) by Sheikholeslami, Chatterjee and Zhang (1998)
  - **CLIQUE** (Clustering In QUEst) by Agrawal, Gehrke, Gunopulos, Raghavan (1998).

Model-Based Clustering Methods

- Use certain models for clusters and attempt to optimize the fit between the data and the model.
- Neural network approaches:
  - The best known neural network approach to clustering is the SOM (*self-organizing feature map*) method, proposed by Kohonen in 1981.
  - It can be viewed as a nonlinear projection from an *m*-dimensional input space onto a lower-order (typically *2*-dimensional) regular lattice of cells. Such a mapping is used to identify clusters of elements that are similar (in a *Euclidean* sense) in the original space.
- Machine learning: probability density-based approach:
  - Grouping data based on probability density models: based on how many (possibly weighted) features are the same.
  - COBWEB (Fisher'87) Assumption: The probability distribution on different attributes are independent of each other --- This is often too strong because correlation may exist between attributes.

Model-Based Clustering Methods (con't)

- Statistical approach*:* Gaussian mixture model (Banfield and Raftery, 1993): A probabilistic variant of *k-means* method.
  - It starts by choosing *k* seeds, and regarding the seeds as means of Gaussian distributions, then iterates over two steps called the *estimation* step and the *maximization* step, until the Gaussians are no longer moving.
  - Estimation: calculating the responsibility that each Gaussian has for each data point.
  - Maximization: The mean of each Gaussian is moved towards the centroid of the entire data set.
- Statistical Approach: AutoClass (Cheeseman and Stutz, 1996): A thorough implementation of a Bayesian clustering procedure based on mixture models.
  - It uses Bayesian statistical analysis to estimate the number of clusters.

Clustering Categorical Data: ROCK

- ROCK: Robust Clustering using linKs, by S. Guha, R. Rastogi, K. Shim (ICDE'99).
  - Use links to measure similarity/proximity
  - Not distance-based
  - Computational complexity: $O(n^2 + nm_m m_a + n^2 \log n)$
- Basic ideas:
  - Similarity function and neighbours: $Sim(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$

    Let T_1 = {1,2,3}, T_2={3,4,5}

$$Sim(T1, T2) = \frac{|\{3\}|}{|\{1,2,3,4,5\}|} = \frac{1}{5} = 0.2$$

See class presentation for algorithm details (on-line)

Data Clustering Outline

- What is cluster analysis?
- What do we use clustering for?
- Are there different approaches to data clustering?
- What are the major clustering techniques?
- What are the challenges to data clustering?

Problems and Challenges

- Considerable progress has been made in scalable clustering methods:
  - Partitioning: k-means, k-medoids, CLARANS
  - Hierarchical: BIRCH, CURE
  - Density-based: DBSCAN, CLIQUE, OPTICS
  - Grid-based: STING, WaveCluster.
  - Model-based: Autoclass, Denclue, Cobweb.
- Current clustering techniques do not address all the requirements adequately (and concurrently).
- Large number of dimensions and large number of data items.
- Strict clusters vs. overlapping clusters.