Chapter 8. Cluster Analysis-II

- Introduction
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Summary

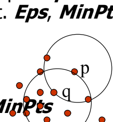
Density-Based Clustering Methods

- 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)

Density-Based Clustering: Background

- Two parameters:
  - ***Eps***: Maximum radius of the neighbourhood
  - ***MinPts***: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid dist(p,q) <= Eps\}$
- 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:

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

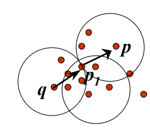


MinPts = 5

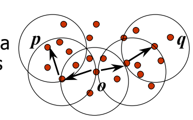
Eps = 1 cm

Density-Based Clustering: Background (II)

- 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*.



Cluster

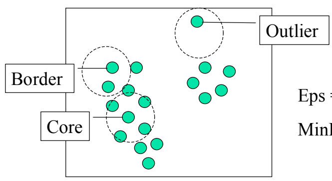
- Cluster: Let D be a database of points. A cluster wrt Eps and MinPts is a non-empty subset of D satisfying the following conditions
  - For all p, q: if p is in C and q is density reachable from p wrt Eps and MinPts, then q is in C (maximality)
  - For all p, q in C: p is density-connected to q wrt Eps and MinPts (Connectivity)

Noise

- Let C1,..Ck be the clusters of the database D wrt the parameters Eps and MinPts(j), j=1,..,k. Then we define noise as the set of points in the database D not belonging to any cluster Cj.

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



Eps = 1cm

MinPts = 5

DBSCAN: The Algorithm

- Arbitrary select a point ***p***
- Retrieve all points density-reachable from ***p*** wrt ***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.

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
  - Produces a special order of the database wrt its density-based clustering structure
  - This cluster-ordering contains info equiv 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

OPTICS: Extension from DBSCAN

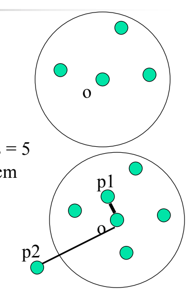
- Complexity: $O(kN^2)$
- Core Distance
- Reachability Distance

MinPts = 5

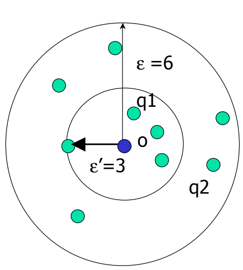
ε = 3 cm

Max (core-distance (o), d (o, p))

r(p1, o) = 2.8cm. r(p2,o) = 4cm



OPTICS Terminology



Core distance (o) = 3

R(q1,o) = 3

R(q2,o)=distance (q2,o)

Core distance

Core distance of an object p

Let *p* be an object from a database *D*; Let ε be a distance value. Le Nε (*p*) be the Eps-neighborhood of *p*; Let MinPts be a natural number; Then core_distance wrt ε and MinPts is :

- undefined if *p* is not a core point
- MinPts-distance(*p*)=distance to its MinPts neighbor

i.e., Core distance is the smallest distance ε′ between p and an object in its ε-neighborhood such that p would be a core object wrt ε′ if this neighbor is contained in Nε(p)

Reachability Distance

- Reachability distance of object *p* wrt object *o* .

Let p and o be objects in D. The reachability distance of p wrt o is defined as:

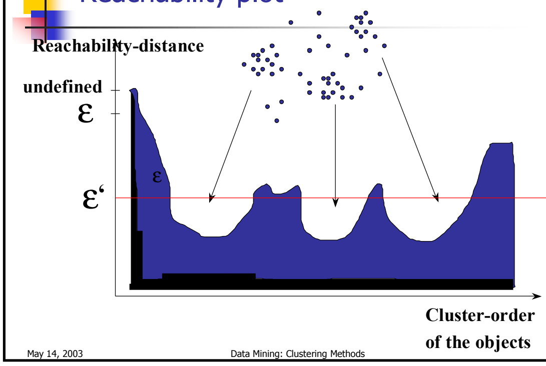
--undefined if o is not a core object

--max (core_distance (o), distance (o,p)) otherwise

r(p,o) is the smallest distance such that p is directly density reachable from o.

- note that r(p.o) can not be smaller than the core distance of o because for smaller distances no object is directly density-reachable from o.

Reachability plot



DENCLUE: using density functions

- DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Major features
  - Solid mathematical foundation
  - Good for data sets with large amounts of noise
  - Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
  - Significant faster than existing algorithm (faster than DBSCAN by a factor of up to 45)
  - But needs a large number of parameters

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Summary

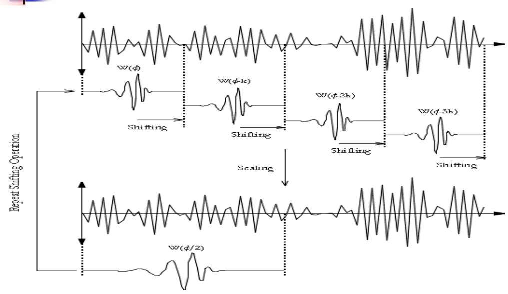
Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
  - STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
  - WaveCluster by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
    - A multi-resolution clustering approach using wavelet method
  - CLIQUE: Agrawal, et al. (SIGMOD'98)

WaveCluster (1998)

- Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
- A multi-resolution clustering approach which applies wavelet transform to the feature space
  - A wavelet transform is a signal processing technique that decomposes a signal into different frequency sub-band.
- Both grid-based and density-based
- Input parameters:
  - # of grid cells for each dimension
  - the wavelet, and the # of applications of wavelet transform.

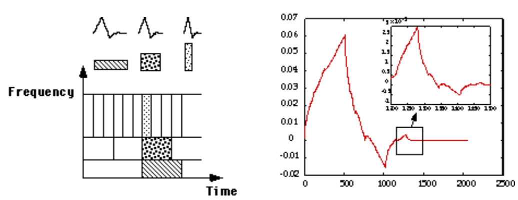
What is a Wavelet (1)?



WaveCluster (1998)

- How to apply wavelet transform to find clusters
  - Summaries the data by imposing a multidimensional grid structure onto data space
  - These multidimensional spatial data objects are represented in a n-dimensional feature space
  - Apply wavelet transform on feature space to find the dense regions in the feature space
  - Apply wavelet transform multiple times which result in clusters at different scales from fine to coarse

What Is Wavelet (2)?



Quantization

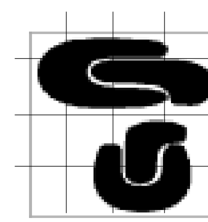
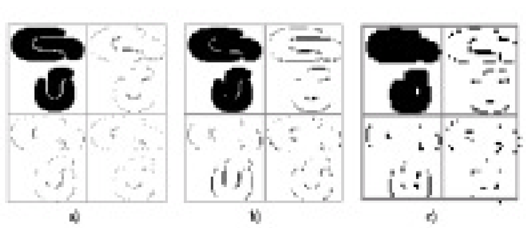


Figure 1: A sample 2-dimensional feature space.

Transformation



WaveCluster (1998)

- Why is wavelet transformation useful for clustering
  - Unsupervised clustering
    It uses hat-shape filters to emphasize region where points cluster, but simultaneously to suppress weaker information in their boundary
  - Effective removal of outliers
  - Multi-resolution
  - Cost efficiency
- Major features:
  - Complexity O(N)
  - Detect arbitrary shaped clusters at different scales
  - Not sensitive to noise, not sensitive to input order
  - Only applicable to low dimensional data

Chapter 8. Cluster Analysis

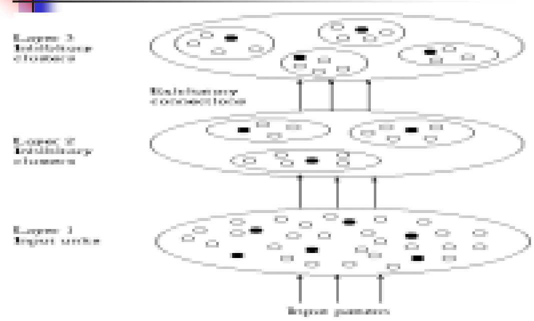
- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Summary

The EM Algorithm

Other Model-Based Clustering Methods

- Neural network approaches
  - Represent each cluster as an exemplar, acting as a "prototype" of the cluster
  - New objects are distributed to the cluster whose exemplar is the most similar according to some dostance measure
- Competitive learning
  - Involves a hierarchical architecture of several units (neurons)
  - Neurons compete in a "winner-takes-all" fashion for the object currently being presented

Model-Based Clustering Methods



Self-organizing feature maps (SOMs)

- Clustering is also performed by having several units competing for the current object
- The unit whose weight vector is closest to the current object wins
- The winner and its neighbors learn by having their weights adjusted
- SOMs are believed to resemble processing that can occur in the brain
- Useful for visualizing high-dimensional data in 2- or 3-D space

Expectation-Maximization

May 14, 2003

Two component mixture model

- Mixture example
- 20 data points
- Generate a delta with prob. \pi, and then depending on the outcome, deliver either Y1 or Y2

$$Y_1 \sim N(\mu_1, \sigma_1^2)$$
$$Y_2 \sim N(\mu_2, \sigma_2^2)$$
$$Y = (1-\Delta).Y_1 + \Delta.Y_2$$
$$\Delta \in \{0,1\}, \Pr(\Delta = 1) = \pi$$

- -0.39, 0.12. 0.94, 1.67, 1.76, 2.44, 3.72, 4.28, 4.92, 5.53, 0.06, 0.48, 1.01, 1.68, 1.80, 3.25, 4.12, 4.60, 5.28, 6.22

Mixture Example (Hastie)

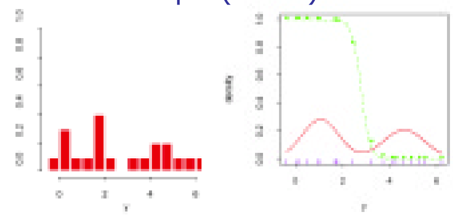


Figure 8.5: Mixture example. Left panel: histogram of data. Right panel: maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation y, as a function of y.

Mixture example

Let $\varphi_\theta(y) \sim N(\theta) = N(\mu, \sigma^2)$

Density of Y $\quad g_Y(y) = (1-\pi)\varphi_{\theta_1}(y) + \pi\varphi_{\theta_2}(y)$

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

log-likelihood: $\quad \ell(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \log\left[(1-\pi)\varphi_{\theta_1}(y_i) + \pi\varphi_{\theta_2}(y_i)\right]$

Parameter estimation

Direct maximization of $\ell(\theta; \mathbf{Z})$ is quite difficult numerically

Consider (unobserved) variables Δ_i taking values 0 or 1

If $\Delta_i = 1$ then Yi comes from model 2, else from model 1.

Suppose we knew the values of Δ_i. Then,

$$\ell_0(\theta; \mathbf{Z}, \mathbf{\Delta}) = \sum_{i=1}^{N}\left[(1-\Delta_i)\log\varphi_{\theta_1}(y_i) + \Delta_i \log\varphi_{\theta_2}(y_i)\right]$$

and the maximum likelihood estimates of μ_1 and σ_1^2 would be sample mean and variance of those data with $\Delta_i = 0$ same for the other case also.

but we do not know the values of deltas. Instead, we use the expected values—also called the responsibility of model 2 for observation *i*:

$$\gamma_i(\theta) = \mathrm{E}(\Delta_i \mid \theta, \mathbf{Z}) = \Pr(\Delta_i = 1 \mid \theta, \mathbf{Z})$$

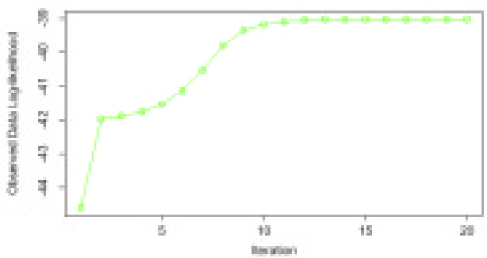
Expectation step: soft assignment of each observation to each model; the current estimates of the parameters are used.

Maximization step: weighted ML estimates to update the estimates of the parameters.

Algorithm 8.1 *EM algorithm for two-component Gaussian mixture.*

1. Take initial guesses for the parameters $\hat\mu_1, \hat\sigma_1^2, \hat\mu_2, \hat\sigma_2^2, \hat\pi$ (see text).
2. *Expectation Step:* compute the responsibilities
$$\hat\gamma_i = \frac{\hat\pi\varphi_{\hat\theta_2}(y_i)}{(1-\hat\pi)\varphi_{\hat\theta_1}(y_i) + \hat\pi\varphi_{\hat\theta_2}(y_i)}, \quad i = 1, 2, \ldots, N. \qquad (8.42)$$
3. *Maximization Step:* compute the weighted means and variances
$$\hat\mu_1 = \frac{\sum_{i=1}^N (1-\hat\gamma_i) y_i}{\sum_{i=1}^N (1-\hat\gamma_i)}, \qquad \hat\sigma_1^2 = \frac{\sum_{i=1}^N (1-\hat\gamma_i)(y_i - \hat\mu_1)^2}{\sum_{i=1}^N (1-\hat\gamma_i)},$$
$$\hat\mu_2 = \frac{\sum_{i=1}^N \hat\gamma_i y_i}{\sum_{i=1}^N \hat\gamma_i}, \qquad \hat\sigma_2^2 = \frac{\sum_{i=1}^N \hat\gamma_i (y_i - \hat\mu_2)^2}{\sum_{i=1}^N \hat\gamma_i},$$
and the mixing probability $\hat\pi = \sum_{i=1}^N \hat\gamma_i / N$.
4. Iterate steps 2 and 3 until convergence.

EM algorithm: data log-likelihood



General EM Algorithm: basic ideas

Let D = {x(1), x(2), …, x(n)} –observed data vectors
H = {z(1), …, z(n)} –hidden variables, z(i) associated with x(i).

e.g., z(i) –class labels for the data.

We can write the log-likelihood of the observed data as

$$\ell(\theta) = \log \Pr(D \mid \theta) = \log \sum_H \Pr(D, H \mid \theta)$$

Geneal EM

Let Q(H) be any probability distribution on the H. Then,

$$\begin{aligned}
\ell(\theta) &= \log \sum_H \Pr(D, H \mid \theta) \\
&= \log \sum_H Q(H) \frac{\Pr(D, H \mid \theta)}{Q(H)} \\
&\geq \sum_H Q(H) \log \frac{\Pr(D, H \mid \theta)}{Q(H)} \qquad \text{Jensen's inequality} \\
&= \sum_H Q(H) \log \Pr(D, H \mid \theta) + \sum_H Q(H) \log \frac{1}{Q(H)} \\
&= F(Q, \theta)
\end{aligned}$$

- Function F(Q, θ) is a lower bound on the function –the likelihood l(θ).
- The EM algorithm alternates between maximizing F w.r.t the distribution Q with the parameters θ fixed, and then maximizing F wrt the parameters θ with the distribution Q=Pr(H) fixed.

$$\boxed{\begin{aligned} &\text{E-Step:} \quad Q^{k+1} = \arg\max_Q F(Q^k, \theta^k) \\ &\text{M-Step:} \quad \theta^{k+1} = \arg\max_Q F(Q^{k+1}, \theta^k) \end{aligned}}$$

EM Algorithm

- It is straightforward to show that the maximum of the E-Step is achieved when $Q^{k+1} = \Pr(H \mid D, \theta^k)$
- This Q can be calculated explicitly for many models.
- For this value of Q the bound becomes tight, i.e., the inequality becomes an equality and $\ell(\theta^k) = F(Q, \theta^k)$
- The maximization in the M-step reduces to maximizing the first term in *F* (since the second term does not depend on θ)

$$\theta^{k+1} = \arg\max_\theta \sum_H \Pr(H \mid D, \theta^k) \log \Pr(D, H \mid \theta^k)$$

EM: Summary

- Useful when oitimizing Q is simpler than optimizing the likelihood l.
- In general, any iterative scheme in which the likelihood of *some* data increases with each step, are often referred to as the EM schemes.

Chapter 8. Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Summary

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, EM
- Current clustering techniques do not address all the requirements adequately
- Constraint-based clustering analysis: Constraints exist in data space (bridges and highways) or in user queries

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
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches
- There are still lots of research issues on cluster analysis, such as constraint-based clustering

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.

Next: Pattern finding and retrieval by content

- Association Rules
- Selected topics in Text, Image and Video Retrieval