

Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Constraint-Based Clustering
7. Outlier Analysis
8. Summary

April 26, 2010

Data Acquisition and Processing

1

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
- Cluster analysis
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into 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

April 26, 2010

Data Acquisition and Processing

2

Clustering Applications: Some Examples

- **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

April 26, 2010

Data Acquisition and Processing

3

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

April 26, 2010

Data Acquisition and Processing

4

Measure the Quality of Clustering

- **Dissimilarity/Similarity metric**
 - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
 - The definitions of **distance functions** are usually rather different for interval-scaled, Boolean, categorical, ordinal ratio, and vector variables
 - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
 - There is usually 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

April 26, 2010

Data Acquisition and Processing

5

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- 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
- Incorporation of user-specified constraints
- Interpretability and usability

April 26, 2010

Data Acquisition and Processing

6

Data Structures

- Data matrix
 - (two modes)

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
 - (one mode)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

April 26, 2010

Data Acquisition and Processing

7

Type of data in clustering analysis

- Interval-scaled variables
- Binary variables
- Nominal and ordinal
- Vector objects

April 26, 2010

Data Acquisition and Processing

8

Distance Measures for Different Kinds of Data

- Numerical (interval)-based:
 - Minkowski Distance
 - Special cases: Euclidean (L_2 -norm), Manhattan (L_1 -norm)
- Binary variables:
 - symmetric vs. asymmetric (Jaccard coeff.)
- Nominal variables: # of mismatches
- Ordinal variables: treated like interval-based
- Vectors: cosine measure

April 26, 2010

Data Acquisition and Processing

9

Interval-valued variables

- Standardize data
 - Calculate the mean absolute deviation:

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

April 26, 2010

Data Acquisition and Processing

10

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*:

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

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects, and q is a positive integer
- If $q = 1$, d is Manhattan distance

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

April 26, 2010

Data Acquisition and Processing

11

Similarity and Dissimilarity Between Objects

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

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$
- Properties
 - $d(i, j) \geq 0$
 - $d(i, i) = 0$
 - $d(i, j) = d(j, i)$
 - $d(i, j) \leq d(i, k) + d(k, j)$

April 26, 2010

Data Acquisition and Processing

12

Binary Variables

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

- A contingency table for binary data
- Distance measure for symmetric binary variables: $d(i, j) = \frac{b+c}{a+b+c+d}$
- Distance measure for asymmetric binary variables: $d(i, j) = \frac{b+c}{a+b+c}$
- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables): $sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$

April 26, 2010

Data Acquisition and Processing

13

Dissimilarity between Binary Variables

Example

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

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

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

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

April 26, 2010

Data Acquisition and Processing

14

Nominal Variables

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

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

April 26, 2010

Data Acquisition and Processing

15

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 $r_{if} \in \{1, \dots, M_f\}$
 - 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

April 26, 2010

Data Acquisition and Processing

16

Vector Objects

- Vector objects: keywords in documents
- Broad applications: information retrieval, biologic taxonomy, etc.
- Cosine measure

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{|\vec{X}| |\vec{Y}|},$$

\vec{X}^t is a transposition of vector \vec{X} , $|\vec{X}|$ is the Euclidean normal of vector \vec{X} ,

- A variant: Tanimoto coefficient

$$s(\vec{X}, \vec{Y}) = \frac{\vec{X}^t \cdot \vec{Y}}{\vec{X}^t \cdot \vec{X} + \vec{Y}^t \cdot \vec{Y} - \vec{X}^t \cdot \vec{Y}},$$

April 26, 2010

Data Acquisition and Processing

17

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, ROCK, CAMELEON
- Density-based approach:
 - Based on connectivity and density functions
 - Typical methods: DBSACN, OPTICS, DenClue

April 26, 2010

Data Acquisition and Processing

18

Major Clustering Approaches (II)

- Grid-based approach:
 - based on a multiple-level granularity structure
 - Typical methods: STING, WaveCluster, CLIQUE
- 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: pCluster
- User-guided or constraint-based:
 - Clustering by considering user-specified or application-specific constraints
 - Typical methods: COD (obstacles), constrained clustering

April 26, 2010

Data Acquisition and Processing

19

Calculation of Distance between Clusters

- **Single link**: smallest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \min(t_{ip}, t_{jq})$
- **Complete link**: largest distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \max(t_{ip}, t_{jq})$
- **Average**: avg distance between an element in one cluster and an element in the other, i.e., $\text{dis}(K_i, K_j) = \text{avg}(t_{ip}, t_{jq})$
- **Centroid**: distance between the centroids of two clusters, i.e., $\text{dis}(K_i, K_j) = \text{dis}(C_i, C_j)$
- **Medoid**: distance between the medoids of two clusters, i.e., $\text{dis}(K_i, K_j) = \text{dis}(M_i, M_j)$
 - Medoid: one chosen, centrally located object in the cluster

April 26, 2010

Data Acquisition and Processing

20

Centroid, Radius and Diameter of a Cluster (for numerical data sets)

- Centroid: the "middle" of a cluster
- Radius: square root of average distance from any point of the cluster to its centroid
- Diameter: square root of average mean squared distance between all pairs of points in the cluster

$$C_m = \frac{\sum_{i=1}^N (t_{ip})}{N}$$

$$R_m = \sqrt{\frac{\sum_{i=1}^N (t_{ip} - c_m)^2}{N}}$$

$$D_m = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (t_{ip} - t_{jq})^2}{N(N-1)}}$$

April 26, 2010

Data Acquisition and Processing

21

Partitioning Algorithms: Basic Concept

- Partitioning method:** Construct a partition of a database D of n objects into a set of k clusters, s.t., min sum of squared distance

$$\sum_{m=1}^k \sum_{t_{mi} \in K_m} (C_m - t_{mi})^2$$

- 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*: Each cluster is represented by the center of the cluster
 - k-medoids* or PAM (Partition around medoids): Each cluster is represented by one of the objects in the cluster

April 26, 2010

Data Acquisition and Processing

22

The *K-Means* Clustering Method

- Given k , the *k-means* algorithm is implemented in four steps:
 - Partition objects into k nonempty subsets
 - Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
 - Assign each object to the cluster with the nearest seed point
 - Go back to Step 2, stop when no more new assignment

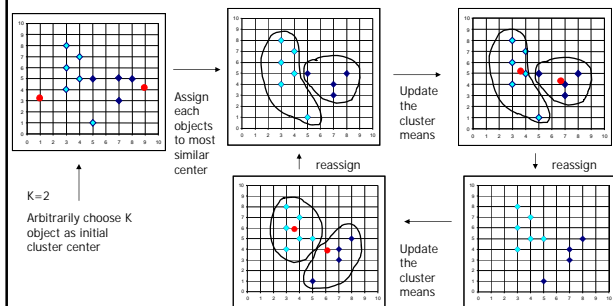
April 26, 2010

Data Acquisition and Processing

23

The *K-Means* Clustering Method

Example



April 26, 2010

Data Acquisition and Processing

24

Comments on the *K-Means* Method

- **Strength:** *Relatively efficient:* $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, $k, t \ll n$
- **Comment:** Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- **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*

April 26, 2010

Data Acquisition and Processing

25

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*
 - 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

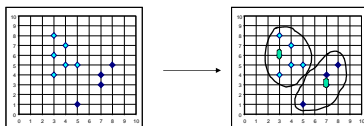
April 26, 2010

Data Acquisition and Processing

26

What Is the Problem of the K-Means Method?

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



April 26, 2010

Data Acquisition and Processing

27

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids)
 - 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*
- *CLARANS*: Randomized sampling

April 26, 2010

Data Acquisition and Processing

28

CLARA (Clustering Large Applications)

- **CLARA** (built 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: deals with larger data sets than *PAM*
- Weakness:
 - 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

April 26, 2010

Data Acquisition and Processing

29

CLARANS ("Randomized" CLARA)

- **CLARANS** (A Clustering Algorithm based on Randomized Search)
- CLARANS draws sample of neighbors 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 may further improve its performance

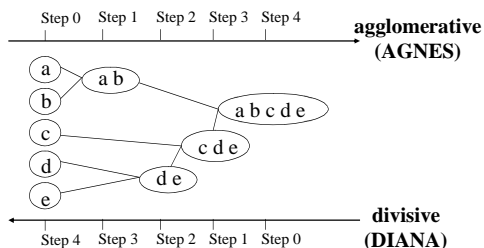
April 26, 2010

Data Acquisition and Processing

30

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



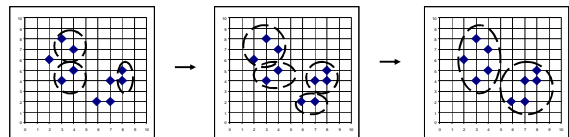
April 26, 2010

Data Acquisition and Processing

31

AGNES (Agglomerative Nesting)

- Introduced in 1990
- Implemented in statistical analysis packages, e.g., Splus
- 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



April 26, 2010

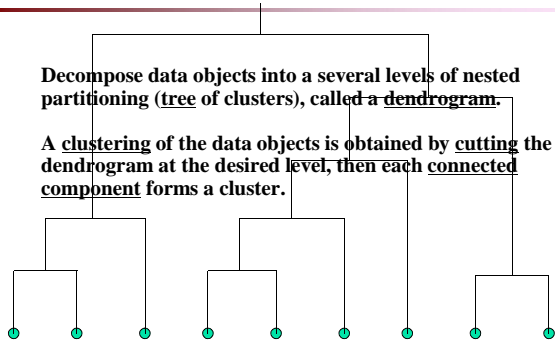
Data Acquisition and Processing

32

Dendrogram: Shows How the Clusters are Merged

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram.

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.



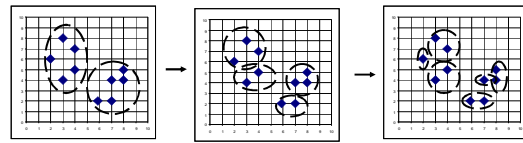
April 26, 2010

Data Acquisition and Processing

33

DIANA (Divisive Analysis)

- Introduced in 1990
- Implemented in statistical analysis packages, e.g., Splus
- Inverse order of AGNES
- Eventually each node forms a cluster on its own



April 26, 2010

Data Acquisition and Processing

34

Extensions to 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 with distance-based clustering
 - BIRCH (1996): uses CF(Clustering Feature)-tree and incrementally adjusts the quality of sub-clusters
 - ROCK (1999): clustering categorical data by neighbor and link analysis
 - CHAMELEON (1999): hierarchical clustering using dynamic modeling

April 26, 2010

Data Acquisition and Processing

35

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
 - OPTICS
 - DENCLUE
 - CLIQUE

April 26, 2010

Data Acquisition and Processing

36

Grid-Based Clustering Method

- Using multi-resolution grid data structure
- Several interesting methods
 - **STING** (a Statistical INformation Grid approach)
 - **WaveCluster**
 - A multi-resolution clustering approach using wavelet method
 - **CLIQUE**
 - On high-dimensional data (thus put in the section of clustering high-dimensional data)

April 26, 2010

Data Acquisition and Processing

37

Model-Based Clustering

- What is model-based clustering?
 - Attempt to optimize the fit between the given data and some mathematical model
 - Based on the assumption: Data are generated by a mixture of underlying probability distribution
- Typical methods
 - Statistical approach
 - EM (Expectation maximization), AutoClass
 - Machine learning approach
 - COBWEB, CLASSIT
 - Neural network approach
 - SOM (Self-Organizing Feature Map)

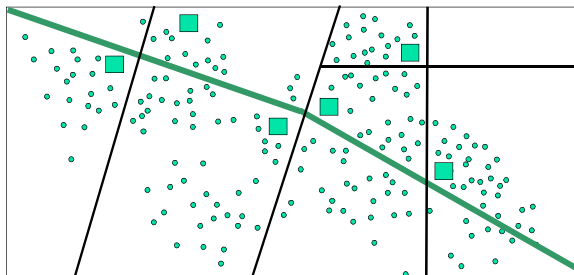
April 26, 2010

Data Acquisition and Processing

38

Why Constraint-Based Cluster Analysis?

- Need user feedback: Users know their applications the best
- Less parameters but more user-desired constraints, e.g., an ATM allocation problem: obstacle & desired clusters



April 26, 2010

Data Acquisition and Processing

39

A Classification of Constraints in Cluster Analysis

- Clustering in applications: desirable to have user-guided (i.e., constrained) cluster analysis
- Different constraints in cluster analysis:
 - Constraints on individual objects (do selection first)
 - Cluster on houses worth over \$300K
 - Constraints on distance or similarity functions
 - Weighted functions, obstacles (e.g., rivers, lakes)
 - Constraints on the selection of clustering parameters
 - # of clusters, MinPts, etc.
 - User-specified constraints
 - Contain at least 500 valued customers and 5000 ordinary ones
 - Semi-supervised: giving small training sets as "constraints" or hints

April 26, 2010

Data Acquisition and Processing

40

What Is Outlier Discovery?

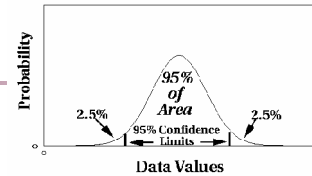
- What are outliers?
 - The set of objects are considerably dissimilar from the remainder of the data
 - Example: Sports: Michael Jordon, Wayne Gretzky, ...
- Problem: Define and find outliers in large data sets
- Applications:
 - Credit card fraud detection
 - Telecom fraud detection
 - Customer segmentation
 - Medical analysis

April 26, 2010

Data Acquisition and Processing

41

Outlier Discovery: Statistical Approaches



- ✗ Assume a model underlying distribution that generates data set (e.g. normal distribution)
- Use discordancy tests depending on
 - data distribution
 - distribution parameter (e.g., mean, variance)
 - number of expected outliers
- Drawbacks
 - most tests are for single attribute
 - In many cases, data distribution may not be known

April 26, 2010

Data Acquisition and Processing

42

Outlier Discovery: Distance-Based Approach

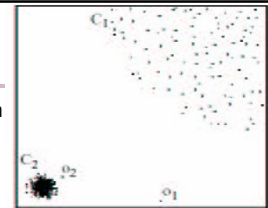
- Introduced to counter the main limitations imposed by statistical methods
 - We need multi-dimensional analysis without knowing data distribution
- Distance-based outlier: A $DB(p, D)$ -outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than D from O
- Algorithms for mining distance-based outliers
 - Index-based algorithm
 - Nested-loop algorithm
 - Cell-based algorithm

April 26, 2010

Data Acquisition and Processing

43

Density-Based Local Outlier Detection



- Distance-based outlier detection is based on global distance distribution
- It encounters difficulties to identify outliers if data is not uniformly distributed
- Ex. C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, o_2
- Distance-based method cannot identify o_2 as an outlier
- Need the concept of local outlier

- Local outlier factor (LOF)
 - Assume outlier is not crisp
 - Each point has a LOF

April 26, 2010

Data Acquisition and Processing

44

Outlier Discovery: Deviation-Based Approach

- Identifies outliers by examining the main characteristics of objects in a group
- Objects that “deviate” from this description are considered outliers
- Sequential exception technique
 - simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
- OLAP data cube technique
 - uses data cubes to identify regions of anomalies in large multidimensional data

April 26, 2010

Data Acquisition and Processing

45

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

April 26, 2010

Data Acquisition and Processing

46