Clustering

Clustering

- Recall
  - Supervised Vs. Unsupervised Learning
  - Unsupervised Learning
    - What have we studied ?
- Clustering is another form of unsupervised Learning

Clustering Vs. Classification

- Similarity
  - Aim to partition data (high-dimensional) into groups/classes/clusters
  - Data items within a group are as similar to each other as possible, but are dissimilar to data items in other groups
- Differences
  - No predefined classes on the basis of which the grouping is done

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
  - **Maximise intra-cluster similarity and minimise inter-cluster similarity**
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

General Applications of Clustering

- 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
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity (within a cluster)
  - low inter-class similarity (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.

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal 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
- Incorporation of user-specified constraints
- Interpretability and usability

Clustering

- How to define a cluster
- How to find if objects are similar or not
- Need to define concept of **distance**
- Not a new field – a branch of statistics
  - Several traditional distance-based algorithms are available
- Dealing with new issues of electronic data
  - New techniques being developed

Distance

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Properties
  - Always positive
  - Distance from x to x = zero
  - Distance from x to y = Distance from y to x
  - Distance from x to y <= distance from x to z + distance from z to y

Distance Measures

- 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 $i = (x_{i1}, x_{i2}, ..., x_{ip})$ 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

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

Distance Measures

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

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

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$
- Also one can use weighted distance, parametric Pearson product moment correlation, or other dissimilarity measures.

Distance Measures

- Categorical Data
  - Sensible distance measure between male and female, between MIT and BCOMP ?
  - Convert all categorical data into numeric data
    - Does this make any sense ?
  - Treat distance between categorical values as a function which has values only 1 and 0
    - Can we compare two objects that have all categorical attributes ?
    - Essentially, how many attribute values are the same

Distance Measures

- Scales of different attribute values
  - GPA varies 0-4, Age varies 18-60
- Problems
  - Total Distance (Sum of the distance between individual attribute values) will be dominated by the age attribute and will be affected very little by the variations in GPA
- Need to normalise
  - E.g. All data varies from 0 to 1

Distance Measures

- What if you want to weight some attributes higher than others ?
- Distance measure might have to incorporate weights

Clustering

- Not always easy to determine how many clusters to expect from a data set
- Maybe based on some domain knowledge
- But that does not mean that a different number of clusters doesn't give you better results
- What is the optimal number ? Open issue
- Many clustering techniques require input of number of clusters to be generated or use a default number

Clustering Methods

- Many different methods
- Selection must be based on requirements
- Consider 2 examples
  - Partition data to minimise the maximum distance between any pair of points in the cluster
  - Partition data to minimise the maximum distance between each point and the point closest to it.

Types of Clustering Methods

- Partitioning Methods
  - Given n objects, make k (<= n) partitions or clusters of data and use iterative relocation. It is assumed each cluster has at least one object and each object belongs to only one cluster.
- Hierarchical Methods
  - *Agglomerative* - Start with each object in a cluster and try to merge similar clusters.
  - *Divisive* -Start with one cluster and then split into smaller clusters

Types of Clustering Methods

- Density-based Methods
  - For each data point in a cluster, at least a minimum number of points must exist within a radius.
- Grid-based Methods
  - Object space is divided into a grid
- Model-based Methods
  - A model is assumed, perhaps based on a probability distribution.

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

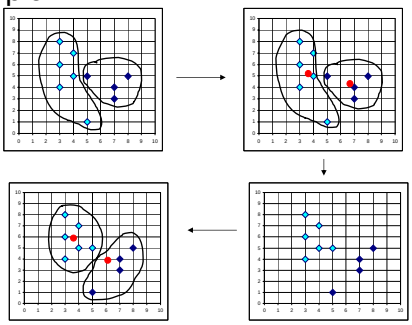
- Given *k*, the *k-means* algorithm is implemented in 4 steps:
  - Choose number of clusters (*k)*
  - Compute *k* random seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
  - Assign each object to a cluster with the nearest seed point. The distance is computed using a distance metric.
  - Once all objects have been assigned, re-compute the mean value for each cluster. This is used as the cluster centre for the next iteration.

The *K-Means* Clustering Method

- Using the new mean value of each cluster, objects are re-assigned to the nearest cluster. In most situations, some objects will move to new clusters, unless the initial choice of the seeds was very good.
- The process continues until no change takes place in the cluster memberships.
- Different starting points will obviously lead to different clusters.

The *K-Means* Clustering Method

- Example



K-Means – An example

Student	Age	Mark1	Mark2	Mark3
A	18	73	75	57
B	18	79	85	75
C	23	70	70	52
D	20	55	55	55
E	22	85	86	87
F	19	91	90	89
G	20	70	65	60
H	21	53	56	59
I	19	82	82	60
J	47	75	76	77

K-Means – An example

Student	Age	Mark1	Mark2	Mark3
A	18	73	75	57
B	18	79	85	75
C	23	70	70	52

- First 3 records be seeds
- Compute distances the 4 attributes
- And use the sum of absolute differences

K-Means – An example

Students	Age	Mark 1	Mark 2	Mark 3	Dist.
D	20	55	55	55	42/76/36
E	22	85	86	87	57/23/67
F	19	91	90	89	66/32/82
G	20	70	65	60	18/46/16
H	21	53	56	59	44/74/40
I	19	82	82	60	20/22/36
J	47	75	76	77	52/44/60

Comments on the *K-Means* Method

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

Comments on the *K-Means* Method

- 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*
  - Not suitable to discover clusters with *non-convex shapes*
  - Does not deal well with overlapping clusters
  - Outliers can pull cluster centres
  - Crisp membership in clusters

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

Evaluating Clusters

- Have members which are very close in a cluster and clusters to be widely apart
- Variance measures are often used – variance within a cluster should be low and between clusters should be high
- Generally a major issue in clustering

In summary...

- Strengths
  - Useful for undirected knowledge discovery
  - Relatively simple
- Weaknesses
  - Difficult to choose distance measures and weightings
  - Sensitive to initial parameter choices
  - Can be difficult to interpret results