Data Mining Summer school

CLUSTER ANALYSIS

Asso. Prof. NGUYEN Tri Thanh

University of Engineering and Technology

Outline

1. Introduction
2. Clustering applications
3. A Categorization of Major Clustering Methods
4. Data representation
5. Flat clustering
6. Hierarchical clustering
7. Cluster evaluation
8. Summary
9. Discussion

Introduction

- Cluster analysis (clustering)
  - Given a set of objects, group similar data according to the characteristics into clusters
- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- How to identify the similarity?
- How to identify the number of clusters?

Introduction

- Clustering is embedded in human naturally
  - Group animals, plants
  - Group students
  - Group customers
  - Facebook groups
  - Interest groups
  - …

Clustering vs Classification

- Clustering
  - No predefined classes (unknown number of clusters)
  - Unlabeled data objects
  - Unsupervised learning
- Classification
  - Predefined classes
  - Labeled data objects
  - Predict/identify the class of an unlabeled object
  - Supervised learning

Clustering applications

- Image Processing and Pattern Recognition
- Spatial Data Analysis
  - Create thematic maps in GIS by clustering feature spaces
  - Detect spatial clusters or for other spatial mining tasks
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

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

Categorization of Clustering Approaches

- 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
- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches (II)

- 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: p-Cluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- Link-based clustering:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

Flat clustering

Data object representation

- A set/vector of features/attributes
  - Person: name, age, sex, job, …
  - Text: set of distinct words

$$dis(d_1, d_2) = \sqrt{\sum_{i=1}^{n} (d_{1i} - d_{2i})^2}$$

- Similarity
  - The distance: the smaller the more similar
  - The similarity: the bigger the more similar

$$sim(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \|d_2\|} = \frac{\sum_{i=1}^{n} d_{1i} d_{2i}}{\sqrt{\sum_{i=1}^{n} d_{1i}^2} \sqrt{\sum_{i=1}^{n} d_{2i}^2}}$$

K-means Clustering

- Number of clusters (k) is known in advance
- Clusters are represented by the centroid of the documents that belong to that cluster

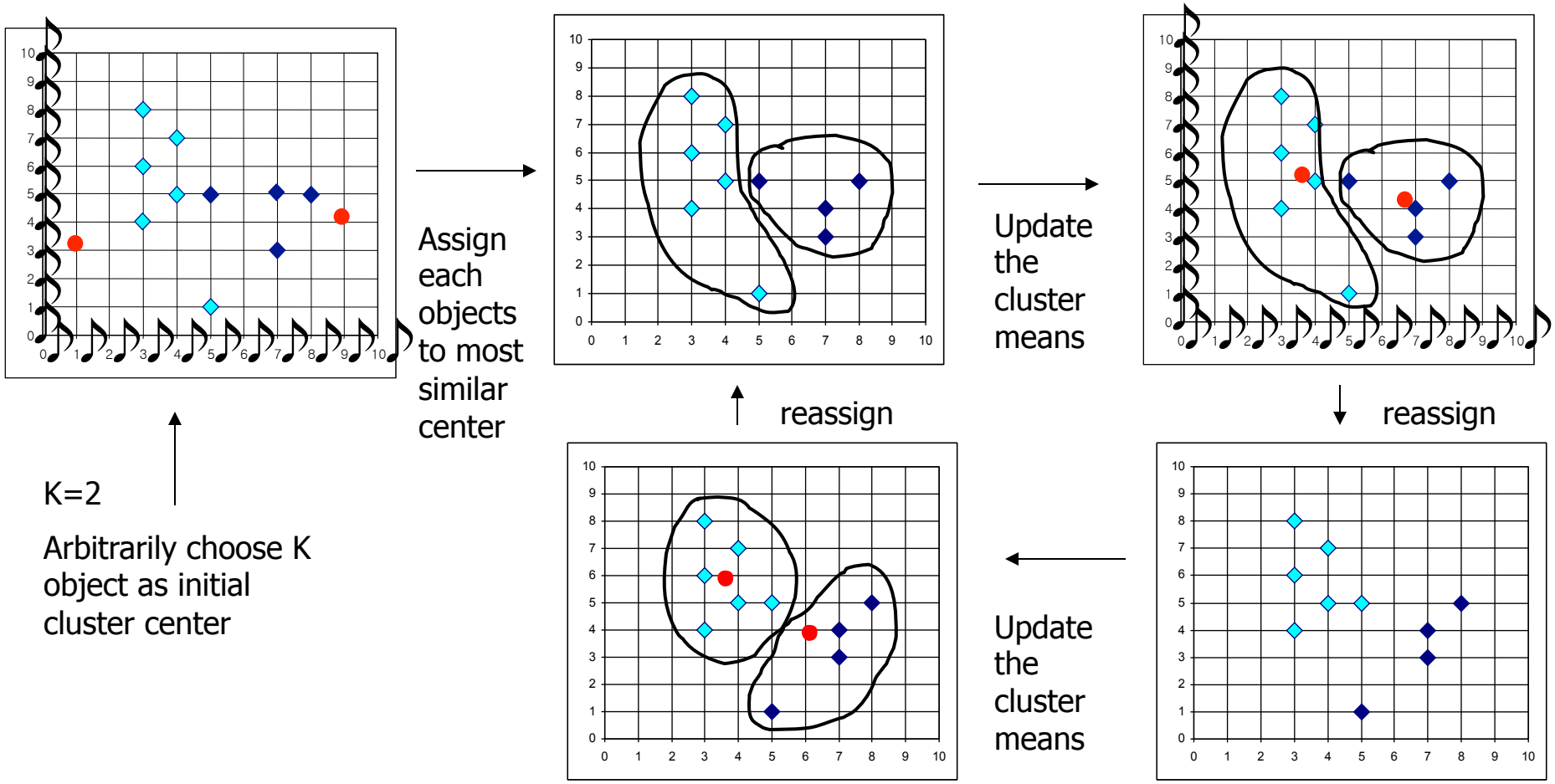
$$c = \frac{1}{\|S\|} \sum_{d \in S} d$$

- Cluster membership is determined by the most similar cluster centroid

1. Select k documents from S to be used as cluster centroids. This is usually done at random.
2. Assign documents to clusters according to their simility to the cluster centroids, i.e. for each document find the most similar centroid and assign that document to the corresponding cluster.
3. For each cluster recompute the cluster centroid using the newly computed cluster members.
4. Go to Step 2 until the process converges, i.e. the same documents are assigned to each cluster in two consecutive iterations or the cluster centroids remain the same.

The *K-Means* Clustering Method

- Example



K-means Clustering Discussion

- In step 2 documents are moved between clusters in order to maximize the intra-cluster similarity
- The clustering maximizes the *criterion function* (a measure for evaluating *clustering quality*)
- In distance-based k-means clustering the criterion function is the *sum of squared errors* (based on Euclidean distance and means)
- For k-means clustering of documents a function based on centroids and similarity is used

$$J = \sum_{i=1}^{k} \sum_{d_j \in D_i} sim(c_i, d_j)$$

K-means Clustering Discussion (cont'd)

- Clustering that *maximizes* this function is called *minimum variance clustering*
- K-means algorithm produces minimum variance clustering, but does not guarantee that it always finds the global maximum of the criterion function
- After each iteration the value of *J* increases, but it may converge to a local maximum
- The result greatly depends on the initial choice of cluster centroids

Sample data

	history	science	research	offers	students	hall
Anthropology	0	0.537	0.477	0	0.673	0.177
Art	0	0	0	0.961	0.195	0.196
Biology	0	0.347	0.924	0	0.111	0.112
Chemistry	0	0.975	0	0	0.155	0.158
Communication	0	0	0	0.780	0.626	0
Computer	0	0.989	0	0	0.130	0.067
Justice	0	0	0	0	1	0
Economics	0	0	1	0	0	0
English	0	0	0	0.980	0	0.199
Geography	0	0.849	0	0	0.528	0
History	0.991	0	0	0.135	0	0
Math	0	0.616	0.549	0.490	0.198	0.201
Languages	0	0	0	0.928	0	0.373
Music	0.970	0	0	0	0.170	0.172
Philosophy	0.741	0	0	0.658	0	0.136
Physics	0	0	0.894	0	0.315	0.318
Political	0	0.933	0.348	0	0.062	0.063
Psychology	0	0	0.852	0.387	0.313	0.162
Sociology	0	0	0.639	0.570	0.459	0.237
Theatre	0	0	0	0	0.967	0.254

K-means Clustering Example (result)

Clustering of CCSU Departments data with 6 TFIDF attributes (*k* = 2)

Bad choice of initial cluster centroids

Iteration	Cluster A	Cluster B	Criterion function
1	{Computer, Political}	{Anthropology, Art, Biology, Chemistry, Communication, Justice, Economics, English, Geography, History, Math, Languages, Music, Philosophy, Physics, Psychology, Sociology, Theatre}	1.93554 (A) + 4.54975 (B) = 6.48529
2	{Chemistry, Computer, Geography, Political}	{Anthropology, Art, Biology, Communication, Justice, Economics, English, History, Math, Languages, Music, Philosophy, Physics, Psychology, Sociology, Theatre}	3.82736 (A) + 10.073 (B) = 13.9003
3	{Anthropology, Chemistry, Computer, Geography, Political}	{Art, Biology, Communication, Justice, Economics, English, History, Math, Languages, Music, Philosophy, Physics, Psychology, Sociology, Theatre}	4.60125 (A) + 9.51446 (B) = 14.1157

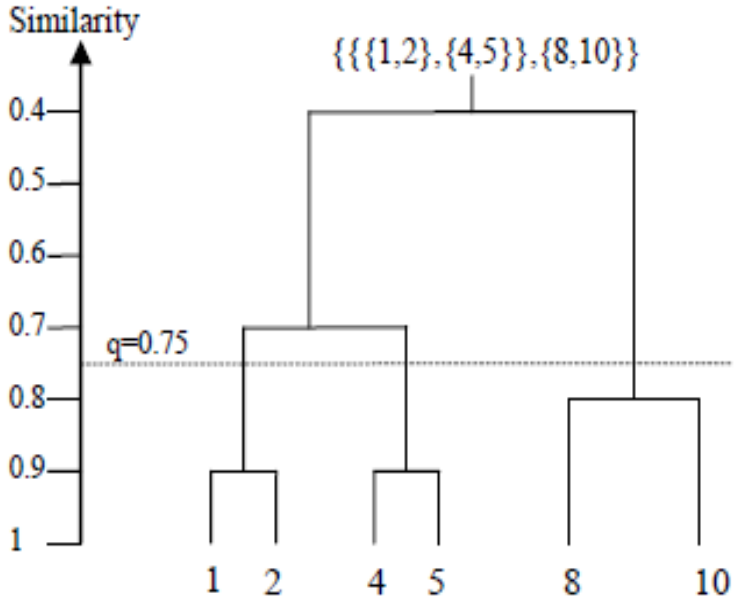
Good choice of initial cluster centroids

Iteration	Cluster A	Cluster B	Criterion function
1	{Anthropology, Biology, Economics, Math, Physics, Political, Psychology}	{Art, Chemistry, Communication, Computer, Justice, English, Geography, History, Languages, Music, Philosophy, Sociology, Theatre}	5.04527 (A) + 5.99025 (B) = 11.0355
2	{Anthropology, Biology, Computer, Economics, Math, Physics, Political, Psychology, Sociology}	{Art, Chemistry, Communication, Justice, English, Geography, History, Languages, Music, Philosophy, Theatre}	7.23827 (A) + 6.70864 (B) = 13.9469
3	{Anthropology, Biology, Chemistry, Computer, Economics, Geography, Math, Physics, Political, Psychology, Sociology}	{Art, Communication, Justice, English, History, Languages, Music, Philosophy, Theatre}	8.53381 (A) + 6.12743 (B) = 14.6612

Hierarchical clustering

Hierarchical Partitioning

- Produces a nested sequence of partitions of the set of data points
- Can be displayed as a tree (called *dendrogram*) with a single cluster including all points at the root and singleton clusters (individual points) at the leaves
- Example of hierarchical partitioning of set of numbers {1, 2, 4, 5, 8, 10}



The similarity measure used in this example is computed as (10-d)/10 where *d* is the distance between data points or cluster centers

Approaches to Hierarchical Partitioning

- *Agglomerative*
  - Starts with the data points and at each step merges the two closest (most similar) points (or clusters at later steps) until a single cluster remains
- *Divisible*
  - Starts with the original set of points and at each step splits a cluster until only individual points remain
  - To implement this approach we need to decide which cluster to split and how to perform the split

Approaches to Hierarchical Partitioning (cont'd)

- The agglomerative approach is more popular as it needs only the definition of a distance or similarity function on clusters/points

Approaches to Hierarchical Partitioning (cont'd)

- For data points in the Euclidean space the *Euclidean distance* is the best choice
- For documents represented as TF-IDF vectors the preferred measure is the *cosine similarity* defined as follows

$$sim(d_1, d_2) = \frac{d_1 \bullet d_2}{\|d_1\| \|d_2\|} = \frac{\sum_{i=1}^{n} d_{1i} d_{2i}}{\sqrt{\sum_{i=1}^{n} d_{1i}^2} \sqrt{\sum_{i=1}^{n} d_{2i}^2}}$$

Agglomerative Hierarchical Clustering

- There are several versions of this approach depending on how similarity on clusters $sim(S_1,S_2)$ is defined (S_1,S_2 are sets of documents)
  - Similarity between cluster *centroids,* i.e. $sim(S_1,S_2)=sim(c_1,c_2)$, where the centroid c of cluster S is

$$c = \frac{1}{\|S\|}\sum_{d\in S} d$$

  - *Maximum similarity* between documents from each cluster (*nearest neighbor clustering)*

$$sim(S_1,S_2) = \max_{d_1\in S_1, d_2\in S_2} sim(d_1,d_2)$$

Agglomerative Hierarchical Clustering (cont'd)

- *Minimum similarity* between documents from each cluster (*farthest neighbor clustering)*

$$sim(S_1, S_2) = \min_{d_1 \in S_1, d_2 \in S_2} sim(d_1, d_2)$$

- *Average similarity* between documents from each cluster

$$sim(S_1, S_2) = \frac{1}{\|S_1\| \|S_2\|} \sum_{d_1 \in S_1, d_2 \in S_2} sim(d_1, d_2)$$

Agglomerative Clustering Algorithm

- *S is the initial set of documents* and *G* is the *clustering tree*
- *k* and *q* are *control parameters* that stop merging
  - when a desired number of clusters (*k*) is reached
  - or when the similarity between the clusters to be merged becomes less than a specified threshold (*q*)

Agglomerative Clustering Algorithm (cont'd)

1. $G \leftarrow \{\{d\} \mid d \in S\}$ (initialize G with singleton clusters, each one containing a document from S)
2. If $|G| \leq k$ then exit (stop if the desired number of clusters is reached)
3. Find $S_i, S_j \in G$, such that $(i, j) = \arg\max_{(i,j)} sim(S_i, S_j)$ (find the two closest clusters)
4. If $sim(S_i, S_j) < q$ then exit (stop if the similarity of the closest clusters is less than q)
5. Remove S_i and S_j from G
6. $G = G \cup \{S_i, S_j\}$ (merge S_i and S_j, and add the new cluster to the hierarchy)
7. Go to 2

For *n* documents both *time* and *space complexity* of the algorithm are $O(n^2)$

Agglomerative Clustering Example 1

Cluster similarity cut off parameter $q = 0$

Average Intracluster Similarity = **0.4257**

```
1 [0.0224143]
   2 [0.0308927]
      3 [0.0368782]
         4 [0.0556825]
            5 [0.129523]
               Art
               Theatre
            Geography
         6 [0.0858613]
            7 [0.148599]
               Chemistry
               Music
            8 [0.23571]
               Computer
               Political
      9 [0.0937594]
         10 [0.176625]
            Communication
            Economics
         Justice
   11 [0.0554991]
      12 [0.0662345]
         13 [0.0864619]
            14 [0.177997]
               History
               Philosophy
            15 [0.186299]
               English
               Languages
         16 [0.122659]
            Anthropology
            Sociology
      17 [0.0952722]
         18 [0.163493]
            19 [0.245171]
               Biology
               Math
            Psychology
         Physics
```

```
1 []
   2 [0.0554991]
      3 [0.0662345]
         4 [0.0864619]
            5 [0.177997]
               History
               Philosophy
            6 [0.186299]
               English
               Languages
         7 [0.122659]
            Anthropology
            Sociology
      8 [0.0952722]
         9 [0.163493]
            10 [0.245171]
               Biology
               Math
            Psychology
         Physics
   11 [0.0556825]
      12 [0.129523]
         Art
         Theatre
      Geography
   13 [0.0858613]
      14 [0.148599]
         Chemistry
         Music
      15 [0.23571]
         Computer
         Political
   16 [0.0937594]
      17 [0.176625]
         Communication
         Economics
      Justice
```

Cluster similarity cut off parameter $q = 0.04$

Average Intracluster Similarity = **0.4516**

Agglomerative Clustering Example 1

Farthest neighbor

$$sim(S_1, S_2) = \max_{d_1 \in S_1, d_2 \in S_2} sim(d_1, d_2)$$

Average Intracluster Similarity = **0.304475**

```
1 [0.098857]
  2 [0.108415]
    3 [0.126011]
      4 [0.129523]
        5 [0.142059]
          6 [0.148069]
            7 [0.148331]
              8 [0.148599]
                9 [0.169039]
                  10 [0.17462]
                    11 [0.176625]
                      12 [0.201999]
                        13 [0.202129]
                          14 [0.223392]
                            15 [0.226308]
                              16 [0.23571]
                                Computer
                                Political
                              Economics
                            Chemistry
                          Anthropology
                        17 [0.245171]
                          Biology
                          Math
                      Communication
                    Physics
                  Psychology
                Music
              18 [0.177997]
                History
                Philosophy
            19 [0.186299]
              English
              Languages
        Art
      Theatre
    Sociology
  Geography
Justice
```

```
1 [0.138338]
  2 [0.175903]
    3 [0.237572]
      4 [0.342219]
        5 [0.57103]
          Art
          Psychology
        6 [0.588313]
          Communication
          Economics
      7 [0.39463]
        8 [0.617855]
          Computer
          Political
        9 [0.622585]
          Biology
          Math
    10 [0.292074]
      11 [0.519653]
        Justice
        Theatre
      12 [0.541863]
        Geography
        Physics
  13 [0.209028]
    14 [0.323349]
      15 [0.56133]
        Anthropology
        Sociology
      16 [0.5743]
        Chemistry
        Music
    17 [0.357257]
      18 [0.588999]
        History
        Philosophy
      19 [0.59315]
        English
        Languages
```

Intracluster similarity

$$sim(S) = \frac{1}{|S|^2} \sum_{d_i, d_j \in S} sim(d_i, d_j)$$

Average Intracluster Similarity = **0.434181**

DIANA clustering algorithm

0. $G \leftarrow \{d_1, d_2, ..., d_n\}$ (initialize G with singleton cluster of all documents); $g \leftarrow G$

1. Find a document d that is the most distinct from the others

$S \leftarrow \{d\}$

2.1. For all $d_i \notin S$, calculate $l_i = avg(\sum_{d_i \notin S} | d_i - d |) - avg(\sum_{d_i \notin S} | d_i - d |)$

2.2. Find d_h having max l_h, if $l_h > 0, S \leftarrow \{d\} \cup \{d_h\}$

3. Repeat until there is no $l_h > 0$. Update G

4. Select the cluster g having the biggest diameter $\sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{m}(d_i - d_j)^2}{m(m\text{-}1)}}$, goto 1

5. Stop when all clusters have a single document

Clustering evaluation

Web Content Mining

- Evaluating Clustering
  - Similarity-Based Criterion Functions
  - Probabilistic Criterion Functions
  - MDL-Based Model and Feature Evaluation
  - Classes to Clusters Evaluation
  - Precision, Recall and F-measure
  - Entropy

Similarity-Based Criterion Functions (distance)

- Basic idea: the cluster center c_i (*centroid* or *mean* in case of numeric data) best represents cluster D_i if it minimizes the sum of the lengths of the "error" vectors $x - c_i$ for all x $\in D_i$

$$J_e = \sum_{i=1}^{k} \sum_{x \in D_i} \left\| x - c_i \right\|^2 \qquad c_i = \frac{1}{\left| D_i \right|} \sum_{x \in D_i} x$$

- Alternative formulation based on *pairwise distance* between cluster members

$$J_e = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{\left| D_i \right|} \sum_{x_j, x_l \in D_i} \left\| x_j - x_l \right\|^2$$

Similarity-Based Criterion Functions (cosine similarity)

- For document clustering the (centroid) *cosine similarity* is used

$$J_s = \sum_{i=1}^{k} \sum_{d_j \in D_i} sim(c_i, d_j) \quad sim(c_i, d_j) = \frac{c_i \bullet d_j}{\|c_i\| \|d_j\|} \qquad c_i = \frac{1}{|D_i|} \sum_{d_j \in D_i} d_j$$

- Equivalent form based on *pairwise similarity*

$$J_S = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{|D_i|} \sum_{x_j, x_k \in D_i} sim(d_j, d_k)$$

- Another formulation based on *intracluster similarity* (used to controls merging of clusters in hierarchical agglomerative clustering)

Avegrage pair-wise similarity

$$J_S = \frac{1}{2} \sum_{i=1}^{k} \frac{1}{|D_i|} \sum_{x_j, x_l \in D_i} sim(d_j, d_l) = \frac{1}{2} \sum_{i=1}^{k} |D_i| sim(D_i)$$

Classes to Clusters Evaluation

- Assume that the classification of the documents in a sample is known, i.e. each document has a class label
- Cluster the sample without using the class labels
- Assign to each cluster the class label of the majority of documents in it
- Compute the *error* as the proportion of documents with different class and cluster label
- Or compute the *accuracy* as the proportion of documents with the same class and cluster label

Classes to Clusters Evaluation (Example)

history		science		research		offers		students		Hall	
14/20		16/20		17/20		14/20		14/20		12/20	
A (11/17)	B (3/3)	B (9/13)	A (7/7)	B (9/12)	A (8/8)	A (8/11)	B (6/9)	B (4/5)	A (10/15)	B (3/5)	A (9/15)
1-A	11-B	2-B	1-A	2-B	1-A	1-A	2-B	8-A	1-A	5-B	1-A
2-B	14-B	5-B	3-A	4-A	3-A	3-A	5-B	9-B	2-B	7-B	2-B
3-A	15-B	7-B	4-A	5-B	8-A	4-A	9-B	11-B	3-A	8-A	3-A
4-A		8-A	6-A	6-A	12-A	6-A	11-B	13-B	4-A	10-A	4-A
5-B		9-B	10-A	7-B	16-A	7-B	12-A	15-B	5-B	11-B	6-A
6-A		11-B	12-A	9-B	17-A	8-A	13-B		6-A		9-B
7-B		13-B	17-A	10-A	18-A	10-A	15-B		7-B		12-A
8-A		14-B		11-B	19-A	14-B	18-A		10-A		13-B
9-B		15-B		13-B		16-A	19-A		12-A		14-B
10-A		16-A		14-B		17-A			14-B		15-B
12-A		18-A		15-B		20-B			16-A		16-A
13-B		19-A		20-B					17-A		17-A
16-A		20-B							18-A		18-A
17-A									19-A		19-A
18-A									20-B		20-B
19-A											
20-B											

Confusion matrix (contingency table)

Actual (classes) \ Predicted (clusters)	Positive	Negative
Positive	*TP*	*FN*
Negative	*FP*	*TN*

TP (True Positive), FN (False Negative), FP (False Positive), TN (True Negative)

$$Error = \frac{FP + FN}{TP + FP + TN + FN} \qquad Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision and Recall

Actual (classes) \ Predicted (clusters)	Positive	Negative
Positive	*TP*	*FN*
Negative	*FP*	*TN*

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Attribute *research*		
Actual (classes) \ Predicted (clusters)	A	B
A	8	3
B	0	9

Precision = 1.00
Recall = 0.73

Attribute *hall*		
Actual (classes) \ Predicted (clusters)	A	B
A	9	2
B	6	3

Precision = 0.60
Recall = 0.82

F-Measure

Generalized confusion matrix for *m* classes and *k* clusters

Classes \ Clusters	1	…	*j*	…	*k*
1	n_{11}	…	n_{1j}	…	n_{1k}
…	…	…	…	…	…
i	n_{i1}	…	n_{ij}	…	n_{ik}
…	…	…	…	…	…
m	n_{m1}	…	n_{mj}	…	n_{mk}

Combining precision and recall

$$P(i,j) = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}} \qquad R(i,j) = \frac{n_{ij}}{\sum_{j=1}^{k} n_{ij}}$$

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j)+R(i,j)}$$

Evaluating the whole clustering

$$F = \sum_{i=1}^{m} \frac{n_i}{n} \max_{j=1,\ldots,k} F(i,j)$$

$$n_i = \sum_{j=1}^{k} n_{ij}$$

$$n = \sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij}$$

Total number of documents

F-Measure (Example)

offers	
n_{11}=8	n_{12}=3
n_{21}=3	n_{22}=6

students	
n_{11}=10	n_{12}=1
n_{21}=5	n_{22}=4

Accuracy(offers)=14/20=0.7

Accuracy(students)=14/20=0.7

offers	
$P(1,1)=0.73,\ R(1,1)=0.73$ $F(1,1)=0.73$	$P(1,2)=0.33,\ R(1,2)=0.27$ $F(1,2)=0.30$
$P(2,1)=0.27, R(2,1)=0.33$ $F(2,1)=0.30$	$P(2,2)=0.67, R(2,2)=0.67$ $F(2,2)=0.67$
$F=\frac{11}{20}0.73+\frac{9}{20}0.67=0.70$	

students	
$P(1,1)=0.67,\ R(1,1)=0.91$ $F(1,1)=0.77$	$P(1,2)=0.2,\ R(1,2)=0.09$ $F(1,2)=0.12$
$P(2,1)=0.33, R(2,1)=0.56$ $F(2,1)=0.42$	$P(2,2)=0.8, R(2,2)=0.44$ $F(2,2)=0.57$
$F=\frac{11}{20}0.77+\frac{9}{20}0.57=0.68$	

Entropy

- Consider the class label as a random event and evaluate its probability distribution in each cluster
- The probability of class *i* in cluster *j* is estimated by the proportion of occurrences of class label *i* in cluster *j*

$$p_{ij} = \frac{n_{ij}}{\sum_{i=1}^{m} n_{ij}}$$

- The entropy is as a measure of "impurity" and accounts for the average information in an arbitrary message about the class label

$$H_j = -\sum_{i=1}^{m} p_{ij} \log p_{ij}$$

Entropy (cont'd)

- To evaluate the whole clustering we sum up the entropies of individual clusters weighted with the proportion of documents in each

$$H = \sum_{i=1}^{k} \frac{n_j}{n} H_j$$

Entropy (Examples)

- A “pure” cluster where all documents have a single class label has entropy of 0
- The highest entropy is achieved when all class labels have the same probability
- For example, for a two class problem the 50-50 situation has the highest entropy of (-0.5 log 0.5- 0.5 log 0.5)=1

Entropy (Examples) (cont'd)

- Compare the entropies of the previously discussed clusterings for attributes *offers* and *students*

$$H(offers) = \frac{11}{20}\left(-\frac{8}{11}\log\frac{8}{11} - \frac{3}{11}\log\frac{3}{11}\right) + \frac{9}{20}\left(-\frac{3}{9}\log\frac{3}{9} - \frac{6}{9}\log\frac{6}{9}\right) = 0.878176$$

$$H(students) = \frac{15}{20}\left(-\frac{10}{15}\log\frac{10}{15} - \frac{5}{15}\log\frac{5}{15}\right) + \frac{5}{20}\left(-\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5}\right) = 0.869204$$

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

Reference

- J. Han and M. Kamber, *Data Mining-Concepts and Techniques*, Morgan Kaufmann, 2006
- Z. Markov and D. T. Larose, *Data mining the web, uncovering patterns in Web content, structure and usage*, John Wiley & Sons, 2007
- Nguyễn Hà Nam, Nguyễn Trí Thành, Hà Quang Thụy, *Giáo trình khai phá dữ liệu*, NXB ĐHQGHN, 2014

Discussion