Cluster Analysis

Group: 7
Course: CSE 537
Instructor: Professor Anita Wasilewska
Presenters: Chenjun Feng 108413298
Yunke Tian 109929662

Sources Cited

[1] Driver, H. E. and A. L. Kroeber (1932) Quantitative expression of cultural relationships. University of California Publications in American Archaeology and Ethnology 31: 211-56.
[2] Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.
[3] Cattell, R. B. (1943). "The description of personality: Basic traits resolved into clusters". Journal of Abnormal and Social Psychology 38: 476–506. doi:10.1037/h0054116
[4] Wasilewska, Anita. (2016). “Introduction to Learning”. The State University of New York at Stony Brook. CSE 537 Spring 2016 Lecture Slides Page 27-28 http://www3.cs.stonybrook.edu/~cse634/16L7learningintrod.pdf
[5] Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.
[6] Sergios Theodoridis & Konstantinos Koutroumbas (2006). Pattern Recognition 3rd ed. p. 635
[7] http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html
[8] McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123
[9] Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352
[10] Zhang, et al. "Graph degree linkage: Agglomerative clustering on a directed graph." 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012
[11] "The DISTANCE Procedure: Proximity Measures". *SAS/STAT 9.2 Users Guide*. SAS Institute. Retrieved 2009-04-26
[12] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering". *The Elements of Statistical Learning* (PDF) (2nd ed.). New York: Springer. pp. 520–528
[13] Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* (American Statistical Association) 66 (336): 846–850
[14] Färber, Ines; Günnemann, Stephan; Kriegel, Hans-Peter; Kröger, Peer; Müller, Emmanuel; Schubert, Erich; Seidl, Thomas; Zimek, Arthur (2010). "On Using Class-Labels in Evaluation of Clusterings"
[15] Li, Xin-Ye; Guo, Li-Jie (2012), "Constructing affinity matrix in spectral clustering based on neighbor propagation", *Neurocomputing* (MIT Press) 97: 125–130

Overview

1. Early History
2. Importance of Clustering
3. Similarity Measure (partially covered by Group 2)
4. K-medoids Clustering
5. Hierarchical Clustering
6. Density-based Clustering (will be covered by Group 14)
7. EM Clustering (covered by Group 2)
8. Pre-clustering

Brief Early History

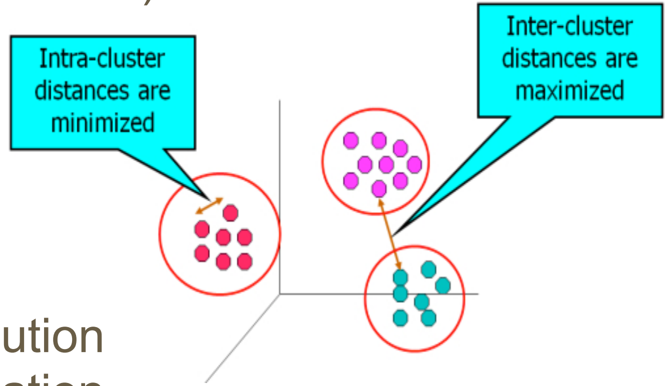
- Originated in anthropology by Driver and Kroeber in 1932[1]
- Introduced to psychology by Zubin in 1938 and Robert Tryon in 1939[2]
- Famously used by Cattell beginning in 1943[3] for trait theory classification in personality psychology
- Widely used in social science, medicine, biology, geography, pattern recognition, and etc.

[1] Driver, H. E. and A. L. Kroeber (1932) Quantitative expression of cultural relationships. University of California Publications in American Archaeology and Ethnology 31: 211-56.
[2] Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.
[3] Cattell, R. B. (1943). "The description of personality: Basic traits resolved into clusters". Journal of Abnormal and Social Psychology 38: 476–506. doi:10.1037/h0054116

The Importance of Clustering

Clustering = Unsupervised learning (no predefined classes)

To group a collection of data objects so that
- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters



Clustering results are used:
- As a stand-alone tool to get insight into data distribution
Visualization of clusters may unveil important information
- As a preprocessing step for other algorithms
Efficient indexing or compression often relies on clustering

Wasilewska, Anita. (2016). “Introduction to Learning”. The State University of New York at Stony Brook. CSE 537 Spring 2016 Lecture Slides Page 27-28 http://www3.cs.stonybrook.edu/~cse634/16L7learningintrod.pdf
Image source: https://apandre.wordpress.com/visible-data/cluster-analysis/

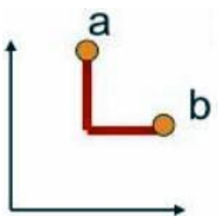
Similarity Measure

Group 2: “Data Visualization” has partially covered this part.

For two objects **X** ($x_1, x_2, x_3, ..., x_i$) and **Y** ($y_1, y_2, y_3, ..., y_i$), we have:

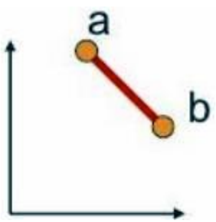
Manhattan Distance =

$$\sum_{i=1}^{n} |x_i - y_i|$$



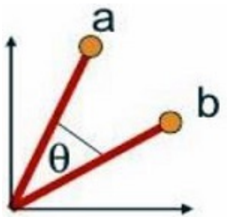
Euclidean Distance =

$$\left(\sum_{i=1}^{n} |x_i - y_i|^2\right)^{1/2}$$



Cosine Similarity =

$$\frac{\langle x, y\rangle}{\|x\|\,\|y\|} \qquad \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$



Pearson Correlation =

$$\frac{\langle x-\bar{x},\ y-\bar{y}\rangle}{\|x-\bar{x}\|\,\|y-\bar{y}\|} \qquad \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2}\sqrt{\sum (y_i-\bar{y})^2}}$$

Li, Xin-Ye; Guo, Li-Jie (2012), "Constructing affinity matrix in spectral clustering based on neighbor propagation", *Neurocomputing* (MIT Press) 97: 125–130

Features of Similarity Measure

Euclidean Distance measures distance between two points.

- Most commonly used
- Not suitable for high dimensional data
- Best choice if data are dense and continuous

Cosine Similarity measures angle between two vectors.

- Value range from -1 to 1
- Invariant to scaling, but sensitive to shifting
- Best choice if data is sparse and asymmetric

Pearson Correlation measures linear relationship between two variables.

- Value range from -1 to 1
- Invariant to both scaling and shifting
- Best choice if data tends to be linearly correlated

Li, Xin-Ye; Guo, Li-Jie (2012), "Constructing affinity matrix in spectral clustering based on neighbor propagation", *Neurocomputing* (MIT Press) 97: 125–130

Partition Clustering (K-means vs K-medoids)

K-means algorithm has been covered in the presentation of Group 2: “Data Visualization”

Drawback of K-means: very sensitive to outliers because such objects dramatically distort the mean value of the cluster

Solution: K-medoids

Using actual objects to represent the clusters, based on the principle of minimizing the sum of general pairwise dissimilarities in each cluster

Kaufman, L. and Rousseeuw, P.J. (1987), Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods, edited by Y. Dodge, North-Holland, 405–416.

K-medoids Clustering (PAM)

PAM = Partitioning Around Medoids

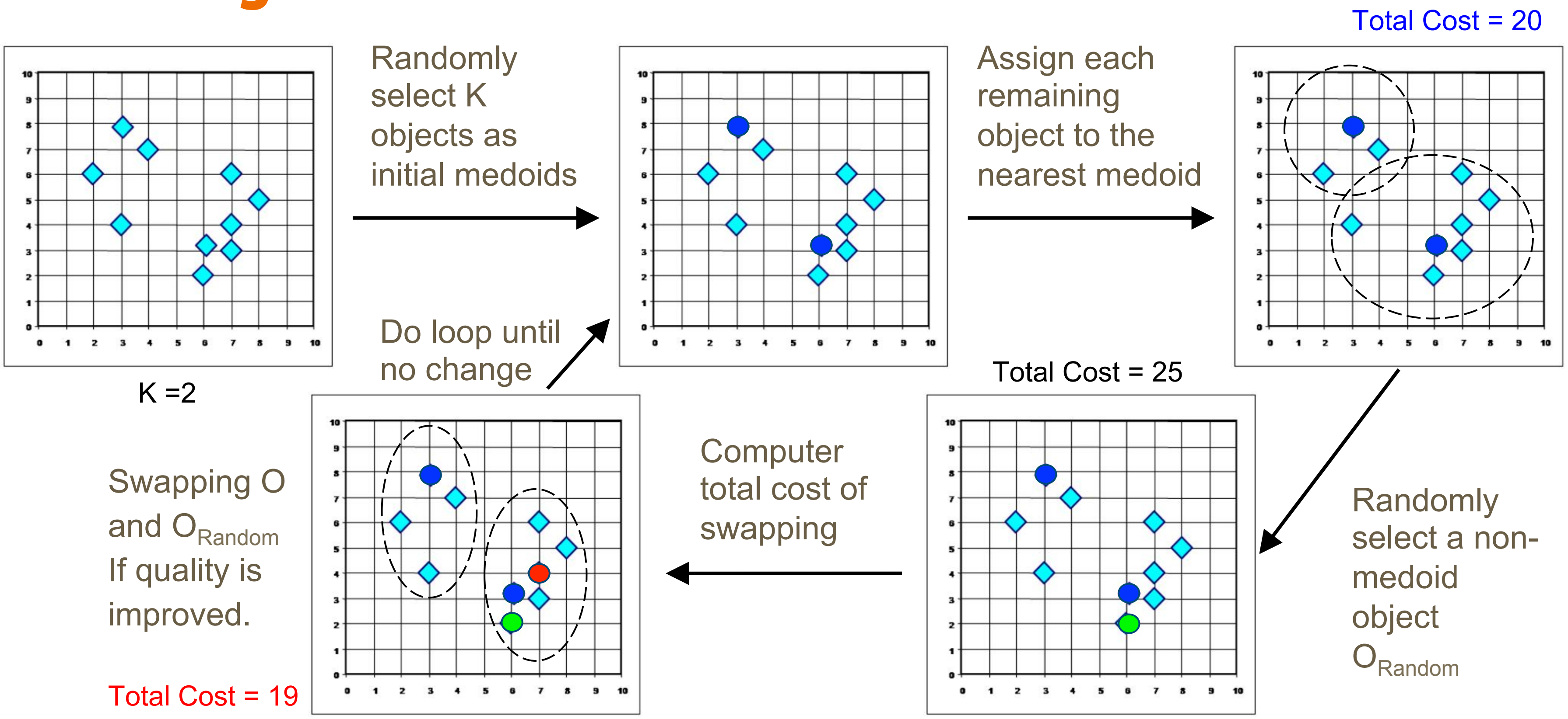
A popular realization of k-medoids clustering

Algorithm:

1. Randomly select K representative objects as initial medoids
2. Associate each data point to the closest medoid
3. Randomly select a non-representative object O_{random}
4. Compute the total cost S of swapping the medoid m with O_{random}
5. If $S < 0$, then swap m with O_{random} to form the new set of medoids
6. Repeat steps 2 - 5 until there is no change

Sergios Theodoridis & Konstantinos Koutroumbas (2006). Pattern Recognition 3rd ed. p. 635.

PAM Algorithm



Advantages & Disadvantages of K-medoids

Advantages:

- It is more robust to noise and outliers as compared to K-means

Disadvantages:

- It requires the specification of K.
- The computation is very costly when data sets are large
  The complexity of each iteration is $O(K(n-K)^2)$, where K is the number of clusters and n is the number of data .

Improvement:

Randomly sample large dataset, then apply PAM algorithm to multiple samples.

http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

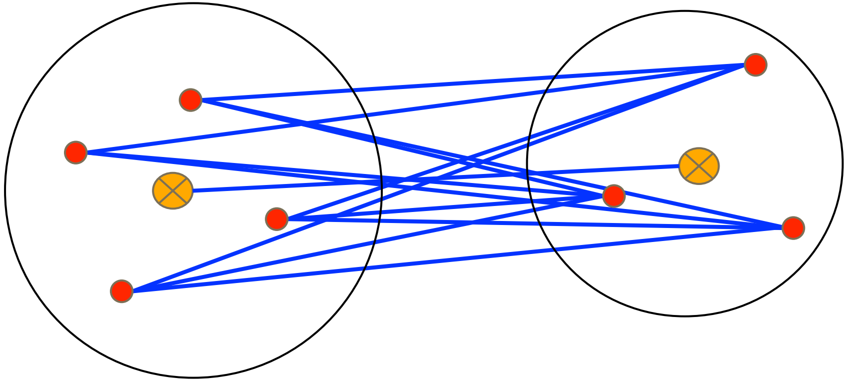
Hierarchical Clustering

Hierarchical Clustering: Creating a hierarchical decomposition of the set of objects using similarity matrix as clustering criteria

Two Main Algorithms: 1. Agglomerative method 2. Divisive method

Similarity Matrix: Linkage methods

- MIN
- MAX
- Group Average
- Distance of Centroid



Rokach, Lior, and Oded Maimon. "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005. 321-352.

Agglomerative Algorithm

Main Steps:

1. Let each data point be a cluster
2. Initialize and compute the similarity matrix

Repeat

3. Merge the two closest clusters
4. Update the similarity matrix

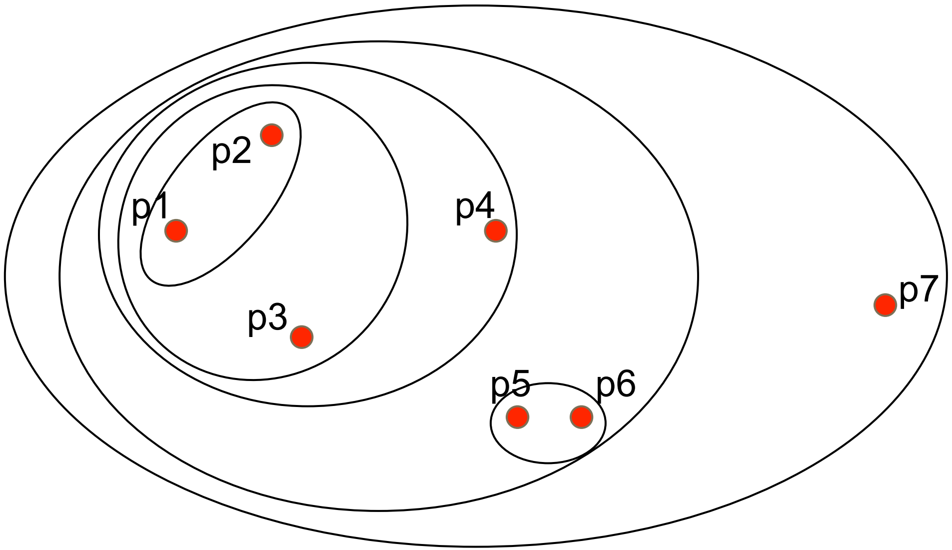
Until only a single cluster remains

5. Draw the dendrogram of the sequences of merges
6. Cut the dendrogram with a certain level to form a certain clustering

Zhang, et al. "Graph degree linkage: Agglomerative clustering on a directed graph." 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012.

Agglomerative Algorithm Example

Use MIN Linkage Method



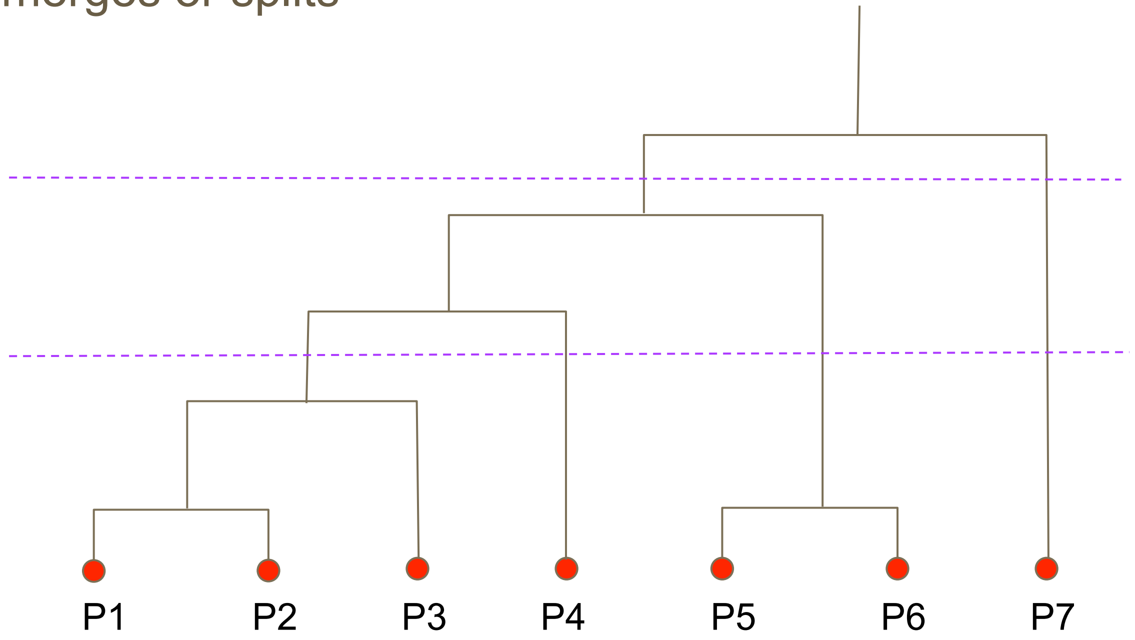
	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Update

Agglomerative Algorithm Example (cont.)

Visualized as a dendrogram: A tree like diagram that records the sequences of merges or splits



{P1, P2, P3, P4, P5, P6}, {P7}

{P1, P2, P3}, {P4}, {P5, P6}, {P7}

Different cuttings generate different clusters!

Similarity Matrix Comparisons

Original Points Two Clusters

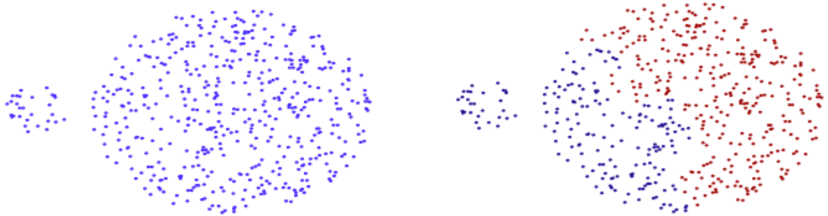
Pro of MIN: No bias towards larger cluster

Original Points Two Clusters

Con of MIN: Sensitive to noises and outliers

Original Points Two Clusters

Pro of MAX: Non-sensitive to noises and outliers



Original Points Two Clusters

Con of MAX: Tends to break & bias to larger clusters

"The DISTANCE Procedure: Proximity Measures". *SAS/STAT 9.2 Users Guide*. SAS Institute. Retrieved 2009-04-26.

Advantages & Disadvantages of Hierarchical Clustering

Advantages:

- There is no need to specify number of clusters.
- Any number of clusters can be obtained using different cutting level.

Disadvantages:

- The computation is very costly. The complexity is $O(N^2 \log(N))$ for Agglomerative, where N stands for the number of data.
- Once two clusters merged, the process cannot be undone.
- It has the problem of either noise sensitivity or large cluster bias

Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2009). "14.3.12 Hierarchical clustering". *The Elements of Statistical Learning* (PDF) (2nd ed.). New York: Springer. pp. 520–528

Density-based Clustering (DBSCAN)

Group 14 will cover this part in their presentation soon.

Distribution-based Clustering (EM Clustering)

Group 2: “Data Visualization” has covered this method.

Pre-clustering for Big Data (Canopy Clustering)

Problem: Traditional clustering algorithms are expensive when the dataset is large.

Solution: Pre-clustering (Canopy)

First stage - using a cheap, approximate distance measure to efficiently divide large data into overlapping subsets, called canopies.

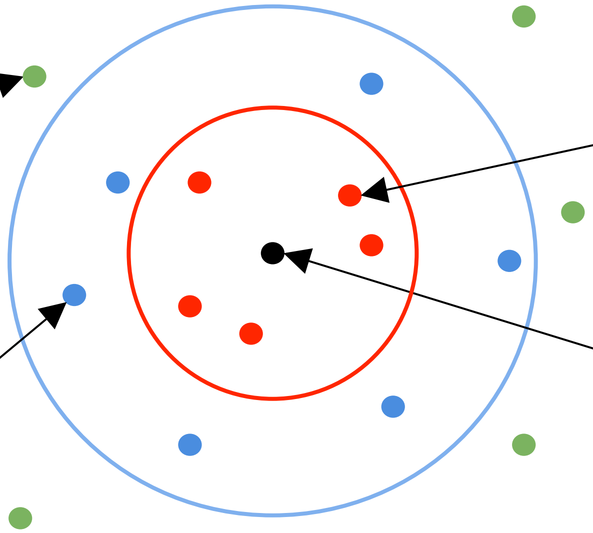
Second stage - only using expensive distance measurements among points that occur in a common canopy

McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123

Canopy Clustering Algorithm (stage 1)

1. Begin with dataset points and with two thresholds T1 (loose distance) and T2 (tight distance), where T1 > T2

3. Points at a distance > T1 don’t belong to canopy A; Points at a distance < T1 belong to canopy A



4. Points at a distance < T2 will be removed and will not be considered as canopy centers in future steps

2. Randomly select a point as the center of a Canopy (A)

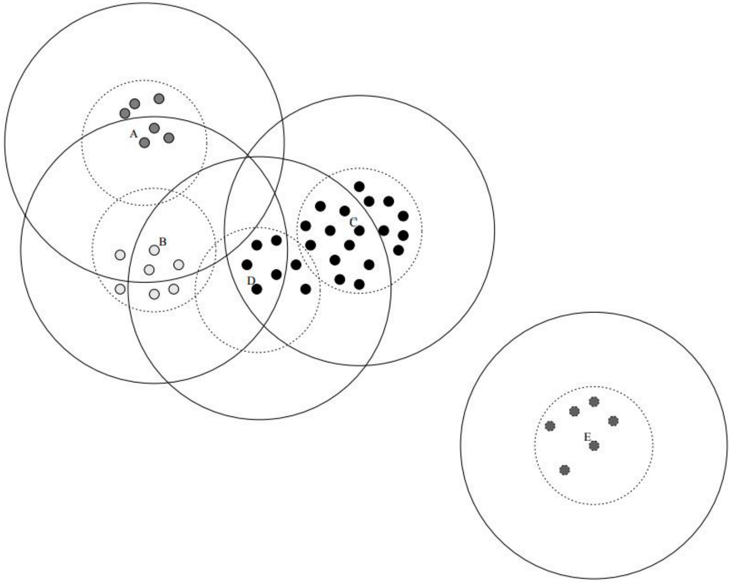
Note: Points at a distance between T2 and T1 belong to canopy A, but can also be canopy centers in future steps

McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123

Canopy Clustering Algorithm (stage 2)

- Repeat step 2 to 4 on the previous slides until no more point can be selected as canopy center

- Expensive distance measurements will only be made between pairs of points in the same canopies, far fewer than all possible pairs in the whole dataset



An example of four data clusters and the canopies that cover them

McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123

Advantages & Disadvantages of Canopy Clustering

Advantages:

- It is fast. Computational complexity:

$$O\left(c\left(\frac{fn}{c}\right)^2\right) = O\left(\frac{f^2}{c}n^2\right)$$

n: number of data points
c: number of canopies
f: average number of canopies covering a data point
fn/c: data points per canopy

- It is widely applicable to many traditional clustering algorithm
- It will not lose any clustering accuracy (may slightly increase accuracy).
- Formerly impossible large clustering problems become practical.

Disadvantages:

- It requires the specification of distance thresholds.
- The weakness of traditional algorithms applied in stage 2 will still exists.

McCallum, A.; Nigam, K.; and Ungar L.H. (2000) "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 169-178 doi:10.1145/347090.347123

Cluster Evaluation

To evaluate whether a clustering is good or bad.

Three aspects:

- Assessing Clustering Tendency: for a data set, test whether a nonrandom structure exists in data
- Determine Number of Clusters: compare the number of resulting clusters with the "optimal" number of clusters
- Measuring Clustering Quality: Extrinsic methods and Intrinsic methods

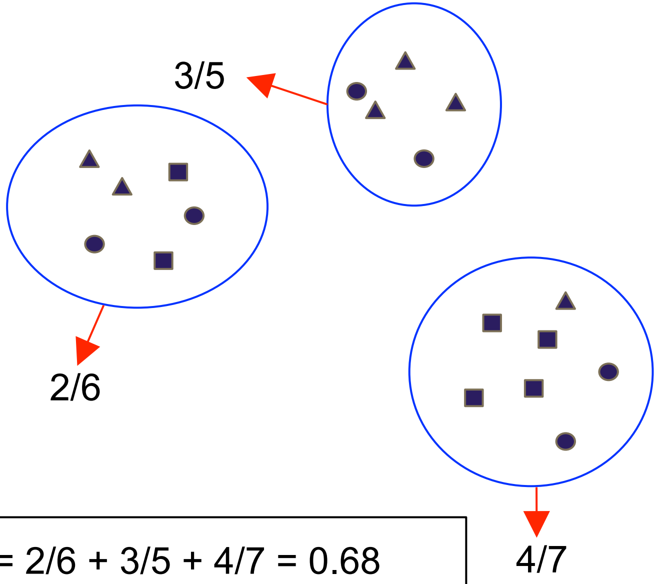
Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* (American Statistical Association) 66 (336): 846–850

Extrinsic Method - Purity Method

This method is used when ideal clustering of objects is known.

$$Purity = \sum_{r=1}^{k} \frac{1}{n} \max_i (n_r^i)$$

points of class
i in cluster *r*



Purity = 2/6 + 3/5 + 4/7 = 0.68

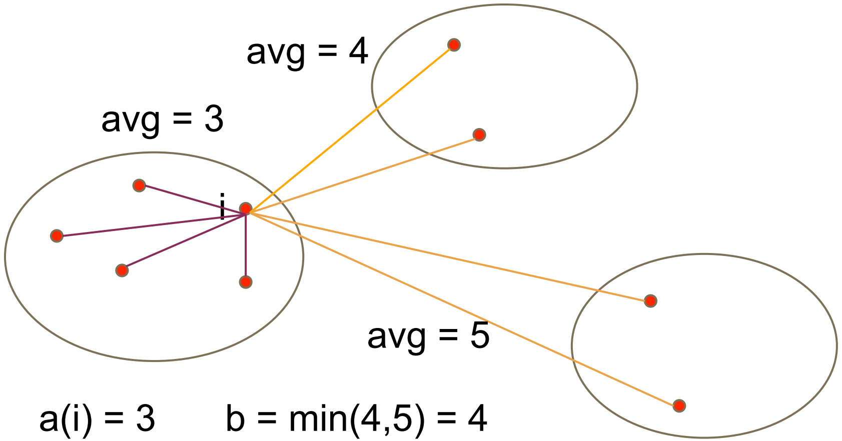
Properties of purity value:

- between 0 and 1
- The more close to 0, the worse
- The more close to 1, the better

Färber, Ines; Günnemann, Stephan; Kriegel, Hans-Peter; Kröger, Peer; Müller, Emmanuel; Schubert, Erich; Seidl, Thomas; Zimek, Arthur (2010). "On Using Class-Labels in Evaluation of Clusterings"

Intrinsic Method - Silhouette Coefficient

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$



b(i) = min(AVGD_BETWEEN(i,k))

a(i) = AVGD_WITHIN(i)

b: how separate i is from other clusters. The larger the better

a: how compact the cluster of i is. The smaller the better

s: between -1 and 1. The larger the better

Rand, W. M. (1971). "Objective criteria for the evaluation of clustering methods". *Journal of the American Statistical Association* (American Statistical Association) 66 (336): 846–850

Summary

1. Cluster analysis was early used in anthropology and psychology in 1930s.
2. Clustering is to minimize intra-cluster similarities and maximize inter-cluster similarities.
3. Manhattan distance, euclidean distance, cosine similarity and pearson correlation are common similarity measures. The property of a dataset determines which one to use.
4. K-medoids clustering (PAM) uses actual objects to represent clusters. It is more robust to outliers than K-means, but the computation is costly.
5. Hierarchical clustering uses similarity matrix to cluster datasets hierarchically. It doesn't require specified number of clusters, but is time costly and not undoable.
6. Canopy clustering is used to pre-cluster large data efficiently. Firstly, it uses a cheap distance measure to get canopies. Secondly, it uses a traditional distance measure for the points only in the same canopy.
7. Cluster evaluation uses extrinsic methods (Purity) and intrinsic methods (Silhouette Coefficient) to assess the goodness of a clustering.

Thank You!