CSC4008 Data Mining

117010279 Ziren WANG

May 2020

7 Cluster Analysis: Basic Concepts & Methods

7.1 Overview

Cluster is a collection of data objects, where similar (or related) to one another within the same group while dissimilar (or unrelated) to the objects in other groups.There can are several usages of clustering as a preprocessing tool:

- **Summarization**: Preprocessing for regression, PCA, classification, and association analysis;
- **Compression**: Image processing: vector quantization;
- **Finding K-nearest Neighbors**: Localizing search to one or a small number of clusters;
- **Outlier detection**: Outliers are often viewed as those "far away" from any cluster.

A good clustering method will produce high quality clusters. i.e. cohesive within clusters while distinctive between clusters. The quality of a clustering method depends on the similarity measure used by the method; its implementation, and its ability to discover some or all of the hidden patterns. They are several considerations for cluster analysis:

- **Partitioning Criteria**: Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable);
- **Separation of Clusters**: Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class);
- **Similarity Measure**: Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity);
- **Clustering Space**: Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering).

7.2 Partitioning Methods

The objective of Partitioning method is to Partition a database D of n objects into a set of k clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i). i.e. $E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - c_i)^2$ is minimized.

7.2.1 K-means Clustering Method

The strength of K-means is efficient ($O(tkn)$, where n is the number of objects , k is the number of clusters and the t is the number of iterations. Generally k, t << n)
On the other hand, the weakness of K-means are 1) Applicable only to objects in a continuous n-dimensional space; 2) Need to specify k, the number of clusters, in advance; 3) Sensitive to noisy data and outliers; 4) Not suitable to discover clusters with non-convex shapes.

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Figure 1: K-means partitioning algorithm

Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

Figure 2: K-medoids partitioning algorithm

7.2.2 K-medoids Clustering Methods

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.

7.3 Hierarchical Methods

There are 2 kinds of methods, namely: agglomerative(AGNES): bottom-up and divisive(DIANA): top-down. We use dendrogram to represent the structure of hierarchy. There are different criteria to define the similarity(distance) between the clusters:

- **Single link**: smallest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = min(t_{ip}, t_{jq})$;
- **Complete link**: largest distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = max(t_{ip}, t_{jq})$;
- **Average**: average distance between an element in one cluster and an element in the other, i.e., $dist(K_i, K_j) = ave(t_{ip}, t_{jq})$;
- **Centroid**: Centroid: distance between the centroids of two clusters, i.e., $dist(K_i, K_j) = dist(C_i, C_j)$;
- **Medoid**: distance between the medoids of two clusters, i.e., $dist(K_i, K_j) = dist(M_i, M_j)$.

Note:
Centroid is the middle of a cluster, defined as: $C_m = \frac{\sum_{i=1}^{N}(t_{ip})}{N}$;
Radius is the square root of average distance from any point of the cluster to its centroid, defined as: $R_m = \sqrt{\frac{\sum_{i=1}^{N}(t_{ip}-c_m)^2}{N}}$;
Diameter is the square toot of average mean squared distance between all pairs of points in the cluster, defined as: $\sqrt{\frac{\sum_{i=1}^{N}\sum_{i=1}^{N}(t_{ip}-t_{iq})^2}{N(N-1)}}$

7.3.1 BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)

BIRCH incrementally constructs a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering. There are two steps: (1): scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data); (2) use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree. BIRCH is (1) sensitive to insertion order of data points; (2)

Clustering feature. Suppose there are three points, $(2,5), (3,2)$, and $(4,3)$, in a cluster, C_1. The clustering feature of C_1 is

$$CF_1 = \langle 3, (2+3+4, 5+2+3), (2^2+3^2+4^2, 5^2+2^2+3^2)\rangle = \langle 3, (9,10), (29,38)\rangle.$$

Suppose that C_1 is disjoint to a second cluster, C_2, where $CF_2 = \langle 3, (35,36), (417,440)\rangle$. The clustering feature of a new cluster, C_3, that is formed by merging C_1 and C_2, is derived by adding CF_1 and CF_2. That is,

$$CF_3 = \langle 3+3, (9+35, 10+36), (29+417, 38+440)\rangle = \langle 6, (44,46), (446,478)\rangle. \quad \blacksquare$$

Figure 3: BIRCH Example

Since we fix the size of leaf nodes, so clusters may not be so natural; (3) Clusters tend to be spherical given the radius and diameter measures.

7.3.2 CHAMELEON: Hierarchical Clustering Using Dynamic Modeling

This method measures the similarity based on a dynamic model. Two clusters are merged only if the interconnectivity and closeness (proximity) between two clusters are high relative to the internal interconnectivity of the clusters and closeness of items within the clusters.
There relative interconnectivity, $RI(C_i, C_j)$, between two clusters, C_i and C_j, is defined as the absolute interconnectivity between C_i and C_j, normalized with respect to the internal interconnectivity of the two clusters, C_i and C_j. That is, $RI(C_i, C_j) = \frac{|EC_{C_i,C_j}|}{\frac{1}{2}(|EC_{C_i}||EC_{C_j}|)}$. Where EC_{C_i,C_j} is the edge cut as previously defined for a cluster containing both C_i and C_j. Similarly, EC_{C_i} (or EC_{C_j}) is the minimum sum of the cut edges that partition C_i (or C_j) into two roughly equal parts.

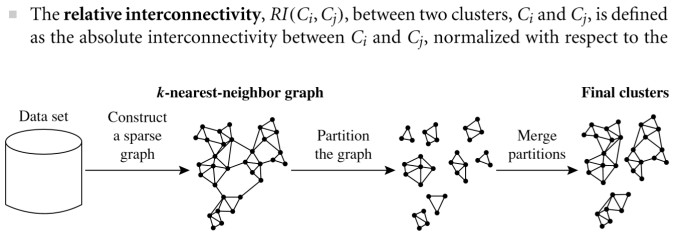


Figure 4: Chameleon: hierarchical clustering based on k-nearest neighbors and dynamic modeling

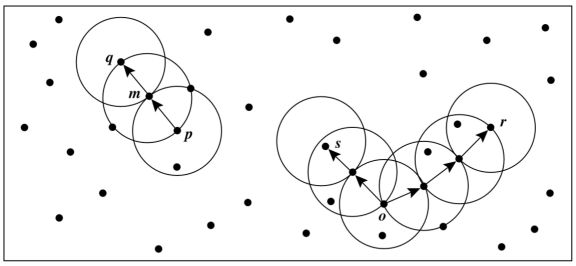
7.4 Density-Based Methods

Major features of DBM: (1) Discover clusters of arbitrary shape; (2) more capable to handle noise; (3) One scan; (4)Need density parameters as termination condition. i.e. Eps: Maximum radius of the neighbourhood and MinPts: Minimum number of points in an Eps-neighbourhood of that point. By considering those 2 parameters, we formally define: $density = \frac{\#neighbors}{\epsilon^2\pi}$

7.4.1 DBSCAN

7.4.2 OPTICS

To overcome the difficulty in using one set of global parameters in clustering analysis, a cluster analysis method called OPTICS was proposed. OPTICS does not require the user to provide a specific density threshold. The



Density-reachability and density-connectivity in density-based clustering. *Source:* Bas

Figure 5: Density-Reachable and Density-Connected

Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

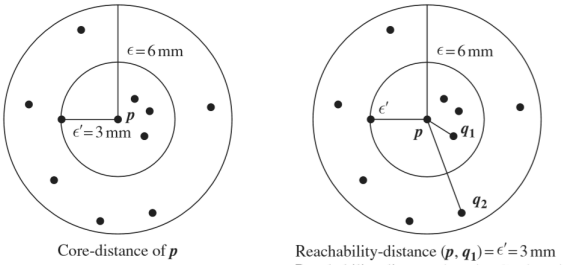
Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                     add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

Figure 6: DBSCAN algorithm

cluste rordering can be used to extract basic clustering information (e.g., cluster centers, or arbitrary-shaped clusters), derive the intrinsic clustering structure, as well as provide a visualization of the clustering.



Core-distance of p

Reachability-distance $(p, q_1) = \epsilon' = 3$ mm
Reachability-distance $(p, q_2) = dist\,(p, q_2)$

Figure 7: OPTICS terminology

7.5 Grid-Based Methods

7.5.1 Strength and Weakness

Strength:

- Automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces;
- Insensitive to the order of records in input and does not presume some canonical data distribution;
- Scales linearly with the size of input and has good scalability as the number of dimensions in the data increases.

Weakness: The accuracy of the clustering result may be degraded at the expense of simplicity of the method.

7.5.2 Several Algorithms

- STING (a STatistical INformation Grid approach) by Wang, Yang and Muntz (1997)
- WaveCluster (A multi-resolution clustering approach using wavelet method) by Sheikholeslami, Chatterjee, and Zhang (VLDB'98)
- CLIQUE (Both grid-based and subspace clustering) by Agrawal, et al. (SIGMOD'98)

7.6 Evaluation of Clustering

7.6.1 Assessing Clustering Tendency

The **Hopkins Statistic** is a spatial statistic that tests the spatial randomness of a variable as distributed in a space. Given a data set, D, which is regarded as a sample of a random variable, o, we want to determine how far away o is from being uniformly distributed in the data space. We calculate the Hopkins Statistic as follows:

1. Sample n points, $\boldsymbol{p_1}, \ldots, \boldsymbol{p_n}$, uniformly from D. That is, each point in D has the same probability of being included in this sample. For each point, $\boldsymbol{p_i}$, we find the nearest neighbor of $\boldsymbol{p_i}$ $(1 \leq i \leq n)$ in D, and let x_i be the distance between $\boldsymbol{p_i}$ and its nearest neighbor in D. That is,

$$x_i = \min_{\boldsymbol{v} \in D} \{dist(\boldsymbol{p_i}, \boldsymbol{v})\}. \tag{10.25}$$

2. Sample n points, $\boldsymbol{q_1}, \ldots, \boldsymbol{q_n}$, uniformly from D. For each $\boldsymbol{q_i}$ $(1 \leq i \leq n)$, we find the nearest neighbor of $\boldsymbol{q_i}$ in $D - \{\boldsymbol{q_i}\}$, and let y_i be the distance between $\boldsymbol{q_i}$ and its nearest neighbor in $D - \{\boldsymbol{q_i}\}$. That is,

$$y_i = \min_{\boldsymbol{v} \in D, \boldsymbol{v} \neq \boldsymbol{q_i}} \{dist(\boldsymbol{q_i}, \boldsymbol{v})\}. \tag{10.26}$$

3. Calculate the Hopkins Statistic, H, as

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}. \tag{10.27}$$

"*What does the Hopkins Statistic tell us about how likely data set* D *follows a uniform distribution in the data space?*" If D were uniformly distributed, then $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} x_i$ would be close to each other, and thus H would be about 0.5. However, if D were highly skewed, then $\sum_{i=1}^{n} y_i$ would be substantially smaller than $\sum_{i=1}^{n} x_i$ in expectation, and thus H would be close to 0.

Figure 8: Hopkins Statistic Analysis

7.6.2 Determining the Number of Clusters

- Empirical method: # of clusters $\approx \sqrt{\frac{n}{2}}$ for a dataset of n points;
- Elbow method: Use the turning point in the curve of sum of within cluster variance w.r.t the # of clusters;

- Cross validation method: Divide a given data set into m parts. Use m – 1 parts to obtain a clustering model. Use the remaining part to test the quality of the clustering. For any k ¿ 0, repeat it m times, compare the overall quality measure w.r.t. different k's, and find of clusters that fits the data the best.

7.6.3 Measuring Clustering Quality

- **Extrinsic**: supervised, i.e., the ground truth is available. Compare a clustering against the ground truth using certain clustering quality measure. Eg. BCubed precision and recall metrics;
- **Intrinsic**: unsupervised, i.e., the ground truth is unavailable.Evaluate the goodness of a clustering by considering how well the clusters are separated, and how compact the clusters are. Eg. Silhouette coefficient

As for extrinsic methods, a measure Q on clustering quality is effective if it satisfies the following four essential criteria:

- **Cluster homogeneity**: the purer, the better;
- **Cluster completeness**: should assign objects belong to the same category in the ground truth to the same cluster;
- **Rag bag**: putting a heterogeneous object into a pure cluster should be penalized more than putting it into a rag bag (i.e., "miscellaneous" or "other" category);
- **Small cluster preservation**: splitting a small category into pieces is more harmful than splitting a large category into pieces.