CSE 444
(Pattern Recognition Sessional)
Clustering

Iyolita Islam

Department of Computer Science and Engineering
Military Institute of Science and Technology

Last Updated: September 19, 2018

Outline

Basic Sequential Algorithm Scheme

- Initially, the number of clusters is not known.
- New clusters are created as the algorithm evolves.
- Each new vector is assigned either to an existing cluster or a new one depending on its distance from already formed ones.

Representation of BSAS

- dataset, X $= \{x_1, x_2, ..., x_N\}$
- d(x, c) = dissimilarity between the vector, x and the cluster, c
- θ = threshold of dissimilarity
- q = maximum number of allowable clusters
- m = the number of current clusters

BSAS - Algorithm

- m = 1
- $C_m = \{x_1\}$
- For i = 2 to N
  - Find $C_k : d(x_i, C_k) = min_{1 \leq j \leq m}$ d(x_i, C_j)
  - if (d$(x_i, C_k) > \theta$) AND $(m < q)$ then
    - * m = m + 1
    - * $C_m = \{x_i\}$
  - -Else
    - * $C_k = C_k$ U $\{x_i\}$
    - * Where necessary, update representatives, R_{C_k}.
  - -End {if}
- End {For}

BSAS - Algorithm Steps

- Initially, m = l
- Take the first vector, x_1 into C_m, $C_m = \{x_1\}$
- Take the next vector and measure the dstance from the existing clusters.
    - Find $C_k : d(x_i, C_k) = min_{1 \leq j \leq m}$ d(x_i, C_j)
- Pick the minimum one, check the threshold and q.
    - if (d$(x_i, C_k) > \theta$) AND $(m < q)$
        - m = m + l
        - $C_m = \{x_i\}$
    - else
        - $C_k = C_k$ U $\{x_i\}$
        - Where necessary, update representative R_{C_k}.
- Repeat from the Step 02 for Rest of the data.

Example - BSAS

Example

Apply BSAS for the following data vectors and information:

- $x_1 = (2, 5)$ $x_4 = (2, 2)$ $x_7 = (1, 1)$
  $x_2 = (6, 4)$ $x_5 = (1, 4)$ $x_8 = (2, 1)$
  $x_3 = (5, 9)$ $x_6 = (5, 4)$
- Feature order: $x_8, x_6, x_1, x_5, x_2, x_4, x_3, x_7$
- $\theta = 2.5$
- q = 6

BSAS Queries

What factors affect the result of BSAS?

- Feature order
- θ
- q

What are the disadvantages of BSAS? - self study

Modified Basic Sequential Algorithm Scheme

- Modified version of BSAS
- It has two phases:
  - Total cluster determination
  - Clustering of unassigned vectors

Representation of MBSAS

- dataset, X $= \{x_1, x_2, ..., x_N\}$
- d(x, c) = dissimilarity between the vector, x and the cluster, c
- θ = threshold of dissimilarity
- q = maximum number of allowable clusters
- m = the number of current clusters

MBSAS - Algorithm

Cluster Determination

- m = l
- $C_m = \{x_1\}$
- For i = 2 to N
  - Find $C_k : d(x_i, C_k) = min_{1 \leq j \leq m}$ d(x_i, C_j)
  - if (d$(x_i, C_k) > \theta$) AND $(m < q)$ then
    - m = m + l
    - $C_m = \{x_i\}$
  - End {if}
- End {For}

MBSAS - Algorithm (Continued)

Assignment of vectors

- For i = 1 to N
  - if x_i has not been assigned to a cluster, then
  - Find $C_k : d(x_i, C_k) = min_{1 \leq j \leq m}$ d(x_i, C_j)
    - $C_k = C_k$ U $\{x_i\}$
    - Where necessary, update representatives, R_{C_k}.
  - End {if}
- End {For}

Example - MBSAS

Example

Apply MBSAS for the following data vectors and information:

- $x_1 = (2, 5)$ $x_4 = (2, 2)$ $x_7 = (1, 1)$
  $x_2 = (6, 4)$ $x_5 = (1, 4)$ $x_8 = (2, 1)$
  $x_3 = (5, 9)$ $x_6 = (5, 4)$
- Feature order: $x_8, x_6, x_1, x_5, x_2, x_4, x_3, x_7$
- $\theta = 2.5$
- q = 6

MBSAS Queries

What factors affect the result of MBSAS?

- Feature order
- θ
- q

What are the advantage/ disadvantage of MBSAS? - self study

K-means Algorithm

- It uses "Centroid" concept (same as representative).
- A vector is considered to be in a particular cluster, if it is closer to the centroid of that clusters than that of the others.

Representation of K-means

- dataset, $X = \{x_1, x_2, ..., x_N\}$
- k = the number of clusters

K-means - Algorithm Steps

1. Initialize the centroids- $C_1, C_2, C_3, ..., C_k$
2. Calculate the distance between each point and centroids.
3. Assign the data point to the nearest cluster based on the minimum distance from centroid.
4. Recalculate the centroids.

   For example, if C_i has S members, then,

   Centroid of $C_i = (\frac{x_1+x_2+...+x_s}{s} , \frac{y_1+y_2+...+y_s}{s})$
5. Recalculate the distance between every data vector and centroid.
6. if there are changes in clusters repeat from Step 02.
7. if there is no adjustment, STOP.

Example - K-means

Example

Apply K-means for the following data vectors and information:

- $x_1 = (2, 5)$ $x_4 = (2, 2)$ $x_7 = (1, 1)$
  $x_2 = (6, 4)$ $x_5 = (1, 4)$ $x_8 = (2, 1)$
  $x_3 = (5, 9)$ $x_6 = (5, 4)$
- $k = 2$

K-means Queries

What factors affect the result of K-means?

- k

What are the disadvantages of K-means? - self study


Thank You