بازشناسی الگو

درس ۱۵

خوشه‌بندی: الگوریتم‌های ترتیبی

Clustering: Sequential Algorithms

کاظم فولادی

دانشکده مهندسی برق و کامپیوتر

دانشگاه تهران

http://courses.fouladi.ir/pr

خوشه‌بندی: الگوریتم‌های ترتیبی

۱

مقدمه

CLUSTERING ALGORITHMS

❖ **Number of possible clusterings**

Let $X=\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$.

Question: In how many ways the N points can be assigned into m groups?

Answer:

$$S(N,m)=\frac{1}{m!}\sum_{i=0}^{m}(-1)^{m-1}\binom{m}{i}i^N$$

➢ Examples:

$$S(15,3)=2\,375\,101$$
$$S(20,4)=45\,232\,115\,901$$
$$S(100,5)=10^{68}\,!!$$

Example 12.1

Assume that $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$. We seek to find all possible clusterings of the elements of X in two clusters. It is easy to deduce that

$$L_2^1 = \{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}\}$$

and

$$L_2^2 = \{\{\boldsymbol{x}_1\}, \{\boldsymbol{x}_2\}\}$$

Taking into account (12.1), we easily find that $S(3, 2) = 2 \times 1 + = 3$. Indeed, the L_3^2 list is

$$L_3^2 = \{\{\boldsymbol{x}_1, \boldsymbol{x}_3\}, \{\boldsymbol{x}_2\}\}, \{\{\boldsymbol{x}_1\}, \{\boldsymbol{x}_2, \boldsymbol{x}_3\}\}, \{\{\boldsymbol{x}_1, \boldsymbol{x}_2\}, \{\boldsymbol{x}_3\}\}$$

Especially for $m = 2$, (12.2) becomes

$$S(N, 2) = 2^{N-1} - 1 \tag{12.3}$$

(see Problem 12.1). Some numerical values of (12.2) are [Spat 80]

- $S(15, 3) = 2375101$
- $S(20, 4) = 45232115901$
- $S(25, 8) = 690223721118368580$
- $S(100, 5) \simeq 10^{68}$

❖ A way out:

➢ Consider only a small fraction of clusterings of X and select a "sensible" clustering among them

- Question 1: Which fraction of clusterings is considered?
- Question 2: What "sensible" means?

- The answer depends on (1) the specific clustering algorithm and (2) the specific criteria to be adopted.

خوشه‌بندی: الگوریتم‌های ترتیبی

۲

دسته‌بندی الگوریتم‌های خوشه‌بندی

MAJOR CATEGORIES OF CLUSTERING ALGORITHMS

- Sequential:

  A single clustering is produced.

  One or few sequential passes on the data.

- Hierarchical:

  A sequence of (nested) clusterings is produced.

  - Agglomerative
    - Matrix theory
    - Graph theory
  - Divisive
  - Combinations of the above (e.g., the Chameleon algorithm.)

❖ Cost function optimization.

For most of the cases a single clustering is obtained.

- ➢ Hard clustering (each point belongs exclusively to a single cluster):
  - Basic hard clustering algorithms (e.g., *k*-means)
  - *k*-medoids algorithms
  - Mixture decomposition
  - Branch and bound
  - Simulated annealing
  - Deterministic annealing
  - Boundary detection
  - Mode seeking
  - Genetic clustering algorithms
- ➢ Fuzzy clustering
  (each point belongs to more than one clusters simultaneously).
- ➢ Possibilistic clustering
  (it is based on the *possibility* of a point to belong to a cluster).

- Other schemes:
  - Algorithms based on graph theory (e.g., Minimum Spanning Tree, regions of influence, directed trees).
  - Competitive learning algorithms (basic competitive learning scheme, Kohonen self-organizing maps).
  - Subspace clustering algorithms.
  - Binary morphology clustering algorithms.
  - …

خوشه‌بندی: الگوریتم‌های ترتیبی

۳

الگوریتم‌های خوشه‌بندی ترتیبی

SEQUENTIAL CLUSTERING ALGORITHMS

❖ The common traits shared by these algorithms are:

- One or very few passes on the data are required.
- The number of clusters is not known a-priori, except (possibly) an upper bound, q.
- The clusters are defined with the aid of
  - An appropriately defined distance $d(\underline{x}, C)$ of a point from a cluster.
  - A threshold Θ associated with the distance.

➢ Basic Sequential Clustering Algorithm (BSAS)

- $m = 1$ \{number of clusters\}
- $C_m = \{\underline{x}_1\}$
- **For** $i = 2$ **to** N
  - Find $C_k : d(\underline{x}_i, C_k) = \min_{1 \le j \le m} d(\underline{x}_i, C_j)$
  - **If** $(d(\underline{x}_i, C_k) > \Theta)$ **AND** $(m < q)$ **then**
    - $m = m + 1$
    - $C_m = \{\underline{x}_i\}$
  - **Else**
    - $C_k = C_k \cup \{\underline{x}_i\}$
    - *W*here necessary, update representatives (*)
  - **End** {if}
- **End** {for}

(*) When the mean vector $\underline{m}_C$ is used as representative of the cluster C with n_c elements, the updating in the light of a new vector $\underline{x}$ becomes

$$\underline{m}_C^{new} = (n_C \, \underline{m}_C + \underline{x}) / (n_C + 1)$$

➢ Remarks:

- The order of presentation of the data in the algorithm plays important role in the clustering results. Different order of presentation may lead to totally different clustering results, in terms of the number of clusters as well as the clusters themselves.

- In BSAS the decision for a vector $\underline{x}$ is reached prior to the final cluster formation.

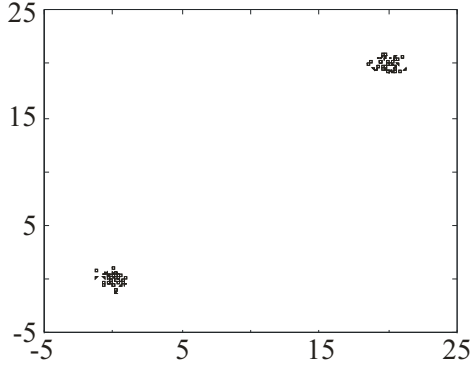
- BSAS perform a single pass on the data. Its complexity is $O(N)$.

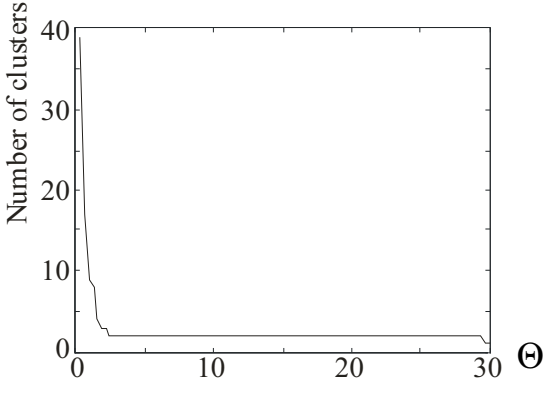
- If clusters are represented by point representatives, compact clusters are favored.

➢ Estimating the number of clusters in the data set:

Let *BSAS*(Θ) denote the *BSAS* algorithm when the dissimilarity threshold is Θ.

- **For** Θ = *a* **to** *b* **step** *c*
  - Run *s* times *BSAS* (Θ), each time presenting the data in a different order.
  - Estimate the number of clusters m_Θ, as the most frequent number resulting from the *s* runs of *BSAS* (Θ).
- Next Θ
- Plot m_Θ versus Θ and identify the number of clusters *m* as the one corresponding to the widest flat region in the above graph.





➢ MBSAS, a modification of BSAS

In BSAS a decision for a data vector $\underline{x}$ is reached prior to the final cluster formation, which is determined after all vectors have been presented to the algorithm.

- MBSAS deals with the above drawback, at the cost of presenting the data twice to the algorithm.
- MBSAS consists of:
  - A **cluster determination phase** (first pass on the data),
    which is the same as BSAS with the exception that no vector is assigned to an already formed cluster. At the end of this phase, each cluster consists of a single element.
  - A **pattern classification phase** (second pass on the data),
    where each one of the unassigned vector is assigned to its closest cluster.

➢ Remarks:

- In MBSAS, a decision for a vector $\underline{x}$ during the pattern classification phase is reached taking into account all clusters.
- MBSAS is sensitive to the order of presentation of the vectors.
- MBSAS requires two passes on the data. Its complexity is $O(N)$.

Modified Basic Sequential Algorithmic Scheme (MBSAS)

- *Cluster Determination*
- $m = 1$
- $C_m = \{\boldsymbol{x}_1\}$
  - For $i = 2$ to N
  - Find C_k: $d(\boldsymbol{x}_i, C_k) = \min_{1 \le j \le m} d(\boldsymbol{x}_i, C_j)$.
  - If $(d(\boldsymbol{x}_i, C_k) > \Theta)$ AND $(m < q)$ then
    - $m = m + 1$
    - $C_m = \{\boldsymbol{x}_i\}$
  - End {if}
- End {For}

Pattern Classification

- For $i = 1$ to N
  - If $\boldsymbol{x}_i$ has not been assigned to a cluster, then
    - Find C_k: $d(\boldsymbol{x}_i, C_k) = \min_{1 \le j \le m} d(\boldsymbol{x}_i, C_j)$
    - $C_k = C_k \cup \{\boldsymbol{x}_i\}$
    - Where necessary, update representatives
  - End {if}
- End {For}

➢ The MaxMin algorithm

Let W be the set of all points that have been chosen to form clusters up to the current iteration step. The formation of clusters is carried out as follows:

- **For each** $\underline{x} \in X - W$ determine $d_{\underline{x}} = \min_{\underline{z} \in W} d(\underline{x}, \underline{z})$
- Determine $\underline{y}$: $d_{\underline{y}} = \max_{\underline{x} \in X - W} d_{\underline{x}}$
- **If** $d_{\underline{y}}$ is greater than a prespecified threshold **then**
  - this vector forms a new cluster
- **else**
  - the cluster determination phase of the algorithm terminates.
- **End** {if}

After the formation of the clusters, each unassigned vector is assigned to its closest cluster.

➢Remarks:

- The MaxMin algorithm is more computationally demanding than MBSAS.
- However, it is expected to produce better clustering results.

❖ Refinement stages

The problem of closeness of clusters: "*In all the above algorithms it may happen that two formed clusters lie very close to each other*".

➢ A simple merging procedure

- (*A*) Find C_i, C_j $(i < j)$ such that $d(C_i,C_j) = \min_{k,r = 1,\ldots,m,\ k \neq r} d(C_k,C_r)$
- **If** $d(C_i,C_j) \leq M_1$ then \{ M_1 is a user-defined threshold \}
  - Merge C_i, C_j to C_i and eliminate C_j.
  - If necessary, update the cluster representative of C_i .
  - Rename the clusters C_{j+1}, …, C_m to C_j, …, C_{m-1}, respectively.
  - $m = m - 1$
  - Go to (*A*)
- **Else**
  - Stop
- **End** {if}

❖ The problem of sensitivity to the order of data presentation:

"*A vector* $\underline{x}$ *may have been assigned to a cluster* C_i *at the current stage but another cluster* C_j *may be formed at a later stage that lies closer to* $\underline{x}$"

➢ A simple reassignment procedure

- **For** $i = 1$ **to** N
  - Find C_j such that $d(\underline{x}_i, C_j) = \min_{k=1,\ldots,m} d(\underline{x}_i, C_k)$
  - Set $b(i) = j$ \{ *b(i)* is the index of the cluster that lies closest to $\underline{x}_i$ \}
- **End** {for}
- **For** $j = 1$ **to** m
  - Set $C_j = \{\underline{x}_i \in X : b(i) = j\}$
  - If necessary, update representatives
- **End** {for}

❖ A two-threshold sequential scheme (TTSAS)

- The formation of the clusters, as well as the assignment of vectors to clusters, is carried out concurrently (like BSAS and unlike MBSAS)
- Two thresholds Θ_1 and Θ_2 ($\Theta_1 < \Theta_2$) are employed
- The general idea is the following:
  - **If** the distance $d(\underline{x},C)$ of $\underline{x}$ from its closest cluster, C, is greater than Θ_2 **then**:
    - A new cluster represented by $\underline{x}$ is formed.
  - **Else** if $d(\underline{x},C) < \Theta_1$ **then**
    - $\underline{x}$ is assigned to C.
  - **Else**
    - The decision is postponed to a later stage.
  - **End** {if}

The unassigned vectors are presented iteratively to the algorithm until all of them are classified.

➢ Remarks:

- In practice, a few passes (≥2) of the data set are required.
- TTSAS is less sensitive to the order of data presentation, compared to BSAS.

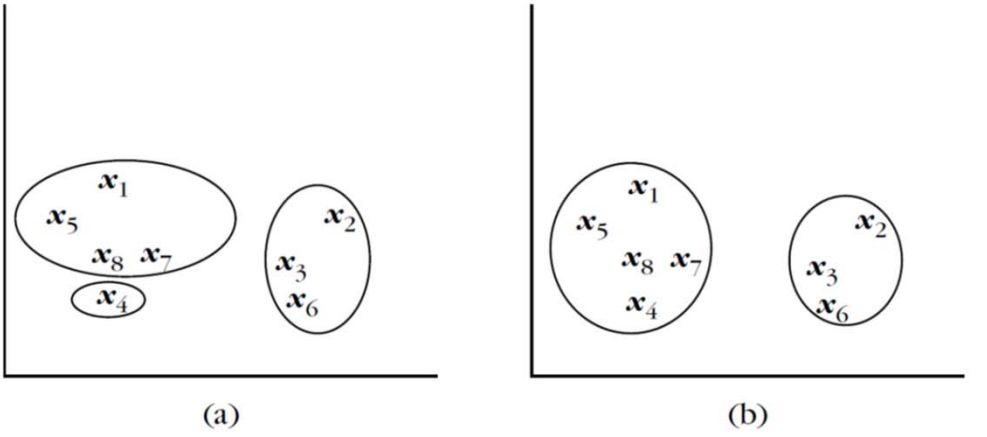


FIGURE 12.3

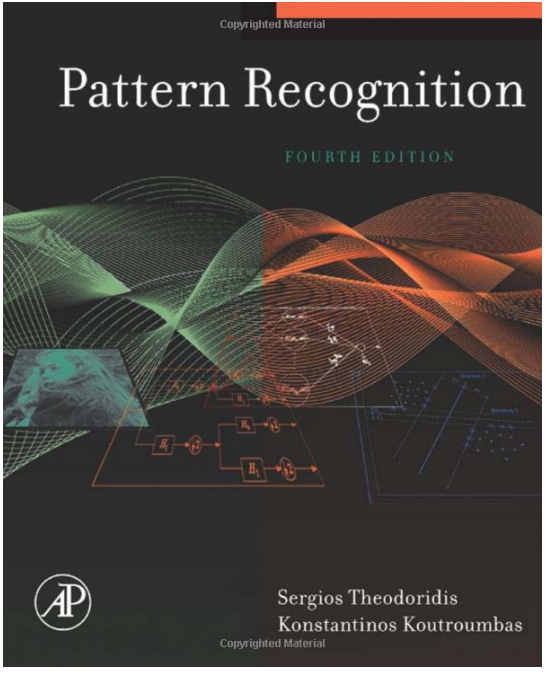
(a) The clustering produced by the MBSAS. (b) The clustering produced by the TTSAS.

خوشه‌بندی: الگوریتم‌های ترتیبی

۴

منابع

منبع اصلی



S. Theodoridis, K. Koutroumbas,
Pattern Recognition,
Fourth Edition, Academic Press, 2009.

Chapter 12

CHAPTER

Clustering Algorithms I: Sequential Algorithms 12

12.1 INTRODUCTION

In the previous chapter, our major focus was on introducing a number of proximity measures. Each of these measures gives a different interpretation of the terms *similar* and *dissimilar*, associated with the types of clusters that our clustering procedure has to reveal. In the current and the following three chapters, the emphasis is on the various clustering algorithmic schemes and criteria that are available to the analyst. As has already been stated, different combinations of a proximity measure and a clustering scheme will lead to different results, which the expert has to interpret.

This chapter begins with a general overview of the various clustering algorithmic schemes and then focuses on one category, known as sequential algorithms.

12.1.1 Number of Possible Clusterings

Given the time and resources, the best way to assign the feature vectors $\boldsymbol{x}_i$, $i = 1, \ldots, N$, of a set X to clusters would be to identify all possible partitions and to select the most sensible one according to a preselected criterion. However, this is not possible even for moderate values of N. Indeed, let $S(N, m)$ denote the number of all possible clusterings of N vectors into m groups. Remember that, by definition, no cluster is empty. It is clear that the following conditions hold [Spat 80, Jain 88]:

- $S(N, 1) = 1$
- $S(N, N) = 1$
- $S(N, m) = 0, \quad \text{for } m > N$

Let L_{N-1}^{k} be the list containing all possible clusterings of the $N-1$ vectors into k clusters, for $k = m, m-1$. The Nth vector

- Either will be added to one of the clusters of any member of L_{N-1}^{m}
- Or will form a new cluster to each member of L_{N-1}^{m-1}

627