Pattern Recognition Lecture
“Clustering: Sequential Algorithms”

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
www.pr.informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen, Germany

Introduction

Categories of Clustering Algorithms

Sequential Clustering Algorithms

Overview

1. Introduction
2. Categories of Clustering Algorithms
3. Sequential Clustering Algorithms

Overview

Introduction

Categories of Clustering Algorithms

Sequential Clustering Algorithms

1. Introduction
2. Categories of Clustering Algorithms
3. Sequential Clustering Algorithms

Introduction

- Different combinations of a proximity measure and a clustering scheme will lead to different results, which the expert has to interpret.
- This lecture gives an overview of various clustering algorithmic schemes and then focuses on one category, known as sequential algorithms.

Number of Possible Clusterings (1)

- The best way to assign the feature vectors $\mathbf{x}_{i=1,\dots,N}$ of a set X to clusters would be to identify all possible ways to partition the feature space and to select the most sensible one according to a preselection criterion.

- However, this is not possible even for moderate values of N.

Number of Possible Clusterings (2)

- Let $S(N, m)$ denote the number of all possible clusterings of N vectors into m clusters. Remember that, by definition, no cluster is empty. So, $S(N, 1) = 1$, $S(N, N) = 1$, and $S(N, m) = 0$ for $m > N$.
- Let L^k_{N-1} be the list containing all possible clusterings of the $N - 1$ vectors into k clusters, for $k = m, m - 1$. A next vector $\mathbf{x}_N$ will either be added to one of the clusters of any member of L^m_{N-1}, or will form a new cluster to each member of L^{m-1}_{N-1}.

Number of Possible Clusterings (3)

- Thus, we may write

$$S(N,m) = mS(N-1,m) + S(N-1,m-1)$$

- And finally,

$$S(N,m) = \frac{1}{m!}\sum_{i=0}^{m}(-1)^{m-i}\left(\begin{array}{c} m \\ i \end{array}\right) i^N$$

E. g., $S(100,5) \approx 10^{68}$.

- Example for three vectors and two clusters $S(3,2) = 3$.

Overview

1 Introduction

2 Categories of Clustering Algorithms

3 Sequential Clustering Algorithms

Categories of Clustering Algorithms (1)

Sequential Algorithms

These algorithms produce a single clustering. In most of them, the feature vectors are presented to the algorithm once or a few times. The final result is, usually, dependent on the order in which the vectors are presented to the algorithm.

Hierarchical Algorithms

- *Agglomerative Algorithms.* These algorithms produce a sequence of clusterings of decreasing number of clusters at each step. The clustering produced at each step results from the previous one by merging two clusters into one.
- *Divisive Algorithms.* These algorithms act in the opposite direction; that is, they produce a sequence of clusterings of increasing m at each step. The clusterings produced at each step results from the previous by splitting a single cluster into two.

Categories of Clustering Algorithms (2)

Algorithms Based on Cost Function Optimisation

This category contains algorithms in which “sensible” is quantified by a cost function J. Most of these algorithms use differential calculus and produce successive clusterings while trying to optimise J. There are some subcategories:

- *Hard and crisp clustering algorithms*, where a vector belongs exclusively to a specific cluster.
- *Fuzzy clustering algorithms*, where a vector belongs to a specific cluster up to certain degree
- *Probabilistic clustering algorithms* follow the Bayesian classification scheme and each vector $\mathbf{x}$ is assigned to a cluster C_i for which $P(C_i|\mathbf{x})$ is maximum.
- *Possibilistic clustering algorithms* measure the possibility for a feature vector $\mathbf{x}$ to belong to a cluster C_i.
- *Boundary detection algorithms*. Instead of determining the clusters by the feature vectors themselves, these algorithms adjust iteratively the boundaries of the regions where clusters lie.

Categories of Clustering Algorithms (3)

Other Clustering Algorithms

- Branch and bound clustering algorithms
- Genetic clustering algorithms
- Stochastic relaxation methods
- Valley-seeking clustering algorithms
- Competitive learning algorithms
- Algorithms based on morphological transformation techniques
- Density-based algorithms
- Subspace clustering algorithms
- Kernel-based methods
- ...

Overview

1 Introduction

2 Categories of Clustering Algorithms

3 Sequential Clustering Algorithms

BSAS - Assumptions (1)

BSAS - Basic Sequential Algorithmic Scheme

Assumptions

- All vectors are presented to the algorithm only once.
- The number of clusters is not known a priori in this case.

BSAS - Basic Idea

- Let $d(\mathbf{x}, C)$ denote the distance between a feature vector $\mathbf{x}$ and a cluster C.
- The user-defined parameters required by the algorithmic scheme are the threshold of dissimilarity, Θ, and the maximum allowable number of clusters, q.
- The basic idea of the algorithm is the following: each vector is assigned either to an existing cluster or to a newly created cluster, depending on its distance from the already formed ones.

BSAS - Algorithm

Let m be the number of clusters that the algorithm has created until now. Then the algorithmic scheme may be stated as:

```
• $m = 1$
• $C_m = \{\mathbf{x}_1\}$
• For $i = 2$ to $N$
    • Find $C_k : d(\mathbf{x}_i, C_k) = \min_{1 \leq j \leq m} d(\mathbf{x}_i, C_j)$
    • If $(d(\mathbf{x}_i, C_k) > \Theta)$ AND $(m < q)$ then
        • $m = m + 1$
        • $C_m = \{\mathbf{x}_i\}$
    • Else
        • $C_k = C_k \cup \{\mathbf{x}_i\}$
        • Where necessary, update representatives (see next Slide)
    • End {if}
• End {For}
```

Different choices of $d(\mathbf{x}, C)$ lead to different algorithms.

BSAS - Updating Representatives

- When C is represented by a single vector $\mathbf{m}_C$

$$d(\mathbf{x}, C) = d(\mathbf{x}, \mathbf{m}_C)$$

- If the mean vector is used as a representative, the updating may take place in an iterative fashion, that is,

$$\mathbf{m}_{C_k}^{\mathrm{new}} = \frac{(n_{C_k^{\mathrm{new}}} - 1)\mathbf{m}_{C_k}^{\mathrm{old}} + \mathbf{x}}{n_{C_k^{\mathrm{new}}}}$$

  where $n_{C_k^{\mathrm{new}}}$ is the cardinality of C_k after the assignment of $\mathbf{x}$ to it.

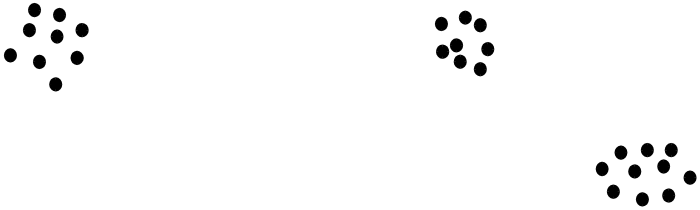
BSAS - Conclusions

- The order in which the vectors are presented to the BSAS plays an important role in the clustering results. Different presentation ordering may lead to totally different clustering results, in terms of the number of clusters as well as the clusters themselves.

- Another important factor affecting the result of the clustering algorithm is the choice of the threshold Θ.

- If the number q of the maximum allowable clusters is not constrained, we leave it to the algorithm to decide about the appropriate number of clusters (see next Slide).

BSAS - Example



Three clusters are formed by the feature vectors. When q is constrained to a value less than 3, the BSAS algorithm will not be able to reveal them.

BSAS - Determining the Number of Clusters

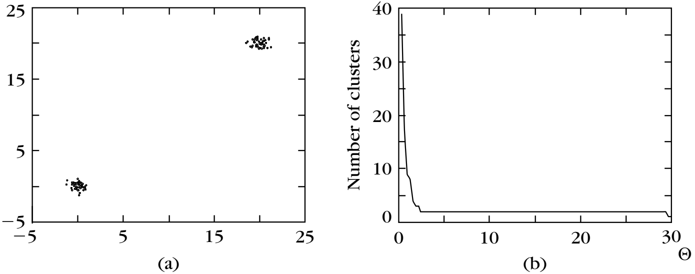
If BSAS(Θ) denotes the BSAS algorithm with a specific threshold of dissimilarity Θ, the algorithm for determining the number of clusters looks like follows:

- For $\Theta = a$ to b step c
    - Run s times the algorithm BSAS(Θ), each time processing 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

The values a and b are the minimum and maximum dissimilarity levels among all pairs of vectors in X, that is

$$a = \min_{i,j=1,\ldots,N} d(\mathbf{x}_i, \mathbf{x}_j) \quad \text{and} \quad b = \max_{i,j=1,\ldots,N} d(\mathbf{x}_i, \mathbf{x}_j)$$

Number of Clusters vs. Threshold



(a) The data set. (b) The plot of the number of clusters versus Θ. It can be seen that for a wide range of values of Θ, the number of clusters, m, is 2.