IHB0090
Clustering
Assignment

Moritz Hangen

2024/04/08

Contents

1 Task

1.1 Chapter 8 of "Introduction to Data Mining"

Study chapter 8, *"Cluster Analysis: Basic Concepts and Algorithms"*, from the book *"Introduction to Data Mining"*. Go deep into K-means clustering; no need to go deep into Agglomerative Hierarchical Clustering, just basic understanding; no need to read 8.4 DBSCAN; from 8.5 Cluster Evaluation, study the overview 8.5.1 and The Silhouette Coefficient.

1.2 Slide Presentation

Together with the book, go through also the slides I prepared - more figures to help with the topic. K-means as the centre of focus is at the end.

1.3 Article Review

Read the article *"The use and reporting of cluster analysis in health psychology: A review"* by Clatworthy et al. Figure out, what type of information and why has to be presented while performing and reporting cluster analysis. How many items you got?

1.4 Example Articles

Example articles (Article 01, 02, 03) are all using clustering. Select one from the three articles and explain why and how clustering is performed. No need to explain the main aim of the publication, just the aim of clustering. Sometimes clustering can be the main aim.

1.5 Clustering Reporting Analysis

Analyze clustering reporting in your selected Article, indicate which items were reported and which method was used. Also which items were not reported - the shortcomings in reporting.

1.6 Submission and Hours

Upload answers to point 5 & 6 to Moodle and indicate the total working hours.

2 Results

2.1 Article Review

Found measures:

- **Computer program:** Common practice. Different software on different computers might lead to varying results.
- **The similarity measure:**Pattern vs. Elevation –>Significant impact on resulting clusters. Based on expected outcome, different methods shall be used.
- **The cluster method:** Agglomerative methods vs. Partitioning methods; Choice will lead to varying results of varying quality based on the method used.
- **The procedures used to determine the number of groups in the data:** Selection of number of clusters should be based on factual criteria or a standardized method. Methods may be subjective, as prior expectations may influence the decided number of clusters.
- **The validation procedures:** There is the possibility, that clusters may be found where there are none (e.g. homogenous data). Thus, it has to be shown that the clusters are stable (clusters created from sample data roughly represent the real data) and provide value to the research.

2.2 Example Articles

For analysis the first article was selected.
Main purpose why clustering was used: To analyze functional connectivity (FC) and cluster it into two (optimal amount of clusters according to cluster analysis) groups:

- Frequent and sparsely connected state
- More frequent and less interconnected state

How clustering was performed:

- Data collection (three-dimensional brain volume imaging)
- Data preprocessing (Sliding window approach)
- K-Means clustering (K = 2)
- Verification

2.3 Clustering Reporting Analysis

In the first article, these are the results for the reporting criteria:

- **Computer program:** The Data Processing & Analysis for Brain Imaging toolbox; Statistical Parametric Mapping software.
- **The similarity measure:** Manhattan distance.
- **The cluster method:** No information of agglomerative or partitioning. Shortcoming in reporting.
- **The procedures used to determine the number of groups in the data:** Similarity of FC matrices between time windows: 2 clusters.
- **The validation procedures:** Repeating clustering analysis through correlation distance function. This confirmed the use of 2 clusters.

2.4 Submission and Hours

Total time: 4h.