Lab 6: Clustering - demo + Case study

08.05.2019 / prepared together with M.Deckert

Laboratories are aimed at achieving practical experiences with using clustering algorithms.

The lab include two parts: practical experiences with basic software tools and than carrying out more advanced case study with a new data set - which is identified together with the instructor.

Part 1 - practical experiences with basic algorithms and software tools

1. Use K-Means algorithm to group the following points:

Point	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

2. Create an arff file from the set of learning examples given in Exercise 1.
3. Run SimpleKMeans clustering algorithm in WEKA software. Compare the result with the one from Exercise 1.
4. A mannual calculation example – Try to group the following numbers using the Single Linkage and Complete Linkage method: 2,5,9,15,16,18,25,33,33,45 (for simplicity and easier calculation- it is one attribute only).
5. Test AHC algorithm on CARS data in Statistica software.
   - Open CARS.txt file in STATISTICA software.
   - Set 'Nazwy przypadków' (ang. Cases names) with 'Menedżer nazw przypadków' (ang. Cases names menager) and set 'Zmienna:' (ang. Variable) to the name of first variable containing names of the cars. Next, delete the column with the variable containing names of the cars.
   - Go to 'Statystyka'(ang. Statistics) tab.
   - Expand 'Wielowymiarowe'(ang. Multidimensional) and choose 'Analiza skupień' (ang. Clustering). Next, choose 'Aglomeracja' (ang. agglomeration).
   - Choose 'Zmienne' (ang. Variables) and select all 5 numerical variables.
   - On 'Więcej' (ang. More) tab change 'Grupuj' (ang. group) to 'Przypadki (wiersze)' (ang. cases in rows). You can also change agglomeration method ('Metoda aglomeracji:) and distance metric ('Miara odległości'), which will be used. To run the analysis, press 'Ok' button.
   - To see the result of agglomeration as a dendrogram choose 'Poziomy, hierarchiczny wykres drzewa' (ang. Horizontal, hierarchical dendrogram). More results can be found on a 'Więcej' (ang. More) tab.

Part 2 - a case study with a new data set

Gain experiences to become a data analysts with the following objectives:

- Try to discover a meaningful structure in your data and provide its description + interpretation
- Show that you can use appropriately at least two clustering algorithms
- Find an interesting data problem:
    1. Acquire the good data set
    2. Preprocess data (transform from the raw format into an appropriate file format)
    3. Decide on the attribute space (if necessary, you should try to select the most relevant attributes)
- Apply clustering algorithms to these data and tune their parameters
- Interpret results and provide a characteristic description of clusters (based on their representatives such as centroids)

More details are given in the attached pdf presentations.