UNIVERSITÄT MANNHEIM

Data and Web Science Group
Prof. Dr. Heiko Paulheim
B6, 26 – B022
68159 Mannheim

Data Mining I

Exercise 2: Cluster Analysis

2.1. Analyzing the Customer Data Set

1. Import the *customers* data set into RapidMiner. The customer data set is provided on the website as an Excel file.
2. Cluster the dataset using K-Means clustering (Using the `k-Means` operator). Experiment with different K values. Which values do make sense? What does the clustering tell you concerning your product portfolio? What does the clustering tell you concerning your marketing efforts in different regions?
3. Cluster the data set using Agglomerative Hierarchical Clustering (Using the `Agglomerative Clustering` operator). What does the dendrogram tell you concerning your customer groups?
4. Flatten the hierarchical clustering so that you get 3 or 4 customer groups. Name these groups with appropriate labels.

2.2. Analyzing the Students Data Set

1. Aggregate the *students* data set (from Exercise 1) by student and calculate the average mark and the average number of attended classes.
2. Cluster the data set using the K-Means algorithm.
   Does one attribute dominate the clustering? What can you do about this? Assign suitable labels to your clusters.
3. Cluster the data set using Agglomerative Hierarchical Clustering. Experiment with different settings for calculating the cluster similarity. What is a good setting?
4. What does the dendrogram tell you about the distances between the different groups of students?

2.3. Clustering the Iris Data Set

1. Cluster the Iris data set (from the Sample Repository of your RapidMiner installation) using different algorithms and parameter settings.
2. Does it make sense to normalize the data before applying the algorithms?
3. Try to choose an algorithm and parameter setting that reproduces the original division into the three different species.

2.4 Clustering the Geo Data Set

1. Within the geo data set (provided on the website) the coordinates (x & y) of housings of inhabitants of an area are collected. Have a look at the data and visualize it using the Plot View
2. Cluster the data using k-Means (k=3). Do the clusters represent the original areas?
3. Apply DBSCAN and play around with the epsilon. Can you reproduce the original areas using this cluster algorithm?

2.5. Clustering the Zoo Data Set

1. The Z*oo* data set describes 101 animals using 18 different attributes. The data set is provided on the website as an ARFF file. Import the *Zoo* data set into your local repository using the operators “Read ARFF” and “Store”.
2. Cluster the data set using Agglomerative Hierarchical Clustering. Experiment with different parameter settings in order to generate a nice species tree.
3. Can you assign some appropriate species names to the clusters at the upper levels?