Data Mining Fall 2017

Project 4: Cluster Analysis

Due: see Canvas
Points: 100

Please submit your report in PDF format and code (if necessary) in a separate file.

Introduction

We will work one last time with federal employment data.

Write a report covering in detail all steps of the project. The results have to be reproducible using your report. Carefully describe every assumption and every step in your report. Also, mention any program/code/additional data that you are using for your analysis.

Follow the CRISP-DM Framework for your Report

3. Data Preparation [30 points]

- Describe which features you want to use for clustering and why. [20]
- What is the scale of measurement of the features and what are appropriate distance measures? [10]

4. Modeling [50 points]

- Perform cluster analysis using several methods (at least k-means and hierarchical clustering) using different feature subsets. [30]
- How did you determine a suitable number of clusters for each method? [10]
- Use internal validation measures to describe and compare the clusterings and the clusters (some visual methods would be good). [10]
- Can you use a feature as the ground truth (e.g., the continent or the rank) and perform external validation? [exceptional work]

5. Evaluation [10 points]

- Describe your results. What findings are the most interesting?

Exceptional Work [10 points]