Data Analytics 874
Post-Block Assignment 1: Data Clustering and Cluster Analysis
Department of Industrial Engineering

Deadline: 21 December 2020, 23:59

Instructions

The focus of this assignment is on data clustering and cluster analysis. For the purposes of this assignment, note the following instructions:

- You may select any of the options provided in this document.
- Assignments are to be completed by students individually.
- Submit your own work.
- You will submit a pdf document addressing all of the aspects listed in the specifications below. Name your pdf document `????????PBA1_874.pdf`, where the question marks are replaced with your student number.
- Make sure that your pdf document has a title page containing the assignment number (and an indication of the selected option) as the title, and your initials, surname and student number.
- Submit your pdf document no later than the deadline.

1 Option 1: Structured Data Clustering and Cluster Analysis

We will continue with our analysis of the abalone data set. However, this time, you are provided with a cleaned data set, `abaloneCleaned.xlsx`, which you can download from Sunlearn. The input features are as follows:

- Gender
- Length
- Diameter
- Height
- Whole weight
- Shucked weight

- Viscera weight
- Shell weight

The target feature is named Rings, and refers to the number of rings found in the shell. The number of rings indicates the age of the abalone.

Note that the data set has only the data quality issues corrected. No data transformations have been done to get the data set ready for data clustering.

1.1 Does Clustering Make Sense?

Your first step is to determine if clustering of this dataset makes sense. In other words, you need to illustrate if the dataset does contain underlying clusters, or if the data provided is simply random points. In your pdf file, under an appropriate heading, explain what you have done to find an answer to this question, provide an answer to this question, and substantiate that answer with evidence. (10)

1.2 Data Clustering

You have to decide which data clustering algorithm you will use to cluster the abalone data. Your report has to include sections where you address the following aspects:

- Clustering algorithm: Give a short description of the algorithm selected with a short motivation of why you have selected this algorithm. (5)
- Optimal number of clusters: Determine the optimal number of clusters. You have to describe the process used to determine the optimal number of clusters, and provide empirical evidence in support of your decision. (10)
- Features used: Provide a list of features used, and provide motivations for exclution of any feature. (5)
- Data transformation: Discuss any data transformations that you have applied to the data prior to clustering. (10)
- Cluster descriptive statistics: Provide descriptive statistics for the different clusters. In addition also give the custer centroids and the number of instances that belong to each cluster. Then, also determine the class (i.e. number of rings) that occur the most in each cluster. (10)

1.3 Cluster Visualization

For this section, you are going to do a number of different visualizations of the clusters, as follows:

- Provide a cluster plot to visualize the clusters formed. (5)
- Give a SPLOM to illustrate the clusters for all pairs of descriptive features. (10)
- Draw descriptive feature histograms with reference to each of the target levels (number of rings). (10)
- Draw a parallel coordinate plot to illustrate the formed clusters. (5)

1.4 Cluster Analysis

For this section, you are going to explore as much as possible about the different abalone age groups. Provide the following:

- From the cluster descriptive statistics, see if you can identify any rules to discern among the different clusters (age groups). (10)
- From the SPLOM that you have provided in the previous section, identify:
  - the descriptive features that result in good separation of clusters (5)
  - for those clusters that can be well separated, try to identify rules to characterise the clusters (10)
- With reference to the histograms, which features do you believe are the most important in classifying abalone? Provide motivations for your answers. (10)
- Show how the parallel coordinate plot that you have drawn support your answer above on the most important features. Also, from this plot, indicate any features that can possibly be removed. Provide motivations. (10)

2 Option 2: Clustering Non-Stationary Data

The focus of this option is on clustering of non-stationary data, i.e. data where observations change over time, or data streams where data arrives over time. For this assignment, you will have to do additional research to explore such data clustering approaches, and you will have to implement and compare two such approaches on one dataset.

To help you with the research part, find uploaded to SUNlearn the following documents:

- L Rokach, *A Survey of Clustering Algorithms*, uploaded as dataMiningHandbook.pdf under the reading material folder for topic 2 on SUNlearn.
- K Georgieva, *A Computational Intelligence Approach to Clustering of Temporal Data*, uploaded as Georgieva.pdf under the reading material folder for post-block assignment 1.
- AJ Graaff, AP Engelbrecht, *Clustering Data in An Uncertain Environment using Artificial Immune System*, uploaded as GraaffEngelbrecht.pdf under the reading material folder for post-block assignment 1.
- NB Roa, L Travé-Massuyès, VH Grisoles-Palacio, *DyClee: Dynamic clustering for tracking evolving environments*, uploaded as Rao_etal.pdf under the reading material folder for post-block assignment 1.
- A King, *Online k-Means Clustering of Nonstationary Data*, uploaded as King.pdf under the reading material folder for post-block assignment 1.

You have to submit are report wherein you have to address the following:

- Discuss the issues for clustering caused by non-stationary data. (20)
- Provide a review of approaches to non-stationary data clustering and how the issues above are addressed. (20)
- In more detail, describe the two approaches that you have selected. (15)
- Describe the data set used, and any transformations that have been done. Also explain how the dataset wlll help to exlore the issues listed above. (20)
- Provide empirical results to compare the performance of the two clustering approaches. First provide the empirical process that you have followed, to include the performance measures used. Then provide and discuss the results. Which of the two approached do you consider to be the best approach? (50)

3 Option 3: Time Series Clustering

The focus of this topic is on time series clustering. You will have to explore literature on time series clustering, to determine different approaches to time series clustering. You will then select one time series clustering algorithm, and will evaluate its effectiveness in clustering financial time series.

To help you with the research part, find uploaded to SUNlearn the following documents:

- L Rokach, *A Survey of Clustering Algorithms*, uploaded as dataMiningHandbook.pdf under the reading material folder for topic 2 on SUNlearn.
- X Wang, KA Smith, R Hyndman, D Alahakoon, *A Scalable Method for Time Series Clustering*, uploaded as WangSmithHyndmanAlahakoon.pdf under the reading material folder for post-block assignment 1.
- J Paparrizos, L Gravano, *Fast and Accurate Time-Series Clustering*, uploaded as PaparrizosGravano.pdf under the reading material folder for post-block assignment 1.
- P Roelofsen, *Time Series Clustering*, uploaded as Roelofsen.pdf under the reading material folder for post-block assignment 1.
- S Aghabozorgi, AS Shirkhorshidi, TY Wang, *Time-Series Clustering - A Decade Overview*, uploaded as AghabozorgiShirkhorshidi.pdf under the reading material folder for post-block assignment 1.
- TW Liao, *Clustering of Time-Series Data - A Survey*, uploaded as Liao.pdf under the reading material folder for post-block assignment 1.

You have to submit are report wherein you have to address the following:

- Discuss different approaches to time series clustering. (20)
- Select a time series clustering algorithm for clustering financial time series. Motivate your choice, and describe the algorithm in detail. (15)
- Describe the time series used, and any data transformations applied to the time series. (20)
- Describe your empirical process, to also include the performance measures that you have used. (20)
- Present and discuss the results for the chosen clustering algorithm, and provide a means to validate the clustering results. (50)