Homework

Name: Aris Podotas

University: National and Kapodistrian University of Athens

Program: Data Science and Informaion Technologies

Specialization: Bioinformatics - Biomedical Data

Lesson: Clustering Algorithms

Date: November 2024

Contents

1 Preface

Before any of the further analysis a look at the files provided in the report are all the files that were originally sent with the description along with a file named *main.m* that implements the Matlab code that generates all the solutions and a file named *holder.pdf* that contains this report.

The *main.m* file will also contain comments explaining what goal each function has.

2 Feeling the data

2.1 Missing data

It is stated in the description of the project that: Only the pixels with nonzero class label will be taken into consideration in this project. This will be considered a form of missing data.

How will we handle this missing data? This dataset is adequately large for pruning of missing values to occur. A variable that copies the original data and is then filtered for missing data will be made. The missing data needs to be handled first in this analysis since it is explicitly stated that missing values will not be considered and thus any representation of the data before removing these values will be improper.

An interesting note is that the output is 13.908 data vectors. The full data had 22.500 which means that we had missing data for 8.592 feature vectors (zero label missing data).

Is this the only missing data in our sample? Not necessarily, because values for the features could contain ***NaN*** or ***missing***. Once we remove the missing data we know we have (zero labels) we apply a check for other missing data and handle it. Our approach for other missing data should be different than before (zero label) since zero label missing data is explicitly stated for removal, however with other missing data values we can imploy other missing data handling. In this dataset there are no ***NaN, missing*** fields (see appropriate function in the *main.m* file). Should there have been we would know that the missing values would have been within one of the *spectral bands* and we could have handled it with substitution of the *mean* or *median* of the feature that was missing so as to not remove the whole data vector. Both the labels and the values were checked.

2.2 Data type

This segment is about the values our data takes (discreet or continuous). Actually we cannot look at the data itself for, large datasets like the one provided do not get visualized in the variable previewer. The way this will be overcome is by utilizing the knowledge of the dataset. We have a-priori knowledge for the features of the dataset due to the nature of the problem. Each feature takes continuous values within a *spectral band.*

It is generally a good idea to view the distributions of the data and the features. This will not be possible because of visualization issues associated with the number of features.

Actually the problem of representation plagues this dataset because none of the outputs would fit neatly into a format on the page so all outputs are only available withing the varibale browser in Matlab.

2.3 Cross corelation

Representing multiple features with just one that correlates to a high degree would reduce the computational complexity of the data clustering to a great degree.

3 Feature selection/transformation

4 Selection of the clustering algorithm

5 Execution of the algorithms

6 Characterization of the clusters

7 Citations