Hands-on clustering #1:

Multi-spectral image segmentation: MARS

Joris Guérin

October 2021

Abstract

The objective of this hands-on exercise is to use in practice the different concepts studied during this course on clustering. Image segmentation is a standard problem where unsupervised classification is used. Here, we want to apply clustering to the pixels of a multi-spectral image representing the surface of Mars. The goal is to identify different "types" of pixels corresponding to different geological compositions, thus generating a geological map of Mars.

1 Data

Files For this study, you will need two data files:

- mars.csv
- mask.csv

Description The data in the "mars.csv" file represent a hyper-spectral image of the surface of Mars. Visible and near-infrared imagery is a key tele-detection technique, used to study planets thanks to satellites with on-board spectrometers. In march 2014, the OMEGA equipment collected more than 310 Go of raw images. It mapped Mars surface with a resolution between 300 and 3000 meters, depending on the height of the space ship (Bibring et al., 2005). For each pixel, the spectral response between 0.36 and 5.2 μm was collected and sampled accross 255 channels. The objective is to characterize the geological composition of Mars surface and in particular to distinguish between different classes of silicates, minerals, oxydes, carbonates and frozen regions.

These data are represented by an image of 300 x 128 pixels. A vector of 255 spectral values (variables, characteristics, features) is associated to each of the 38,400 pixels (instances, individuals).

2 Objectives

According to experts, there are K=5 geological classes to identify on the map. The objective is to operate an automated unsupervised classification of the pixels, in order to segment the image into 5 distinct categories (colors). Before conducting the segmentation, an exploratory approach is conducted to better understand the data and decide how it should be pre-processed. Then, different clustering algorithms are applied and compared on the MARS dataset.

3 Data exploration

3.1 Read the data and understand the formatting

1. Read the "mars.csv" data using the *read_csv* method of the *pandas* library.
2. Display a summary of the data using the *describe* method.

3. How many rows and columns do the data have? What does it mean in terms of number of data points? Dimensionality of the data?

3.2 Data preprocessing

1. Draw the histograms of six different dimensions (wave length) selected at random. They represent the distributions of the values of a given dimension across the different pixels.
   (a) What can you say regarding the discriminative power of single dimensions?
   (b) The symmetry of the data?
2. Draw the box plots of all different wave lengths. They should all appear on the same graph.
   (a) What can you say regarding the spread of the different dimensions?
   (b) The symmetry of the data?
   (c) Outlier values?
3. Should the data be transformed before applying clustering? Why?
4. Conduct the required data transformations/preprocessing. Visualize the new distribution of the transformed data (histograms, box plots), what can you say?

3.3 Dimensionality reduction

1. Using the appropriate scikit-learn package, conduct a Principal Component Analysis decomposition of the reduced MARS data.
2. Visualize the explained variance as a function of the number of dimensions selected. *We can use the pca.explained_variance_ratio_ argument from the scikit learn PCA implementation.*
3. Plot the *Variable factor map* for the first two PCA dimensions. What can you say?
4. Plot the *Individuals factor map* for the first two PCA dimensions. What can you say?
5. How many dimensions should be used to conduct a cluster analysis?

4 Clustering

4.1 K-means

1. How can we choose the number of clusters K to apply K-means on these data?
2. Conduct one of the two methods? What K should be chosen?
3. In practice, the experts tell us that the number of different geological compositions in this image is 5. Apply K-means on the selected PCA dimensions, with this value of K.
4. Display the clusters in the PCA *Individual factor map.*
5. Display the 300x128 image where the pixels represent the clusters. We can use the *reshape* function.
6. Plot the curves representing the values of the wave lengths of the cluster centers.
7. Repeat the process using the complete data instead of the data reduced with PCA. Compute and display the confusion matrix of the two classification. Compare the two results using external metrics (normalized mutual information, Fowlkes Mallows). Are the clusters obtained similar?

4.2 Agglomerative Clustering

1. What is a good number of clusters for Agglomerative Clustering? *We can use only the PCA reduced data and consider only a random subset of the data to decrease computation time*

2. Apply Agglomerative Clustering on the MARS dataset, visualize the results.

4.3 Gaussian Mixture Models

1. What is a good number of clusters and a good covariance matrix constraint to train a Gaussian Mixture Model on this data? *We can use only the PCA reduced data and consider only a random subset of the data to decrease computation time*

2. Apply GMM on the MARS dataset, visualize the results.

3. Repeat the above procedure but with the number of clusters fixed to 5.

5 Comparison of clustering algorithms

A classification of the pixels was provided by domain experts and can be found in the *mask.csv* file. Although it might not be true, here, we consider the colors in this file to be ground truth.

1. Visualize this ground truth file on the 2D PCA projection. Is it different from the different clusterings obtained? Why?

2. Using external validation metrics and confusion matrix, evaluate which clustering algorithms, parameters, data preprocessing, etc. gives the best clustering results. *In particular, we can study the influence of using PCA, the number of classes, the linkage methods, the covariance constraint, etc.*

3. When conducting an automated procedure to select the optimal number of clusters, some found that 5 is not always the best choice with respect to the approaches presented in this course. Can we improve the results by using a different number of clusters then the one provided by the experts? If yes, try to explain why.

Acknowledgments

This document was adapted from the following github repository: `https://github.com/wikistat/Exploration/blob/master/Mars/ML-Tutorial-Mars.ipynb`.

References

Jean-Pierre Bibring, Yves Langevin, Aline Gendrin, Brigitte Gondet, François Poulet, Michel Berthé, Alain Soufflot, Ray Arvidson, Nicolas Mangold, John Mustard, et al. Mars surface diversity as revealed by the omega/mars express observations. *Science*, 307(5715):1576–1581, 2005.