Entropy Based Fuzzy Clustering

Prepared for

Dr. Farid

Eng. Rawan Abubakr Bakry

Eng. Rana Mohamed

Authors

20200270 Diaa Eldiyn Essam Zaki

20200468 Mohamed Kamal Elsayed

24^{th} Dec 2023

Introduction:

In this report, we utilized PCA for dimensionality reduction on an initial high-dimensional dataset. By removing extraneous dimensions from the data, PCA enabled more efficient fuzzy clustering in the reduced feature space. We applied an entropy-based fuzzy clustering algorithm to partition the dimensionally reduced data. Evaluating the effectiveness of PCA preprocessing along with comparisons to traditional fuzzy clustering methods are also presented. The combination of these two unsupervised techniques provides an exploratory workflow for analyzing the clustering tendencies and intrinsic structure within complex, real-world datasets.

Objectives:

- Search about one unlabeled dataset
- Apply PCA technique from scratch on it.
- Using different Q Matrix and selecting the best one that do not change the original feature vectors too much and at the same time reduce the dimensionality.
- Report your trails and comment on the result you got from applying PCA.
- Apply the unsupervised learning technique you chose on your dataset before and after applying PCA with the best Q matrix.
- Update your report with your comment on the effect of using PCA.

Results:

- Using PCA with 2 dimensions helps to make data more separable compared to using the most important feature which is sepal width and sepal height as shown in the picture.
- Still some intersection of data which result from original plot between versicolor and virginica.

- Effect of Using PCA:
  - Clustering the original data led to the formation of two clusters, which aligns with logical expectations. This is due to the close proximity of the versicolor and virginica classes, as opposed to the setosa class.

- Utilizing PCA with the optimal Q matrix containing two eigenvectors maintained two clusters observed in the original data case. However, the centroids and outlier instances differed from those in the original dataset.