Spectral Clustering

Dror Ayalon (dda290)

CUSP-GX-5006 Machine Learning for Cities (NYU)
Assigment # 4

Abstract

This report show a comparison between three clustering techniques: Spectral Clustering, K-means, and Hierarchical Clustering. Given the dataset that was used for this report, the most accurate techniques were found to be Spectral Clustering (when using 'lobpcg' as the eigen solver, and considering 25 neighbors when constructing the affinity matrix), and Hierarchical Clustering (when computing the linkage using the 'cosine' method, and using an 'average' as a linkage cretirion).

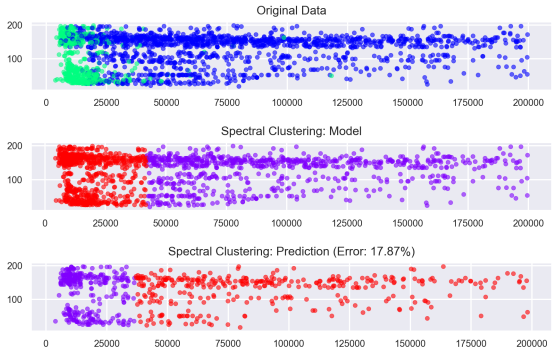

Figure 1: Raw data VS. model genrated by a Spectral Clustering algorithm VS. the most accurate prediction by the Spectral Clustering algorithm

1 Methods and Data Sets

1. The dataset used for this study is 'manhattan-dof.csv', which was made available to us by NYU. The dataset includes 2645 samples. The attributes that were used from this dataset are the following:
    - BldClassif - Building class. Used as a cluster indicator for validation purposes.
    - GrossSqFt, MarketValueperSqFt - Indipendent variables that were used to generate the prediction model.
    - I seems that the data that was very hard to cluster, and in most cases, the model that was generated by the algoritms, which are being described below, was not accurate in describing the raw data.
2. The data was cleaned to remove outlies.
3. The entire study was done using Python3 and the machine learning Python package scikit-learn (http://scikit-learn.org/).
4. To validate the results, a cross-validation process was used, based on the 'train_test_split' method of the scikit-learn package. During the process, 5 batches of data were generated. 4 of which were used as training sets and 1 was used as a validation set. The random split process was done 50 times for each classfication method. Moreover, a Bootstraping pro-

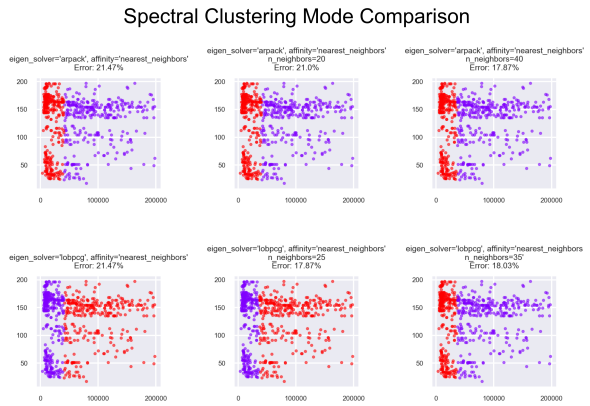

Figure 2: A comparison between different configurations of the Spectral Clusteing algorithm.

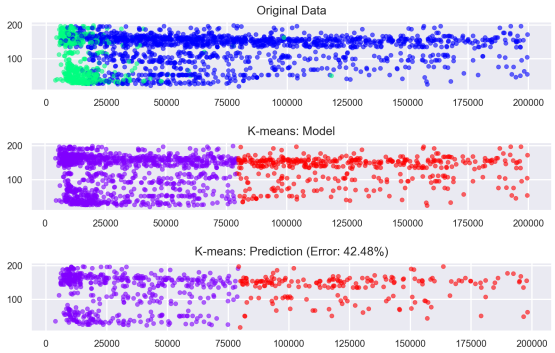

Figure 3: Raw data VS. model genrated by a K-means Clustering algorithm VS. the most accurate prediction by the K-means Clustering algorithm

cedure was used to improve the restuls of a few of the classification algorithms. This topic will be discussed below.

5. The following scikit-learn algorithms were used to the generate the results for this study:

   - sklearn.cluster.SpectralClustering - Was used to cluster the data using a Spectral Clustering algorithm. This algorithm, using when using 'lobpcg' as the eigen solver and considering 25 neighbors when constructing the affinity matrix, was found to be effective an relatively accurate (comparing to other results) in modeling the clusters in the raw data, and therefore, the prediction model was relatively accurate (see figure 1).
   - sklearn.cluster.KMeans - Was used to cluster the data using a K-means algorithm. This algorithm was performing poorly in generating an accurate model of the raw data (probably because of the randomality of the data). Therefore, even-though the prediction is relatively accurate comparing to the model, it is still far away from the raw data (see figure 5).
   - sklearn.cluster.AgglomerativeClustering - Was used to cluster the data using an Hierarchical algorithm. Even-though this algorithm, when computing the linkage using the 'cosine' method and using the 'average' as a linkage cretirion, showed identical results to those of the Spectral Clustering algorithm, it generated a model that is quite different from the one that was generated by the Spectral Algorithm (see figure 2).

2 Results

- The result of running a comparison between all the tested algorithms shows a similarity in accuracy between the Spectral Clustering algorithm and the Hierarchical Clustering algorithm (tied at 17.87% miss clustered samples, see figure 6).
- Since the data that was used was not very easy to cluster, it is easy to assume that different dataset could yield different results.

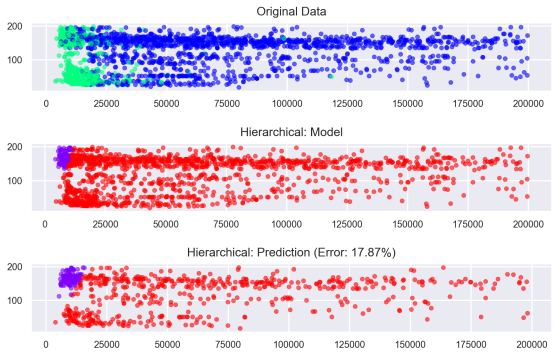

Figure 4: Raw data VS. model genrated by a Hierarchical Clustering algorithm VS. the most accurate prediction by the K-means Clustering algorithm

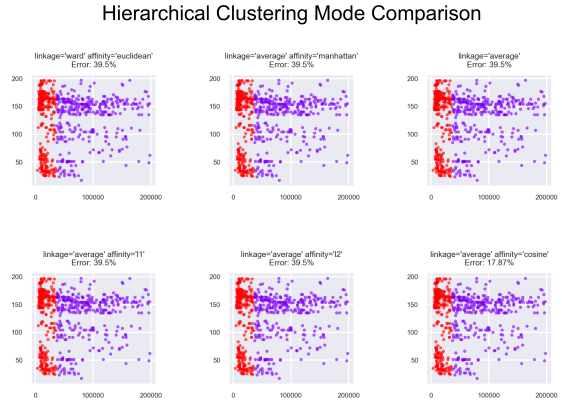

Figure 5: A comparison between different configurations of the Hierarchical Clusteing algorithm.

Spectral Clustering

	Error Rate
Eigen_solver = arpack n_neighbors = 10	21.47%
Eigen_solver = arpack n_neighbors = 20	21.00%
Eigen_solver = arpack n_neighbors = 40	17.87%
Eigen_solver = lobpcg n_neighbors = 10	21.47%
Eigen_solver = lobpcg n_neighbors = 25	17.87%
Eigen_solver = lobpcg n_neighbors = 35	18.03%

Hierarchical Clustering

	Error Rate
linkage = ward affinity = euclidean	39.50%
linkage = average affinity = manhattan	39.50%
linkage = average affinity = euclidean	39.50%
linkage = average affinity = l1	39.50%
linkage = average affinity = l2	39.50%
linkage = average affinity = cosine	17.87%

Figure 6: A comparison between error rates of different clustering methods.