MID-SEMESTER ASSIGNMENT: COMPARISON OF CLUSTERING ALGORITHMS

Akiva Bruno Melka
332629393

ABSTRACT

With the emergence of powerful computation tools and the need to analyze very large data sets, clustering has become a fundamental issue. Clustering consists of organizing data sets into subsets of instances presenting similar characteristics and separate those that are different. As such, several algorithms have been developed, each with its own particularities. Among the many challenges involved in clustering are the data processing and dimension reduction to obtain a more practical data set. The choice of the number of clusters and the most appropriate algorithm is also crucial to the process. Finally, several metrics have been developed to measure the efficiency of the clustering. The purpose of this assignment is to present a comparison between some of those algorithms and draw conclusions. We apply this panel of methods to three distinct data sets.

`https://github.com/Akivamelka/unsupervised_mid_semester`

1 INTRODUCTION

Clustering and classification are both fundamental tasks in Data Mining. Classification is used mostly as a supervised learning method, clustering for unsupervised learning (some clustering models are for both). The goal of clustering is descriptive, that of classification is predictive Since the goal of clustering is to discover a new set of categories, the new groups are of interest in themselves, and their assessment is intrinsic.

The process of clustering requires several sub tasks. The first, and probably most essential, is the pre-processing. Since most data sets are raw, they contain a lot of noise (instances with no real relevancy to the data set). Those instances have to be removed in a systematic way. The features of the data set require also analysing. Some, such as the "id"of the instance, provide absolutely no valuable information for the clustering since each instance has a distinct "id". Other features are extremely correlated and are, therefore, redundant and will not add to the clustering. On the contrary, they will appear as an handicap and weight heavily on the process. To resolve this issue, the data scientist has to study each of them, compute their correlation and decide which ones to keep. Another, more systematic, way to reduce the number of features to the essential ones is Dimension reduction. Finally, features have to be separated into those that are numerical and categorical. We further discuss how to deal with each kind.

Once the data set has been properly "cleaned", begins the clustering. We here present several clustering algorithms, some based on proximity, others based on density. Those algorithms are also described in the Methods section. The main issue is to determine the optimal number of clustering. We will also describe how to estimate this number. After the clustering, we will again use Dimension reduction for the visualization. Finally, we will evaluate the efficiency of each algorithm.

2 METHODS

2.1 DATA PROCESSING

To compare different algorithms, we applied them on three different data sets with different specificity (size, number of features,...). We also had to apply various transformation to the data sets to make them more manageable.

2.1.1 DATA SETS

- **Diabetes.** The first data set we used represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. It includes over 100,000 instances. Detailed description of all the attributes is provided in Table 1 from (1).

  We observed that some features (”weight”, ”medical speciality” and ”payer code”) were missing a lot of data so we removed them. The ”encounter id” was also removed since each instance has a unique number and therefore does not provide any information for the clustering. We also removed ”gender” and ”race” for the clustering and used those feature as classes for the external validation discussed further. Finally, we removed the instances with missing data and all those with the same ”patient nbr” so that the sample would be in independent.
- **Click stream for online shopping.** The second data set contains information on click stream from online store offering clothing for pregnant women. Data are from five months of 2008. The data set contains 14 features described in (2) and over 165,000 instances.

  The ”country” feature was removed to be used as the class for the external validation. We also removed ”year” since all the instances are in 2008 and, therefore, this feature does not provide information for the clustering. We removed ”page 2 (clothing model)” for simplicity purposes.
- **Online Shoppers Purchasing Intention.** The third data set consists of over 12,000 instances over 18 features. Description of the features can be found in (3).

  The features ”Weekend”, ”Revenue”, and ”VisitorType” were removed and used as classes. We observed that the three ”Duration” features were extremely correlated to their attribute so we removed the attribute and kept the duration

2.1.2 NUMERICAL FEATURES AND SCALING

For the numerical features, different approaches are to be considered. Depending on the amplitude of the values and their distributions, scaling can be necessary. For instance, a feature with relatively high value would be regarded during the dimension reduction as influential compared to others since its variance would be high, when, in reality, it is not the case. Therefore, it requires scaling to fit the other features. For instance, we scaled the ”Duration” features in the third data set and many in the first data set. On the other hand we might want, sometimes, to keep the values.

The choice of scaling depends on the data. Since our data was not normally distributed, we mainly used a Min Max scaling.

2.1.3 CATEGORICAL FEATURE AND ONE-HOT ENCODING

For the categorical features, there are also several approaches possible. If the categories are strings, we replaces them by numbers. If they were numbers and there is relevancy to their order, we kept them as is or scaled them if necessary, again using the Min Max scaling. For those with no relevancy to their order, we encoded them using One Hot encoding to enhance the Dimension reduction and the clustering. This means that we replaced a feature with n categories into n virtual features with 0 or 1 whether the instance was indeed in the category or not.

2.2 DIMENSION REDUCTION

The dimension of the data set is the number of features it contains, which can be lower than the initial number of features if we removed some of them or higher if we used encoding since it replaces one feature with n features, n being the number of categories in the feature.

Since the dimension can be relatively high, it is more efficient for the clustering algorithm that we reduce it as much as possible (also for computing time issues) with losing as little information as possible. Therefore, we computed the number of components necessary to still express 95% of the variance of the initial data set.

We applied two algorithms for the reduction: PCA and Kernel PCA with two types of Kernel (cosine and RBF). Depending on the data set, results were better with one or the other algorithm.

We also used Dimension reduction after the clustering with only 2 components for visualization purposes. We added one more algorithm, t-SNE, which converts similarities between instances into joint probabilities and minimize the Kullback-Leibler divergence of those probabilities. The figures are presented in the Results section.

2.3 CLUSTERING

As mentioned above, the purpose of this assignment is to compare clustering algorithm. We chose to compute K-means, Gaussian Mixture, Fuzzy C, DBSCAN, Hierarchical and Spectral clustering.We do not get into the details of each algorithm since this was covered in class.

To determine the most efficient number of clusters, we use the Elbow method, which is a heuristic method that looks at the sum of squared distances of instances to their closest cluster center. We also use the Silhouette score, which is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. It ranges from -1 (worst) to 1 (best).

To measure the performance of the clustering we also use the Silhouette score for internal validation and the The Fowlkes-Mallows index for external validation. It is defined as the geometric mean between the precision and recall. I ranges between 0 and 1. Another measure of external validation is to compute the Mutual Information between the labels obtained after clustering and an external class that we choose. For instance, "country" in the second data set or "gender" in the first.

3 RESULTS

In this section, we present the results obtained from the different clustering algorithms (see Figures). We also present the measure of performance. Variations of clustering depending on the pre-processing we applied and the dimension reduction algorithm are in the **Supplement Material**.

4 DISCUSSION

We generally observe similarities between the different algorithms. Differences are due to the structure of the algorithm. For instance K-means uses distance to centers so if the clusters are centered and regrouped , the clustering will work properly, whereas the Gaussian mixture will not cut as properly. Even if different algorithms lead to similar clusters, we infer from the differences that the choice of the algorithm can influence the results. We observe on our data sets that the Hierarchical clustering and K-means yield usually very similar results. DBSCAN is quire erratic and needs a very fine tuning of the parameters.

The external validation is not conclusive. This is explained by the fact that in the first data set, if we use the gender as the class, patient are apparently indifferently male or female so the clustering does not reflect this criteria. In the second and third data sets, we used respectively the country and the revenue as class but in the second data set, a vast majority of the customers were from Poland and spread over all clusters. Same thing for the third data set where the revenue is the same for 80% of the instances.

We also observe major differences depending on the pre-processing. Apart from the differences observed across algorithm, it seems the main criteria leading to a net clustering is how the data set was previously "cleaned". It would seem indeed appropriate to conduct a deeper study with probably anomaly detection to remove the noisy instances and auto-encoding.

REFERENCES

[1] B. Strack, J. DeShazo, C. Gennings, J. Olmo, S. Ventura, K. Cios, and J. Clore, Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records, BioMed. research international **2014**, (2014).

[2] M. Łapczyński and S. Bialowas, Discovering Patterns of Users' Behaviour in an E-shop-Comparison of Consumer Buying Behaviours in Poland and Other European Countries, Studia Ekonomiczne **151**, 144 (2013).

[3] C. Sakar, S. Polat, M. Katircioglu,and Y. Kastro, Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks Neural Computing and Applications **31**, 6893 (2019).

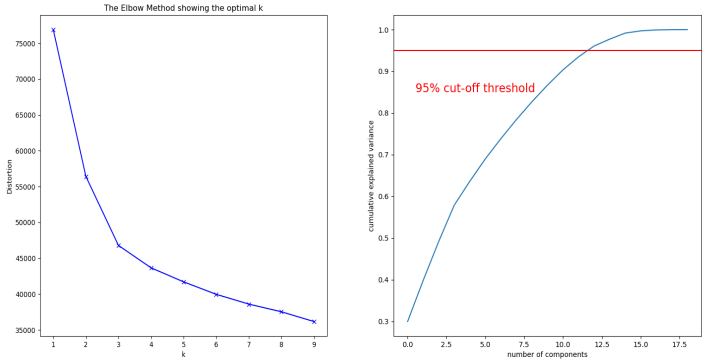

Figure 1: Diabetes. Left plot - Elbow method for K-means. The optimal number of clusters is 3. Right plot - Number of components for the PCA dimension reduction to get 95% of the variance is 13.

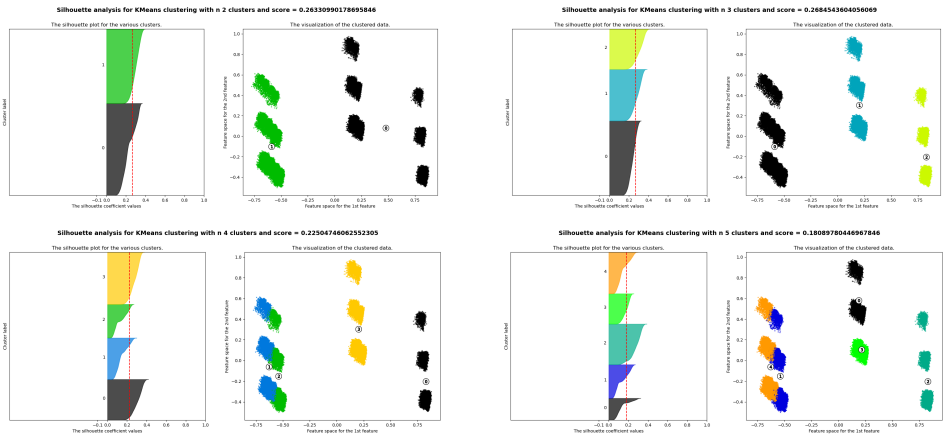

Figure 2: Diabetes. Silhouette scores and profile plotting using K-means. The highest score is for 3 clusters.

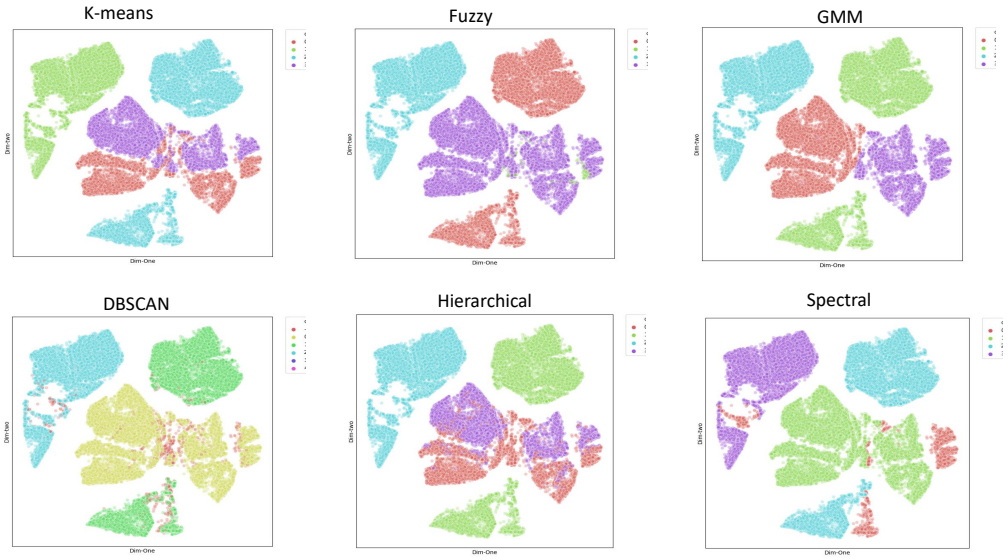

Figure 3: Diabetes. t-SNE representations of the clustering with 3 clusters for the different algorithms.

	Fowlkes-Mellow	Silhouette	Mutual Info
K-means	0.36377	0.35884	0.00023
Fuzzy	0.29727	0.41987	0.00026
GMM	0.35321	0.35965	0.00041
DBSCAN	0.25233	0.41238	0.00025
Hierarchical	0.33443	0.35809	0.00023
Spectral	0.38274	0.38903	0.00026

Figure 4: Diabetes. Comparison of different scores for the algorithms.

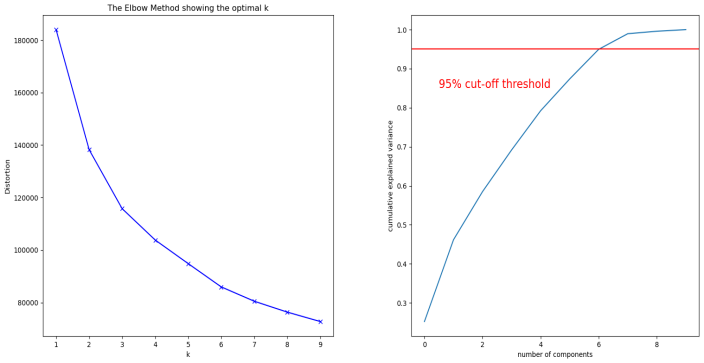

Figure 5: Clothing shopping. Left plot - Elbow method for K-means. The optimal number of clusters is 4. Right plot - Number of components for the PCA dimension reduction to get 95% of the variance is 7.

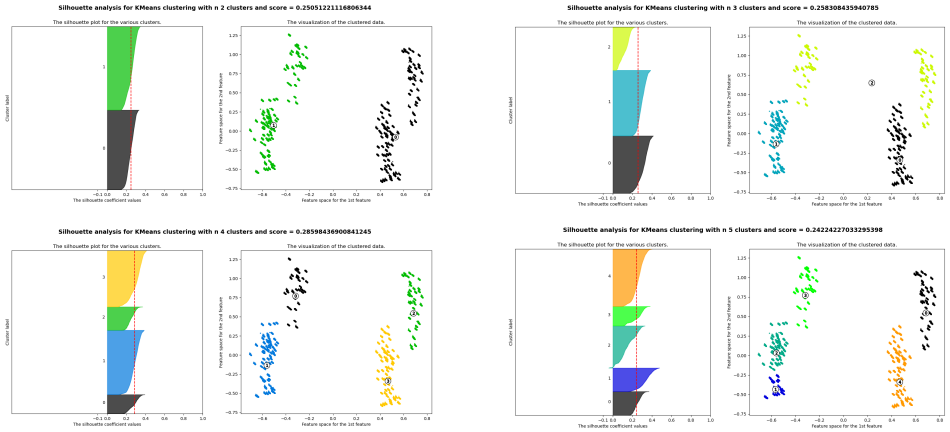

Figure 6: Clothing shopping. Silhouette scores and profile plotting using K-means. The highest score is for 4 clusters.

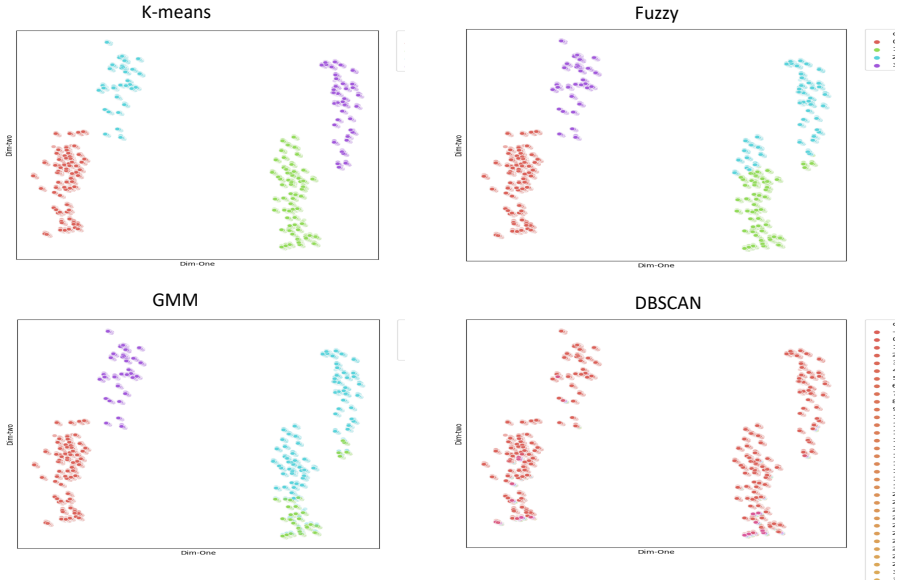

Figure 7: Clothing shopping. PCA representations of the clustering with 5 clusters for the different algorithms.

	Fowlkes-Mellow	Silhouette	Mutual Info
K-means	0.67374	0.40434	0.00006
Fuzzy	0.64947	0.38786	0.00107
GMM	0.49072	0.41342	0.00009
DBSCAN	0.26337	0.36779	0.00523
Hierarchical	0.60102	0.39824	0.00136
Spectral	0.63360	0.38876	0.00100

Figure 8: Clothing shopping. Comparison of different scores for the algorithms.

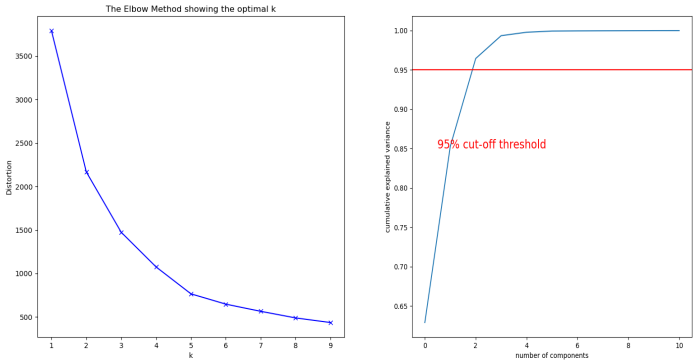

Figure 9: Online shopper intention. Left plot - Elbow method for K-means. The optimal number of clusters is 5. Right plot - Number of components for the PCA dimension reduction to get 95% of the variance is 3.

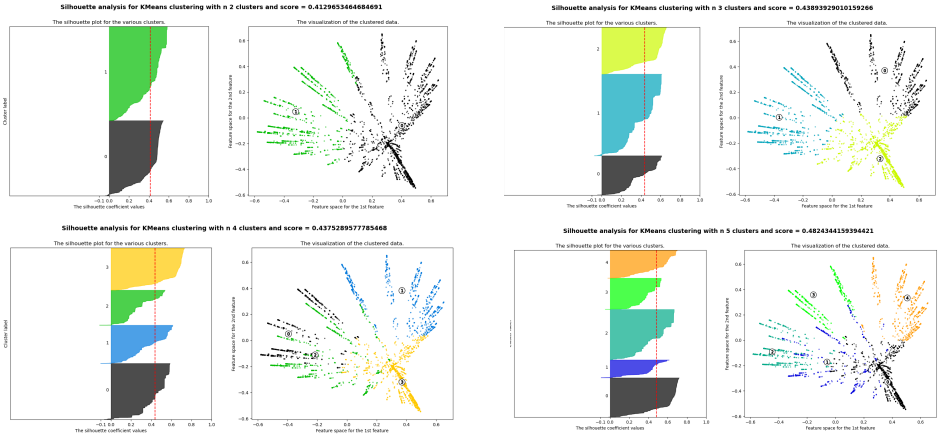

Figure 10: Online shopper intention. Silhouette scores and profile plotting using K-means. The highest score is for 5 clusters.

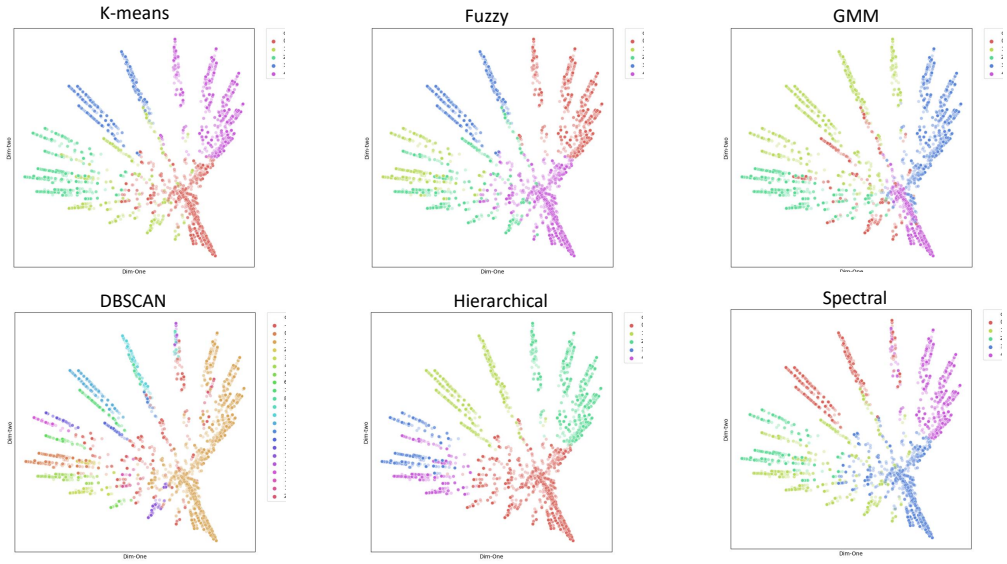

Figure 11: Online shopper intention. Kernel PCA representations of the clustering with 5 clusters for the different algorithms.

	Fowlkes-Mellow	Silhouette	Mutual Info
K-means	0.67374	0.40434	0.00006
Fuzzy	0.64947	0.38786	0.00107
GMM	0.49072	0.41342	0.00009
DBSCAN	0.26337	0.36779	0.00523
Hierarchical	0.60102	0.39824	0.00136
Spectral	0.63360	0.38876	0.00100

Figure 12: Online shopper intention. Comparison of different scores for the algorithms.