Unsupervised learning

Ofek Yerushalmi, BSc Student, Bar Ilan University

March 6, 2022

Abstract

Clustering techniques are frequently used to analyze census data and obtain meaningful large scale groups. some of the most used techniques are k-means, DBSCAN, GMM, Mean Shift, Spectral clustering and Hierarchical clustering algorithm. In recent years, similar tasks has been done frequently. [2] Nevertheless, evaluation studies comparing all of those approaches are rare and usually inconclusive. In this paper an experimental approach to this problem is adopted. Using the US 1990 data set, we compared all of the above clustering methods and concluded that the most appropriate method for this information is k-means. In addition, during the article we will look at the effect of external variables like age, sex, origin, and Yearwrk. We have come to the conclusion that the external variable which connects to the clusters in the best way is Yearwrk. Finally, we examine abnormal samples, show their low association with external variables, and present the information divided into clusters using the cluster method we selected and visualise the clusters associated with the external variables. git link.

1 Introduction

As part of the homework in an unsupervised learning course, we were asked to analyze information from the United States Population Census which was done in 1990. In the following article I am going to analyze the information from the Population Census and divide it into clusters. First, I will delete from my information a number of external variables so that I can check if they associates with the clusters. Next, I will find anomalies in the information, and check if the anomalies information also associates with the external variables, and finally I will present the information in a way that best shows the relationship between the external variables and the clusters. Clustering methods have been applied before on population data while using k-means and other methods [1]. The term cluster analysis encompasses a wide group of algorithms. The main goal of such algorithms is to organize data into meaningful structures. This is achieved through the arrangement of data observations into groups based on similarity.[2]

The main objective of this paper is to evaluate the performance of several clustering methods in the clustering problem, under specific conditions.

2 Methods

2.1 Preprocessing

First, we'll import the relevant libraries into our notebook. I have previously downloaded and stored the data, and loaded it into the notebook. I have removed the external variables dAge, dHispanic, iYearwrk and iSex from the data, and applied one-hot encoding for all the categorial variables. Since the information is very large, and it is not possible to run the algorithms on all the information in a reasonable time, I chose to take a sample from the

information. with Kmeans, DBSCAN, GMM, Hierarchical clustering i choose a data with size 200000 samples and with Mean shift and Spectral clustering i choose a data with 20000 samples (memory and CPU limitation). In each cluster method I repeated the tests several times to verify the same results and checked this by t-test method.

2.2 Clustering Methods

I have chosen to use several clustering methods: Kmeans, DBSCAN, GMM, Mean shift, Spectral clustering and Hierarchical clustering.

Kmeans is a clustering algorithm that identifies clusters of similar counties based on their attributes. The Kmeans algorithm allows the user to specify how many clusters to identify. In this instance, I ran on all the number of clusters in the range of 2 to 25.

DBSCAN is a density-based clustering non-parametric algorithm and is one of the most common clustering algorithms and also most cited in scientific literature. DBSCAN requires eps parameter (the maximum distance between two samples for one to be considered as in the neighborhood of the other.) and a min sample (The number of samples (or total weight) in a neighborhood for a point to be considered as a core point). In this instance, I ran on epsilons between 2 - 50.

GMM (or Gaussian mixture models) involves the mixture (i.e. superposition) of multiple Gaussian distributions. Gaussian mixture models can handle even very oblong clusters. GMM requires n-components, The number of mixture components. In this instance, I ran on all the number of components in the range of 2 to 25.

Mean shift clustering algorithm is a centroid-based algorithm that helps in various use cases of unsupervised learning. it requires Bandwidth, which used in the RBF kernel. If not given, the bandwidth is estimated using sklearn.cluster.estimate-bandwidth.

Spectral Clustering is a growing clustering algorithm which has performed better than many traditional clustering algorithms in many cases. It treats each data point as a graph-node and thus transforms the clustering problem into a graph-partitioning problem. it requires n-clusters, The dimension of the projection subspace. In this instance, I ran on all the number of n-components in the range of 2 to 25.

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. it requires n-clusters, the number of clusters to find. In this instance, I ran on clusters between 2-25.

2.3 Estimating Quality of Clustering

After applying the clustering methods on the data, I had to choose the best clustering method for our information. To do so, I needed to know how well an object has been classified with each clustering method with each hyper parameter setup. therefor, I used the sillhoutte metric.

Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.

For choosing the best method and hyperparameter setup, I produced for every such setup a plot visualizing the distribution of sillhoutte values of all the samples. The chosen setups were setups that had high average sillhoutte values while having thin and short tails on the left side representing underperforming samples.

2.4 Association of clusters with external variables

I have examined the association of the clusters with the external variables using mutual information. Mutual information between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

2.5 Anomalies

Anomaly detection is the process of identifying unexpected items or events in data sets, which differ from the norm. And anomaly detection is often applied on unlabeled data which is known as unsupervised anomaly detection.

In this case, we were asked to find anomalies in the data. In order to indetify outliers, I used cluster based anomaly detection method. among the best clustering methods, I selected the outliers by selecting the samples with the lowest sillhoutte metric. This is done by selecting the threshold value at the base of the left tail in the distribution graph of the silhouette. In order to test whether anomalies are associated with any of the external variables I used mutual information.

2.6 Dimensionality reduction

We will be using principal component analysis (PCA) to reduce the dimensionality of our data. This method decomposes the data matrix into features that are orthogonal with each other. The resultant orthogonal features are linear combinations of the original feature set. You can think of this method as taking many features and combining similar or redundant features together to form a new, smaller feature set.
in order to propose a visualization that best characterize the clusters associated with the external variables, I reduced the dimensions of the information into 2. And I presented the cluster according to the cluster method, and next to it the data which divided into the external variables. In many cases, the relationship between the cluster and the external variable can be clearly seen.

3 Results

3.1 Clustering

For each cluster method, I printed the silhouette and the silhouette distribution graph.

for Kmeans, the best number of clusters were 5, and the silhouette distribution graph is: (sillhoutte score for 5 clusters is 0.28 and for 25 clusters is 0.083)

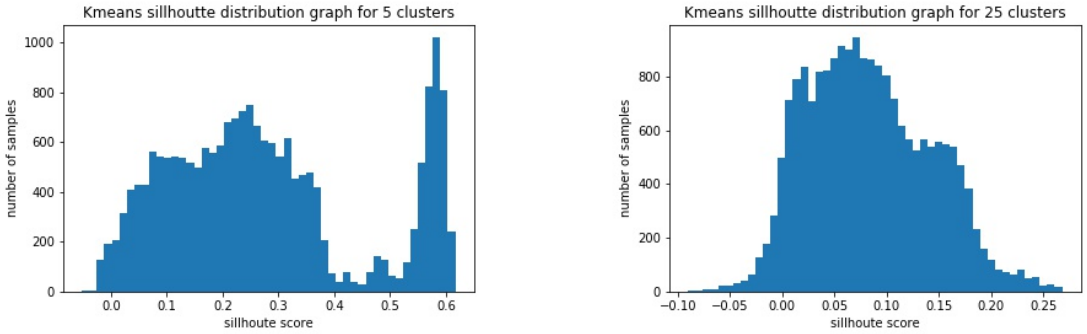

Figure 1: Silhoutte distribution of two Kmeans clustering instances, with 5 clusters (on the left) vs 25 clusters (on the right). One can see that in case of 5 clusters, not only the average sillhoutte value was highest, but also the negative tail representing big portion of misclassified samples was smaller

For DBSCAN, I applied the clustering method on the data with a wide range of parameters, and in all of them I got a bad result regarding the sillhoutte and the distribution the sillhoutte. For example: (sillhoutte score for eps = 6 is 0.02 with only 2 clusters)

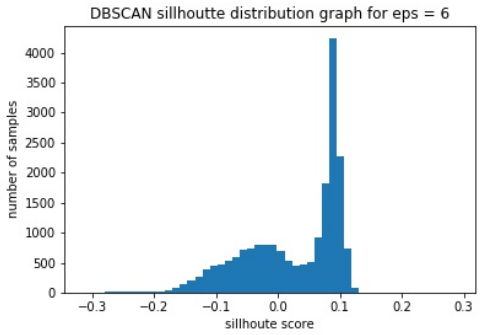

Figure 2: Silhoutte distribution of the best DBSCAN clustering results. We can we both low average value and negative tail

for GMM, the best number of clusters were 5, and the silhouette distribution graph is: (sillhoutte score for 5 clusters is 0.19 and for 25 clusters is 0.05)

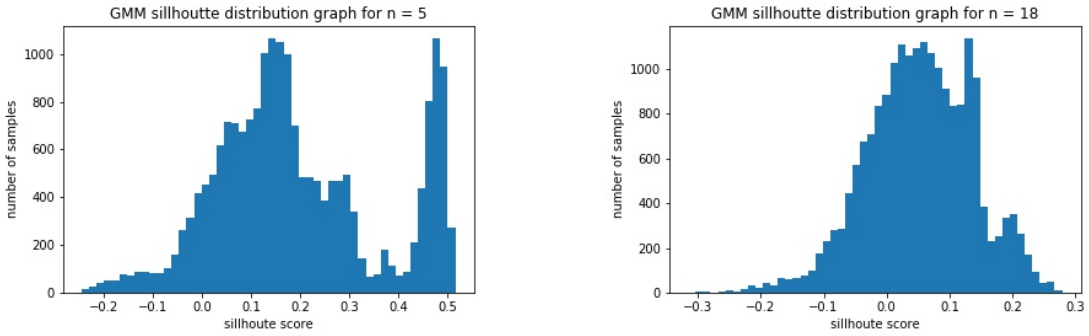

Figure 3: Silhoutte distribution of the some GMM clustering results. The average Sillhoute for 5 clusters (left) was best, compared to the case of 18 clusters, for example (right)

For Spectral clustering, the best number of clusters were 5, and the silhouette distribution graph is: (sillhoutte score for n = 6 clusters is 0.17 and for n = 18 clusters is -0.12)

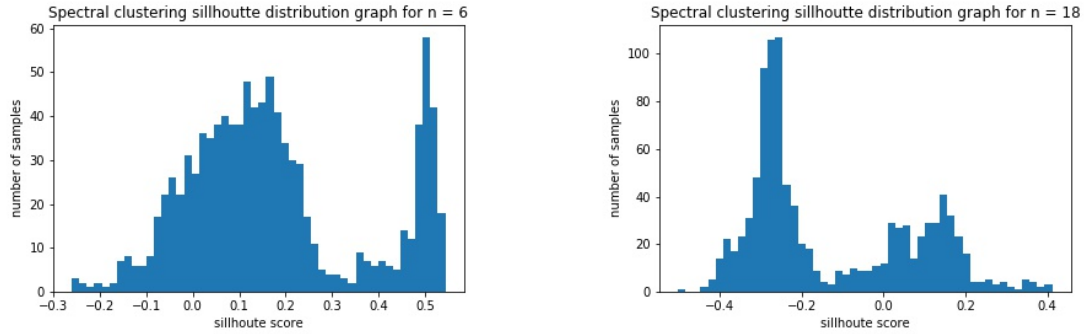

Figure 4: Silhoutte distribution of the some Spectral clustering results. The average Sillhoute for n = 6 (6 clusters) clusters (left) was best, compared to the case of n = 18 (14 clusters) clusters, for example (right)

For Mean Shift clustering, the best number of clusters is 2, and the silhouette distribution graph is: (with 0.35 sillhoutte score)

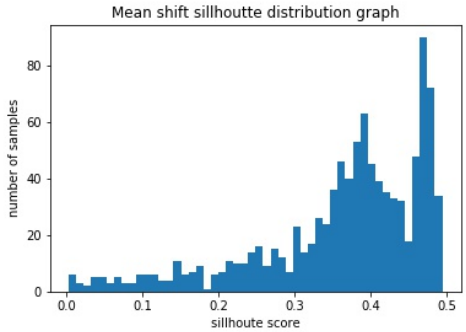

Figure 5: Silhoutte distribution of the best Mean Shift clustering results. althogth we have a good avg sillhoutte score, we have only 2 clusters.

After analyzing the graphs of the silhouette distributions between the different clustering methods, and taking into account the silhouette results, I came to the conclusion that the cluster method most suitable for this data is GMM for 5 clusters (with 0.2 sillhoutte score) and Mean Shift (for 0.35 sillhoutte score)

3.2 Association of clusters with external variables

I have examined the association of the clusters with the external variables using mutual information, and aclculated the average mutual information for each method. I applied mutual information on each clustering method and the result were:

For Kmeans:
mutual inforamtion for dAge is 0.626
mutual information for dHispanic is 0.001645
mutual information for iYearwrk is 1.0957
mutual information for iSex is 0.04693
avg = 0.44265

For GMM:
mutual inforamtion for dAge is 0.6206
mutual information for dHispanic is 0.0062
mutual information for iYearwrk is 1.096814
mutual information for iSex is 0.016270
avg = 0.44248

For Hierarchical clustering:
mutual inforamtion for dAge is 0.6075
mutual information for dHispanic is 0.00315
mutual information for iYearwrk is 1.1134
mutual information for iSex is 0.0519
avg = 0.4405

For Meanshifit:
mutual inforamtion for dAge is 0.0
mutual information for dHispanic is 8.6042e-16
mutual information for iYearwrk is 4.44e-16
mutual information for iSex is 0.0
avg = 3.2e-16

For Spectral clustering:
mutual inforamtion for dAge is 0.198
mutual information for dHispanic is 0.00619
mutual information for iYearwrk is 0.6820
mutual information for iSex is 0.01890
avg = 0.226
According to the numbers, we can clearly see that the external variable which is best associated with the clusters is iYearwrk. And the clustering method best associates with the outside variables is kmeans. In addition, because the mutual information results for Mean Shift came out so bad, we can say that this is not the appropriate clustering method, and now we are left with kmeans.

4 Anomalies

In order to indetify outliers,I used cluster based anomaly detection method. among the best clustering methods,I selected the outliers by selecting the samples with the lowest sillhoutte metric. Thisis done by selecting the threshold value at the base of the left tail in the distributiongraph of the silhouette.
Therefor, I have plotted the sillhoutte distribution graph for Kmean:

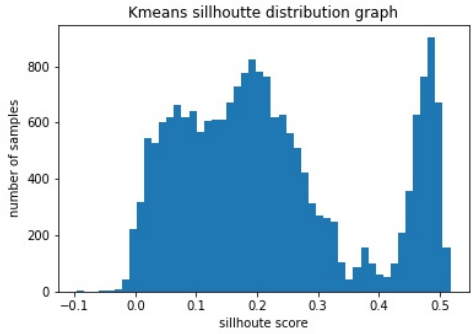

Figure 6: Kmeans anomalies detection

As we can see in the figure 6, samples with sillhoutte value bellow -0.02 are abnormal. In figure 7 we can see the visualisation of the abnormal cases:

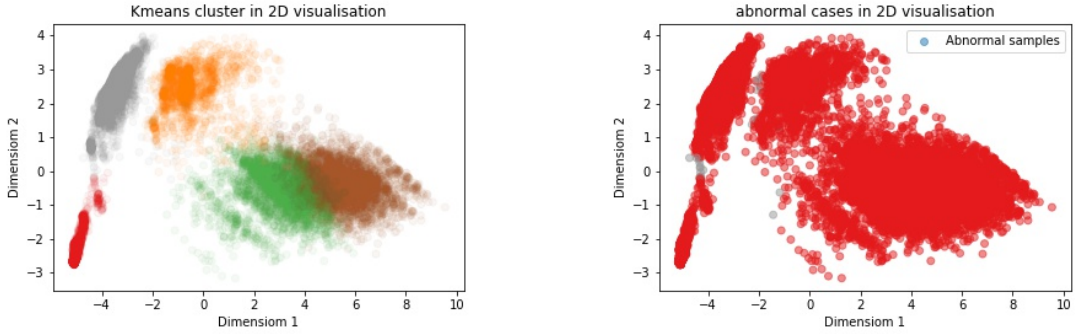

Figure 7: Comparison of the clustering made using K-means method with 5 clusters (left) to the abnormal cases using a -0.01 sillhoutte threshold using the same cluster. The X and Y axis of the visualization are based on first two PCA components

In order to test whether anomalies are associated with any of the external variables I calculated mutual information between the external variables and the Kmeans cluster outliers result.
The results were:
For dAge 0.00099
For dHispanic 0.00099e-05
For iYearwrk 0.001394
For iSex 1.293325e-06
After analyzing the result, we can say that generally there is a low association of the abnormality to the external variables. Among the 4 external variables, we can say that iYearWrk and dAge are most associated with abnormality.

5 Dimensionality reduction

In order to visualise the clustering association with the external variables I have reduced the dimensionality to 2 using PCA. In figures 8 - 11 we can see that the association is most strong, as proposed by the mutual information metric, between the k-means cluster and the iYearwork variable.

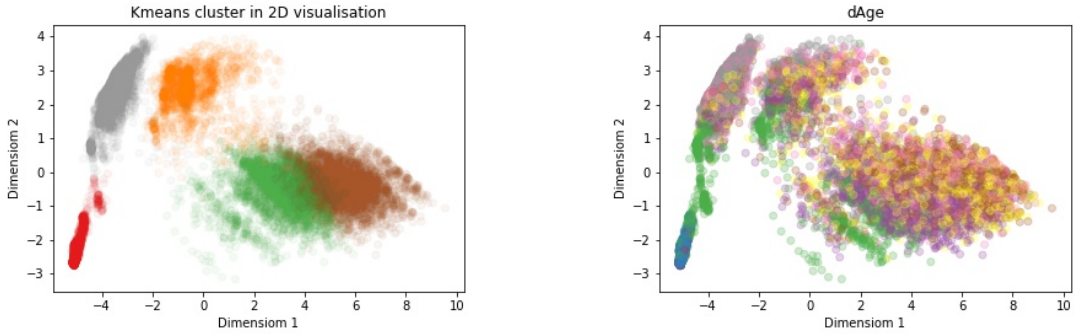

Figure 8: Associating the K-means 5 clusters clustering (left) to the dAge external variable (right)

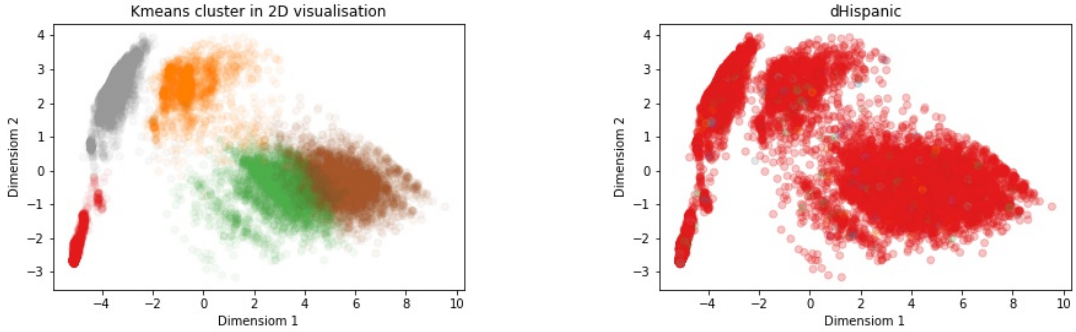

Figure 9: Associating the K-means 5 clusters clustering (left) to the dHispanic external variable (right)

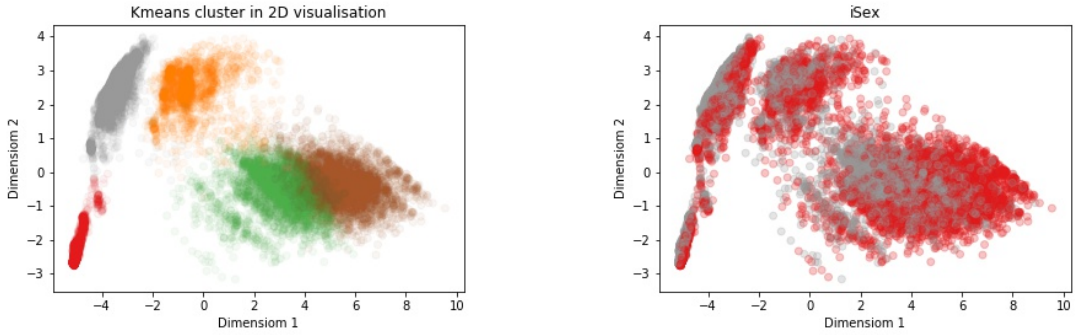

Figure 10: Associating the K-means 5 clusters clustering (left) to the iSex external variable (right)

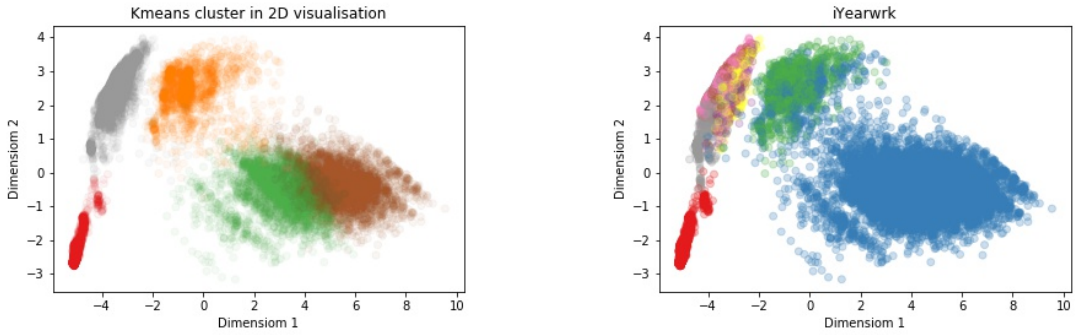

Figure 11: Associating the K-means 5 clusters clustering (left) to the iYearwrk external variable (right)

6 Discussion

According to the results presented it can be argued that k-means is the best clustering method for this data set. By accessing the models created within SKlearn, we were able to improve the explainability of our modelling and draw actionable conclusions. Using these techniques, we have been able to better understand the essential characteristics of the US 1990 cencus and segment the electorate into groupings accordingly. In addition, as we have seen in the study, the external variables iSex and dHispanic do not connect well with the clusters we have reached. It can be interpreted that one's gender and origin do not have much effect on where one end up in life, as opposed to Yearwrk and age. Such an interpretation may contradict some social stigmas that claim that women are less successful or that people of eastern origin are less intelligent. Moreover, After writing the article and analyzing the various clustering methods on a data set of this type, I can say that the most suitable clustering method for this kind of data set is Centroid-based clustering (and in particular k-means) and that there is no such thing as a wrong-chosen algorithm – some of them are just more suitable for the particular dataset structures. I have come to the conclusion that the most glorious and innovative algorithm will not always fit our data, and sometimes the simple one is actually the best.

References

[1] Carl Anderson (2016) *Clustering the US population: observation-weighted k-means*, Towards Data Science

[2] Fernando Bação (2016) *Clustering census data: comparing the performance of self-organising maps and k-means algorithms*, ALMADA, Portugal

[3] Han Man Eitan Sela (2018) *Analyze US census data for population segmentation using Amazon SageMaker*

[4] Yugesh Verma (2021) *Hands-On Tutorial on Mean Shift Clustering Algorithm*, EVELOPERS CORNER

[5] Susan Li (2019) *Anomaly Detection for Dummies*, Medium

[6] Laurea Magistrale (2017) *Esplorazione di indici di qualità per configurare gli algoritmi di clustering*, POLITECNICO DI TORINO