Using PCA, k-Means and hierachical clustering for unsupervised learning in the clustering of pulsars

Joel Kischkel

July 17, 2020

Abstract

During their lifespan stars go through several stages. Massive stars become pulsars which are fast rotating neutron stars. Because each pulsar sends out a unique impulse profile it is possible to survey them. The HTRU2 data set is a collection of pulsars and non pulsar objects in a eight-dimensional space. With the Principal Component Analysis it is possible to reduce the number of dimensions. With the reduced dimensions and the k-means algorithm it is possible to perform clustering. An alternative approach is using Hierachical Clustering. Since the data set is large a visualization with the dendrogram is not useful. The chosen number of clusters is determined by, the data set it self, because an increasing of clusters cannot be intepret correctly.

1 Introduction

The first pulsar was found 1968. Because of the unique impulse a pulsar is sending out scientists first thought they had discovered an extraterrestrial civilization. [1] Pulsars are born from stars. At the end of its lifetime a star starts to expand.

At this stage most of the nuclear fuel is burned and the gravitation compresses the core which increases the heat. The star will blow off its atmosphere. The remaining part of the star is its core of very high density combined with a small radius. Our sun for example will be as big as the earth at the end. This type of star is called white dwarf. [2] If the mass of the white dwarf is more than 1.4 sun masses, the core starts to collapse and a neutron star is born. The gravitational collapse causes a "neutron porridge" in the core. This star is characterized by a diameter of approximately 10 km. [3] Pulsars are rapidly rotating neutron stars. Through the fast rotation a jet of particle is released of the pulsar. In Figure 1 a pulsar with its characteristic beam on the magnetic axis is shown. On the earth we measure an impulse in regular intervals. [4] This pulse shape or profile is individual for each pulsar. [5] The goal of this paper is to use unsupervised learning methods to find pulsars in the HTRU2 data set.

2 Dataset and Methods

2.1 Description of the dataset

The HTRU2 data set contains samples of pulsar candidates. The data set is puplicy available at the UCI Machine Learning Repository. [6] It contains 17,898 samples where 1,639 are pulsars and 16,259 are other objects. The data set contains 9 feature:

1. Mean of the integrated profile
2. Standard deviation of the integrated profile
3. Excess kurtosis of the integrated profile
4. Skewness of the integrated profile
5. Mean of the DM-SNR curve

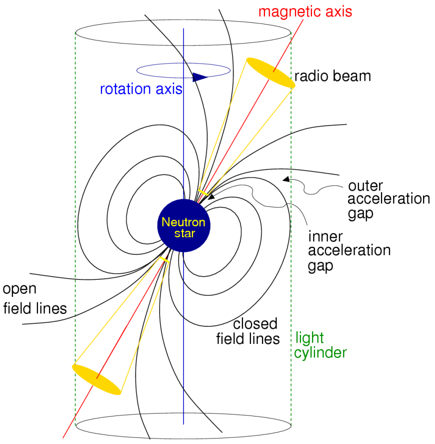


Figure 1: Pulsar (Source: https://www.cv.nrao.edu/course/astr534/Pulsars.html))

6. Standard deviation of the DM-SNR curve
7. Excess kurtosis of the DM-SNR curve
8. Skewness of the DM-SNR curve
9. Class

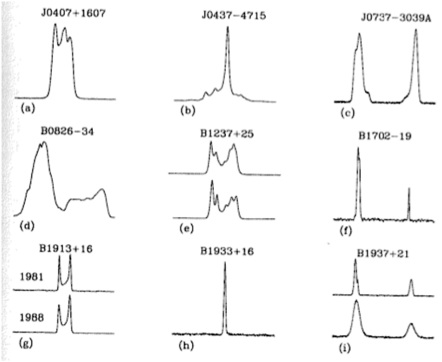


Figure 2: Integrated profile of pulsars (Source: Handbook of Pulsar Astronomy (Loheimer and Kramer Page 9))

The first four variables are statistics from the integrated profile of the pulsar. The figure 2 shows the integrated profile of the pulsar. It measures the radio emission of the pulsar. Through the beam of emissions on the earth is periodically an increase of radio radiation registered. [7] The remaining four features are the statistics for the DM-SNR Curve. When the signal of the pulsar is travalling through the interstellar medium lower frequencies of the signal are slow down while high frequencies arrive earlier on the earth. Without any correction of this effect the signal would be smeared out. The Figure 3 shows the dipersion measure (DM) delay through different frequencies. The lower part of it shows the corrected effect and summing up the frequencies to a single profile. [8] SNR stands for signal to noise ratio. It measures how good a signal is. A high SNR value indicates that the signal is stronger than the noise. [9] The DM-SNR curve in the HTRU2 data set combines both feature. The last feature is the class. A value of 0 means it is not a candidate while of 1 means it is. The data set contains no missing values.

The data is very complex. The figure 4 shows different combination of variables. For simplicity we just use the feature number instead of the name. Red is the color which correspondence to Class 0 while green to Class 1. We see that there is no combination of two features which would allow us to separate the classes complete. Mostly we have on one side pulsars and then on the other side non pulsars and in the middle a mix between both. A good example is feature 5 and 6. The mean and the standard derivation of the DM-SNR curve creates a negative parabola formed figure which mixes inside both classes, impossible to separate clearly. To overcome this problem, we will use the Principal Component Analysis to reduce the number of dimensions.

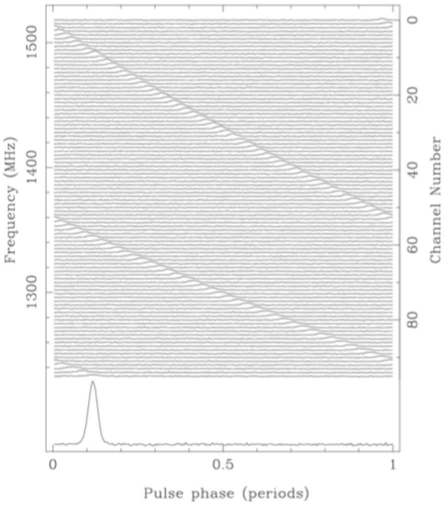


Figure 3: Dispersion measure of a pulsar (Source: Handbook of Pulsar Astronomy (Loheimer and Kramer Page 20))

2.2 Methods

For the analysis of the data we are using a copy of the original data without the Class feature. We perform a dimensional reduction with the Principal Component Analysis. With k-Means and hierachical clustering we will cluster the data into groups. These groups represent the Class feature in the original data. Each method is shortly described and compared with the others. We are using the Class feature to compare the results with the original data.

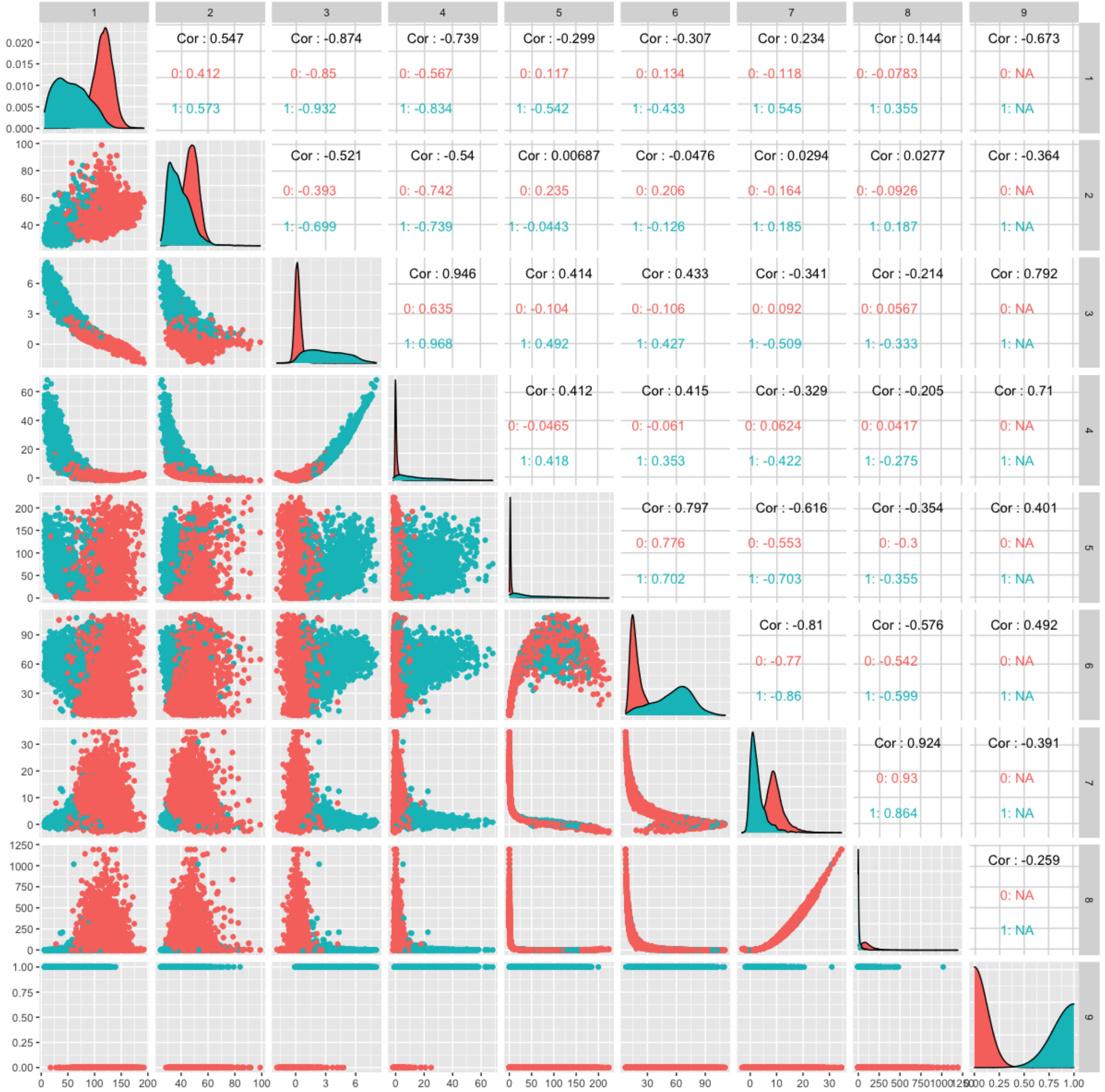
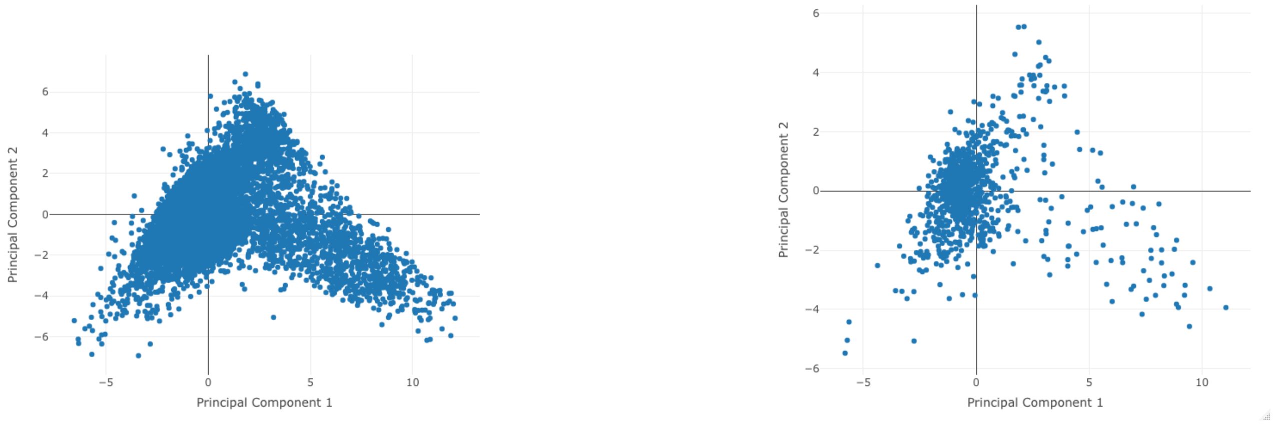


Figure 4: Pulsar data set

3 Principal Component Analysis

The Principal Component Analysis (PCA) will allow us to reduce the number of dimensions without losing too many information. This will allow us to plot the data in a meaningful low dimensional space. The principal component analysis finds so called principal components which are a linear combination of the original variables We seek for each principal component the maximized variance. The principal components are uncorrelated. [10] Following the approach as described before we could plot the data, but we lost many information because we have just considered two variables. Before we perform the PCA we have to scale the variables. This is necessary because the variables were measured in different units. Unscaled data could lead that large values of some elements causes a very large variance. [11] Figure 5a shows the result of the PCA with two components. The plot builds an arrow. It is easier to interpret the subset of the data set as seen in Figure 5b. On the left side of the arrow exists a cluster with lot of points close to each other. The right side contains points which are closer away from each other. Therefore we can assume that these two arms of the arrow represent a class.

For a better visualization we use the Class feature of the original data set to colorize the result. Red points are associated with Class 0 while blue are with 1. In Figure 6a we see that each arm contains one value of cluster and they merge in the middle. A better result can be seen in Figure 6b. The subset allows us to see that the points are mostly seperated. That is a big difference compared to Figure 5b. Here blue points occured in the middle of the red points. But we see also that on

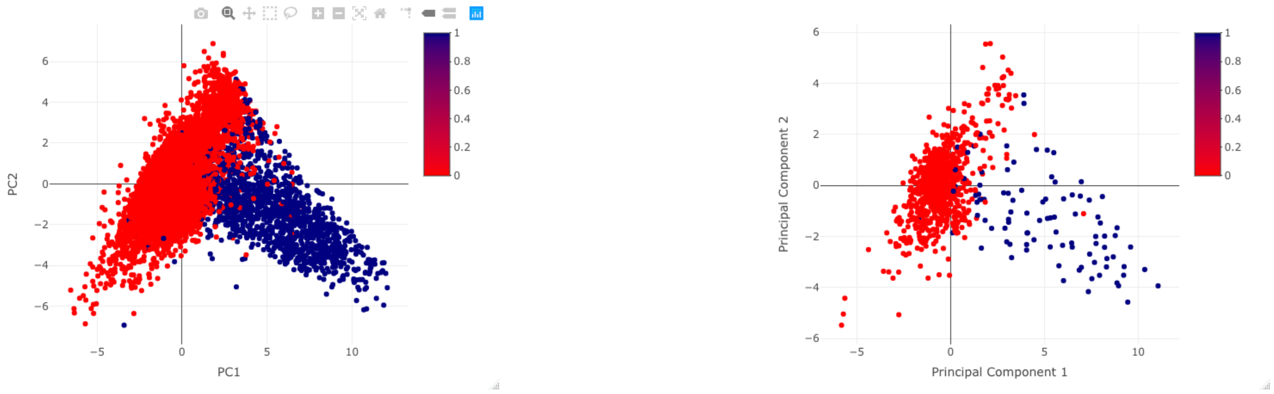


(a) Full data

(b) Subset

Figure 5: PCA result with full and subste of data

the frontier the seperation is not possible and points occure in the other group. A better result we could optain by using more components. Table 1 shows the variance for each component. With two components we can explain 0.78 of the variance. An additionl component would add another 0.10 to it, so that we can explan 0.88 of the total variance. The figure 7 shows the result in a three dimensional space. Even we just a subset of the data it is visually difficult to find differences to previous plots. For the forther analysis we will use two and three principal componets. This allows use to visualize the result.



(a) Full data

(b) Subset

Figure 6: PCA result with full and subste of data colorized

PC1	0.516755838
PC2	0.268075637
PC3	0.101168620
PC4	0.057178103
PC5	0.032278365
PC6	0.019984913
PC7	0.002555242
PC8	0.002003281

Table 1: Variance of the principal components

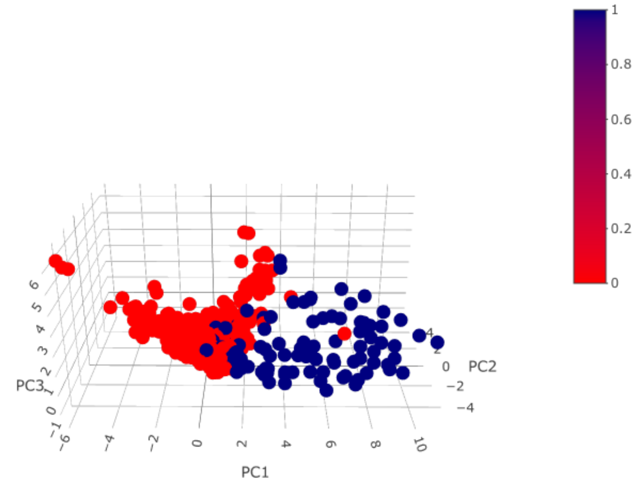


Figure 7: Result of the 3 dimensional PCA plotted with a subset colorized

4 k-Means clustering

In our original data data set we have the feature Class which tells us if it is a pulsar or not. In practice we have no label that tells us if this object is a pulsar or not. We use now the k-Means algorithm to find cluster which allows use to distinguish into pulsars and other objects. The algorithm works in the way that before starting you define a number of clusters. Inside the cluster the k-Means seeks to minimze the Euclidian distance. At the beginning each data point is associated with a random cluster. In several operation the centeroid, also called the mean, is computed and data points with the closed distance are assigned to the next centeroid. [11] We know that with our data set we can find two types of stellar objects. One type are pulsars and the other group are other stellar objects. We conclude that we need to find two clusters. To cluster them we are using the reduced data set from the pca. We perform the algorithm with the full and the subset of the principal component analysis result. As algorithm we use the approach of Hartigan and Wong. This algorithm not just assign a data point to the closest centeroid, it minimizes as well the squares of errors within each cluster. [12] We use this algorithm because we discovered that in higher dimension the traditional k-Means algorithm did not work very well. To ensure that the algorithm will not run forever we set the maximum number of iteration to 100. We plot the result to see what clusters the algorithm found. In Figure 8 we see the result of the clustering. The first thing we notice is that the frontier between the two clusters is sharp. In figure 6a we see that there was more interaction on the border between the two classes. Red represent Class 0 while blue Class 1. Again we see that the two arms are clustered to one class. For a better visualization we again just use a subset of the data. We see in the Figure 9a subset that there is a clear separation between the groups. We added the color green to point out this values which are wrong clustered. If we compare it with the original classes we see that partly the algorithm was not able to cluster correctly. This happens on the frontier between the two cluster and when single values appear in the other cluster. We know that an additional dimension adds more variance to our reduced data. Figure 9b shows the three-dimensional graph of the data. Again, green points are misclustered. To see the differences between the two dimensions we check the size. In Table 2 we see that an additional dimension increases the number of founded pulsars. In the original data are 1,639 pulsars. In two dimension the kmeans clusters 1,219 while in an additional third dimension they are 2,029.

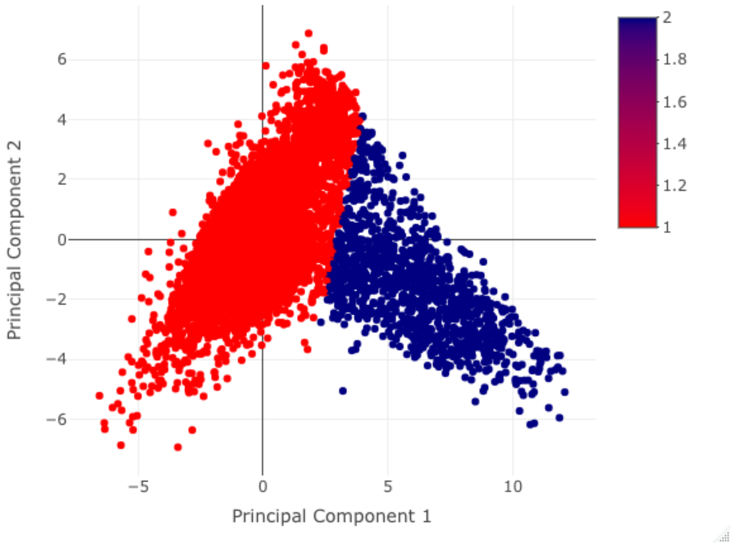
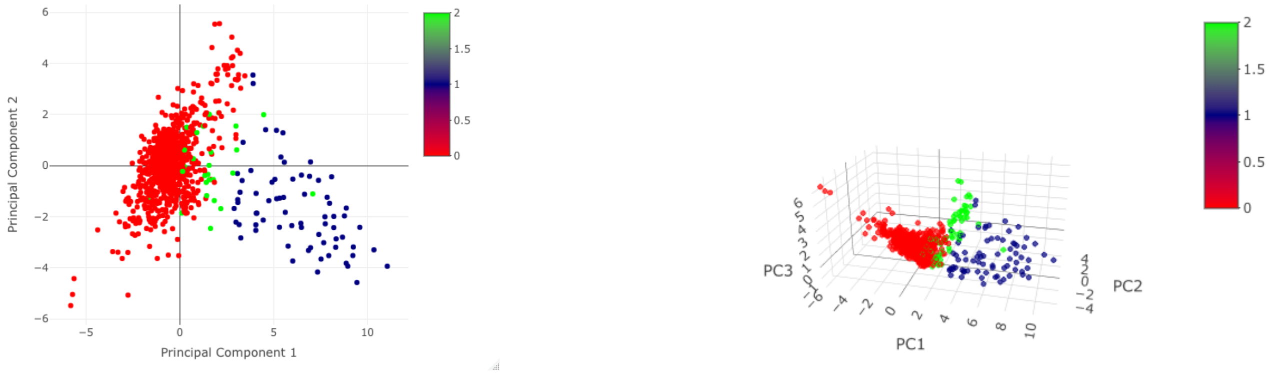


Figure 8: k-means clustering on the full data set



(a) k-means clustering on the subset

(b) Subset

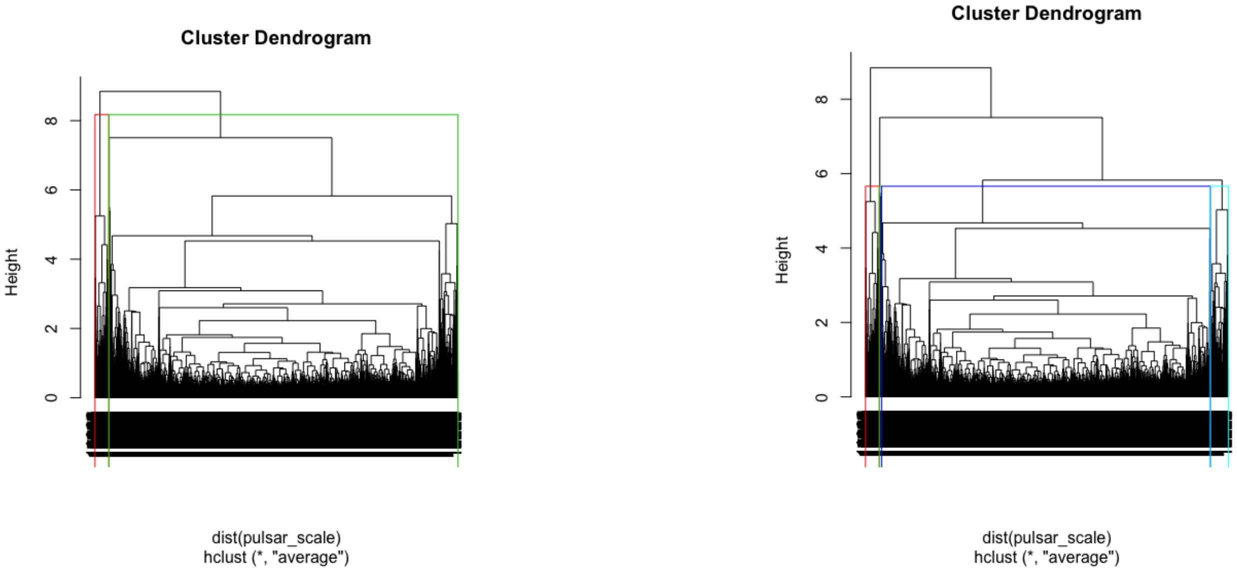
Figure 9: k-means clustering on the subset in three dimension

	0	1
2 dimensional	16679	1219
3 dimensional	15869	2029
Original Data	16259	1639

Table 2: variance of the principal components

5 Hierarchical clustering

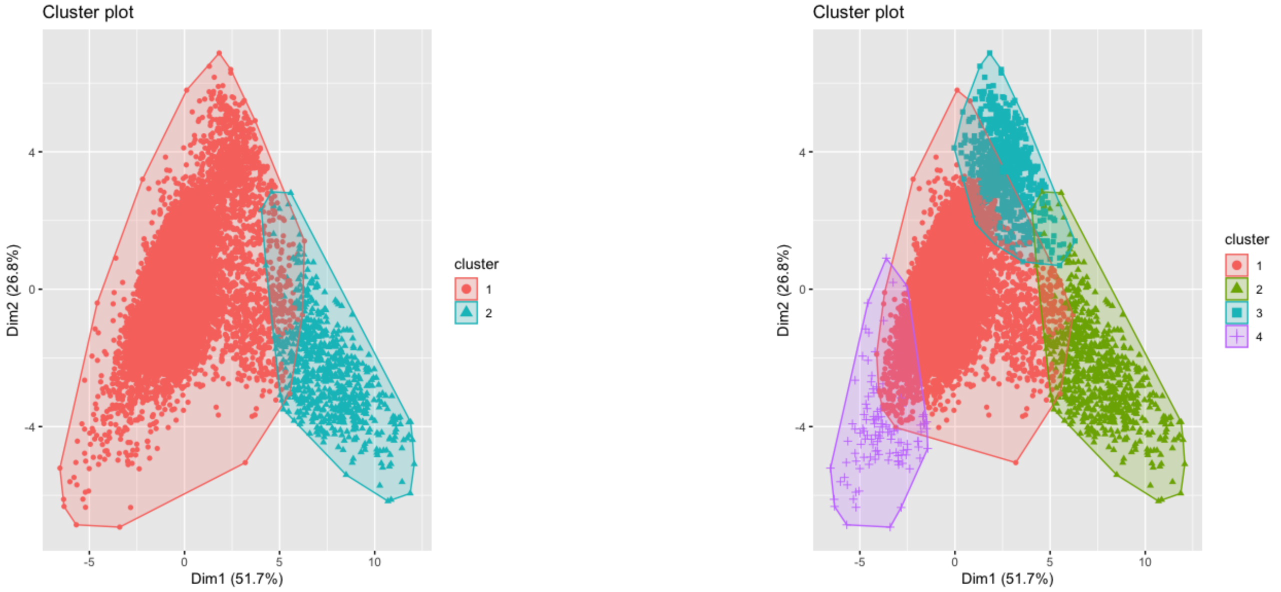
The k-Means algorithm was able to cluster most of the pulsars but failed in areas where it was difficult to separate the data. We are using now with the hierarchical clustering an alternative approach to find maybe more useful cluster. The hierarchical cluster does not start with a specific number of clusters. Instead the algorithm finds similarities between values and merges then similar values together into a group. Again, similar groups will be merged into bigger groups. [11] For the hierarchical cluster we are not using the reduced data set. Instead we are used a the original data set which is scaled and without the Classes feature. For measuring between the values or clusters we are using the average distance. Figure 10a shows the dendrogram generated by the cluster analysis. To make the interpretation more simple we plotted two clusters into the dendrogram. Because the dataset contains so many different values, in the lower part the dendrogram cannot interpret anymore. If we start from the top, we see that we can divide the data into two groups. A better visualization can be seen if we plot the data into two dimensions. The Figure 11a correspondence with the dendrogram. The clusters are the same as seen before in the k-Means clustering. Notice that the classes after the computation are now 1 and 2. Where 1 correspondence to 0 and 2 to 1. If we increase the number of clusters to four, we see the following dendrogram in Figure 10b. Again, a better visualization is done if we plot the data into two dimensions. Figure 11b have four cluster. Cluster 1 and 2 we have seen before. New are cluster 3 and 4. We can conclude that cluster 4 is part of 1 and with this in a cluster of non-pulsars. The problem is cluster 3. We know through our original data that this is a mix of pulsars and stars. But with our chosen methods we cannot say that these stars are in Class 1 or 0.



(a) Hierachical Clustering with two cluster marked in boxs

(b) Hierachical Clustering with four cluster marked in boxs

Figure 10: Dendrogram of pulsar data set with differend cluster marked as boxes



(a) Two cluster in a two-dimensional space

(b) Four cluster in a two-dimensional space

Figure 11: Visualization of the cluster debending on the number of clusters chosen from the dendrogram

6 Conclusion

With the Principal Component Analysis, we were successful in reducing the number of dimensions. Because we wanted to visualize the result, we used two and three principal components. With the reduced data set we could perform k-Means clustering. The problem is that with this we could not correctly cluster on the frontier and outliers. An increasing number of clusters did not improve the result. For this result are two reasons responsible. First our data is based on just two classes. Additional cluster are difficult to interpret since we can just say this is a pulsar or not. A more detailed data set would may allow us also to build sub cluster of pulsars. This could lead to a better clustering. The second reason are the chosen number of dimensions for the k-Means clustering. While in a two-dimensional data the k-Means found less data inside the pulsar cluster than in the original data, in the three-dimensional data it found more. The hierarchical cluster analysis is not a good way to find cluster in this specific dataset. The first reason is that the computational time is much higher than in the k-Means approach. Second because the dataset is large, the dendrogram becomes large as well and with this difficult to interpret.

References

[1] APS News. This Month in Physics History February 1968: The Discovery of Pulsars Announced. `https://www.aps.org/publications/apsnews/200602/upload/feb06.pdf`, 2006. Access: 2020-25-06.

[2] Donald G. York, Owen Gingerich, Shuang-Nan Zhang, editor. *The Astronomy Revolution*. CRC Press, 2012.

[3] Hannu Karttunen, Pekka Kröger, Heikki Oja, Markku Poutanen, Karl Johan Donner, editor. *Fundamental Astronomy*. Springer, 2017.

[4] NASA. Immagine the Universe - Neutron Stars. `https://imagine.gsfc.nasa.gov/science/objects/neutron_stars1.html`. Access: 2020-25-06.

[5] Gerrit Verschuur. *The Invisible Universe - The Story of Radio Astronomy*. Springer, 2007.

[6] Robert Lyon. HTRU2 Data Set. `https://archive.ics.uci.edu/ml/datasets/HTRU2#`, 2017. Access: 2020-23-06.

[7] Duncan Lorimer, Michael Kramer. *Handbook of Pulsar Astronomy.* Cambridge University Press, 2005.

[8] John M. Ford. *Pulsar Search Using Supervised Machine Learning.* PhD thesis, Nova Southeastern University, 2017. Access: 2020-01-07.

[9] Wikipedia. Signal-to-noise ratio. `https://en.wikipedia.org/wiki/Signal-to-noise_ratio`. Access: 2020-01-07.

[10] Markus Ringnér. What is principal component analysis? *Nat. Biotechnology*, 26(3):303 – 304, 2008.

[11] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning - with Application in R.* Springer Science + Business Media New York, 2013.

[12] Laurence Morissette and Sylvain Chartier. The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15 – 24, 2013.