FINAL 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 are 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 those techniques on a data set that describes a sample of pulsar candidates. With the clustering tools we demonstrate that pulsar are not per se a category of star, according to the measurement features used, but rather "anomalies" that do not follow the distribution of other ordinary stars.

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

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 by applying anomaly detection. The features of the data set require also analysing. Some provide absolutely no valuable information for the clustering. 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. features also often have to be scaled to ensure that each provide the same amount of information. Finally, a more systematic, way to reduce the number of features to the essential ones is Dimension reduction.

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. After the clustering, we will again use Dimension reduction for the visualization. Finally, we will evaluate the efficiency of each algorithm using statistical tests.

The data set studied describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (HTRU2) (1). Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter (see (2) for more uses). Machine learning tools are now being used to automatically label pulsar candidates to facilitate rapid analysis. Classification systems in particular are being widely adopted. Here the legitimate pulsar examples are a minority positive class, and spurious examples the majority negative class. We here propose to demonstrate

that pulsars are exceptions and their characteristics do not follow the distribution of other stars. As such, the best clustering to describe pulsars are not the regular algorithms but rather those used to detect anomalies.

2 METHODS

2.1 DATA PROCESSING

The data set contains 16,259 spurious examples and 1,639 real pulsar examples. These examples have all been checked by human annotators. Each candidate is described by 8 continuous variables, and a single class variable. The class labels used are 0 (negative) and 1 (positive). The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables (mean, standard deviation, kurtosis, and skewness) that describe an longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve (see (3) for more details). We will separate the class from other features and use it as an external validation to test the efficiency of our clustering. We will also analyze the features and their relevancy.

2.1.1 NUMERICAL FEATURES AND SCALING

Our data set only contains numerical features and no categorical features. For those 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, the kurtosis and skewness have small values relatively to the mean or standard deviation. We use MinMax and Standard scaling (see Results section).

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 apply the PCA 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 AND ANOMALY DETECTION

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

Data sets often contain outliners. Those are noisy instances that do not really fit in the general characterization of the data set and, therefore, usually disturb the clustering and would need to be removed. In order to identify those samples, we use two algorithm: DBSCAN and Isolation Forest (see the Results section).

2.4 STATISTICAL TESTS

We first perform internal validation. We compute the average 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). It gives a measure of well clustered is the data set disregarding the true labels.

We also compare clustering algorithms. We use again the silhouette score but this time for each sample and compute an ANOVA test between each set. If the null hypothesis (all have the same mean) is rejected (p-value smaller than a given threshold), we choose the algorithm with the highest average silhouette score and compare it one by one with the others. Indeed, the ANOVA test only determines if the samples have the same mean but if it is not the case, it does not tell us which one is the highest, so we need to determine if the algorithm with the highest mean actually performs better than all the others.

Finally, we perform external validation. we use the Fowlkes-Mallows index. It is defined as the geometric mean between the precision and recall. It ranges between 0 and 1. Another measure of external validation is to compute the Mutual Information between the labels obtained after clustering and the external class.

3 RESULTS

As described above, we first identify the relevant features. The data set is comprised of 8 features. The four first ones are the distribution parameters obtained from the integrated profile that has been averaged in both time and frequency. The second set are similarly obtained from the DM-SNR curve. We plot the data set using PCA 2-dimension representation and observe (Fig. 1) that only the first set of variables (plot b) is relevant and provides a clear view of the two categories (pulsar and non-pulsar). The set from the DM-SNR curve does not provide information with regard to the clustering (whether by itself on plot c or all together on plot a). We therefore performed the rest of the study only with the first set. We also used t-SNE 2-dimension representation but this representation is less significant so we did not present it here (see Supplemental Material)

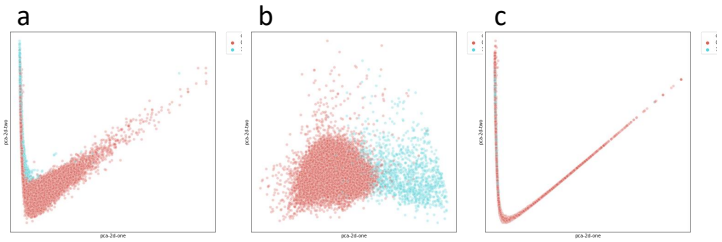

Figure 1: 2-dimension representation of the data set using PCA. Plot (a) - entire data set. Plot (b) - only the parameters from the longitude-resolved signal. Plot (c) - only the parameters from DM-SNR curve. Plot (b) gives a much clearer representation.

We applied scaling to the remaining features and we also performed dimension reduction to the new data set using the PCA algorithm. We obtained that the number of features expressing 95% of the variance is 2. Those operations did not improve the representation of the data set (see Supplemental Material).

We then applied the clustering algorithms described above (see Fig. 2). Since K-means, Fuzzy and Hierarchical are proximity based, they draw a straight "cut" in the group of regular stars and therefore perform poorly, while GMM provides a better clustering. The average silhouette scores for each algorithms are given if Fig. 4.

Our claim is that, with respect to the integrated profile, pulsars are actually outliners. They do not really form a cluster but rather are in the "fat tails" of the distribution. Therefore, we applied algorithms usually used to exclude outliners such as DBSCAN and Isolation Forest (see Fig. 3). As an internal validation, the silhouette score obtained from those two algorithms were higher than the others but even more interestingly, the Mutual Information is also higher confirming that those algorithm have a better fit of the actual classes and that pulsars are outliners.

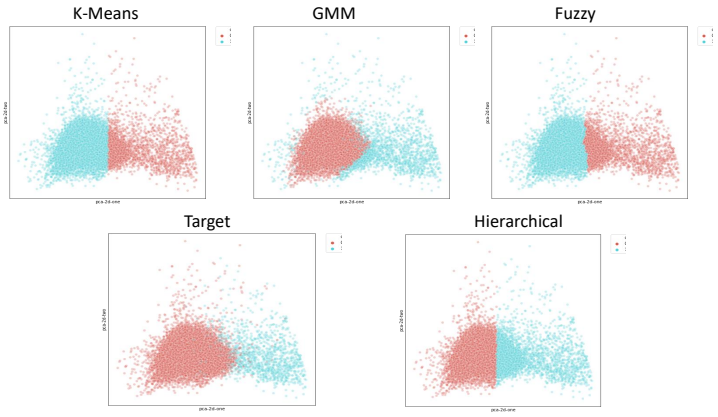

Figure 2: 2-dimension representation of the data set using PCA after clustering with each algorithm. The target plot corresponds to the original classes.

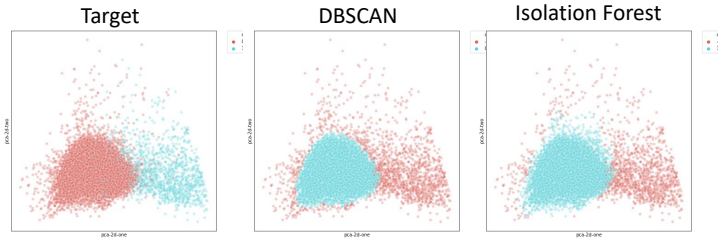

Figure 3: 2-dimension representation of the data set using PCA after clustering with each anomaly detection algorithms. The target plot corresponds to the original classes.

To corroborate our claim that Isolation Forest performs better than the other, we computed the p-value from an ANOVA test over the silhouette score obtained for each sample for each algorithm. This p-value is extremely low (10^{-100}), hence rejecting the null hypothesis that the algorithms have the same average silhouette score. We also computed the p-values of t-tests between the silhouette score of each sample obtained with the Isolation Forest algorithm against all other algorithms and, again, these p-values are extremely small. This is due to the fact that the labels are indeed different and Isolation Forest performs better, but also because the size of the data set is quite large.

We intuitively used 2 clusters since the classes in the original data set are binary (pulsar or non-pulsar). Nevertheless, another interesting result is that, for the "regular" algorithms, when using 3 clusters, although the silhouette score and the Fowlkes-Mallows score decrease, the Mutual Information increases. However, it is still smaller than with Isolation Forest (see Supplemental Material).

4 Discussion

We generally observe similarities between the different "regular" algorithms. Differences are due to the structure of the algorithm. For instance K-means uses distance to centers so if since the clusters are not centered, the clustering does not work properly, whereas the Gaussian mixture will perform better. Even if different algorithms lead to similar clusters, we infer from the differences that the choice of the algorithm can influence the results.

We demonstrated here two important features for the detection of pulsars. The first one is that the variables obtained from the integrated profile are much more relevant that the ones obtained from the DM-SNR curves. Another conclusive aspect of our study is that pulsar are outliners and, as such, are more easily identified with algorithms such as DBSCAN or Isolation Forest.

Algorithm	Silhouette score	Fowlkes Mallows score	Mutual Information
K-means	0.58943719	0.881541112	0.121477044
GMM	0.5763643	0.924011406	0.132899171
Hierarchical	0.553889093	0.833684311	0.107251765
Fuzzy	0.541966592	0.808113458	0.100105293
DBSCAN	0.615578521	0.942762407	0.147090287
Isolation Forest	0.662514096	0.956161988	0.155003216

Figure 4: Table of the different scores obtained with each algorithm. On the last line, we observe that Isolation Forest performs better than all other algorithms.

REFERENCES

[1] Lyon, Robert J and Stappers, BW and Cooper, Sally and Brooke, JM and Knowles, Joshua D, Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society **459** 1104 (2016)

[2] Roberts, Nora and Lorimer, DR and Kramer, M and others, Handbook of pulsar astronomy, **4** (2005)

[3] Lyon, Robert James, Why are pulsars hard to find?, PhD Thesis, University of Manchester, (2016)