Draft version July 17, 2024
Typeset using LaTeX default style in AASTeX631

Galaxy Clustering and Machine Learning*

Srivi Balaji[1] and Sana Kohli[1]

[1]*Department of Computer Science, The University of Texas at Austin*

ABSTRACT

Galaxies are fundamental units of the universe and studying their clustering and super clustering, or lack thereof, can provide key insights into the large-scale structure of the universe. This research explores the higher-order structure of galaxies by employing clustering algorithms, specifically K-means, to identify clusters of galaxies. Leveraging data from the Sloan Digital Sky Survey (SDSS), we analyze right ascension, declination, and redshifts of over 2.79 million galaxies. Our findings contribute to the evolving field of galaxy clustering, offering insights into the large-scale structure of the universe. The prospect of deriving superclusters by re-applying clustering algorithms on identified clusters opens possibilities for deeper exploration into the hierarchical organization of galaxies. Current computational limits as well as the ambiguity in the field are described in an attempt to outline potential future developments to advance the area of research.

1. INTRODUCTION

The advancement of astronomical technology has enabled the rapid growth of data collection, greatly increasing information on galaxies in the universe and the need for computational power.

A significant contributor to the galaxy clustering field was George Abell. His paper, "Abell Catalog of Rich Clusters of Galaxies," details the use of optical data to cluster thousands of galaxies into over 2700 galaxy clusters (Abell 1958). The continuation of this work can be credited to Jan Hendrik Oort and Neta Bahcall. Oort's work focused on the discovery of superclusters (Oort, 1983). This concept led to the theory of a conceptual large-scale structure, which was established by Bahcall (Bahcall, 1988).

Today, our view of clustering galaxies has developed from manual processes to highly advanced automated ones, allowing machines to identify patterns unseen to the human eye. Our research delves into examining higher-order galaxy clustering through clustering algorithms to create models of the large scale structure of the universe. Collecting data from the Sloan Digital Sky Survey (SDSS), we utilize the K-means clustering algorithm to identify clusters of galaxies. This question offers the chance to explore the intersection of astronomical research with the rapid advancement of machine learning, contributing to the understanding of the large-scale structure of the universe.

2. THE CLUSTERING ALGORITHM

We followed the process of obtaining our data, determining the most apt units to use for physical space, and identifying the number of clusters to build the clustering model.

2.1. *Obtaining the Data*

To inform our clustering algorithm, the Sloan Digital Sky Survey (SDSS), specifically Data Release 17 with a focus on the catalog with spectroscopic classifications, provided access to accurate values of right ascensions, declinations, and redshifts for galaxies in the universe. To query the data, we utilized the SciServer platform to access a dataset containing 2,790,253 rows. The SDSS utilized standard astronomical right ascension and declination in J2000 format.

For our analysis, we used Python on Jupyter Notebook and pandas was utilized along with matplotlib to plot various facets of the data, and compare columns. We used the scikit-learn package to run the K-Means algorithm on our data.

2.2. *Preliminary Analysis*

* Written in 2024

Our analysis consisted of deriving multiple statistics from the data, and interpreting these numbers to better understand the shape and significance of the dataset, as well as its limitations and potential biases.

The units of right ascension were given in degrees, and 50% of the galaxies were within a range of 130^{o} to 224^{o}, with the mean being 173^{o}. Furthermore, the units of declination were also measured in degrees, and 50% of the galaxies were within a 6^{o} to 37^{o} range. The third input was redshift, which is a galaxy's recessional velocity in kilometers per second. In order to facilitate spatial galaxy clustering, the data underwent conversion from celestial coordinates to a Cartesian system, converting redshift into distance in megaparsecs in a process discussed in the next section. The range of galaxies obtained from the SDSS dataset fell within a maximum distance of 8.6 megaparsecs, with a mean of 1.8 megaparsecs. The middle 50% of galaxies consisted of distances from 0.68 megaparsecs to 2.6 megaparsecs.

2.3. *Unit Conversion*

The data that was first obtained from the SDSS gave the units of right ascension and declination in celestial coordinates with the J2000 epoch. Additionally, as described before, the redshift of a galaxy is simply its recessional velocity divided by the speed of light, so redshift is essentially unit-less in terms of physical space. To cluster on the spatial characteristics of the galaxies, the units were converted into megaparsecs.

With redshift, Hubble's Law provides the mechanism to convert to megaparsecs. Through using velocity is equal to Hubble's constant multiplied by proper distance as well as using Hubble's constant in km s-1 Mpc-1, the proper distance for each galaxy can be calculated. Furthermore, with right ascension, declination, and proper distance, the equatorial celestial coordinates can be converted to a Cartesian system. In this system, the X-axis points towards the vernal equinox point, the Y-axis points towards a declination of 0 degrees and right ascension of 6 hours, and the Z-axis points towards a declination of 90 degrees, along Earth's north polar axis. After converting the right ascension and declination from degrees to radians, the x, y, and z values can be calculated using the following formulas.

$$x = dcos(declination)cos(rightascension)$$

$$y = dcos(declination)sin(rightascension)$$

$$z = dsin(declination)$$

With these values in megaparsecs, the galaxy data can be better visualized in physical space.

2.4. *K-means Cluster Determination*

With the use of K-means as our clustering algorithm, one of the requirements to fit the data is having a pre-defined number of clusters. Thus, this required us to determine the number of clusters that are most representative of the galaxies included within our dataset. Among other techniques, one of the primary techniques to explore this revolves around examining a plot of the number of clusters versus mean squared error or cost. In this case, cost can be considered as the average distance of each galaxy from its cluster center. As the number of clusters increases, the cost naturally decreases as there are more clusters closer to each galaxy. However, the point where there is an "elbow" in the plot, or the place where the most significant cost change happens before the graph begins to flatten out, is considered as the most apt number of clusters without overfitting.

By first sorting our dataset by proper distance in megaparsecs, we were able to run the algorithm on a wide range from 1 to 200 clusters. Then, identifying 90 clusters as the "elbow" point, we extrapolated this to around 2250 clusters for the entire dataset of galaxies.

3. ALGORITHM OUTPUT AND RESULTS

In evaluating our clustering algorithm, we relied on statistics to ensure that clusters were evenly distributed and accurately represented galaxies in close proximity. Because galaxies are clustered together based on proximity to each other as well as their common distance, sorting by distance and running the K-means algorithm on these smaller subsets allows for the algorithm to assign more weight to galaxies having similar proper distances, rather than creating clusters that span through megaparsecs in the universe.

3.1. *Comparison to Abell's Catalog*

Comparing our results to Abell's original rich clusters of galaxies catalog also provided a valuable benchmark for evaluating the performance of our model. Abell's catalog contained approximately 2700 clusters, as compared to our

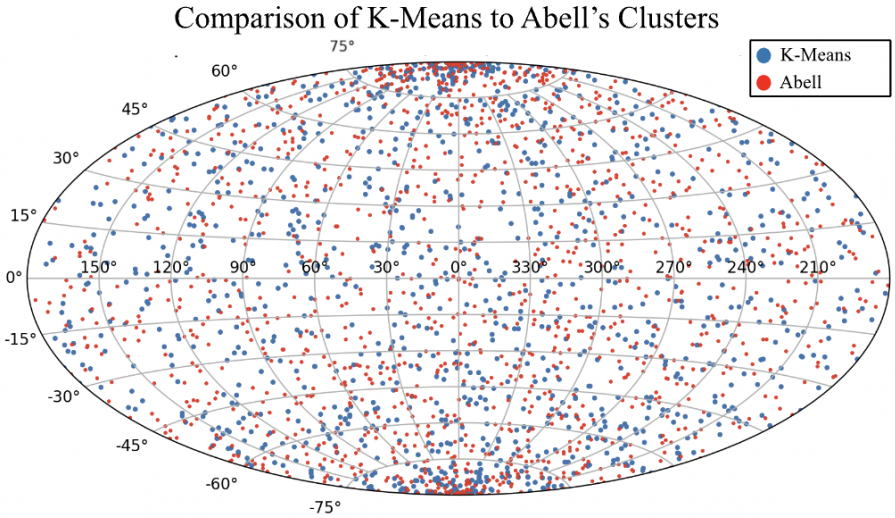

Figure 1. A comparison of the clusters obtained from the K-means algorithm plotted against the clusters Abell determined on an Aitoff projection.

2250 clusters. It measured some key statistics of the clusters, including richness groups. For our clusters, the mean of galaxies was approximately 1240 with a minimum of 25 and a maximum of 23593 galaxies. The 25% quartile was 876, the 50% quartile was 1103, and the 75% quartile was 1387 galaxies. Additionally, the Abell radius, which details the average distance of each galaxy from its cluster center (describing the compactness of a cluster), is stated as approximately 2 megaparsecs. For our clusters, the mean of their Abell radii is closer to 1900 megaparsecs, indicating the possibility of areas where more clusters could be determined.

In Figure 1, our clusters overlaid with Abell's clusters present notable similarities, demonstrating the effectiveness of our algorithm in capturing the distribution of galaxy clusters. Overall, the number and distribution of clusters appear fairly consistent. However, notable differences emerge near the poles, where Abell's catalog records more clusters, suggesting potential biases or limitations in our clustering approach near celestial poles.

In terms of the model's performance, it provides an unbiased approach to determining the distance between galaxy clusters. However, limitations were observed, particularly in achieving uniformity in cluster size and shape. This uneven distribution may impact the interpretation of results and thus needs further exploration to refine the clustering algorithm. Comparing our algorithm's performance to human-performed clustering, we reference the comparison with Abell's clusters. This comparison serves as a valuable validation of our automated approach, highlighting areas of consistency and divergence between algorithmic and manual clustering methodologies.

3.2. *Comparison to Existing Catalogs*

Other researchers have employed various techniques to create galaxy clusters, each focusing on different considerations and methodologies to optimize their results. For example, Yang et al. (2007) developed a catalog based on the SDSS Data Release 4, focusing on identifying centers of potential galaxy groups and determining their characteristic luminosity. Their catalog, consisting of approximately 300,000 groups, was tested for performance using mock galaxy catalogs. This approach provided a comprehensive look at galaxy group dynamics and the underlying dark matter halos, offering insights into the mass and luminosity relationships within the universe. While Yang et al. focused on mass-to-light ratios and halo properties, our methodology emphasizes the spatial distribution and proximity of galaxies by converting celestial coordinates into a Cartesian system and clustering based on these distances.

Berlind et al. (2006) took a different approach, utilizing the friends-of-friends (FoF) algorithm to cluster galaxies from the SDSS redshift survey. This method is designed to recover galaxy systems within the same dark matter halos. Testing with mock catalogs revealed challenges in preserving the halo multiplicity function, size distribution, and velocity dispersion. It's important to note the various trade-offs that come with different techniques, such as computational efficiency and methodological flexibility, with our K-means approach providing a structured, distance-

based clustering that contrasts with the parameter-free FoF method. Ultimately, these comparisons underscore the importance of algorithm selection based on the goals and constraints of the study.

4. CONCLUSION

Our research into higher-order galaxy clustering using the K-means algorithm has revealed significant insights into the universe's structure. Through statistical analysis, we've demonstrated that galaxies can indeed be clustered based on spatial characteristics, challenging previous assumptions about their distribution. Our exploration into determining the optimal number of clusters using the "elbow" method has allowed us to mitigate overfitting, ensuring that our clusters represent meaningful patterns while achieving a balance in cluster configuration for enhanced accuracy.

Despite significant advancements, further progress is needed in galaxy clustering to achieve a precise depiction of the universe's structure. Existing algorithms like K-means lack tailored features for galaxies and require pre- defined cluster numbers. Exploring superclusters through iterative clustering offers insights into hierarchical structures, while utilizing techniques such as K-means on detected clusters reveals organizational relationships. Documenting clusters in a catalog facilitates comparisons with established ones, such as Abell's clusters, guiding further investigations for deeper insights into the universe's intricate dynamics.

Overall, our research contributes to a better understanding of the universe's structure and dynamics through machine learning-driven exploration of higher-order galaxy clustering, promising advancements in cosmology and technology.

REFERENCES

[1]Abell, G. O. 1958, Abell Catalog of Rich Clusters of Galaxies.

[2]Bahcall, N. 1988, Large-Scale Structure in the Universe Indicated by Galaxy Clusters, Annual Review of Astronomy and Astrophysics. 26 631-686

[3]Berlind et al. 2006, Percolation Galaxy Groups and Clusters in the SDSS Redshift Survey: Identification, Catalogs, and the Multiplicity Function, The Astrophysical Journal Supplement Series. 167 1-25

[4]Oort, J.H. 1983, Superclusters, Annual Review of Astronomy and Astrophysics. 21 373-428

[5]Yang et al. 2007, Galaxy Groups in the SDSS DR4. I. The Catalog and Basic Properties, The Astrophysical Journal. 671 153-170