Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm to Identify Galaxy Cluster Members

Mingrui Zhang
Department of Astronomy, Beijing Normal University (Beijing), Beijing 100875, China
541240857@qq.com

Abstract. Galaxies are important structures for studying the universe, and clusters are the physical environment of galaxies. Their study is of great significance for understanding the evolution of galaxies and the distribution of matter. Classification of galaxies into clusters is an urgent subject. How do we classify some observed galaxy data points as clusters? How to ensure the correctness of classification? Based on the results of CoDECS numerical simulation and combining DBSCAN algorithm, this paper attempts to classify the data and compare and explain the results of the three methods. Then, based on the data of Abell 383 cluster, further comparison and analysis of the three methods were made. This research can be a basis on measuring new stars.

1. Introduction
Galaxies are important structures in the study of the universe, and although the universe is homogeneous on a large scale, it is structured in detail. Therefore, the universe can be connected with a network structure. The network structure of the universe is partly a region of rich galaxies, which can be divided into clusters, clusters and superclusters according to the number of gravitationally bound galaxies. Clusters of galaxies are made up of less than 50 galaxies at least as bright as the Milky Way. Clusters of galaxies typically contain hundreds or thousands of galaxies. And the rest of the web of the universe is almost empty. This means that the distribution of matter in the universe is a layered pattern, and by studying different levels of structure, namely the galaxies, clusters and superclusters mentioned above, we can get the state of the universe's evolution.

Observations of the near-field universe show that almost half of the galaxies are in clusters. In the process of evolution, the galactic community enters the rich cluster. Rich clusters are the largest gravitationally bound systems, containing about 10 percent of galaxies. Therefore, galaxy clusters and clusters are the material environment of galaxies, and their study is of great significance for understanding the evolution of galaxies and the distribution of matter.

Cluster analysis, also known as group analysis, is a statistical analysis method to study classification problems and an important algorithm for data mining. There are six common clustering algorithms: k-means (K mean) clustering, mean drift clustering, density-based clustering method (DBSCAN), maximum expectation (EM) clustering using gaussian mixture model (GMM), agglomeration hierarchical clustering, and Graph Community Detection. In astronomy, the commonly
used clustering algorithm for cluster classification is FoF (friends-friends), which classifies points with distances smaller than IFoF into a cluster. This is a density-based clustering algorithm, very similar to the DBSCAN algorithm we will study.

2. Data introduction and processing

2.1. Data introduction

Part of our data comes from the CoDECS project, namely the coupling cosmology Dark Energy simulations project, the project is a large-scale non-standard cosmology n-body numerical simulation, the simulation to follow the law of the evolution of the cold Dark matter particles and baryon, and through the simulation we can study the effect of Dark Energy in the universe structure formation. At this stage, the project includes six different cosmological models.

We according to standard cosmological \( \Lambda \)CDM model, the \( 1024^3 \) total mass is \( 5.84 \times 10^{10}h^{-1}M_\odot \) of dark matter particles, and \( 1024^3 \) total mass is \( 1.17 \times 10^{10}h^{-1}M_\odot \) to simulate the baryons, after simulated evolution, in \( 1(h^{-1}Gpc)^3 \), due to the speed of dark matter particles and galaxies deviation is small, so the available every dark matter particles distribution corresponding to the area of the distribution of galaxies, we randomly selected from these particles after the simulation part, get the right ascension and declination and redshift, used in our data.

In the other part, we used the real data from Harvard University about Abell 383. The Abell cluster table is a star table with about 4000 clusters, among which at least 30 members have a redshift value of \( z > 0.2 \). The data include ascension, declination and redshift.

2.2. Main target

The algorithm that divides the individuals in the data set into several usually mutually disjoint subsets is called "clustering algorithm", in which the individuals belonging to the same class have "features" more similar than those of other classes. The "feature" here can be the nature of the individual itself or the distance between individuals, depending on the method and definition we use.

In this paper, we tried DBSCAN method. The design starting point and working principle of each algorithm are explained below. We need to use these three methods to get the classification results of the data, and then compare with the correct results to choose the optimal algorithm.

We randomly select 2000/300/4000 points from data sources, and obtain their coordinates with respect to the earth based on Hubble's law according to their right ascension, right declination and redshift velocities. Then the positions of 2000/3000/4000 points are input into the clustering algorithm, which is divided into several classes; then we compare the classification results with the correct results, and investigate the accuracy of the algorithm. The distances in the method is Euclidean distances.

![Figure 1. Correct result of the data](image-url)
After completing the classification of the simulated data, we will continue to verify and compare the above methods with all the data of the actual Abell 383 cluster.

3. **DBSCAN analysis**

3.1. **DBSCAN algorithm**
Density-Based Spatial Clustering of Applications with Noise is a clustering algorithm based on data point distribution density. The algorithm can identify the degree of data density and classify the data points in the distribution. At the same time, sporadic data points can be identified as noise, rather than being classified within a class.

DBSCAN requires two parameters: eps and minPts.

The specific steps of the algorithm are as follows:
1. Start at any unvisited point and find all nearby points within the scanning radius eps (including eps) with its distance.
2. If the number of nearby points is greater than or equal to minPts, the current point and its nearby points form a cluster, and the starting point is marked as accessed; otherwise the point is temporarily marked as a noise point.
3. All unmarked visited points in the cluster are treated in the same way.
4. If all points in the current cluster are marked as accessed, the same algorithm is used to deal with the unvisited points and create a new cluster.
5. Repeat steps 1-4 until all points are marked as intra-cluster points or noise points, and the algorithm ends.

From the algorithm steps, it can be seen that DBSCAN does not require the shape of each cluster, but only requires the data points inside each cluster to be "sufficiently dense" (that is, larger than the minimum containing points).

3.2. **Result of the analysis**
Firstly, the random distribution of 2000 points is given (all algorithms below use the same 2000 points, so the distribution of 2000 points is only given here).

![2000 random points](Figure 2)
After being processed by DBSCAN, the result is shown in Figure 3.

![Figure 3. Result of DBSCAN algorithm](image)

It can be found that a total of 5 clusters can be found by reasonably setting DBSCAN parameters, among which the small cluster in the upper right corner can still be found. However, the cluster in the lower left corner is divided into two parts. A closer look reveals that the cluster in the lower left corner has two colors (blue and light purple). We believe that this is due to cluster fragmentation caused by sampling, and it can be assumed that these two clusters belong to the same large, undiscovered cluster. The following are the authentication results of the membership properties of a cluster of 2,000 randomly selected points:

| number   | galaxy center(ra, dec) | mean speed   | speed dispersion | accuracy   | recognized/all |
|----------|------------------------|--------------|------------------|------------|----------------|
| 1627     | (1.570675, -0.000018)  | 35919.132097 | 618.841581       | 0.854369   | 616/721        |
| 246452   | (1.579559, 0.006250)   | 36285.338031 | 63.060225        | 0.541667   | 13/24          |
| 21440    | (1.554077, 1.579559)   | 36160.258990 | 359.307105       | 0.851429   | 149/175        |
| 63719    | (1.587042, -0.013958)  | 36045.412634 | 117.383882       | 0.618421   | 47/76          |
| 175577   | (1.589679, -0.012398)  | 36363.651979 | 50.058224        | 0.769231   | 10/13          |

We can see that the cluster member certification of accuracy is higher, the various clusters accurate certification for the average, two of them is bigger (the member number is 721 and 175) recognition rate is higher, shows that in the area of dense enough, DBSCAN better recognition rate, and for members to the less number of galaxies in the cluster, DBSCAN is recognition rate will have large fluctuations. In this regard, we can adjust the scanning radius and minimum inclusion points of DBSCAN to improve the recognition rate to some extent.

4. Conclusion
To sum up, through the study of simulated data, we believe that DBSCAN algorithm is the most effective and accurate algorithm. By comparing with the correct figure, we can find that the DBSCAN algorithm can accurately identify all classes and eliminate noise interference to a certain extent, which is impossible to be achieved by the KMeans algorithm and the Decision Tree algorithm. DBSCAN method's advantage is its classification standard is classified by density, not limited to the shape of the
classification, and so on these data can get very good result, but according to the real Abell data, we can't confirm that the algorithm must be the best, we need more data to verify the correctness of the algorithm, in addition, for different shapes of each cluster, and in the end we will galaxies classified different target number, we also should be rational to choose the appropriate algorithm, cannot treat as the same.

References
[1] Tiit Sepp, Mirt Gramann. Galaxy structures u2014 groups, clusters and superclusters [J]. International Journal of Geometric Methods in Modern Physics, 2014, 11 (02): 1460010-9.
[2] Yu H, Serra A L, Diaferio A, et al. Identification of galaxy cluster substructures with the Caustic method [J]. Astrophysical Journal, 2015, 810 (1).
[3] Feretti L, Gioia I ~, Giovannini G. Merging Processes in Galaxy Clusters [C] // Merging Processes in Galaxy Clusters. Merging Processes in Galaxy Clusters, 2002.
[4] Baldi M. The codecs project: a publicly available suite of cosmological N-body simulations for interacting dark energy models [J]. Monthly Notices of the Royal Astronomical Society, 2018, 422 (2): 1028-1044.