Galaxy Clusters Reconstruction

A Zarodnyuk, E Trofimova, A Solovyov and D Gradoboev
Laboratory of Methods for Big Data Analysis, HSE University, Moscow, Russia
E-mail: azarodnyuk@hse.ru, etrofimova@hse.ru

Abstract. In the present work, we introduce a machine learning-based approach for galaxy clustering. It requires to determine clusters to provide further galaxies groups’ masses estimation. The knowledge of mass distribution is crucial in dark matter research and study of the large-scale structure of the Universe. State-of-the-art telescopes allow various spectroscopy range data accumulation that highlights the need for algorithms with a substantial generalization property. The data we deal with is a combination of more than twenty different catalogues. It is required to provide clustering of all combined galaxies. We produce a regression on the redshifts with the coefficient of determination $R^2$ equals 0.99992 on the validation dataset with training dataset for 3,154,894 of galaxies ($0.0016 < z < 7.0519$). The application of a modern hierarchical density clustering algorithm for the separation of the groups of the galaxies allows us to obtain a result that is consistent with the work [4].

1. Introduction

During the last few decades, the technique and technology of sky surveys made a giant leap. Nowadays there are a wide variety of telescopes that can survey a sky on a different wavelength: X-ray telescopes, ultraviolet telescopes, optical telescopes, infrared telescopes, submillimetre telescopes, radio telescopes. The information obtained from observations helps scientists to study the evolution and the large scale structure of the Universe. That is why it is so vital right along to update the knowledge of the astronomical objects.

Galaxy clusters are the most massive objects in the Universe held together by gravity. However, the detection of galaxy clusters is not a trivial problem. The thing is we can obtain directly only the coordinates of a galaxy on the celestial sphere – angles of right ascension and declination. To use the redshift as a third coordinate is not correct because the Fingers of God effect (FOG) is where clusters stretch along with a cosmologically significant distance along the line-of-sight. It is caused by a Doppler shift associated with the random peculiar velocities of galaxies bound in structures such as clusters. The large speeds that lead to this effect are related to the gravity of the group by means of the virial theorem, and they change the observed redshifts of the galaxies in the cluster. The deviation from Hubble’s law relationship between distance and redshift is altered, and this leads to inaccurate distance measurements [1,2,3].

The classical method in astrophysics to detect galaxy clusters calls Friends-of-Friends algorithm [11] has been used in cosmology for at least two decades to identify interesting objects and quantify structure in simulations. According to [12] FoF is a special case of the DBSCAN [13] algorithm corresponding to a $MinPts$ parameter of zero.
2. Mathematical statement of the problem
Let $X = \{X\}_{i=1}^N$ be a set of $N$ galaxies and $Y = \{Y\}_{j=1}^M$ be a set of $M$ clusters. Let us have a subset of $X$ is labeled galaxies with increasing subsequence of indexes $\{i_k\}_{k=1}^n (n < N)$ and a subset of $Y$ with increasing subsequence of indexes $\{j_k\}_{k=1}^m (m < M)$ is a set of $m$ clusters. Let $F$ be a class of surjection functions:

$$\forall f \in F \lor \forall X_i \ (i = \{1,..,N\}) \ \exists! f_i : \ f(X_i) = f_i.$$ 

The goal is to find $f \in F$:

$$\forall X_{i_k} \ (k = \{1,..,n\}) \ \exists! Y_{j_l} \ (l = \{1,..,m\}) : \ f(X_{i_k}) = Y_{j_l}.$$ 

3. Data description
The dataset contains 480 features for 4,109,726 of galaxies. The dataset is the composition of the information from SDSS BOSS + eBOSS, 2dfgrs, 6dfgs, Updated Zwicky catalogue, The CfA Redshift Catalogue, Hectospec, LAMOST, LEGA-C, DEEP2, DEEP3, WiggleZ, GAMA DR3, 2dFLenS, DES, UKIDSS, GALEX, SkyMapper, VIKING, KIDS, VST ATLAS, UHS, SDSS, VHS, DECaLS, PANSTARRS, WISE [5].

The most meaningful features include:

- $\text{ra}$, the angle of right ascension (from 0 to 360 degrees)
- $\text{dec}$, the angle of declination (from -90 to +90 degrees)
- $z$, the redshift (positive number)

4. Metrics
A common issue of clustering problems is how to evaluate the performance of the algorithm. This issue arises when a number of clusters are unknown, and there are no ground truth labels to compare sets. In the present research, this problem is vital: we should notice that there are not any well-trusted labels for galaxy clusters. The brands we use as true labels were obtained by FoF in [4]. Here and after we assume, the predicted clusters should correspond to true labels as precise as possible, but we should remember that the true labels are not ground true.

To evaluate the quality of clustering performance as the ground truth, we use labelled galaxies extracted from [4]. The labels are obtained for 400,960 of galaxies in the dataset and 170,617 of them are labelled as a single cluster. And as a measure of the similarity of two clusterings of a set of points, we choose Fowlkes-Mallows index (FMI), Adjusted Rand index (ARI) and Normalized mutual information (NMI).

4.1. FMI
The Fowlkes-Mallows index is defined as follows:

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}},$$

where $TP$ (true positive) is the number of pair of galaxies that belongs in the same clusters in both true labels and predicted labels, $FP$ (false positive) is the number of pair of galaxies that belongs in the same clusters in true labels and not in predicted labels) and $FN$ (false negative) is the number of pair of galaxies that belongs in the same clusters in predicted labels and not in true labels). The score ranges from 0 to 1. A high value indicates a good similarity between two clusters.
4.2. **ARI**

Rand index adjusted for a chance. The Rand Index computes a similarity measure between two clusterings by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true labels. The raw RI score is then "adjusted for a chance" into the ARI score using the following scheme:

$$ARI = \frac{RI - \text{Expected}_{RI}}{\text{max}(RI) - \text{Expected}_{RI}}$$

The adjusted Rand index is a value close to 0 for random labelling independently of the number of clusters and samples. It equals 1 when the clusterings are identical.

4.3. **NMI**

Normalized Mutual Information (NMI) is a normalized measure of the similarity between two labels of the same data $U$ and $V$ (MI):

$$MI(U, V) = \sum_{i=1}^{[U]} \sum_{j=1}^{[V]} \frac{|U_i \cap V_j| N}{|U_i||V_j|} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

Where $|U_i|$ is the number of the samples in cluster $U_i$ and $|V_j|$ is the number of the samples in cluster $V_j$. NMI is a value between 0 (no mutual information) and 1 (perfect correlation). In this function, mutual information is normalized by some generalized mean of $H($true labels$)$ and $H($predicted labels$)$, defined by the average method.

4.4. **Caustics**

As an additional criterion of understanding if the clustering performance is good or not may give us the caustic curves method [7, 9, 10]. Let caustic curve be the dependence of the escape velocity of a galaxy in the cluster on the distance from the cluster’ centre. We may consider caustics as follows: if for the given in-cluster radio a galaxy has an absolute line-of-sight velocity higher than the absolute value of caustic the galaxy is very likely will leave the cluster and it does not gravitationally link with the group.

5. **Proposed method**

For clusterization of galaxies we use HDBSCAN algorithm [8]. We apply clustering on the 3 main parameters: the angle of right ascension (ra), the angle of declination (dec) and the redshift (z).

The merged dataset contains a big amount of missed data. One of the most meaningful parameters, the redshift, contains 23% of unrepresentable data. To reconstruct this missing part of z, we use a regression model, 1

The proposed model is represented by the simple neural network in Figure 1, that consists of 5 fully-connected layers with Rectified linear unit (ReLu) activation function [6].

The following 43 features are chosen for regression task: redshift-related information, aperture photometry through 2, 2.8 5.7, Petrosian magnitudes, Kron magnitudes.

6. **Experiments and results**

There are a lot of single-cluster galaxies in our dataset; thus, it is imbalanced. The clustering results are summarized in Table 1. The most reliable scores are shown by the following metrics: FMI and ARI. The NMI tends to separate all the galaxies to the single clusters: the more groups
Figure 1. Regression neural network architecture.

Table 1. Metrics values on the best clustering performance.

| Metric | Score  |
|--------|--------|
| FMI    | 0.58863|
| ARI    | 0.56737|
| NMI    | 0.95724|

we receive, the higher the value of NMI. That is why we estimate clustering performance by FMI.

The specificity of the problem lies in the fact that repartitioning clusters into subclusters is not as critical as combining several true clusters into a larger one. Figures 2–5 illustrate the errors of the clusterization. Each figure corresponds to one cluster in true labels. Colours show the splitting of the group in predicted labels. Light colours correspond to predicted clusters of size greater than 1. Dark points correspond to predicted clusters of size 1.

Figure 2. True cluster in predicted colors.  
Figure 3. True cluster in predicted colors.

From these pictures, it is seen that the proposed clustering algorithm does not provide the same clusters as true labels. Most of the predicted clusters are almost identical to true clusters,

1 See https://gitlab.com/lambda-hse/galaxies
some of the true clusters were separated by the algorithm, that is not critical to this task.

Figures 6–7 illustrate caustic method to clustering evaluation for blue and orange clusters in Figure 3, respectively. Notice that very few galaxies lies outside caustics [14]. It means that we predicted cluster correctly according to physical laws. The lighter points correspond to greater values of the redshift.

7. Conclusion and perspectives
We propose an approach for the galaxies clusters reconstruction problem. The algorithm applies the composition of the information from more than 25 catalogues. The normalized mutual information of performed clustering is 96%. For further performance evaluation of clustering according to the physical point of view it is planned to use caustics as a base to construct the physics-based clustering metric.

Acknowledgments
We are thankful to A. Ustyuzhanin, I. Chilingarian, I. Zolotukhin, V. Toptun, I. Katkov, A. Afanasiev, I. Zolotukhin, K. Grishin for providing the dataset, their guidance and expertise. The work leading to these results has received funding from Russian Science Foundation under
grant agreement n° 19-71-30020. The research was supported in part through computational resources of HPC facilities at NRU HSE.

References
[1] Jackson J C 1972 A Critique of Rees’s Theory of Primordial Gravitational Radiation (Monthly Notices of the Royal Astronomical Society) 156: 1P–5P. arXiv:0810.3908. Bibcode:1972MNRAS.156P...1J. doi:10.1093/mnras/156.1.1p.
[2] Kaiser N 1987 Clustering in real space and in redshift space (Monthly Notices of the Royal Astronomical Society) 227: 1–21. Bibcode:1987MNRAS.227....1K. doi:10.1093/mnras/227.1.1s
[3] Percival W J, Samushia L, Ross A J, Shapiro C and Raccanelli A. 2011 Redshift-space distortions (Phil. Trans. R. Soc. A) 369, 5058-5067.
[4] Saulder C, van Kampen E, Chilingarian I V, Miesenk S and Zellinger W 2016 The matter distribution in the local Universe as derived from galaxy groups in SDSS DR12 and 2MRS (Astronomy and Astrophysics vol 596) 33, p 33.
[5] VizieR catalogue: https://vizier.cfa.harvard.edu/viz-bin/VizieR-6?-out.form=%2bH-source=J/A%2bA/596/A14/q%2alst_sGrID=%20%20%20%20%2020438, acced 30.10.2020
[6] Nair V, Hinton G E 2010 Rectified Linear Units Improve Restricted Boltzmann Machines (27th International Conference on International Conference on Machine Learning, ICML’10, USA: Omnipress) pp. 807–814, ISBN 9781605589077
[7] Yu H, Serra A L, Diaferio A and Baldi M 2015 Identification of galaxy cluster substructures with the caustic method (The Astrophysical Journal vol 810) 37 p12 doi: 10.1088/0004-637X/810/1/37
[8] McInnes L and Healy J 2017 Accelerated Hierarchical Density Clustering (2017 IEEE International Conference on Data Mining Workshops (ICDMW)) pp. 33-42 arXiv: 1705.07321, doi: 10.1109/ICDMW.2017.12.
[9] Diaferio A 1999 Mass estimation in the outer regions of galaxy clusters (Monthly Notices of the Royal Astronomical Society vol 309) (Pub. Oxford University Press (OUP)) 3, pp. 610–622.
[10] Gifford D, Miller C J, Kern N 2013 Velocity Anisotropy and Shape Bias in the Caustic Technique (The Astrophysical Journal vol 774) (American Astronomical Society) 2 pp 116-129.
[11] Press W H, Davis M 1982 How to identify and weigh virialized clusters of galaxies in a complete redshift catalog (Astrophysical Journal Part 1 vol 259) pp 449-473.
[12] Kwon Y C, Nunley D, Gardner J F, Balazinska M, Howe B and Loebman S. Scalable clustering algorithm for N-body simulations in a shared-nothing cluster https://homes.cs.washington.edu/~magda/papers/kwon-ssdbm10.pdf
[13] Ester M, Kriegel H P, Sander J, Xu X 1996 A density-based algorithm for discovering clusters in large spatial databases with noise (In: Proc. of the 2nd KDD Conf.) pp. 226–231.
[14] Caustics source: https://github.com/giffordw/CausticMass