Clustering algorithm: Sensitivity of mass determination using Abell 3581

Susan Wilson

Supervisor: Dr Oozeer Nadeem Co-Supervisor: Dr Ilani Loubser

NORTH-WEST UNIVERSITY YUNIBESITI YA BOKONE-BOPHIRIMA NOORDWES-UNIVERSITEIT

SKA AFRICA SQUARE KILOMETRE ARRAY

Abstract:

Galaxy clusters have gained impetus with huge amount of multi-wavelength data becoming available online. They are important as they provide a way to study galaxy formation and evolution as well as large scale structure in the Universe. In order to understand the environment and various processes within clusters, we need to characterise these clusters. The richness of the cluster is a crucial parameter which needs to be determined. At the same time this allows us to determine other properties such as velocity dispersion, size and mass. This task is not an easy one and has many possible ways to do so. In this poster a comparison between various methods for clustering are analysed. Using virtual observatory (VO) tools possible candidates are selected and then the KMM algorithm is used to test for multi-modality in the groups to determine the final possible member candidates. The Gaussian Mixing Model (GMM) algorithm is tested for comparison. Another method used is that of hierarchy which relies on catalogs and papers as used by Simbad is also compared. The preliminary results obtained are compared against literature values.

Introduction:

Clustering algorithms can be divided into two main groups, Hierarchical methods and Partitioning methods. Hierarchical methods create a tree decomposition of a database. Partitioning algorithms divide the database into a set of clusters. Two examples of each method will be discussed below with application to Abell 3581, with the IC4374 as the Brightest Cluster Galaxy (BCG).

Hierarchical Methods:

SIMBAD

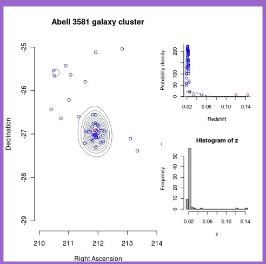

Figure 1: The SIMBAD astronomical database has a hierarchy method which uses information gained from bibliographic references and catalogues. A binned kernel density estimate is applied to the sources in order to estimate the probability density function. The outliers-those lying outside the contours and those with redshifts not in the main group of the histogram- are removed to obtain the member candidates.

Dendrogram

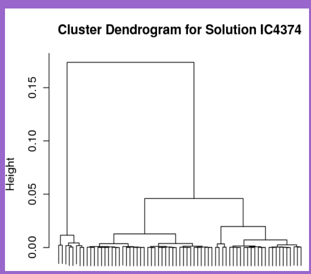

Figure 2: A dendrogram is a tree that splits the dataset into smaller groups until each group contains only one object(Sander+). To determine the number of groups a set level of similarity is chosen and then all lines which cross this level indicate a cluster. The chosen set level of similarity is normally about 50% of the height.

Partitioning Methods:

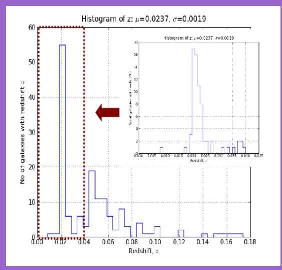

Figure 3: Partitioning algorithms make an initial partition of the database and then use an interactive strategy to make the further partitions, Sander et al. A 3MPc search around the BCG was performed. The data was found from the catalogue Jones+(2009). A histogram of the redshift revealed basic possible groups within this sample. The redshift of the BCG is 0.02 and therefore we select the group in this area. The inset shows this area and the possibility of more than one group. The dendrogram of this group (Figure 2) suggests two possible groups. The KMM & GMM was applied to obtain a better redshift distribution and then a 2MPc search was performed.

KMM

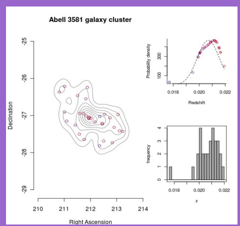

Figure 4: The KMM algorithm works best for homoscedastic data and therefore for the best results from KMM the groups must be checked for different variances using the Levene test. KMM is then run using these different variances to obtain the group of interest and then re-run on this group to check for groups of the same variance. A binned kernel density estimate is applied to the sources and corresponding outliers are removed.

GMM

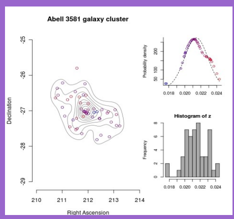

Figure 5: Gaussian mixture modeling (GMM) is a general class of algorithms that KMM belongs too. GMM tests the homoscedastic and heteroscedastic case simultaneously. Unlike KMM, GMM does not specify which group each source belongs to. If the source falls within 1 sigma from the mean it belongs to the group. A binned kernel density estimate is applied to the sources and the outliers are removed to finally obtain the cluster members.

Mass calculation:

For each method: SIMBAD,KMM and GMM- the velocity dispersion, radius and mass were calculated.

The velocity dispersion was calculated using

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} {v_i}^2}{N-1}}$$

Where

$$v_i = c * \frac{z_{mean} - z_{kmm}}{1 + z_{kmm}}$$

The errors on the velocity dispersion was calculated using bootstrapping. Using the virial theorem, the mass and R_200 are given by:

$$M_{200} = \frac{3\sigma^2 R_{200}}{2G}$$

$$R_{200} = 1.73 \frac{\sigma}{1000 km/s} \frac{1}{\sqrt{\Omega_\Lambda + \Omega_0 (1+z)^3}} h^{-1} MPc$$

Methods	No. of Sources	Redshift	Velocity Dispersion km/s	Radius_200 MPc	Mass_200 E13 Solar Masses
SIMBAD	31	0.022	512 ± 54.8	1.25	11.5±3.69
KMM	28	0.021	190 ± 17.3	0.47	0.590 ± 0.16
GMM	47	0.022	389 ± 27.4	0.95	5.06 ± 1.1
Literature		0.020	198 ± 5	1.11	0.662 ± 0.05

Table 1: Shows the various parameters obtained for each clustering methods applied and using the equations given above.

Results/Conclusion:

The various parameters we derived, were used to obtain the mass of Abell 358, which is given in Table 1. Each of the methods give very different results for the cluster mass estimate. Johnstone +(2005) give the radius and redshift using X-ray data obtained from Chandra which basically confines the hot gas around the cluster. The velocity dispersion is obtained from Smith +(2000). The percentage difference between KMM and the literature value is only 3.7% compared to 97% for GMM. The literature values are obtained from spectra and have not been corrected for aperture effects which could result in some error. They also only sampled a few galaxies in the cluster. Piffaretti+(2011) gives R_500 as 0.719 MPc and M_500 as 1.08E14 Solar Masses with data obtained from the MCXC Meta-Catalogue. Comparing our methods with literature values we can rule out SIMBAD as a reliable method as the mass is greater than that for M_500.

Comparing KMM and GMM we can conclude that GMM is more accurate due to the fact that it does not depend on the data being homoscedastic and contains more sources,

A more detailed analysis is being carried out on a sample of 56 sources.

References:

1.Ashman, K.M., Bird, C.M., & Zepf,S.E.,1994.AJ,108:6
2.Gnedin,Oleg,.2012. Email Correspondence
3.Johnstone R.M. et al,.2005.MNRAS,356,237-246.
4.Jones,D.H. et al,.2009.MNRAS,299,683-698.
5.Muratov, A.L. & Gnedin, O.Y,. 2010, ApJ, submitted, arXiv:1002.1325
6.Piffaretti,R. et al,.2011.A&A,534,A109.
7.Sander,J. et al,.1998 Data Mining and Knowledge Discovery archive,Volume 2 Issue 2
8.Scott, D. 1979.Biometrika 66: 605–610. doi:10.1093/biomet/66.3.605.
9.Smith R.J. et al,.2000.MNRAS,313,469-490.
10.This research has made use of the NASA/IPAC Extragalactic Database(NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.
11. This research has made use of the SIMBAD database and Vizier ,operated at CDS, Strasbourg, France

I would like to thank SKA Africa for the funding to complete my Project.