Open University of Catalonia (UOC) Master's Degree in Data Science

MASTER'S THESIS

Area: 4: Data Science

Applying Density-Based Algorithms to Galaxy Groups Catalogs

Unveiling Galaxy Structure with Unsupervised Clustering

Author: Carlos Toro Peñas

Tutor: Laura Ruíz Dern

Professor: -

Madrid, October 6, 2025

Copyright

FINAL PROJECT RECORD

Title of the project:	Applying Density-Based Algorithms to Galaxy Groups Catalogs
Author's name:	Carlos Toro Peñas
Collaborating teacher's name:	Laura Ruíz Dern
PRA's name:	First and last name
Delivery date (mm/yyyy):	MM/YYYY
Degree or program:	Master's degree in Data Sicience
Final Project area:	4: Data Science
Language of the project:	English
Keywords:	Clustering, Galaxy groups, cosmology

Dedication/Quote

Still working on it.

Abstract

This work focuses primary on the application of density-based algorithms to datasets obtained from various surveys, such as the Two-degree Field Galaxy Redshift Survey (2dFGRS) and the Sloan Digital Sky Survey (SDSS). As a result of this application, hyper-parameters tunning and an assessment will be conducted to identify the strengths and weaknesses of these algorithms in actual galactic group detection. In the future, these algorithms may be applied to new surveys and other regions of the sky.

Keywords:Clustering, Galaxy groups, cosmology.

Resumen

Este trabajo tiene como tema central la aplicación de diferentes algoritmos basados en densidad a juegos de datos obtenidos de estudios como Two-degree Field Galaxy Redshift Survey (2dFGRS) y Sloan Digital Sky Survey (SDSS). Como resultado de esa aplicación, se hará un ajuste de híper-parámetros así como una evaluación del desempeño de tales algoritmos para analizar sus fortalezas y debilidades en su habilidad para la detección de cúmulos galácticos catalogados. Futuramente, se podrán realizar aplicaciones de estos algoritmos a nuevos estudios y otras regiones del firmamento.

Palabras clave: Clusterización, cúmulos de galaxias, cosmología.

Contents

List of Figures

Chapter 1

Introduction

1 Justification of interest and relevance

Currently, it has been established that on large scales, the structure of the universe is formed by a vast cosmic web primarily composed of dark matter [1]. The topology of this cosmic web consists of a network of filaments that enclosing large voids [3]. These filaments, comprising mainly of dark matter halos, contain the baryonic matter composed of galaxy clusters and intergalactic matter. The largest and most populated clusters and superclusters of galaxies reside within these dark matter filaments, predominantly at their intersection points, and thus form the largest structures of the visible universe.

Determining the structure of these clusters is therefore crucial for understanding the matter distribution in the universe. This is where clustering algorithms come into play. For reasons that will be discussed further, the density-based algorithms may be the most appropriate for this study.

In this work we want to partially answer how density-based algorithms can be applied to determine how the matter is grouped in the universe.

2 Personal motivation

Given my background and strong interest in astrophysics and cosmology, upon entering the field of Data Science, it is easy to recognize the vast potential for applying the multiple Machine Learning (ML) techniques to these scientific domains. In particular, the study of the large-scale structure of the Universe is, without a doubt, one of the most fascinating topics in science today and where ML methods can find a large number of applications.

3 Goals definition

There is a list of objectives I aim to achieve with this work:

- Generate a visualization map of the data object in this study.
- Apply some density-based algorithm to galaxy and galaxy-groups datasets obtained from SDSS and 2DFGRS.
- Figure out which of these algorithms work better and determine the possible causes.
- Create a validation methods to obtain a hyperparameter tunning that optimize the group detection.
- Detect possible methods to improve this study.

It is also expected to achieve an approximation to actual groups of galaxies through the clusters obtained by the density-based methods.

4 Methodology and project development

Unsupervised algorithms, particularly density-based methods, will be applied to datasets drawn from surveys such as the SDSS and 2DFGRS to generate galaxy clustering models. These datasets are available at [2]:

https://gax.sjtu.edu.cn/data/Group.html

To evaluate the performance of these models, the following criteria will be followed:

- Detected Clusters: Groups successfully classified as clusters (often referred to as True Positives at the group level).
- Undetected Clusters: Groups not found or not identified in the clusters set (equivalent to True Negatives at the group level).
- Cluster Purity Ratio: The proportion of members in a detected cluster that actually belong to the underlying group/structure.
- Cluster Completeness Ratio: The proportion of members of a true underlying group/structure that are successfully included within the detected cluster.
- Misclassified Members: Individual data points (galaxies) belonging to a true group but classified outside of detected cluster. (Often referred to as False Negatives at the individual member level).

- External Data Classified as Members: Individual data points (galaxies) not belonging to a true group but erroneously classified inside a detected cluster. (Often referred to as False Positives at the individual member level).

For this study, R programming language will be used to deploy and run scripts within RStudio environment. Some python scripts can be used as well.

5 Schedule

A Gantt diagram in figure 1.1 shows the different stages of project development. The stages have been grouped on three sets:

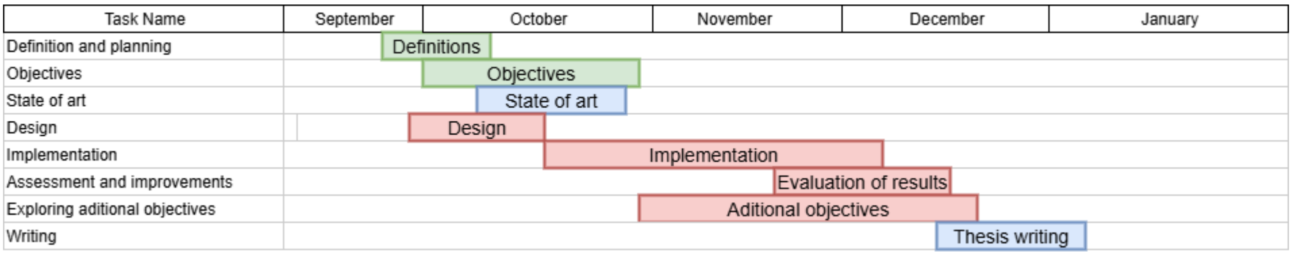

Figure 1.1: Stages of the project.

- The Planning stage (shown in green) involves gathering resources and defining the project's objectives.
- The technical development stage (shown in red) includes design, data processing, method application and outcomes assessment.
- Research and writing stages (shown in blue).

Most of the time there is an overlap of stages, due following:

- Initial stages: starting the project composed of several tasks.
- There are two stages of objectives: one defined at the start of the project and the other during implementation, which may lead to further development.
- Evaluating the outcomes as part of development.

Bibliography

[1] Einasto J. (2014). *Dark Matter And Cosmic Web Story.* New Jersey: World Scientific Publishing Co. Pte. Ltd.

[2] Blanton M. R. et al.(2005). New york university value-added galaxy catalog: A galaxy catalog based on new public surveys. *New York, NY. Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place. p 2562, available at https://doi.org/10.1086/429803.*

[3] Anatole S. et al.(2024). The causal effect of cosmic filaments on dark matter halos. *Lund Observatory, Division of Astrophysics, Department of Physic. Lund, Sweden, p [1-5] available at https://doi.org/10.48550/arXiv.2409.13010.*