University of Sussex
Department of Physics & Astronomy

Using Machine Learning Algorithms to Identify Fast and Slow Rotating Galaxies

Joshua Fenech

Submitted in part fulfilment of the requirements for the degree of
Physics with Astrophysics

Abstract

Text of the Abstract.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Objectives

Motivation and Objectives here.

1.2 Contributions

Contributions here.

1.3 Statement of Originality

Statement here.

Chapter 2

Background Theory

2.1 Introduction

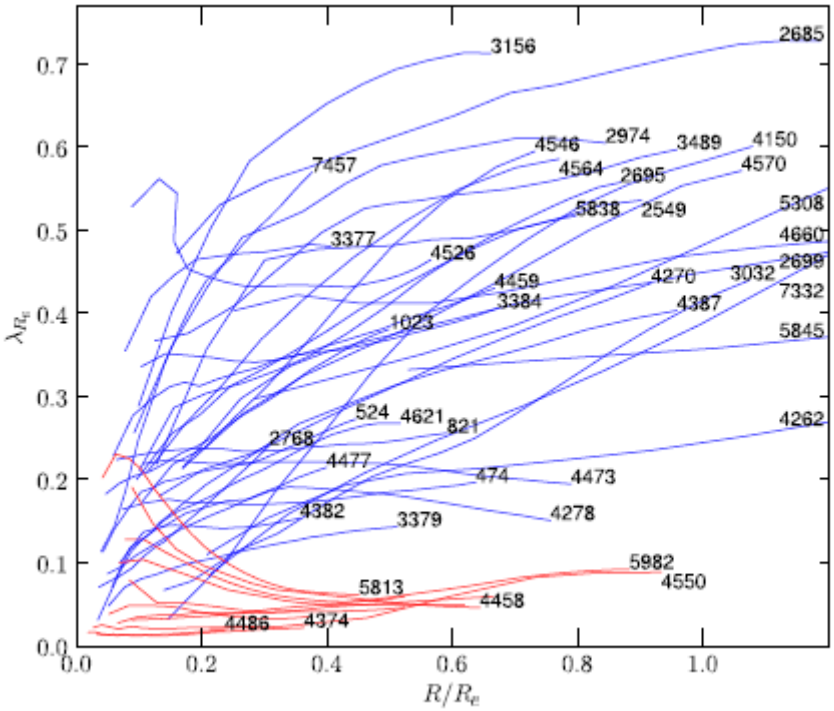
Galaxy morphology has traditionally been classified based on the Hubble sequence, originating from the identification galaxy features from photographic plates. However, this classification is based on visual distinctions and fails to accurately represent early-type galaxies (E's and S0's) and it was argued by [Cap+11] and [Ems+11] that a more telling classification would be based on the spin parameter due to the intrinsic qualitative change in velocity structure exhibited by galaxies, with a threshold of separating slow ($\lambda < 0.1$) and fast (≥ 0.1) rotators, where λ is defined as[Ems+07]:

$$\lambda_R = \frac{\sum_{i=1}^{N_p} F_i R_i |V_i|}{\sum_{i=1}^{N_p} F_i R_i \sqrt{V_i^2 + \sigma_i^2}} \tag{2.1}$$

where F_i, R_i, V_i and $_i$ are the flux, circular radius, velocity and velocity dispersion of the ith spatial bin, the sum running on the N_p bins. This is superior over a velocity dispersion classification, V/σ, which fails when confronted by galaxies with kinematically decouple cores (KDC), "whose angular momentum vector is misaligned with respect to that of the bulk of the galaxy" [MBW10]

The spin parameter is costly to determine due to its reliance on integral field spectroscopy and has therefore only been found for a small number of galaxies: 260 from the ATLAS3D

Figure 2.1: Radial λ_R profiles for the 48 E and S0 galaxies of the SAURON sample. Profiles of slow and fast rotators are coloured in red and blue, respectively. NGC numbers are indicated for all fast rotators and most slow rotators [Ems+11, p.6]



survey and 446 from SAMI. This compares with over 500 million galaxies with photometric data from the Sloan Digital Sky Survey (SDSS) alone [Unk]. Although other classifications do not rely on this parameter, it is still of value in relation to other properties. It is therefore of great value to find an alternative means of identifying the rotation. In order to evaluate the rotation and its effect on observable parameters, it is necessary to consider the current understanding of galaxy morphology. The traditional means of classifying galaxies was based on the Hubble tuning fork that split galaxies into spiral and elliptical galaxies based on their visual morphological appearance. HUBBLE TUNING FORK IMAGE HERE There are several reasons for this distinction. Spirals are generally younger galaxies with a net angular momentum and hence more likely to form discs on a plane coincident with this. FUCKIN REFERENCE There have been a variety of methods of quantifying these classifications, but generally consists of identifying the different components of the galaxy, being the disk and the bulge, and their relative importance. Several (MORE DETAIL HERE) properties of galaxies correlate with their classification, and the subject of this part of the paper will be to briefly describe how this is so (IMPROVE THE ABOVE PARAGRAPH).

2.2 Measured Parameters

The parameters used in modelling the data were:

2.2.1 Variations of the Spin Parameter

The ATLAS3D paper measured λ to 1 effective radius, R_e and to half the effective radius, $R_e/2$, where

$$I(R_e = I_0/e) \tag{2.2}$$

whereas the SAMI paper only measured this for R_e[Cor+16, p. 3].

2.2.2 Sérsic Index of the Single Fit and Bulge Component, n and n_b

The Sérsic profile models the light intensity over the surface of the galaxy in terms of an exponential function as a function of the distance from the centre, R, and the Sérsic index n:

$$I(R) = I_e exp\{-b_n[(R/R_e)^{1/n} - 1]\} \tag{2.3}$$

The range of the Sérsic index covers the full range from steep (i.e. concetrated n»1) to shallow (n$\lesssim$ 1) surface brightness profiles Galaxies

$$\lambda_{Re} = (0.31 \pm 0.01) \tag{2.4}$$

2.3 ATLAS^{3D}

This survey combined According to [Cap+11] this survey focused on a 'volume-limited ($1.16 \times 10^5 Mpc^3$) sample of 260 early-type (elliptical E and lenticular S0) galaxies (ETGs)...The sample consists of nearby (D $<$ 42 Mpc, $|\delta - 29°| < 35°$, $|b| > 15°$) morphologically selected ETG's extracted from a parent sample of 871 galaxies (8 per cent E, 22 per cent S0 and 70 per cent spirals)

brighter than $M_K < -21.5$mag (stellar mass $M_\star \gtrsim 6 \times 10^9 M_\odot$).' ETG's were defined as having de Vaucouleurs T type T > -3.5 and T ≤ -3.5 for S0 and E galaxies respectively, which correlates with the Hubble classes, i.e. lenticulars and ellipticals.

Chapter 3

Methods

3.1 The scikit-learn library

It was chosen to implement algorithms from the scikit-learn module of the python language since it "exposes a wide variety of machine learning algorithms, both supervised and unsupervised, using a consistent, task-oriented interface, thus enabling easy comparison of methods for a given application" [Ped+12]. This allowed several different algorithms to be implemented within the same environment and gain meaningful results with minimal prior coding and analysis. The algorithms initially chosen was decision trees (DT's) since these allow both regression and classification analysis, and are known as 'white boxes' due to their relatively transparent process whereby the mechanics of training could be evaluated more readily.

3.2 Decision Trees

Different machine learning algorithms use different statistical tests in order to evaluate data and make inferences. DT's emulate a logical classification procedure similar to that used in biology to identify species. Starting from the full dataset a series of binary if-then tests are consequentially performed and data split into 2 branches continuously until a final statistical

criterion is satisfied: the sklearn decision tree uses the gini impurity to evaluate the success. The algorithm does this by splitting the data into two branches by choosing the split that minimise the gini index, defined as:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{3.1}$$

arbitrarily choosing a value that splits the parameter space in 2, forming a node and 2 branches. The

Sklearn can use the Gini impurity or the entropy as the determining how to split the tree. The default method of Gini impurity was used here.

$$Q_{left}(\theta) = (x, y)|x_j <= t_m \tag{3.2}$$

$$Q_{right}(\theta) = Q \setminus Q_{left}(\theta) \tag{3.3}$$

The impurity at m is computed using an impurity function H(), the choice of which depends on the task being solved (classification or regression)

$$G(Q, \theta) = \frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta)) \tag{3.4}$$

Select the parameters that minimises the impurity

$$\theta^* = \text{argmin}_\theta \, G(Q, \theta) \tag{3.5}$$

Recurse for subsets $Q_{left}(\theta^*)$ and $Q_{right}(\theta^*)$ until the maximum allowable depth is reached, $N_m < \min_{samples}$ or $N_m = 1$. [SKD]

The impurity measure used is the gini impurity:

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{3.6}$$

The backend of the SKlearn modules were however not evaluated directly but implemented naïvely. SKLearn uses the CART(Classification And Regression Tree) method which builds on the ID3 algorithm originally formulated 1986(PROB DON'T NEED PRECEDING SENTENCE, DOESN'T CONTRIBUTE MUCH). The parameters were initialised to their defaults as described in the documentation[SKD]. Using the modules themselves required elementary use of python to pass the module arrays of necessary values. Implementation proceeded in 2 stages once the data had been suitably formatted. The data was arbitrarily split into 2 groups, a training and a test set, based solely on their position within the results. NEED TO CHECK THEY HAD NO ORDER. For classification, the training set was passed to the module as a list with each item holding the feature value (i.e. Sérsic index) or a list of feature values if more than one feature used, and target variable (i.e. fast/slow rotator classification, spin parameter value). A function was output that incorporated the learned rules which was then applied to the test data set resulting in an array of predicted values that could be measured against the known values. Initially, the algorithm was run as a classification problem due to the more simple analysis of success and errors. This is because it allowed the success to be evaluated in a more elementary fashion, without recourse to analysing error distributions. Decision tree and random forest methods were then applied and the parameters adjusted manually to achieve optimal results. The parameters included:

3.2.1 Parameters

SKLearn allows several parameters to be adjusted manually in order to maximise the efficiency of the models, and these are particular to each. For decision trees, these are as follows:

3.3 Data & Formatting

3.3.1 ATLAS^{3D}

Spectroscopic data (namely Sérsic index of the single fit and bulge component, n and n_b respectively, and D/T, the Disk-to-Total light ratio) was extracted from [Kra+13] whilst the kinematic parameters λ_{Re}, $\lambda_{Re/2}$ and the Fast/Slow rotation classification were extracted from [Ems+11]. The classifier was first trained using the spin parameter λ_{Re} with the FS rotation classification as target variable as a test measure, which successfully predicted 100% of the test set. The algorithm was then trained using n and D/T individually and simultaneously.

Chapter 4

Results

4.1 ATLAS^{3D}

Initially, the algorithm was trained using the λ_{Re} to predict the FS classification since they were related directly according to eq:2.4 (NEED TO FIX REFERENCING EQUATIONS). Promisingly, the classifier was 100% successful in its predictions.

Chapter 5

Conclusion

5.1 Summary of Thesis Achievements

Summary.

5.2 Applications

Applications.

5.3 Future Work

Future Work.

Bibliography

[Ems+07] E. Emsellem et al. "The SAURON project - IX. A kinematic classification for early-type galaxies". In: *Monthly Notices of the Royal Astronomical Society* 379.2 (Jan. 2007), 401fffdfffdfffd417. DOI: 10.1111/j.1365-2966.2007.11752.x.

[MBW10] Houjun Mo, Frank Van den Bosch, and S. White. *Galaxy formation and evolution*. Cambridge University Press, 2010.

[Cap+11] Michele Cappellari et al. "The ATLAS 3D project fffdfffdfffd I . A volume-limited sample of 260 nearby early-type galaxies : science goals and selection criteria 1 I N T RO D U C T I O N". In: 836 (2011), pp. 813–836. DOI: 10.1111/j.1365-2966.2010.18174.x.

[Ems+11] Eric Emsellem et al. "The ATLAS 3D project fffdfffdfffd III . A census of the stellar angular momentum within the effective radius of early-type galaxies : unveiling the distribution of fast and slow rotators". In: 912 (2011), pp. 888–912. DOI: 10.1111/j.1365-2966.2011.18496.x.

[Ped+12] Fabian Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2012), pp. 2825–2830. ISSN: 15324435. DOI: 10.1007/s13398-014-0173-7.2. arXiv: 1201.0490. URL: http://dl.acm.org/citation.cfm?id=2078195%7B%5C%%7D5Cnhttp://arxiv.org/abs/1201.0490.

[Kra+13] Davor Krajnović et al. "The ATLAS3D project - XVII. Linking photometric and kinematic signatures of stellar discs in early-type galaxies". In: *Monthly Notices of the Royal Astronomical Society* 432.3 (2013), pp. 1768–1795. ISSN: 00358711. DOI: 10.1093/mnras/sts315. arXiv: 1210.8167.

[Cor+16] L. Cortese et al. "The SAMI Galaxy Survey: the link between angular momentum and optical morphology". In: *Monthly Notices of the Royal Astronomical Society* 000.August (Aug. 2016), stw1891. ISSN: 0035-8711. DOI: 10.1093/mnras/stw1891. arXiv: 1608.00291. URL: http://arxiv.org/abs/1608.00291%7B%5C%%7D5Cnhttp://mnras.oxfordjournals.org/lookup/doi/10.1093/mnras/stw1891.

[SKD] SKDevelopers. *1.10. Decision Trees*. URL: http://scikit-learn.org/stable/modules/tree.html#tree.

[Unk] Unknown. *Scope of SDSS Survey*. http://web.archive.org/web/20080207010024/http://www.8 Accessed: 2017-02-05.