shikant@lycoming.edu

Galaxy Morphology Classification via K-means Clustering and Support Vector Machines

Kanta Shiromizu

Lycoming College 700 College Pl Williamsport, PA 17701

(Dated: December 4, 2019)

Abstract

I present galaxy morphology classification via k-means clustering and support vector machines (SVM) to classify elliptical galaxies from spiral galaxies by using two data sets, the Galaxy Zoo 1 and the seventh data release of the Sloan Digital Sky Survey (SDSS DR7). I randomly extracted 5,221 elliptical galaxies and 5,221 spiral galaxies for the k-means clustering algorithm, 5,221 elliptical galaxies and 15,641 spiral galaxies for the SVM. The color index, inverse concentration index, and a model magnitude were used as numerical input parameters. Two or three different parameters were got paired, and they were classified by the two machine learning algorithms. The k-means clustering algorithm produced an overall accuracy of 60% to 81%, and the SVM produced 83% - 90% with different pairs of parameters. There was a large difference in the accuracy between the two algorithms, and I confirmed that these were due to the linearly inseparable scattered samples. One of the important features of spiral galaxies, the color profile, was also confirmed from the color index, and the distribution of color profiles is broadly spread from reddish to bluish. In this paper, I will start with the introduction section in which I will explain my motivation and brief background for this research. Then, I will explain about galaxy morphology, data sets, parameters, and machine learning algorithms. Finally, I will introduce my results and conclusions.

PACS numbers: 95.80.+p

Keywords: Galaxy Morphology Classification/ Galaxy Zoo/ SDSS DR7/ K-means Clustering/ SVM

I. INTRODUCTION

Astronomy is experiencing rapid growth in data size and complexity, and machine learning algorithms have become increasingly popular among astronomers [1]. Many researchers have worked on astronomical research using machine learning algorithms. For example, Michelle Lochner worked on photometric supernova classification by using several machine learning algorithms, such as k-nearest neighbors and artificial neural networks [2]. Modern astronomers are required to learn skills to handle a large amount of data so that they can facilitate discoveries.

A. Motivation

There are many galaxies in the universe. Researchers at the University of Nottingham estimated the total number of galaxies in the universe to be two trillion galaxies, which is ten times higher than that would be seen in an all-sky survey at Hubble Ultra-Deep Field depth [1]. Astronomers expect James Webb Space Telescope (JWST) [2] will detect the unstudied galaxies in the early universe including the early stellar galaxies, galaxies that produced light 13 billion years ago [3]. My long-term goal is to study undiscovered galaxies in the early universe and find a clue which explains the origin of our galaxy, the earth, and our existence. As a first step toward my goal, I chose to study galaxy morphology classification to understand the fundamental features of galaxies and find the hidden features of galaxy morphology. I used Galaxy Zoo 1 data and the seventh data release of the Sloan Digital Sky Survey (SDSS DR7) and classified elliptical and spiral galaxies with two machine learning algorithms, k-means clustering and SVM, using a Python library called scikit-learn [4][5].

B. Brief Background

Many pilot studies have been conducted in galaxy morphology classification, and researchers have produced pretty high accuracy, which is close to 100%. For example, Zhu et al. applied deep convolutional neural networks to a sample of 28,790 galaxy images from Galaxy Zoo 2 dataset, and approximately 95% of the overall accuracy of the test set was produced [6]. Barchi et al. worked on galaxy morphology classification and produced an accuracy of 90% with k-means clustering and 97% with SVM [7]. Since this area has been

studied well by other researchers, I focused on following things rather than just producing higher accuracy.

1. Finding unstudied combinations of parameters that have the potential to produce high accuracy.

2. Making high dimensional models with multiple parameters

3. Trying as many machine learning algorithms as possible so that I can improve computational skills.

4. Getting familiar with writing scientific papers to prepare for my future career as a professional astronomer.

My first goal is based on the fact that there are few thousands of combinations of parameters, and it is very important for astronomers to choose good combinations. In order to accomplish the goal, I needed to understand the basic background of galaxy morphology deeply. I also wanted to try machine learning with high dimensional pairs of parameters which are rarely studied by other researchers, and I believe there is a lot of potentials that produces high accuracy. The third and fourth objectives are not for scientific sake but for personal investments to become a professional astronomer.

II. GALAXY MORPHOLOGY

In general, a galaxy consists of hundreds of millions or billions of stars and contains large quantities of interstellar gas and dust [8]. However, there is an enormous variety of galaxies with different sizes, masses, stellar contents, and structures [9]. The first classification scheme of galaxies was proposed by Edwin Hubble in 1926, and he originally classified galaxies into two groups: elliptical galaxies and spiral galaxies [10]. Though his classification scheme is considered to be too simple to describe the details of galaxies, it still holds basic ideas of galaxy morphology classification. After Hubble proposed the classification scheme, several researchers including himself have improved it with various approaches, such as adding details of inner structures and other types of galaxies [11]. Thanks to their research, we are now able to distinguish galaxies into several classes, such as elliptical galaxies, lenticular galaxies, spiral galaxies, merging galaxies, etc. In my research, I used elliptical galaxies

and spiral galaxies and made models that can classify one from another at high accuracy. Since many parameters are needed to describe a certain galaxy type [12], I will introduce only parameters that are used as input parameters. I cited visible-wavelength images of an elliptical galaxy (M87) and a spiral galaxy (NGC 1232) in Figure 1, and these are helpful to understand the features of galaxies.

(a) M87 elliptical galaxy (NASA/ESA)

(b) NGC 1232 spiral galaxy (ESO/IDA)

FIG. 1: Comparison of an elliptical and a spiral galaxy

A. Elliptical Galaxies

As you can see from Figure 1a, an elliptical galaxy shows a much redder overall color and smooth texture. To explain the reason that an elliptical galaxy possesses reddish color, I will first explain the relationship between a star and its color. An elliptical galaxy is composed of older stars, and a long time has passed since the galaxy was born. Also, the mass of a star relates to its color. High-mass stars proceed their nuclear fusion reactions faster than low-mass stars and decrease their temperature after the reaction. On the other hand, the temperature of a low-mass star is constantly low during all life stages. Since lower temperature produces redder color, elliptical galaxies emit much redder color due to the existence of later life-stage of high-mass stars and low-mass stars whose temperatures are low [13].

B. Spiral Galaxies

Spiral galaxies have spiral arms in their outer structures and a central bulge in their centers. These features are not seen in elliptical galaxies. It is observationally confirmed that spiral arms are active star-forming regions and consist of younger stars. However, many spiral galaxies also emit redder color from the center, because the central bulge consists of older stars. Therefore, spiral galaxies have both a profile of red and blue colors. The color balance depends on the composition of a spiral galaxy, and some of them have strong redder colors due to the strong emission from stars in the center.

III. DATA SETS

Galaxy Zoo 1 contains pre-classified galaxy classes, and SDSS DR7 contains a large number of parameters regarding features of galaxies. From the Galaxy Zoo 1 data, I extracted a CSV file that contains 667,946 galaxies whose galaxy types are grouped into elliptical, spiral, and uncertain galaxies. For the SDSS DR7 data, I used the SQL Search System on SkyServer to extract data. I selected several parameters, the magnitude of five filter bands, the Petrosian radius, and the model magnitude, modelMag. I converted the magnitude of five filter bands to color index and the Petrosian radius to the inverse concentration index, which will be discussed in Section IV.

A. Galaxy Zoo

Galaxy Zoo is one of the biggest citizen science projects participated by amateur volunteers from all over the world. In the last 12 years, various galaxy zoo projects have been conducted, such as Galaxy Zoo 2 and Galaxy Zoo: Hubble. In this research, I used data from the original Galaxy Zoo Project, Galaxy Zoo 1, which ran from July 2007 until February 2009 [4]. More than 100,000 volunteers participated in this project and classified galaxies images of SDSS into six classes - elliptical, clockwise spiral, anticlockwise spiral, edge-on, star/don't know, or merger. I extracted a CSV file that contains 667,946 galaxies with several columns, questions asked to participants and galaxy classes confirmed by the votes from the questions. If a galaxy collects more than 80% of all votes to an elliptical (spiral) galaxy, the galaxy will be labeled as "elliptical" ("spiral"). Otherwise, the galaxy will be

labeled as “uncertain."

B. SDSS DR7

Sloan Digital Sky Survey (SDSS) began its operation in 2000 and has produced both imaging and spectroscopic data. The survey uses a 2.5 m telescope at Apache Point Observatory in New Mexico [14]. The telescope equips a large-format mosaic CCD camera that can obtain five different optical bands (u, g, r, i, and z) and two digital spectrographs to get the spectra of astronomical objects [15]. The optical bands allow us to observe the colors of astronomical objects by subtracting magnitudes of different bands from each other, like g-u and r-g [16].

In June 2001, the SDSS released its Early Data Release (EDR), consisting of 462 square degrees of imaging data and 54,008 spectra of objects, to the general astronomical community. People were able to extract data from several websites such as the Space Telescope Science Institute and National Astronomical Observatory of Japan [17]. As of today, the fifteenth data release (SDSS DR15) is the latest data that is accessible [18].

In my research, I used the seventh data release (SDSS DR7) [5], because it includes spectra data of galaxies in Galaxy Zoo 1 data. For the data extraction from the SDSS DR7, I used the SQL Search on the SDSS SkyServer, which allows us to extract only parameters and galaxies that we need [19]. Then, I merged the SDSS DR7 with Galaxy Zoo 1 data by jointing with object IDs of galaxies (ex. 587731187277627676). 10,442 galaxy data (elliptical: 5,221 / spiral: 5,221) are used in k-means clustering, and 20,862 galaxies (elliptical: 5,221 / spiral: 15,641) are used in SVM. The parameters will be discussed in Section V.

IV. PARAMETERS

A. Color Index

As I mentioned in Section II, knowing the colors of galaxies is one of the best approaches to classify galaxy morphology. In this section, I will first explain the magnitude and color filters that are used in SDSS data and introduce the color index. There are mainly two types of magnitude to determine the brightness of astronomical objects, which are the apparent magnitude and absolute magnitude. The apparent magnitude is the number that measures

the brightness of a star or galaxy seen from the earth, and absolute magnitude expresses the intrinsic brightness of an object. When I use the word "magnitude" in this paper, it always indicates the apparent magnitude.

The magnitude is calculated with the following equation:

$$m = -\log_{2.51} \frac{F}{F_0} \tag{1}$$

Where F is the radiant flux, which is the amount of energy arriving at the Earth per unit time from an astronomical object. F_0 is the standard for the magnitude system, and the radiant flux of Vega, F_{Vega}, is used in SDSS.

The values of each of the five color filters (u, g, r, i, and z) represents the magnitude of each color. The green (g) and red (r) are magnitudes of optical lights, and ultraviolet (u) and two infrared wavelengths (i and z) are electromagnetic waves that are invisible to people. Suppose g=10, r=0, and therefore g-r=10 is observed from a galaxy, the ratio of red flux and green flux is calculated from the equation (1) by putting the value of $g - r$ into m, which becomes $10 = -log_{2.51}\frac{F}{F_{Vega}}$. We call the difference of magnitudes of two color filters as "color index." Please keep in mind that apparent magnitudes decrease as flux increases, thus a star with a higher g-r value is redder than a star with a lower g-r value. From this theory, we can classify elliptical galaxies (redder) from spiral galaxies (usually bluer).

B. Inverse Concentration Index

The concentration index of the light distribution of a galaxy is known to correlate with the galaxy morphology, and the emission of light from elliptical galaxies is more concentrated in the center than that from spiral galaxies [20]. I used the inverse concentration index, which is defined by the ratio of the half light Petrosian radius to the 90% light Petrosian radius [21]. To understand the inverse concentration index, I will explain the Petrosian radius and Petrosian flux.

In general, It is difficult to measure the flux of a galaxy, because it does not have the same radial surface brightness throughout its structure. The Petrosian radius and flux make the measurement more easily by introducing the idea of the local intensity I(R) and average intensity within a radius I_a(R). The average intensity is calculated with the following

equation:

$$I_a(R) = \frac{\int_0^R I(r) 2\pi r dr}{\pi R^2} \tag{2}$$

This equation allows us to know how much average intensity of light is emitted from within a radius R. The local intensity I(R) is the intensity at radius R and the local intensity at the Petrosian radius satisfies the following equation.

$$I(R_P) = \eta \frac{\int_0^{R_P} I(r) 2\pi r dr}{\pi R_P^2} \tag{3}$$

This equation tells that the local intensity equals some value η times the average intensity at a certain radius, and we call this radius the Petrosian radius. Ideally, we want to use η = 1 for simplicity, but the Petrosian radius turns out to be very small when we choose it. Therefore, we need to choose a smaller value of η so that we can define the Petrosian radius farther from the center of a galaxy and make a better measurement. The SDSS assigned the value η as 0.2.

Petrosian flux F_P is defined as the sum of all the flux within κ times the Petrosian radius.

$$F_P = \int_0^{\kappa R_P} I(r) 2\pi r dr \tag{4}$$

Again, SDSS doesn't use $\kappa = 1$ but $\kappa = 2$ for a practical reason. In general, the surface brightness of a galaxy is pretty high at the Petrosian radius, and we cannot obtain accurate data due to the bad signal-to-noise ratio. Therefore, the SDSS chose $\kappa = 2$ so that the Petrosian flux includes more of the galaxy's light and improves the signal-to-noise ratio.

The Petrosian half-light radius R_{50}; the radius which contains half of the Petrosian flux, and the Petrosian ninety-percent radius R_{90}; the radius which contains 90% of the Petrosian flux are introduced to calculate the inverse concentration index C.

$$C = \frac{R_{50}}{R_{90}} \tag{5}$$

If a flux concentrates on the center region, the galaxy should have a smaller R_{50}, larger R_{90} and thus smaller C than the galaxy the flux equally comes from the entire structure. Since the flux of an elliptical galaxy concentrates on near the center, the inverse concentration should be smaller than spiral galaxies.

C. Model Magnitude: modelMag

There are three types of model magnitudes, devMag, expMag, and modelMag. The modelMag uses the better of devMag and expMag in red band as a matched aperture to calculate the flux in all bands. The devMag is associated the de Vaucouleurs model fit, and the surface brightness is calculated with the following equation.

$$I(r) = I_0 \exp\left(-7.67\left(\frac{r}{r_e}\right)^{\frac{1}{4}}\right) \tag{6}$$

Where r_e is the radius of the isophote containing half of the total luminosity of the galaxy. The expMag is associated with the exponential model fit, and the surface brightness is calculated with the following equation.

$$I(r) = I_0 \exp\left(-1.68\frac{r}{r_e}\right) \tag{7}$$

We choose one of the two models of higher likelihood in the red filter, and the model is applied to the other bands while allowing only the amplitude to vary. The magnitude obtained from this process is termed modelMag. The modelMag enables us to measure unbiased colors of galaxies. In particular, the model colors have a higher signal-to-noise ratio for faint galaxies. I used this modelMag as a third parameter.

V. MACHINE LEARNING

Machine learning is a tool for data analysis that creates automatic analytical models based on similarities, patterns, rules, etc. Machine learning algorithms are mainly divided into two groups, supervised machine learning algorithms and unsupervised machine learning algorithms. Supervised machine learning algorithms are trained with a set of input features and targets, and the validity of the trained model is confirmed with testing data. Then, we use the model to predict unknown data based on their features. Linear regression, support vector machines, and neural networks are the examples of the algorithms [22][23][24]. On the other hand, unsupervised learning algorithms discover similarities of input data whose labels are undefined and assign to several labels based on the similarities. These algorithms are mainly divided into three groups, clustering, dimensionality reduction, and anomaly detection [22][23]. For example, the k-means clustering method groups each sample into

several clusters by calculating the distance between each sample to a centroid of a cluster. I used the k-means clustering and SVM in this research.

A. K-means Clustering

As I mentioned in the previous section, k-means clustering is one of the unsupervised learning algorithms. Each cluster has a centroid, and each sample is assigned to the nearest cluster. There are two methods to choose the initial centroids. One way is to randomly place the initial centroids, another way is to place them far away from each other using the k-means++ algorithm. In general, the latter method leads to better and more consistent results than the former one. I used the k-means++ to select the initial centroids, even though the randomly assigned centroids also worked well in my models. The initialization process using the k-means++ and entire k-means clustering algorithm can be summarized as follows [23][25].

1. Initialize an empty set M to store the k centroids being selected.
2. Randomly choose the first centroid $\mu^{(j)}$ from the input samples and assign it to M.
3. For each sample $x^{(i)}$ that is not in M, find the minimum squared distance $d(x^{(i)}, M)^2 = \sum_{j=1}^{m(dimension)}(x_j - y_j)^2$ to any of the centroids in M.
4. Use a weighted probability distribution $\frac{d(\mu^{(p)},M)^2}{\sum_i d(x^{(i)},M)^2}$ in order to randomly select the next centroids $\mu^{(p)}$.
5. Repeat steps 2 and 3 until k centroids are chosen.
6. Assign each sample to the nearest centroid $\mu^{(j)}$, j $\in$ 1,..., k.
7. Move the centroids to the center of the samples that were assigned to it.
8. Repeat steps 6 and 7 until the cluster assignments do not change or a user-defined tolerance or maximum number of iterations is reached.

In general, we need to do find the best suitable values of "k" by using several useful methods, such as silhouette analysis and elbow method [23]. However, I selected k=2 without using these methods, because my goal is to separate data into two classes, elliptical and spiral galaxies.

B. Support Vector Machines

The support vector machines (SVM) algorithm is one of the supervised learning algorithms and widely used in various research areas. The main goal of SVM is to maximize the margin between the support vectors of each class, which are the closest samples to the opposite class. I will explain the theory of SVM using several equations. Figure 2 is useful to understand concepts behind equations.

There are three important terms to explain the margin maximization. They are the negative hyperplane, positive hyperplane, and decision boundary. The two hyperplanes lie on the support vectors of each class. The main goal of SVM is to maximize the distance between the positive hyperplane and negative hyperplane. Each hyperplane is expressed as follows:

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{pos}} = 1 \tag{8}$$

$$w_0 + \boldsymbol{w}^T \boldsymbol{x_{neg}} = -1 \tag{9}$$

$\boldsymbol{w}$ is a weight vector, and $\boldsymbol{x}$ is data point. If we subtract the two linear equations, equation (8) and equation (9), from each other and normalize the resulting equation by dividing by the length of the vector, $||\boldsymbol{w}|| = \sqrt{\sum_{j=1}^{m} {w_j}^2}$, the distance between the positive and negative hyperplane (margin) is obtained as a follow.

$$\frac{\boldsymbol{w}^T}{||\boldsymbol{w}||}(\boldsymbol{x_{pos}} - \boldsymbol{x_{neg}}) = \frac{2}{||\boldsymbol{w}||} \tag{10}$$

In addition there is a condition that has to follow when the algorithm calculates the margin. If $y^i = 1(x^i \in \text{“}A\text{”}class)$ and $y^i = -1(x^i \in \text{“}B\text{”}class)$, the condition is expressed as $y^i(w_0 + \boldsymbol{w}^T \boldsymbol{x}) \geq 1$. The condition tells that all negative samples should be in one side of the negative hyperplane, and all positive samples should fall behind positive hyperplane. In short, the basic concept of SVM algorithm is to maximize this margin under this condition.

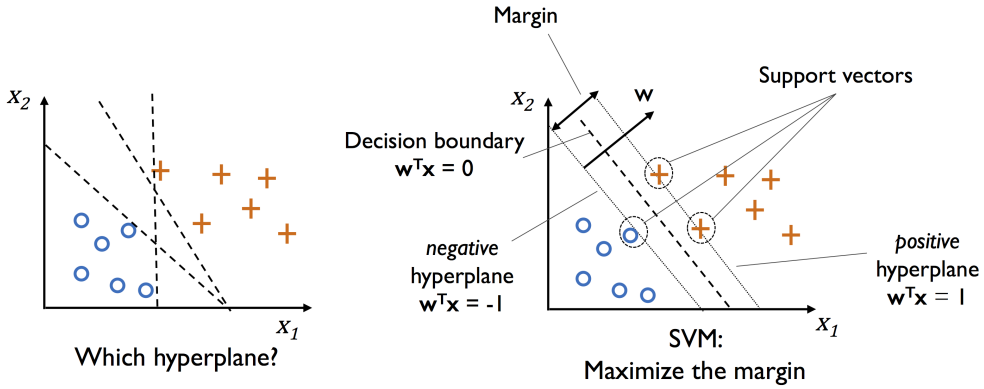


FIG. 2: Margin Maximization [23]

The reason I chose the SVM is that it enables us to classify linearly inseparable data using the kernel methods. The methods create nonlinear combinations of the original features to project them onto a high-dimensional space via a mapping function, ϕ, so that the data becomes linearly separable. For example, as Figure 3 shows, a two-dimensional dataset can be transformed into a new three-dimensional feature space, and the classes become separable linearly with the following projection:

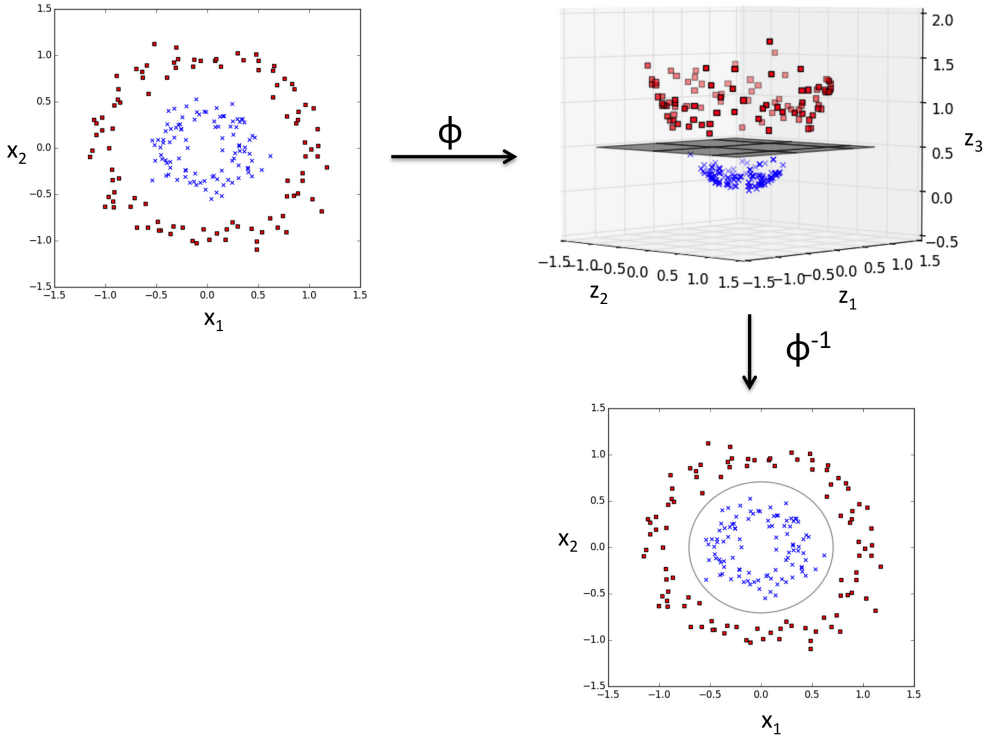


FIG. 3: Kernel method for nonlinearly separable 2-D dataset [23]

This transformation only requires to replace the dot product $x \cdot y$ by $\phi(x) \cdot \phi(y)$ [26]. However, the dot product of ϕ is computationally expensive when we use high-dimensional data. The kernel function: $\kappa(x, y) = (\phi(x) \cdot \phi(y))$ is introduced in order to solve this problem. Several kernel tricks make computation easier and faster. I chose the polynomial kernel and the Gaussian kernel, each kernel is described as equation (11) and equation (12).

$$\kappa(x, y) = (x, y)^d \tag{11}$$

$$\kappa(x, y) = \exp\left(-\gamma ||x - y||^2\right) \tag{12}$$

Here, d is a dimension of data, and $\gamma = \frac{1}{2\sigma^2}$ is a free parameter that we can optimize. For example, when d=2 and x, y $\subseteq R^2$, the polynomial kernel becomes

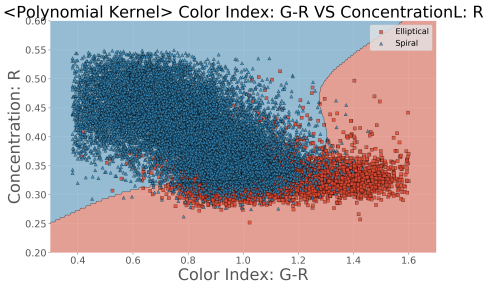
$$(x, y)^2 = \left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right)^2 = \left(\begin{pmatrix} {x_1}^2 \\ \sqrt{2}x_1x_2 \\ {x_2}^2 \end{pmatrix} \cdot \begin{pmatrix} {y_1}^2 \\ \sqrt{2}y_1y_2 \\ {y_2}^2 \end{pmatrix}\right) = (\phi(x) \cdot \phi(y)) \tag{13}$$

In short, the kernel tricks enable the algorithm to do calculations much faster and easier. Although SVMs work for both linear and nonlinear problems via the kernel trick, we need to tune various parameters during a coding process to produce good accuracy. For example, the γ in the Gaussian kernel can be understood as a cut-off parameter for the Gaussian sphere [23]. when we increase the value of γ, the influence or reach of the training samples get increased, which leads to a tighter decision boundary. In contrast, If we decrease the number value of γ, the decision boundary of the Gaussian Kernel SVM model will be soft and smooth. In my research, I used both kernels to classify nonlinear scatter plots of samples.

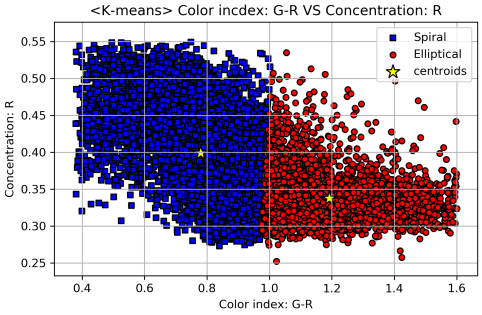
VI. RESULTS

I worked on k-means clustering and SVM with 6 different pairs of parameters. The pairs are $(U - G, G - R), (Con < G >, Con < R >), (G - R, Con < G >), (G - R, Con < R >), (U - G, G - R, Con < R >), and (modelMag < G >, modelMag < R >)$. U-G and G-R are color indices, and Con<G> and Con<R> are inverse concentration indices of green and red bands. modelMag<G> and modelMag<R> are the model magnitudes of the green and red band. The visualized result of a pair, $((G - R), Con < R >)$, is shown in Figure 4. The

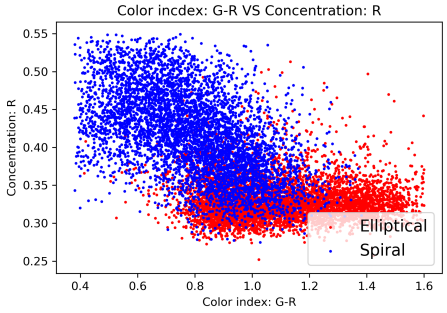
classification with Polynomial SVM was performed better than k-means clustering. Table I shows that other pairs of parameters also produced constantly higher accuracy with SVMs. In contrast, most pairs could not produce more than 80% of accuracy with the k-means clustering, and it is considered to be occurred due to the linearly inseparable data.



(a) Polynomial SVM



(b) K-measn clustering



(c) Scatter plots with actual labels

FIG. 4: Visualization of results $(G-R, Con<R>)$

TABLE I: Summary of the results

	(U-G, G-R)	(Con<G>, Con<R>)	(G-R, Con<R>)	(G-R, Con<G>)	(U-G, G-R, Con<R>)	(modelMag<G>, modelMag<R>)	Average
K-means	80%	81%	75%	75%	76%	60%	75%
Linear SVM	84%	88%	88%	89%	89%	87%	88%
Polynomial SVM	83%	87%	89%	90%	90%	87%	88%
Gaussian SVM	83%	87%	89%	90%	90%	87%	88%
Average	83%	86%	85%	86%	86%	80%	85%

VII. CONCLUSIONS

A. Summary of Results

I obtained an overall accuracy of 60% to 81% from the k-means clustering algorithm. I confirmed that k-means clustering does work for galaxy morphology as long as samples are linearly separable. On the other hand, each SVM algorithm produces high accuracy, 83% to 90%, with all pairs of parameters. The different results between k-means clustering and SVM algorithms are understandable since samples are not linearly separable, and the kernel trick allowed the samples to be classified correctly.

As you can see from the x-axis of Figure 4c, the color index of spiral galaxies are spread from lower values to higher ones, which means some spiral galaxies indicate redder color. This fact confirms the fundamental feature of spiral galaxies about the color profile which I showed in Section II. Normally, spiral galaxies have bluer color in their spiral arms and redder color in their centers due to the different compositions of stars in both regions. Therefore, some spiral galaxies produce much redder color than bluer colors.

Finally, 3-dimensional models, $(U-G, G-R, Con < R >)$, produced almost the same accuracy as other 2-dimensional models did. To confirm whether high dimensional models can produce higher accuracy than 2-dimensional models, I need to work on other high dimensional pairs of parameters.

B. Future Works

I plan to explore further combinations of parameters that would produce higher overall accuracy. Getting 90% of overall accuracy with K-means clustering and 95% of overall accuracy with SVM is a goal. At some point, I will extract imaging data from SDSS and try the convolutional neural networks to analyze the morphology patter of elliptical and spiral galaxies.

Through senior research, I realized the importance of data science. Even with the organized data sets of SDSS DR7, I struggled to choose the parameters due to the lack of my knowledge about the parameters. Therefore, I would like to keep studying galaxies and become a researcher who can understand the background of parameters of various datasets.

I will start my master's program at Leiden University in the Netherlands in the special-

ization of Research Astronomy (MSc). Besides the course works, I will work on two different research projects and write up one master's thesis during a two-year program, so I have to prepare for the program so that I can complete the master's program successfully. Finally, I end the conclusion by listing several things from what I have learned, and which I would like to value on for my future research life.

1. Spending time in reading pilot studies to make sure that planning studies have not been studied yet by other researchers. I could not keep the originality very well in the senior research, but instead, I learned a lot of important skills including extraction of astronomical data using SQL, data preparation and analysis with Python, public speaking, and writing a paper with LaTex.

2. Constantly setting work schedules so that I can see where I am and where I should go. During this senior research, I experienced that the situation had changed many times as my research proceeded. For example, It took a longer time than I had expected to extract data from SDSS DR7 because I had to use the SQL system, which was not familiar with me,d to choose required parameters. Therefore, it is important to make a plan or strategy for both the short and long term in order not to lose a direction.

3. Discussing research with colleagues and professors. I often found myself focusing on my research too much and not spending time in discussion. I think that making discussion with colleagues and professors is one of the efficient approaches to make ideas or solve issues. In graduate school, I would like to talk with people about own research like many theoretical physicists do.

Acknowledgments

I thank Dr. Christopher Kulp for providing me solid feedback and encouraging me to accomplish my research. I also thank my colleagues in PHYS 448 who supported me via discussions and peer reviews. Finally, I appreciate my family for allowing me to study in

the states.

---

[1] C. J. Conselice, A. Wilkinson, K. Duncan, and A. Mortlock, The Astrophysical Journal **830**, 83 (2016).

[2] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel, J. B. Hutchings, P. Jakobsen, S. J. Lilly, K. S. Long, et al., Space Science Reviews **123**, 485 (2006).

[3] *Early stellar galaxies*, URL `https://www.universiteitleiden.nl/en/research/research-projectsscience/early-stellar-galaxies`.

[4] C. Lintott, K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, et al., Monthly Notices of the Royal Astronomical Society **410**, 166 (2010).

[5] K. N. Abazajian, J. K. Adelman-McCarthy, M. A. Agüeros, S. S. Allam, C. A. Prieto, D. An, K. S. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, et al., The Astrophysical Journal Supplement Series **182**, 543 (2009).

[6] X.-P. Zhu, J.-M. Dai, C.-J. Bian, Y. Chen, S. Chen, and C. X. Hu, Astrophysics and Space Science **364**, 1 (2018).

[7] P. Barchi, F. da Costa, R. Sautter, T. Moura, D. Stalder, R. Rosa, and R. de Carvalho, arXiv preprint arXiv:1705.06818 (2017).

[8] M. S. Longair, *Galaxy formation* (Springer, 2008).

[9] W. C. Keel, *The road to galaxy formation* (Springer Science & Business Media, 2007).

[10] E. P. Hubble, The Astrophysical Journal **64** (1926).

[11] S. Bergh and S. Van den Bergh, *Galaxy morphology and classification* (Cambridge University Press, 1998).

[12] P. Brosche, Astronomy and Astrophysics **23**, 259 (1973).

[13] D. D. Clayton, *Principles of stellar evolution and nucleosynthesis* (University of Chicago press, 1983).

[14] D. G. York, J. Adelman, J. E. Anderson Jr, S. F. Anderson, J. Annis, N. A. Bahcall, J. Bakken, R. Barkhouser, S. Bastian, E. Berman, et al., The Astronomical Journal **120**, 1579 (2000).

[15] J. Gunn, M. Carr, C. Rockosi, M. Sekiguchi, K. Berry, B. Elms, E. De Haas, Ž. Ivezić, G. Knapp, R. Lupton, et al., The Astronomical Journal **116**, 3040 (1998).

[16] R. Lupton, M. R. Blanton, G. Fekete, D. W. Hogg, W. O'Mullane, A. Szalay, and N. Wherry, Publications of the Astronomical Society of the Pacific **116**, 133 (2004).

[17] C. Stoughton, R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. Connolly, D. J. Eisenstein, J. A. Frieman, G. Hennessy, et al., The Astronomical Journal **123**, 485 (2002).

[18] D. S. Aguado, R. Ahumada, A. Almeida, S. F. Anderson, B. H. Andrews, B. Anguiano, E. A. Ortíz, A. Aragón-Salamanca, M. Argudo-Fernández, M. Aubert, et al., The Astrophysical Journal Supplement Series **240**, 23 (2019).

[19] A. S. Szalay, J. Gray, A. R. Thakar, P. Z. Kunszt, T. Malik, J. Raddick, C. Stoughton, and J. vandenBerg, arXiv preprint cs/0202013 (2002).

[20] W. Morgan, Publications of the Astronomical Society of the Pacific **70**, 364 (1958).

[21] I. Strateva, Ž. Ivezić, G. R. Knapp, V. K. Narayanan, M. A. Strauss, J. E. Gunn, R. H. Lupton, D. Schlegel, N. A. Bahcall, J. Brinkmann, et al., The Astronomical Journal **122**, 1861 (2001).

[22] D. Baron, arXiv preprint arXiv:1904.07248 (2019).

[23] S. Raschka and V. Mirjalili, *Python machine learning* (Packt Publishing Ltd, 2017).

[24] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems* (" O'Reilly Media, Inc.", 2017).

[25] D. Arthur and S. Vassilvitskii, in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (Society for Industrial and Applied Mathematics, 2007), pp. 1027–1035.

[26] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, IEEE Intelligent Systems and their applications **13**, 18 (1998).