Classification of Stars and Galaxies Using A Support Vector Machine Approach

Wilson Lin
Earlham College
Wlin12@earlham.edu

ABSTRACT

The rate of astronomical data collection is increasing and for the data to be used in scientific analysis, data is required to be quickly and effectually classified. The purpose of this work is to implement a linear support vector machine to discern galaxies from stars. Data from the Sloan digital sky survey is used for both training and testing data.

Using just a linear classifier, a successful classification rate of around 70% was achieved with a training set much smaller than the testing set.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Classifier evaluation

General Terms

Machine Learning, Astronomy

Keywords

Support Vector Machines, Object Classification

1. INTRODUCTION

Astronomy, like other areas of science, is observing a rapid increase in the quantity of available data. Today, the Chandra X-ray Observatory, the Hubble Space Telescope, the Sloan Digital Sky Survey, and the Two Micron All Sky Survey among others, are all rigorously assembling an increasing colossal archive of data. The Sloan digital sky survey currently has around 100 TB of raw data and produces around 200GB of data a night while the Large Synoptic Survey Telescope, which is currently under construction and set to start operations on 2019, is estimated to produce around 30 Terabytes per night [1]. As the progression of technology continues, this rate of data collection will only increase.

In addition to optical information, where the visible light from celestial objects is documented, we are receiving information from a greater range of the electrical spectrum such as radio, infrared, ultraviolet, x-ray, and gamma ray information.

The classification of celestial objects was often done by using the perceived shape of the object, with point sources being identified as stars and elongated objects as galaxies. This works well when the objects are bright and readily visible, however as the brightness decreases, it becomes erroneous to rely on the shape. For example, objects such as ultra-compact dwarf galaxies are often misclassified as stars due to their high mass/light ratio [2, 3].

For such objects, other properties must be used. ColorâĂŞcolor and colorâĂŞmagnitude diagrams are often used for this purpose [4]. Color-color diagrams plot the difference of two magnitudes at different wavelengths, which makes them independent of distance.

Due to the influx of data outpacing the current ability to classify it, astronomers are looking for new methods of classification. One unique way is the âĂIJZooniverseâĂİ, which is an online crowd sourcing project where around a million individual users contribute to science and classification. For example, one of their projects, âĂIJgalaxy zooâĂİ users are shown an image of a galaxy and then answer questions to determine its classification. However, data collection rate is only increasing and becoming more complex. An automated system for classifying objects would allow us to double check current classifications and to classify future data.

An automated algorithm should be able to take into account

the multiple Observations in each of the different recorded wavelengths. Each wavelengths carries important information about the properties of the objects. To obtain the most reliable classification, instead of using a predefined set of features, an automated algorithm should learn the most significant features among the large number of measured ones using a training set, and use the features for the classification task.

One solution to this is a Neural Network. Neural networks do not require any specific knowledge of the data to be used. Neural networks have been shown to be capable of extracting reliable information and patterns from large amounts of data utilizing any amount of dimensions required. They also begun to make strides is astronomy in object detection [5] as well as star/galaxy classification using raw image data and the brightness of pixels to determine galaxy morphology [6].

My goal is to implement a support vector machine(SVM) algorithm that can automatically classify objects once trained on a data set. Like neural networks, support vector machine can utilize data contain any number of dimensions, allowing full use of the available data. They are a newer machine learning model and provide a alternative to neural networks. To train a support vector machine, we give it a series of positive and negative data points, and it separates the data to the best of its ability using a hyper plane. The classification of objects in astronomy is one of the most basic problems, and SVMs show a promising classification strategy.

2. SUPPORT VECTOR MACHINES

The main goal of a SVM is to calculate the most optimal separating decision line, plane, or hyper plane separating two sets of any number dimensional data points. A training data set, consisting of positive and negative instances of the object that we want to detect, is used to provide the SVM with the data points to separate. The SVM calculates the hyper plane between the classes of objects by maximizing the margin, or the distance, between the two classes of data. In addition to linear classifiers, SVMs can generate non-linear classifiers through mapping data points to higher dimensions. These groups of algorithms are called kernel methods. Due to time constraints only a linear classifier was implemented in this paper.

2.1 Applications

SVMs have been growing in popularity in the recent years. SVMs are often used for handwriting recognition [8], voice recognition [9, 10], and have begun making strides in the medical field in areas such as cancer tissue classification [11] and diagnosis [12]. In astronomy, SVMs have been used to identify supernova in astronomical imagery [13] and quantify the morphologies of galaxies based on a dozen of its properties, such as luminosity and redshift [14]. Recently Svms have not only been used to classify objects but also to predict the characteristics of specific objects. For example, the amount of redshift [15]. SVMs offer a new technique that may be widely used in many different areas of astronomy.

The SVM algorithm was implemented to build a linear classifier for photometric data to predict whether or not a source is a star or galaxy.

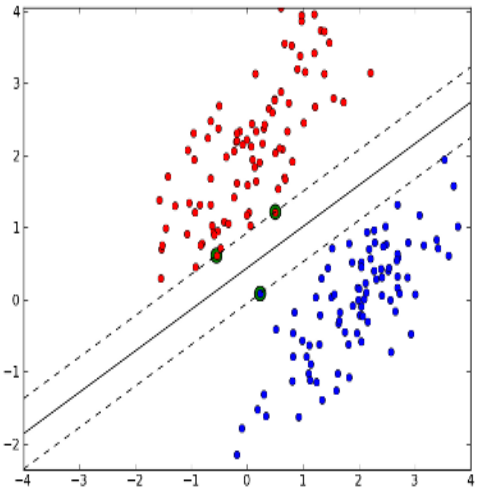

Figure 1: [7] An SVM separates the two classes of data by maximizing the margin or distance between the two training data sets. A new unknown data point can be classified depending on what side of the classifying line that it lands on.

3. DATA

Data from the Sloan Digital Sky Survey was used due to their comprehensive documentation and of their wealth of data. Apendex A shows the spame of the data was was used plotted on a graph.

The SDSS telescope is a 2.5 m f/5 modified RitcheyâĂŞChrÃľtien wide-field altitude-azimuth telescope located at the Apache Point Observatory (APO) at Sunspot, New Mexico. The telescope images the sky by scanning along great circles at the sidereal rate. The telescope is also equipped with two double fiber-fed spectrographs, permanently mounted on the image rotator. Imaging is done in pristine observing conditions (photometric sky, image size FWHM) and spectroscopy is done during less ideal conditions. All observing will be done in moonless sky. Besides the 2.5 m telescope, the SDSS makes use of three subsidiary instruments at the site. The photometric telescope (PT) is a 0.5 m telescope equipped with a CCD camera and the SDSS filter set. Its task is to calibrate the photometry. [16]

The five filters in the imaging array of the camera, u, g, r, i, and z, (rougly standing for ultraviolet, green, red, infrared, and z stands for nothing) have effective wavelengths, in Angstroms, of 3550, 4770, 6230, 7620, and 9130 respectively. [17].

4. RESULTS

A linear SVM algorithm was implemented and applied to training and testing data. When predicting 45,000 data points, with a low quantity of training data (around 250 data points), the program shows around 50% success rate, no better then guessing. However, increasing of the amount available training data, subsequently increases the successful

prediction rate up to around 70% at 2000 data points.

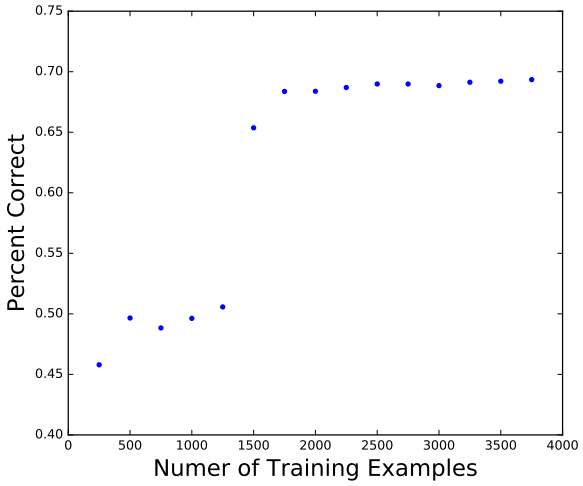

Figure 2: The implementation program results when used on 45,000 test data points with different number of training data points (250-4,000 points at 250 point intervals).

This shows that a relatively small number of training examples can used to predict the classification of a vastly larger data set.

5. CONCLUSIONS

The availability of data is increasing. To best utilize this new influx of data, a efficient method of processing and categorization is required. The results of this study paper show that it is possible in the foreseeable future to categorize astronomical data using SVMs.

6. REFERENCES

[1] Emad Soroush. *Multi-versioned Data Storage and Iterative Processing in a Parallel Array Database Engine*. PhD thesis, 2014.

[2] Steffen Mieske and Pavel Kroupa. An extreme imf as an explanation for high m/l ratios in ucds? the co index as a tracer of bottom-heavy imfs. *The Astrophysical Journal*, 677(1):276, 2008.

[3] EA Evstigneeva, MJ Drinkwater, CY Peng, M Hilker, R De Propris, JB Jones, S Phillipps, MD Gregg, and AM Karick. Structural properties of ultra-compact dwarf galaxies in the fornax and virgo clusters. *The Astronomical Journal*, 136(1):461, 2008.

[4] Igor V Chilingarian and Ivan Yu Zolotukhin. A universal ultraviolet–optical colour–colour–magnitude relation of galaxies. *Monthly Notices of the Royal Astronomical Society*, 419(2):1727–1739, 2012.

[5] S Andreon, G Gargiulo, G Longo, R Tagliaferri, and N Capuano. Wide field imagingâĂŤi. applications of neural networks to object detection and star/galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 319(3):700–716, 2000.

[6] David Bazell and Yuan Peng. A comparison of neural network algorithms and preprocessing methods for star-galaxy discrimination. *The Astrophysical Journal Supplement Series*, 116(1):47, 1998.

[7] Support vector machines in python, url = http://www.mblondel.org/journal/2010/09/19/support-vector-machines-in-python/.

[8] Claus Bahlmann, Bernard Haasdonk, and Hans Burkhardt. Online handwriting recognition with support vector machines-a kernel approach. In *Frontiers in handwriting recognition, 2002. proceedings. eighth international workshop on*, pages 49–54. IEEE, 2002.

[9] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[10] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[11] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[12] Effendi Widjaja, Wei Zheng, and Zhiwei Huang. Classification of colonic tissues using near-infrared raman spectroscopy and support vector machines. *International journal of oncology*, 32(3):653–662, 2008.

[13] Raquel Romano, Cecilia R Aragon, Chris Ding, et al. Supernova recognition using support vector machines. In *Machine Learning and Applications, 2006. ICMLA'06. 5th International Conference on*, pages 77–82. IEEE, 2006.

[14] D Rouan, L Tasca, G Soucail, O Le Fèvre, et al. A robust morphological classification of high-redshift galaxies using support vector machines on seeing limited images-i. method description. *Astronomy & Astrophysics*, 478(3):971–980, 2008.

[15] Dan Wang, Yanxia Zhang, and Yongheng Zhao. Support vector machines for photometric redshift estimation from broadband photometry. *Data Science Journal*, 6:S474–S480, 2007.

[16] Donald G York, J Adelman, John E Anderson Jr, Scott F Anderson, James Annis, Neta A Bahcall, JA Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal*, 120(3):1579, 2000.

[17] M Fukugita, T Ichikawa, JE Gunn, M Doi, K Shimasaku, and DP Schneider. The sloan digital sky survey photometric system. *The Astronomical Journal*, 111:1748, 1996.

APPENDIX

A. PLOT OF DATA

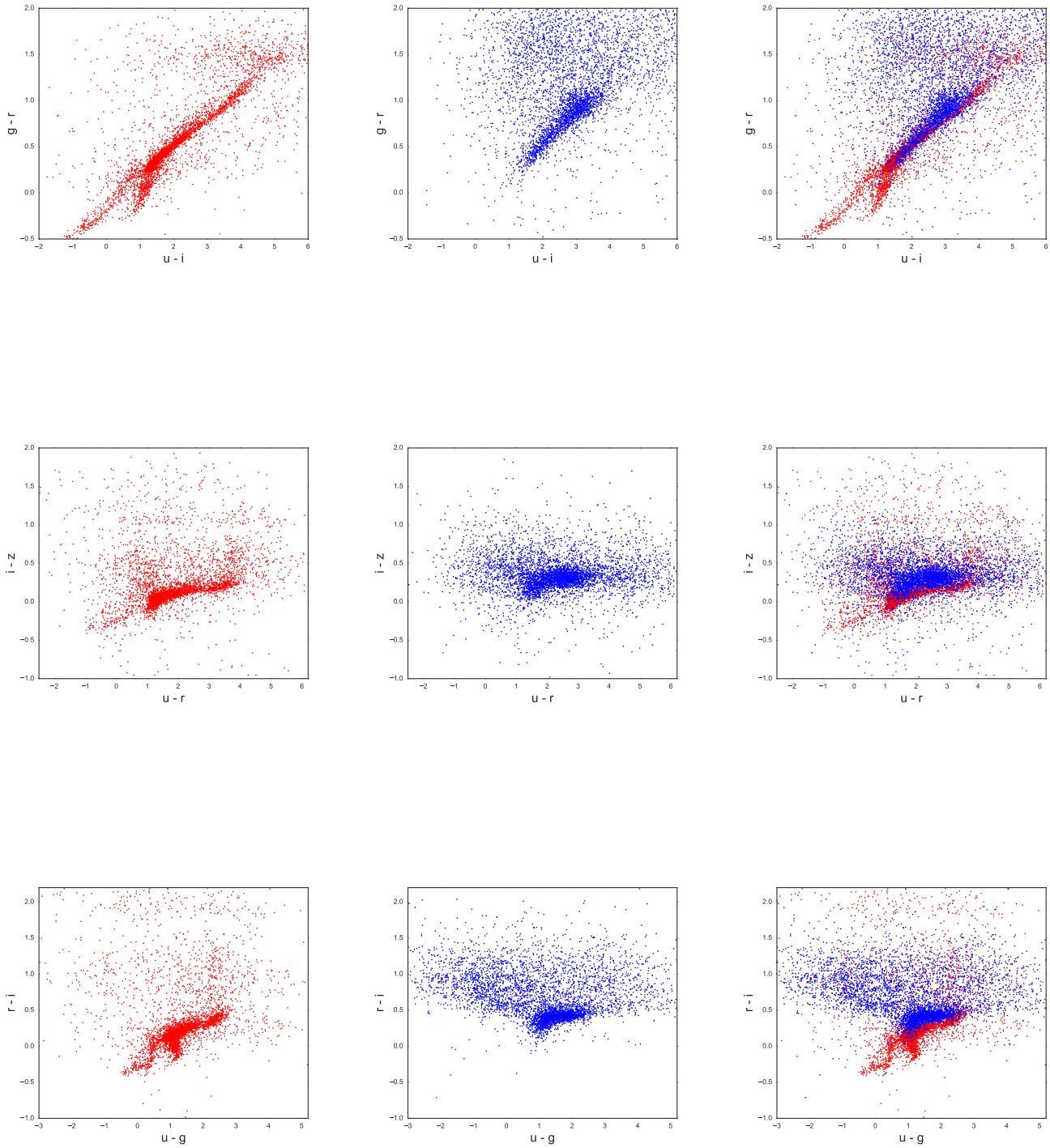

Figure 3: A few color-color plots describing the data set that was obtained from Sloan. The graph on the left show the wavelength differences of stars and the middle graphs show galaxies. The right are the previous graphs overlayed. The data does not look linearly separable, and thus it would be difficult for a linear classifier to show a high success rate for such data. However, it look separable enough to be better than random guessing. The graphs are plotted with python's pyplot package.