# USING MACHINE LEARNING FOR AUTOMATIC CLASSIFICATION OF CLASSICAL CEPHEIDS

A Thesis

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Dallas Kidd

May 2015

USING MACHINE LEARNING FOR

AUTOMATIC CLASSIFICATION OF CLASSICAL CEPHEIDS

_____
Dallas Kidd

APPROVED:

_____
Dr. Ricardo Vilalta, Chairman
Department of Computer Science

_____
Dr. Shishir Shah
Department of Computer Science

_____
Dr. Klaus Kaiser
Department of Mathematics

_____
Dean, College of Natural Sciences and Mathematics

# Acknowledgements

I would like to thank Dr. Vilalta for his encouragement and extraordinary teaching methods that capture the imagination of the students and inspire us to think creatively when solving problems. Dr. Vilalta is a patient, caring person, and he is the reason I am able to graduate as quickly as I have, especially considering the fact that my undergraduate degree was not in computer science. I would also like to thank him for providing this project idea because I have always loved astronomy and am excited that I am able to contribute to the field through computer science and machine learning.

I would also like to thank the Computer Science Girls organization and Dr. Shah, the organization's advisor, for giving me the inspiration and sense of community that I needed in order to be successful in the computer science program. It was an honor to be the Vice President and then the President of the organization while working on my master's degree.

Similarly, I would like to thank Dr. Huang for the support of the GAANN Fellowship, which supported me financially much of the way through and allowed me to pursue my studies with fewer distractions and invest more time in the Computer Science Girls organization.

In addition, I would like to thank my significant other, Bryan Grandy, as without his support, it would have been nearly impossible for me to embark on this new career direction and experience the amount of success that I have.

USING MACHINE LEARNING FOR

AUTOMATIC CLASSIFICATION OF CLASSICAL CEPHEIDS

_____

An Abstract of a Thesis

Presented to

the Faculty of the Department of Computer Science

University of Houston

_____

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

_____

By

Dallas Kidd

May 2015

# Abstract

With the increasing amounts of astronomical data being gathered, it is becoming more crucial for machine learning techniques to be employed for star classification. Classical Cepheid variable stars can be grouped into several classes, such as fundamental-mode, first-overtone, and second-overtone. Each class has distinctive features, and the light curves of the stars can be analyzed for these features in order to be used in automatic classification. Here, we focus on developing a number of features to be used with the following machine learning methods: Multilayer Perceptron, Naïve Bayes, J48 Decision Trees, and Random Forest. We use the OGLE (Optical Gravitational Lensing Experiment) datasets of Classical Cepheid variable stars in the Large Magellanic Cloud and the Small Magellanic Cloud. Our findings indicate that the Multilayer Perceptron is an excellent method for approaching this problem, as it outperformed the other machine learning methods. We also identify a number of useful features using Information Gain and Gain Ratio. Specifically, the newly developed features to measure symmetry had high classification power.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Problem Statement

Astronomy data collection is increasing rapidly, surpassing human ability to manually process it. In order to sift through the massive amounts of both incoming data and the data already collected, we must develop new automation techniques that are efficient and accurate. New discoveries are waiting to be made that will rely on these methods.

Classification is crucial for many applications within astronomy. Determining whether an observed entity is a supernova or a Cepheid variable star or a rare microlensing event is one of the first steps to learning more about the stars. There are a number of challenges associated with automatic classification, but increasing our ability to successfully employ machine learning methods will increasingly benefit us as we collect vast amounts of astronomy data.

## 1.2 Contribution

In this thesis, the focal point is classification of Cepheid variable stars through the use of newly developed features and machine learning methods. Although Cepheid variable stars only make up a very small portion of the astronomy data collected worldwide, they are important astronomical objects. Also, the research serves as a useful way to explore feature development, choosing the best machine learning methods for astronomy problems, and overcoming challenges faced by computer scientists and astronomers.

Specifically, this thesis includes a variety of features that have not been documented in previous literature, such as approaches for measuring the symmetry of

light curves. The goal was not to use previously documented features but to develop new features based on manually examining light curves and observing differences between star classes. These features were developed without many prior expectations for how they would perform. Instead, the approach was to perform a series of experiments and use measures like Information Gain and Gain Ratio in order to evaluate the features.

Many of the techniques and thought-processes that are applied to this problem can also be used on other astronomy problems and even problems outside of astronomy. Machine learning and computer science as a whole touches so many pieces of our lives, which means every contribution in computer science has the potential for many new, unexpected uses in the future.

## 1.3 Thesis Organization

The thesis will be organized into a series of chapters, starting with this introduction chapter. Background information and related work will be presented in Chapter 2 in order to lay the foundation knowledge required for understanding the concepts presented here. Chapter 3 focuses on methodology. In particular, Chapter 3 discusses the feature development process. Chapter 4 explains how the features from Chapter 3 are used in machine learning experiments and the results. The utility of features is also discussed in Chapter 4. Chapter 5 includes a discussion of experiment results, including an exploration of reasons for why some stars were misclassified in the experiments. Chapter 6 concludes the thesis with a discussion of limitations and approaches for future work, as well as a quick recap of what was presented in the thesis.

# Chapter 2

# Background and Related Work

## 2.1 Astronomy Background

### 2.1.1 Data Growth and Automation Need

New scientific instruments and sensor networks are emerging and generating astronomy data streams that are turning into an issue of Petascale computing, and utilizing this new information rapidly will create opportunities for new discoveries. This means we would be observing hundreds of thousands of transient events every night [11]. Transients can be defined as “all genuine non-moving objects that brighten by a certain amount” [15]. The *Gaia* mission and the Large Synoptic Sky Survey will be collecting over a billion periodic variables [6, 7]. But in order to actually use this huge amount of new information, especially in a real-time manner, it will be essential to use automatic processes. This could have an enormous impact across astronomy, such as by warning us about asteroids that could be hazardous or by detecting extrasolar planets with microlensing flares [11]. The imaging technology being used in astronomy has changed observation methods to be more like “making movies of the sky” rather than taking static snapshots sporadically [10]. Furthermore, the amount of data to be analyzed doubles every 12-18 months [12].

In order to meet the challenges associated with the increasing amount of information, researchers are taking a wide variety of approaches. Currently, there is technology available like robotic telescopes that gather, transport, process, photometer, and store data autonomously. It is beneficial to automate the process because machines

are faster than humans, experimentation can be done more easily because it only involves rerunning code instead of asking people to do it, it is deterministic and repeatable, and much more calibration is possible [8]. More about the current challenges will be explored later in this thesis.

### 2.1.2 Variable Stars and Cepheids

Variable stars, which include the classical Cepheids that will be discussed here, change brightness over time. Analyzing various aspects of their light curves yields interesting information that is useful for classification. Light curves are plots that show the relationship between magnitude and time for stars [20]. It has been said that "the time domain is rapidly becoming one of the most exciting new research frontiers in astronomy" [11]. With the large amounts of information that currently are being accumulated and will accumulated in the future, it is clear that new ways of analyzing star information must be devised.

Cepheids play a unique part in history because their behavioral patterns have helped astronomers gain an understanding of the size of the universe, thanks to Henrietta Leavitt who, while plotting and examining the stars' light curves, discovered there is a relationship between their period and luminosity. Brighter mean magnitude stars have longer periods. A little later, Harlow Shapley used what she had discovered in addition to the absolute magnitude of a Cepheid (his discovery) in order to figure out distances in our galaxy. Also, Edwin Hubble used these discoveries for calculating distances to galaxies nearby [4, 5].

Classical Cepheids, which are also known as Delta Cephei stars, type I Cepheids, and Population I Cepheids are very large, bright stars [2], usually 3-9 times the mass of our Sun [5] that "brighten and fade with clockwork regularity" [3] that is akin to

"breathing." As the stars go through these "bulk pulsations" [5], they expand and contract, and they are brightest when they are the smallest (densest) [3].

Classical Cepheids can oscillate or pulsate in several different modes -- fundamental, first-overtone, second-overtone, or a combination of these. Fundamental-mode Cepheids often have periods that are a few days long, and their light curves tend to be asymmetric, although there are some exceptions [2]. The fundamental-mode of oscillation is when the entire star's surface is either completely moving outwards or completely moving inwards at the same time [5]. The first-overtone Cepheids have much more symmetric light curves, and they tend to have smaller amplitudes in their light curves than the fundamental-mode Cepheids [2]. The first-overtone pulsators can also be said to be oscillating in the "first harmonic" mode, and this occurs when one of the star's regions that girdles the surface (like the equator) all moves outwards at the same time as other parts (like the poles) all move inwards. This causes the star to oscillate with a higher frequency. "Second harmonic" and "third harmonic" occur when there are more "separate patches of the stellar photosphere" that "move simultaneously outwards while adjacent patches are simultaneously moving inwards" [5]. Second-overtone pulsators are rare and have "nearly sinusoidal light curves" and small amplitudes [1, 2, 16], below 0.1 mag" [2]. Second-overtone pulsators are also interesting because they are able to be used "as an independent test of pulsational and evolutionary models" [1]. Triple-mode Cepheids are even rarer than second-overtone Cepheids [2]. Multimode pulsators are "very valuable" due to the fact that "each mode gives independent constraints on stellar parameters" [16].

The stars experience oscillation because the star has a "cyclic predominance of gas pressure wanting to make the star expand followed by gravity wanting to make it contract." Over time, oscillation systems lose energy gradually and stop oscillating

eventually (like a pendulum), often due to frictional forces. But the stars continue to oscillate because each has an ionization layer, which acts like a heat valve [5].

### 2.1.3 OGLE, LMC, and SMC

OGLE, the Optical Gravitational Lensing Experiment, is a wide-field sky survey, and the original motivation for gathering the data was to find microlensing events. The strategy with OGLE is to monitor about 200 million star brightness levels in the Magellanic Clouds and in the Galactic bulge for years, thus gathering large amounts of photometric measurements. There are several phases of the project [1].

A subset of the OGLE data will be used in the experiments performed in this thesis and includes classical Cepheids in the Large Magellanic Cloud (LMC) and the Small Magellanic Cloud (SMC). The OGLE classical Cepheid data for the LMC and SMC was gathered with a 1.3-meter Warsaw telescope in Las Campanas Observatory, Chile, which is operated by the Carnegie Institute of Washington [1, 16]. Also, in order to obtain the OGLE data now available to researchers, it was necessary for a massive search to be undertaken to find Cepheids in the Magellanic Clouds, which required a number of processing steps. For example, when looking for Cepheids in the LMC, “tens of thousands” of light curves were selected for visual inspection [1]. In addition, when searching for Cepheids in the SMC, approximately “6 million stars were subjected to a Fourier-based frequency analysis” [16]. *VI* photometry for both the LMC and SMC was obtained through the use of Difference Image Analysis (DIA) technology [1, 16]. This “is able to perform in dense stellar fields considerably better photometry than the traditional PSF-fitting programs” [1].

The Large Magellanic Cloud (LMC) is our nearest non-dwarf neighboring galaxy. Thus, it is “one of the most fundamental extragalactic targets of modern astrophysics” [1]. Also, at the time of the data collection, OGLE contained the “largest sample of

classical Cepheids detected to date in the LMC and, likely, in any other environment" [1]. In total, 3,361 classical Cepheids are included. Specifically, the sample includes 1,848 fundamental-mode (F), 1,228 first-overtone (1O), 14 second-overtone (2O), 61 double-mode F/1O, 203 double-mode 1O/2O, 2 double-mode 1O/3O, and 5 triple-mode Cepheids [1].

The Small Magellanic Cloud (SMC) is of particular interest because of its part in history. As mentioned in Section 2.1.2, Leavitt worked with Cepheids and discovered the period-luminosity relationship. But it was actually SMC Cepheids that she used to make the discovery [16]. In the SMC OGLE Cepheid data, there are a total of 4,630 variables, which is the largest set identified in any galaxy at the time. Of those, 2,626 are fundamental-mode, 1,644 are first-overtone, 83 are second-overtone, 59 are double-mode F/1O, 215 are double-mode 1O/2O, and 3 are triple-mode Cepheids [16].

The LMC and SMC are very similar, but a primary difference between them is the distribution of Cepheids amongst different Cepheid groups. It appears that there are 2-3 populations of classical Cepheids in the LMC and the SMC. They have different periods and luminosities, and they most likely also have differences in metal abundances and ages. It is also known that "there are many more short-period and fainter Cepheids in each pulsation mode" in the SMC, which can be explained by differing metal abundances in the galaxies. The SMC is a metal-poor environment, but the LMC is more metal-rich. Also, only a few anomalous Cepheid candidates were discovered in the SMC, unlike the LMC. In addition, "spatial distribution of interstellar matter in the SMC is more homogenous than in the LMC" [16].

For each LMC and SMC star, the following information is provided in the data: Cepheid ID, intensity I- and V-band mean magnitude, period in days, uncertainty of the period in days, maximum brightness time in JD, I-band amplitude, and four Fourier

coefficients (R_21, phi_21, R_31, and phi_31) [1, 16]. JD stands for Julian Day or Julian Date, which is commonly used in astronomy to keep track of time in a way that avoids dealing with leap years and other complications [20]. Also, as described in [1], light curves were “fitted by a Fourier series of the order depending on the shape and scatter of the light curve” in order to determine mean luminosities, amplitudes, and Fourier parameters. Using Fourier coefficients is common for “quantitative description of the structure of Cepheid light curves” [1]. Also, for each star, observations are recorded and the following information is given: date and time of observation in JD, the magnitude of the star at the observation time, and the uncertainty of the magnitude recorded [1, 16].

The OGLE dataset has some variations within it, and sometimes individual data values are missing for stars for a number of possible reasons. For example, some of the Cepheids do not have I-band data. One of the reasons this occurs is if the CCD saturation limit is exceeded [1]. Also, the V-band information is sparser than the I-band information [7]. Another issue is that star photometry can be influenced by “strong reddening” or crowding and are marked as “uncertain” in the data [16]. But the LMC and SMC information is still very complete and is very useful for star classification.

### 2.1.4 Challenges

There are a number of challenges that computer scientists face when working in the astronomy field. The amount of data to be processed is rapidly growing, but much of the gathered information has known problems. Scarcity of resources is also a problem; it includes both scarcity of human resources and scarcity of equipment like telescopes. Related to this, the curse of dimensionality poses a significant challenge. It can be difficult to sift through huge amounts of information to identify rare events and to accurately identify legitimate objects of interest, especially with the requirement of being able to process and respond quickly to astronomical events.

Astronomy datasets often suffer from problems like sparsity and heterogeneity, and data may have inconsistent measurement issues [9, 10, 11]. Heterogeneity can come in many forms, such as “differences in cadence, observing region, flux noise, detection limits, and number of observed epochs per light curve” [7]. These differences can make it hard to develop classifiers that perform well on unseen datasets [7]. An example of a measurement issue is if there were a problem with a detector, a measurement in a certain filter could be missing [11]. Similarly, information collected is not always of the same quality for various reasons. For example, the pre-launch Kepler Input Catalog data discussed in [14] needs improvement, but the NASA Kepler Mission light curves dataset was much more useful for classification and identifying new class members.

Noise may be also be present in the data, and the noise may falsely look like transient events, which will make accurate classification even more difficult [8, 11]. Much of the signal classification issue, which identifies if an event is real or not, has been addressed with artificial neural networks and support vector machines, eliminating up to 95% of the image artifacts [11]. Another issue related issue is that rare events are particularly difficult to detect, especially in noisy data. This problem strongly affects microlensing, which is achromatic, time-symmetric, does not repeat, and is “outnumbered by stellar variability by at least a factor of 10 000” [9].

Contextual information in astronomy can be very helpful for classification. For example, if, through analyzing a transient light curve, it is unclear whether it is a blazar, supernova, or a cataclysmic variable star, using contextual information can show that it is most likely to be a supernova if we know that the transient is close to a galaxy. But contextual information can be difficult to use in a manner that can be processed by machines, and new methods are being developed [11]. Finding new ways of integrating

this kind of knowledge with machine learning methods can reduce ambiguity currently present in many classification problems.

Scarcity of resources is a very real, constant problem in astronomy. It is expected that available assets for follow-up observations will not increase to match the increasing amount of data [15]. Because of this, it is necessary to ask what use of resources would yield the maximum information gain and what cost is worth the potential for making a discovery [11, 15]. There are several approaches to handling scarce resources. One way is to focus on real-time processing and follow-up [8], even though this brings challenges of making sure interesting events are not missed but there are few false alarms [11]. This is important because so many potentially interesting events are not receiving follow-up observations, and this can become a significantly worse problem in the future as more information is collected [15]. Another way to deal with scarcity of resources is to preprocess data so that only a reduced amount of data must be examined, as done in [9]. This is related to the curse of dimensionality problem. One approach to deal with it is to only retain features that have strong discriminatory power for classification [12]. There are a number of strategies to select useful features, and some of them will be discussed later.

It is also important to note that developing training and test sets can have associated costs, such as manual labeling of datasets. Different datasets have different costs associated with them, depending on factors like the signal-to-noise ratio for light curves, sparsity, amount of available archival data to help determine classes, and whether or not a star is available for spectroscopic follow-up [6]. There are also different dataset labeling approaches, such as human-scanned, artificial-source constructed, and ground-truth derived. The human-scanned method has humans tediously label possibly thousands of candidates by hand. The artificial-source constructed method involves

inserting artificial events into raw data that must be reasonable likenesses of real events. The ground-truth derived approach removes vagueness and human-scanning non-repeatability problems but can also rely on human labels, as well [8].

Another problem that is a serious issue when working with astronomy data is that the distributions of the training and test data can be different. This creates sample selection bias, which can make it much harder to select the best model when employing automated supervised methods because it biases approaches like 10-fold cross-validation. Normally it would be assumed that the distributions are similar when running machine learning algorithms; but, if they are not, then it is essential that approaches are used to compensate for this [6]. In [6], the following three approaches are taken to combat sample selection bias: importance weighting, co-training, and active learning. They had the best success with active learning.

### 2.1.5 Benefit to Other Fields

Astronomy is only one of many fields being strongly affected by the rapidly growing amount of data becoming available, and so many of the challenges facing astronomy also face other fields. Therefore, the advances that are made in this field may be able to be applied in new, unexpected ways over time. For example, fields like environmental monitoring and security could benefit from real-time data analysis. Developing automated systems driven by machine learning algorithms that efficiently utilize scarce resources is challenging but could have vast benefits in more than just astronomy [11]. For example, with the security field, it is not difficult to draw a comparison between searching for intruders and finding rare microlensing events. The key similarity is the rarity, and it is important to not have false alarms in both fields.

## 2.2 Machine Learning Background

### 2.2.1 Machine Learning Methods

A variety of machine learning methods have been used for classification of stars. Different ways to overcome the challenges faced in astronomy classification have also been explored. Quite a bit of success has already been seen in the astronomy field through the use of computer science techniques, but we are still only at the beginning of incorporating computer science methods into astronomy.

Bayesian methods for classification of events can be useful for dealing with heterogeneity and sparsity in astronomy data, as well as the "curse of dimensionality" [11]. In [11], a small dataset was used with a prototype Bayesian Network model in order to identify six distinct classes. Also in the paper, Naïve Bayes is described as being very useful for dealing with high-dimensional data.

Decision trees can be useful once the light curve information has been converted to a uniform set of feature vectors, as done in [11] with three star classes. But trees can be very sensitive to feature changes, which can significantly change the tree structure [17]. This is significant also for the Random Forest method because that method is a collection of trees.

Gaussian Process Regression is used in [10] with a modeling-based approach in order to deal with challenges presented by imperfect light curves. This method is particularly useful, as it is efficient and, unlike other methods, it can build in censoring when necessary. Thus, it can handle cases when information is missing because an object falls below the detection limit [10]. In [10], feature selection is performed by using the Random Forest method and observing if performance decreases with respect to the Gini Index as features are removed. Identifying features (or predictors) for classification and

then determining the utility of those features is key for successful classification using machine learning methods. This can be accomplished in a variety of ways.

Also in [10], the following five machine learning methods were used: Standard Linear Discriminant Analysis, Recursive Partitioning, Support Vector Machines, Neural Networks, and Random Forest. A combination of classifiers can also be used because different classifiers may be useful for different circumstances. This may be particularly useful for real-time classification of transients [15]. Similarly, using a sequence of neural networks to form a neural network cascade is used in [9] for complex pattern recognition in order to remove variable stars and supernovae based on their light curves in order to identify microlensing events. Neural networks are also used in [19], including the Multilayer Perceptron, which is a class of supervised neural networks. Specifically, the Multilayer Perceptron is a feedforward network, and it is trained using the backpropagation algorithm. Neural networks are particularly useful because no prior knowledge needs to be incorporated in order to use them [19].

### 2.2.2 Training and Testing

Determining methods for developing training and test sets for machine learning experiments is a crucial step, and there are numerous challenges and approaches associated with it. First, it is necessary to avoid using same data for training and testing because then the classification rate would be inflated [10] and overfitting would occur [11]. Instead, we must develop a training set of labeled data and then a separate test set of unlabeled data to use with machine learning methods [7]. This can be done by using two totally different datasets, as done in [7], or it can be done by splitting one dataset into partitions as described below.

One method of dividing the dataset for training and testing is to use 10-fold cross-validation, as used in [11]. In this method, the dataset is divided into 10 partitions

randomly. Ten iterations are performed with one partition retained each time for testing. The other partitions are used for training. At each iteration, a different partition is held out of the training set [11].

Another way of handling this problem is demonstrated in [10] where the dataset was randomly split into ⅔ for training and ⅓ for testing. It is explained in [10] that using one random split was sufficient because they were interested in the relative classification performance and because the sample size was relatively large. But sometimes these methods are not viable options, such as when the dataset is very small, and different approaches must be taken.

# Chapter 3

# Methodology

## 3.1 Overview

In the process of working on this thesis, a variety of features were developed, and they were incorporated throughout several iterations of Python 2.7 code. Python was chosen because it can be useful for rapid prototyping, development, and experimentation.

Several types of features were developed. The simplest features are the raw data features, which directly use the data provided in the OGLE datasets and may perform some basic calculations. There are also fitted curve features, which are less prone to outlier issues. Many features mentioned here have both the raw data version and the fitted curve version for comparison purposes. In addition, two rather different approaches to calculating symmetry were taken. The first calculates the differences between the integral of both halves of the fitted curve. The second approach sums the distances to the center of the light curve from pairs of points along the curve.

The approaches taken in this thesis for developing features were mostly inspired by manually inspecting the light curves and researching the differences between the classes of Cepheids. The newly developed features are not necessarily established features used for classification in astronomy. For example, the second approach to measuring symmetry is an original concept that has not been tested by others. The reason these new approaches are taken in this thesis is because feature

development in the astronomy field is still in the beginning phases, and having new approaches to test is important.

## 3.2 Raw Data Features

Raw data features can be very useful for classifying stars. Although, one of the issues with using these types of features is that outliers can be a problem. For example, if measurement errors occurred while recording the data, those errors would be reflected in the raw data features, which can affect classification accuracy. But approaches can be taken to combat this issue.

Class labels were provided with the dataset. Therefore, for all training and evaluation of performance of the testing set, it was simple to identify which stars were classified correctly and incorrectly.

In order to capture the overall trend of the data, for each star, the minimum and maximum magnitude were used as features. It was important to use the minimum and maximum magnitudes because the different star classes are known to often have different amplitudes. For example, first-overtone Cepheids tend to have smaller amplitudes than fundamental-mode Cepheids [5], as mentioned earlier. Thus, the minimums and maximums should often be different. The minimum phase and maximum phase of each star were also included but mainly just for testing purposes, in order to see if an interesting relationship existed that was unexpected; however, that is unlikely because the phase values generally range from 0 to 1. The reason that the phase ranges from 0 to 1 is because, as directed in the documentation provided with the OGLE data, the process for calculating the light curves is to read in the star's file and subtract the JD and the max JD and then divide by the period to get the

phase. Next, it is necessary to discard the integer fraction of the phase so the values range from 0-0.999.

The mean magnitude and mean phase of each star were also used as features. The idea behind using means is to help avoid outliers. Minimum and maximum values can be thrown off greatly due to outliers, but means are more robust. Outliers might still be a problem, such as when only very sparse information is available, as is the case for the V-band, but the features are still worth exploring and might be useful in conjunction with other, more varied features. When deciding to use these means as features, it was unknown whether or not they would have good classification power.

One issue worth mentioning here is that using these raw data features may not be as effective in this thesis as would be possible because I-band and V-band information was treated in the same manner. What this means is that for each star, two instances are being used for classification, one with I-band information and one with V-band information. Magnitude will be different between the I- and V-band, though, so training a classifier with this information could be conflicting. In the future, a different approach may be taken.

In order to clearly visualize the above described raw data features, a star plot diagram is included on the following page. In the diagram, the blue points are the raw data points, the measurements provided for the star OGLE-LMC-CEP-007. The star class shown here is first-overtone. The I-band information is depicted, which is the denser band of data. The maximum magnitude is clearly shown, but the minimum magnitude is not so clear. With this particular star, there is one outlier present that makes the minimum magnitude look slightly lower than it should. The

reason that it is apparent that this data point is an outlier is because the minimum magnitude should really be occurring at either the minimum or maximum phase, but this data point occurs closer to the middle of the phase.

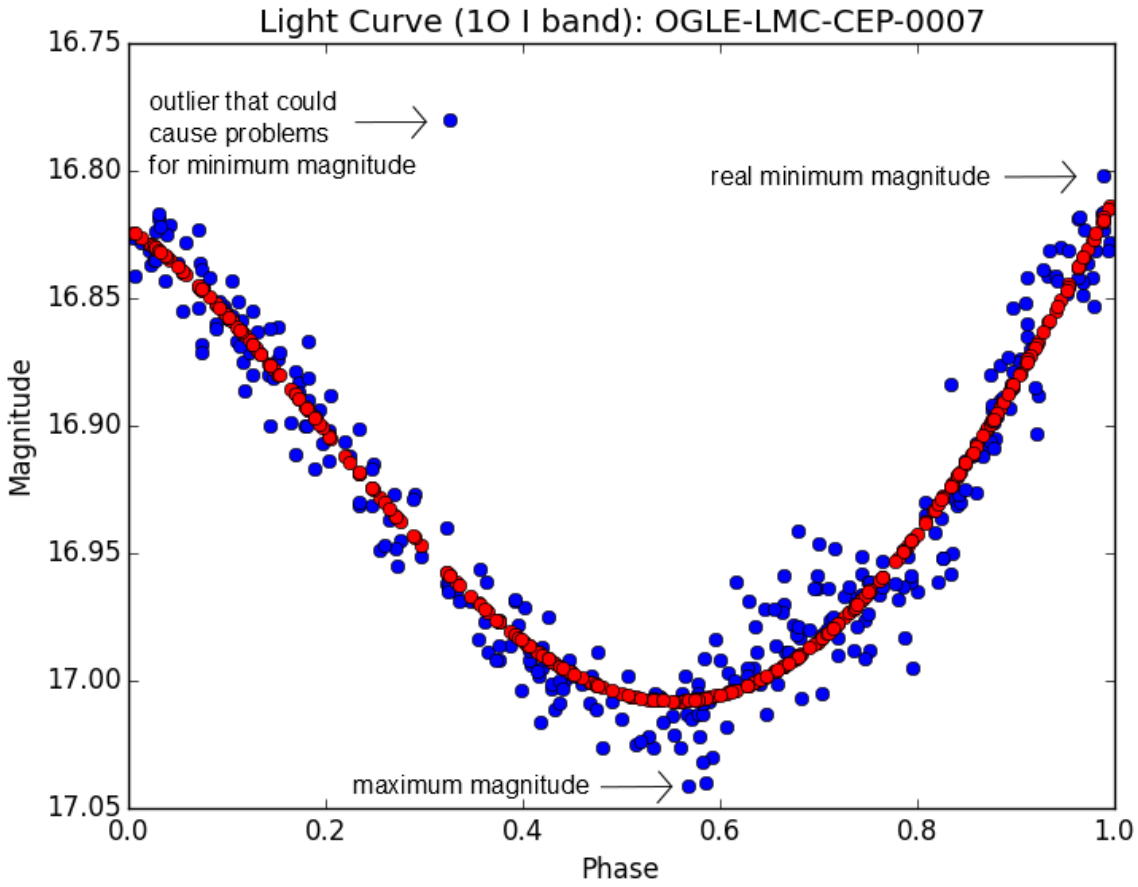


*Figure 3.2: Raw Data Features Example*

As can be seen above, the maximum magnitude of the star is very clear, but the minimum magnitude value may have an outlier issue that slightly skews the feature data. It is important to note that the y-axis decreases towards the top of the graph because this is a standard for light curve plotting.

In addition to the raw data features already described, the x-value (phase value) of the maximum magnitude was incorporated as a feature. Although this feature may sometimes be subject to the outlier problem, it is useful to include the phase data for the

peak magnitude value. One reason for this is because it can help identify if the light curve is centered or not, which may be a measure of symmetry.

The OGLE dataset includes Fourier coefficients and star periods, as well, so these were included as features. The following Fourier coefficients were given in the dataset: R_21, phi_21, R_31, and phi_31. As explained in Section 2.1.3 and in [1], Fourier coefficients are good descriptors of light curves.

Also, inspired by the work in [18], the log of the period given in the dataset was used as a feature, and so was the given mean magnitude. The given mean magnitude may be slightly different from the calculated one mentioned earlier that is calculated with the Python script.

Overall, it is important to have a variety of types of features for machine learning methods in order to capture the most information possible. It is especially important to include many features when testing to see which features are worth retaining. The following section describes the fitted curve features, which may be better at handling data outliers.

## 3.3 Fitted Curve

In order to deal with the outlier issue, a line was fitted to the data using Python's *polyfit* function. A degree of 4 was chosen after some experimentation. Degree of 3 did not fit all of the star's data well enough, but degree of 5 did not seem to be much of an improvement over degree of 4. It seemed prudent to go with the smallest degree possible that fit the data well.

The coefficients of the fitted curve were incorporated as features, in addition to a number of other features involving the fitted curve, several inspired by [18]. The coefficients are numbered from 1-5 and from left to right. The fitted curve version of one

of the raw data features, the x-value (phase value) of the maximum magnitude, was included. Using the fitted curve for this feature helps avoid the outlier issue but still captures relevant information. Also, the depth of the light curve was calculated by subtracting the minimum magnitude from the maximum magnitude on the fitted curve. In addition, two more features were developed by creating two straight fitted lines. The slopes of these straight fitted lines are used as features. The diagram below illustrates this feature.

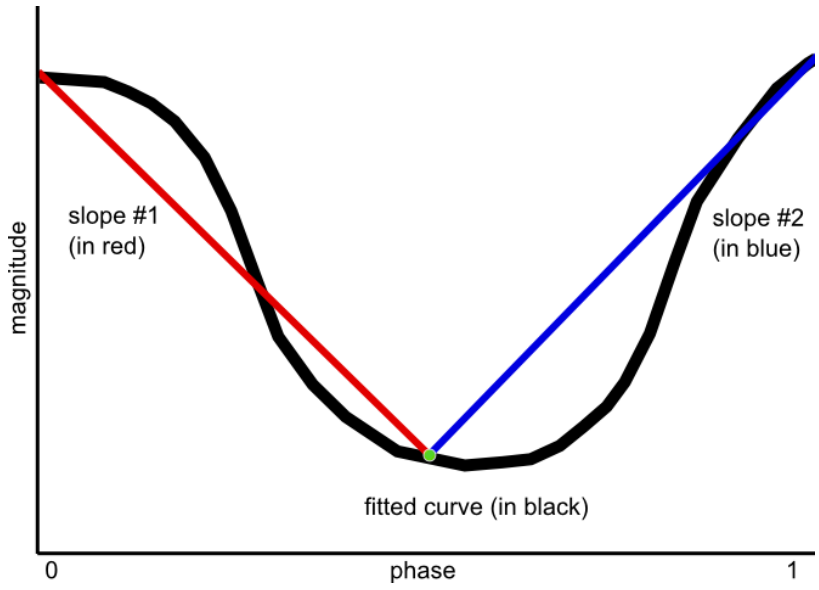


*Figure 3.3: Straight Fitted Lines Example*

In the above image, it is shown that the first straight fitted line is in red and is on the left side of the line fitted to the raw star data. The second straight fitted line is in blue and is on the right side. The slopes of the red and blue line are used for features for each of the stars.

The fitted curve was also used in order to calculate skewness and harmonic mean with Python's SciPy library. In the case of harmonic mean, the flattened fitted curve was used. For skewness, it was used as a preliminary approach to determining symmetry.

Skewness can be very useful for predictions in astronomy, such as for determining whether a source is eclipsing or non-eclipsing [17].

## 3.4 Symmetry, First Approach

Symmetry of light curves is a good way to distinguish between the three classes of stars discussed. Fundamental-mode Cepheids tend to be rather asymmetrical, but first-overtone and second-overtone stars are more symmetrical.

The first approach to assessing symmetry was to use Python's *quad* function in order to calculate the integral of the fitted curve. Specifically, the integral for both halves of the light curve are calculated, and the differences between the two halves are used as features. This procedure is shown in the diagram below and explained more in depth afterwards.

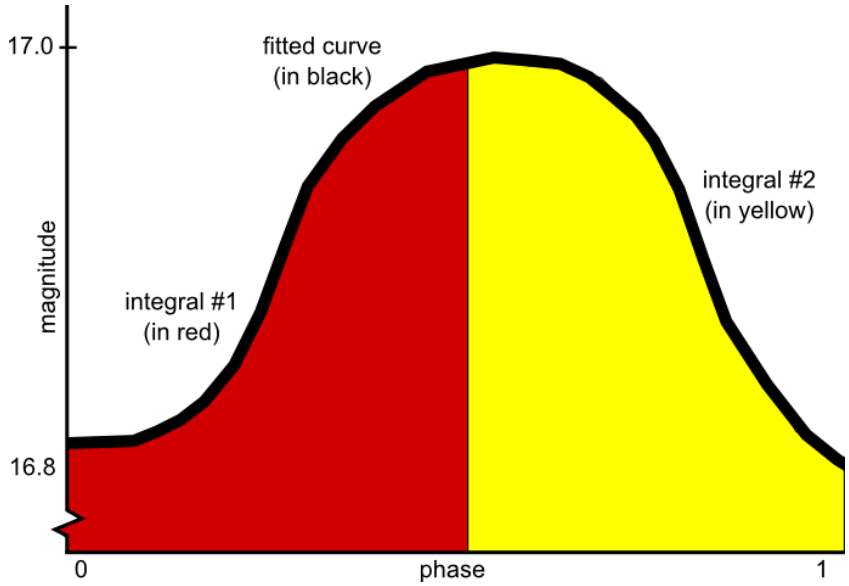


*Figure 3.4: Symmetry Diagram for Approach #1*

In the above diagram, it is shown that two integrals are taken, one for each half of the light curve. The left one is shown in red, and the right one is shown in yellow. Notice that the y-axis focuses on only 16.8 to 17.0, as those are a typical y-minimum and y-

maximum, respectively; but the integral goes all the way down to 0. After the integrals are taken, the differences between them are calculated and used as features. Specifically, the first feature is the difference between the integrals. The second feature is the difference between the absolute error values returned by the integral function. The error difference was included as an extra feature, just in case an interesting relationship occurred.

## 3.5 Symmetry, Second Approach

Because symmetry is so important in distinguishing the classes of Cepheids, it seemed best to focus more on trying to capture that information effectively. In order to do this, the middle of the fitted curve is calculated by iterating through 50 points on the fitted light curve and determining the maximum y-value (the maximum magnitude), which is the approximate middle of the light curve. The reason that 50 points on the light curve were used is because it seemed effective in practice for determining a fairly accurate middle point.

Then, the two halves of the light curves are compared. Starting at the middle of the light curve and working outwards in increments of 0.02, the points on each side of the light curve are directly compared to each other. The distance between these sets of points is taken incrementally, and the differences are summed. The total differences between all the pairs of points are summed in order to represent the symmetry. A lower sum indicates a more symmetrical light curve.

The following diagram illustrates this concept in a simplified way. The center of the light curve is represented by a dotted blue line. Each pair of points is either green or red, and the distance from the point to the center is represented by a straight line of the same color as the point. The number of pairs of points in the

diagram does not exactly match the actual light curve calculations done, as this is just for simple illustration. The important thing here is that the points were determined by a fixed amount along the light curve, not necessarily their y-values.

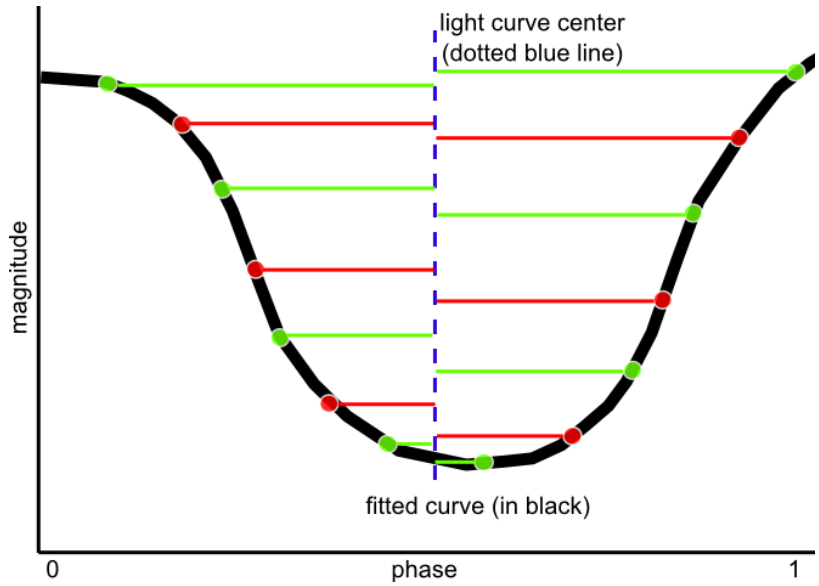


*Figure 3.5: Symmetry Diagram for Approach #2*

The above diagram illustrates a way to approach calculating symmetry that is rather different from the integral function differences used in the first approach. The idea behind using multiple ways to approach symmetry is partially to find out which is best but also to attempt to capture the widest amount of information possible. It could be that using both measures of symmetry is important for accurate Cepheid classification. Having more features available for experimentation is important, and poorly performing features can be ruled out later.

# Chapter 4

# Experiments and Results

## 4.1 Experiment Setup and Background

### 4.1.1 The Data

For the purpose of these experiments, two subsets of the Optical Gravitational Lensing Experiment (OGLE) data are used. This includes Classical Cepheids in the Large Magellanic Cloud (LMC) and Classical Cepheids the Small Magellanic Cloud (SMC). As mentioned previously, the LMC consists of 1,848 fundamental-mode, 1,228 first-overtone, and 14 second-overtone Cepheids. The SMC consists of 2,626 fundamental-mode, 1,644 first-overtone, and 83 second-overtone Cepheids. In the initial experiments, stars with missing data are skipped, so the numbers are a little lower. In the rest of the experiments, very few stars are skipped because missing values are allowed.

Depicted on the next page is OGLE-LMC-CEP-0012, a fundamental-mode Cepheid. In the light curve plot, it can be seen that the V-band is noticeably sparser. The light curve is clearly asymmetrical, as is expected with fundamental-mode Cepheids. The blue dots are the raw data points (the observations) provided in the dataset for this star, and the red dots are the fitted curve points. It can be seen that the fitted curve has a stronger presence in the I-band where the information is denser and a weaker presence in the V-band, where the information is sparser. This is especially noticeable in the V-band near the maximum phase.

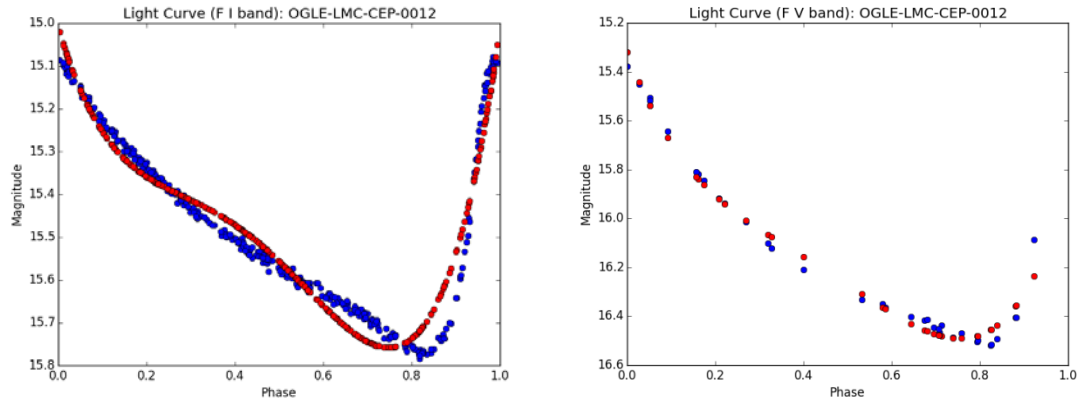


*Figure 4.1.1a: Light Curve of Fundamental-Mode Cepheid*

An example of OGLE-LMC-CEP-0006, a first-overtone Cepheid, is below. The V-band is also much sparser here, but the star below exhibits the characteristic symmetrical nature of first-overtone light curves.

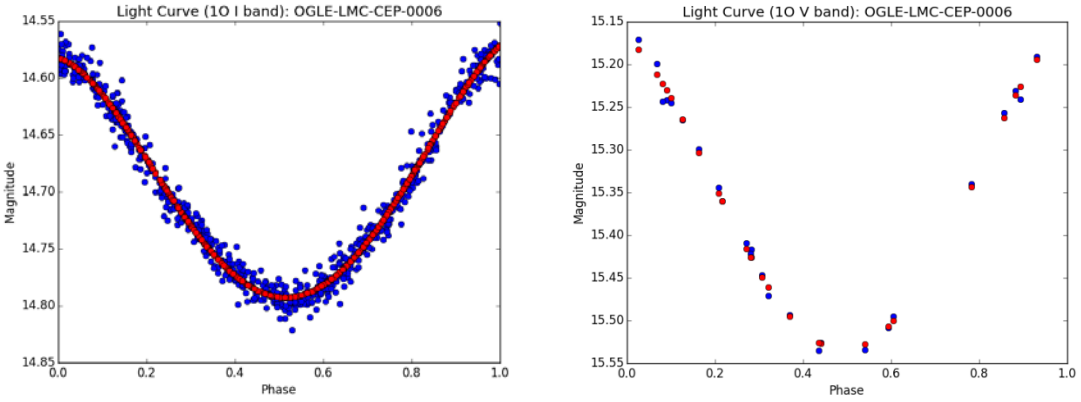


*Figure 4.1.1b: Light Curve of First-Overtone Cepheid*

It is easiest to see in the I-band information for the star above that the light curve is symmetrical, but it is still noticeable for the V-band. Here, the V-band has more data near the minimum and maximum phase values than in the previous example, but there is still a noticeable lack of information around the phase value of 0.7.

Similarly, the second-overtone Cepheid below has a symmetrical light curve, but the amplitude is very small.

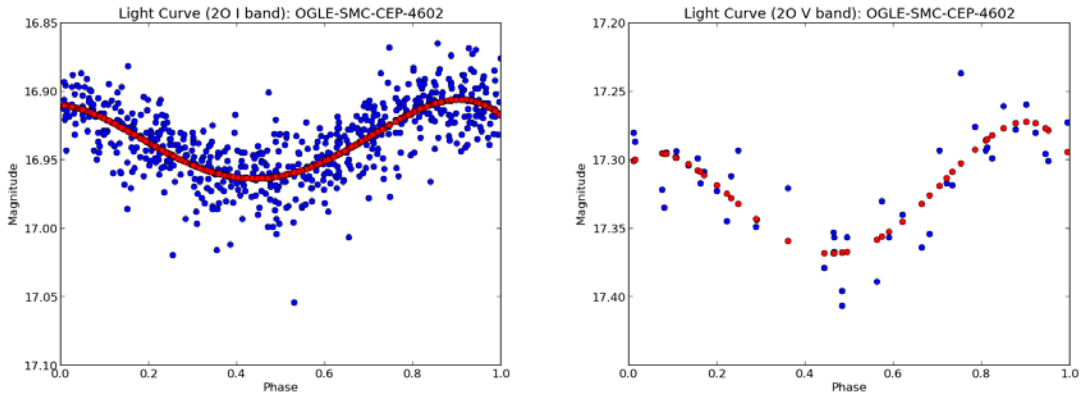


*Figure 4.1.1c: Light Curve of Second-Overtone Cepheid*

In the example above, the typical second-overtone Cepheid is demonstrated. It has the characteristic small amplitude and a fairly symmetrical light curve.

Using these distinctive characteristics of the Cepheid classes as features for the machine learning algorithms will be crucial for accurate classification. If it can be successfully identified that a symmetrical light curve tends to be first-overtone or second-overtone and that the tiniest amplitudes tend to be second-overtone, then the classifiers should do well.

### 4.1.2 Weka

For running machine learning methods and analyzing class distributions, Weka 3.6.9 was used. Weka is an open source Java-based data mining software [13]. It can be used via the command-line or with the graphical interface and has many machine learning algorithms available. It also has methods of determining the value of features, such as Information Gain and Gain Ratio. In order to run experiments in Weka, an ARFF (Attribute-Relation File Format) needs to be generated. In this case, Python was used to generate the ARFF with the feature data separated by commas and the Cepheid variable stars separated by line breaks.

### 4.1.3 Missing Data

In the initial set of experiments, only fundamental-mode and first-overtone Cepheid stars were used, and any stars with missing data were discarded. But once second-overtone Cepheid stars were incorporated, it became clear that this approach would not be effective. This is partially because less information was available for second-overtone stars, but it was also because there were so few second-overtone stars included in the dataset. Thus, every star was important to keep for training or testing, and a method was devised to handle missing data.

As Weka was to be used for running the machine learning algorithms, the method of incorporating missing data needed to be compatible with Weka. A question mark was inserted in place of the missing feature data in the ARFF file in order to signal to Weka that the feature information was missing for that star.

### 4.1.4 Machine Learning Methods

For the machine learning methods, the Weka default settings were used. For example, for the Multilayer Perceptron, a learning rate of 0.3 and a momentum value of 0.2 were used. For J48 Decision Trees, the confidence factor used for pruning was 0.25 was used, and 2 minimum instances per leaf were specified. For Random Forest, 10 trees were specified to be generated, and a random number seed of 1 was chosen. No parameters needed to be specified for Naïve Bayes.

### 4.1.5 General Background

Many experiments were performed in order to determine the effects of the developed features on classification accuracy. Initial experiments focused on fundamental-mode and first-overtone Cepheid variable stars and only included the Large Magellanic Cloud (LMC). Also, since there were plenty of stars in the dataset for these two classes, training

and testing could be performed in multiple ways. After these two groups of experiments were performed, a group of revised experiments was also completed. In these, only the LMC fundamental-mode and first-overtone Cepheids were used (as in the initial experiments). But all 28 features were used, unlike in the initial experiments.

Additionally, measures were used in order to determine the utility of features because later experiments use subsets of the features in order to compare accuracies across machine learning methods. This is in line with avoiding the curse of dimensionality problem mentioned in Chapter 2.

## 4.2 Determining the Utility of Features

To determine the utility or value of features in an unbiased manner, the following two measures were used: Information Gain (IG) and Gain Ratio (GR). These methods were run through Weka, and both measures had fairly similar results. The results are discussed in the table on the following page. The features are ordered in rows by the IG ranking. The second column is the IG value. Next is the GR ranking, which is similar to the IG ranking, but not exactly the same. Then the column after that is the GR value. The feature number and the feature description are in the right-most columns.

The majority of the rankings agree between IG and GR. The first four features, which are the best, are ranked in the same order in both evaluation methods. The best feature is feature number 16, which is one of the Fourier coefficient measures given in the OGLE dataset. The second most highly ranked feature, number 25, is one of the slopes of the fitted straight line. The third, number 15, is a measure of symmetry. Specifically, the second approach to measuring symmetry, as explained in Section 3.5, performed well here. The fourth highest ranking feature, number 22, is the depth of the light curve.

| *IG Rank* | *IG Value* | *GR Rank* | *GR Value* | *Feature Number* | *Feature Description* |
|---|---|---|---|---|---|
| 1 | 1.1593 | 1 | 0.739 | 16 | Fourier coefficient R_21 |
| 2 | 1.0767 | 2 | 0.6900 | 25 | Second slope of fitted straight line |
| 3 | 0.9403 | 3 | 0.5940 | 15 | Symmetry approach #2 |
| 4 | 0.9229 | 4 | 0.5900 | 22 | Depth calculated by high minus low magnitude on fitted curve |
| 5 | 0.8784 | 14 | 0.4530 | 13 | Coefficient4 of fitted curve |
| 6 | 0.8660 | 5 | 0.5780 | 24 | First slope of fitted straight line |
| 7 | 0.8644 | 7 | 0.5570 | 23 | X-value of the maximum magnitude on the fitted curve |
| 8 | 0.8323 | 6 | 0.5680 | 12 | Coefficient3 of fitted curve |
| 9 | 0.7660 | 8 | 0.5400 | 11 | Coefficient2 of fitted curve |
| 10 | 0.7633 | 9 | 0.5390 | 10 | Coefficient1 of fitted curve |
| 11 | 0.6654 | 11 | 0.4660 | 9 | X-value (phase) for maximum magnitude of raw data points |
| 12 | 0.6527 | 12 | 0.4630 | 8 | Symmetry integral difference #2 (error calculation) |
| 13 | 0.6527 | 13 | 0.4630 | 7 | Symmetry integral difference #1 |
| 14 | 0.4674 | 10 | 0.4790 | 20 | Log of the period given in dataset |
| 15 | 0.4356 | 15 | 0.3030 | 27 | Skewness of fitted curve |
| 16 | 0.1587 | 16 | 0.2330 | 14 | Coefficient5 of fitted curve |
| 17 | 0.1568 | 17 | 0.2140 | 4 | Minimum magnitude of raw data points |
| 18 | 0 | 18 | 0 | 6 | Minimum phase of raw data points |
| 19 | 0 | 19 | 0 | 2 | Mean phase of raw data points |
| 20 | 0 | 20 | 0 | 3 | Max magnitude of raw data points |
| 21 | 0 | 21 | 0 | 5 | Max phase of raw data points |
| 22 | 0 | 22 | 0 | 28 | Harmonic mean of fitted curve |
| 23 | 0 | 23 | 0 | 17 | Fourier coefficient phi_21 |
| 24 | 0 | 24 | 0 | 18 | Fourier coefficient R_31 |
| 25 | 0 | 25 | 0 | 26 | Mean magnitude of fitted curve |
| 26 | 0 | 26 | 0 | 21 | Mean magnitude given in the dataset |
| 27 | 0 | 27 | 0 | 19 | Fourier coefficient phi_31 |
| 28 | 0 | 28 | 0 | 1 | Mean magnitude of raw data points, calculated |

*Table 4.2: Information Gain and Gain Ratio for All Features*

The last 11 features in the table do not have any measureable worth, according to Information Gain and Gain Ratio. Also, features ranked sixteenth and seventeenth have very low worth, according to both measurements. Features ranked fourteenth and fifteenth by Information Gain have low values, but notice that feature number 20 is ranked as tenth by Gain Ratio instead of as fourteenth. For more information on the features and how they were developed, please refer back to Chapter 3.

## 4.3 Initial Experiments

### 4.3.1 Explanation

For the initial experiments, only fundamental-mode and first-overtone Cepheids from the LMC were used. I-band and V-band information for each star were treated as separate stars for simplicity of building training and test sets. The V-band is sparser but still contains enough information for good classification. Also, for these initial experiments, all stars that have missing data were excluded from the training and test sets. A total of 21 features were used with the machine learning methods. From the table in Section 4.2, features 1-20 were used and so was feature 23. It is important to note that features 22 through 25, which ranked very highly in the Information Gain and Gain Ratio measurements, were not used in these initial experiments because they had not yet been incorporated. Features 26-28, which performed poorly here, were not included for the same reason.

In order to run the experiments, the first step taken in the Python script (which generates the Weka ARFF for classification) is to calculate and plot the light curves of each star. The actual data points are plotted and used in calculations, but a line is also fitted using Python's *polyfit* function with a degree of 4. (A degree of 3 was used in experiments not detailed here, but some of the degree 3 fitted curves did not adhere well

to the data.) Next, features are extracted from the light curve and given star information, as described previously. Lastly, these features are written to the ARFF.

In Weka, the ARFF is loaded, and machine learning algorithms can be run. Four machine learning methods were run that are worth discussing here, and three methods of dividing training data versus test data were used. Specifically, the following machine learning methods were used: Multilayer Perceptron, Naïve Bayes, J48 Decision Trees, and Random Forest. For dividing the training and test sets, the following three methods were used: (1) 10-fold cross-validation, (2) 70% train and 30% test, and (3) 20% train and 80% test.

### 4.3.2 Results

As mentioned, initial experiments were performed with 21 features, the LMC galaxy, and fundamental-mode and first-overtone star classes. Four machine learning methods and three ways of dividing the training and test data were used. The results appear in the table below.

| | *Multilayer Perceptron Accuracy %* | *Naïve Bayes Accuracy %* | *J48 Decision Trees Accuracy %* | *Random Forest Accuracy %* |
|---|---|---|---|---|
| *10-fold Cross-Validation* | 98.2643 | 92.7680 | 97.8847 | 98.1016 |
| *70% Train, 30% Test* | 97.5889 | 92.3448 | 97.7095 | 97.8300 |
| *20% Train, 80% Test* | 97.0847 | 93.2429 | 96.9492 | 96.9492 |

*Table 4.3.2: Accuracies for Machine Learning Methods in Initial Experiments*

As can be seen in the above table, 10-fold cross-validation had the highest accuracies for most of the methods, the exception being Naïve Bayes. Also, the Multilayer Perceptron performed the best overall, but Random Forest was very close with J48 right behind it. Naïve Bayes clearly had the lowest performance.

## 4.4 Further Experiments

### 4.4.1 Explanation

The next set of experiments included all 28 features. Also, second-overtone Cepheids and Cepheids from the SMC were added. In order to incorporate second-overtone stars, it was unrealistic to use the previously discussed methods for dividing up the training and testing data. This is because there were thousands of stars available for the previous experiments with fundamental-mode and first-overtone stars, but there were not even 100 second-overtone stars in the OGLE LMC and SMC data combined. Thus, in order to deal with this problem, a tiny training set was developed.

The tiny training set was constructed by including the first seven of each class of star, so 21 from the LMC and 21 from the SMC for 42 total stars. But like in previous experiments, the I-band and V-band were treated separately. Specifically, there are 83 instances in the training data. The training data would have included 84 instances, but the first-overtone star OGLE-SMC-CEP-0003 is missing V-band data. The test set consists of 14,795 total instances, which is about 7,400 stars. Only 194 of those instances are second-overtone. First-overtone includes 5,746 instances, and fundamental-mode includes 8,855 instances.

A series of experiments were performed with the 28 features and subsets of those features using the following machine learning methods in Weka: Multilayer Perceptron, Naïve Bayes, J48 Decision Trees, and Random Forest. Also, experiments were performed with both galaxies, as well as the LMC galaxy alone and the SMC galaxy alone in order to compare the classification potential of the features across the two galaxies. Furthermore, some experiments included only the fundamental-mode and first-overtone Cepheids, while other experiments included all three classes, including second-overtone Cepheids.

In total, 216 experiments were performed with each experiment being a combination of features, galaxies, classes, and machine learning methods.

As mentioned, there were several subsets of features used for the experiments. Some experiments used all of the features, and this will be referred to as the Complete Set or Set 1. But others left out features that did not perform well in terms of Information Gain and Gain Ratio measurements. A base set of features to be removed was developed and included the following feature numbers: 1, 2, 3, 5, 6, 17, 18, 19, 21, 26, and 28. The set that only removed these base features will be referred to as the Base Set or Set 2. Set 3 removed the base features plus feature number 27. Set 4 removed the base features plus features number 4 and 14. Set 5 is the Minimal Set, and it removed the base features plus 4, 14, and 27. Sets 1-5 are the Primary Sets and include the first 120 experiments.

Four additional sets were created that removed features that performed reasonably well but not the best, and those are to be referred to as the Secondary Sets. They make up the last 96 experiments. The Secondary Sets were only created to demonstrate that removing too many features is actually detrimental to the performance of the machine learning methods. Set 6 includes Set 5 (the Minimal Set), plus it also removes feature number 20, the log of the star period. Set 7 includes Set 6 plus also removes features 7 and 8, the features that use the integral of the fitted curve to determine symmetry. Set 8 includes Set 7 plus also removes feature number 9, the x-value (phase) for the maximum magnitude determined with the raw data, not the fitted curve. Lastly, Set 9 includes Set 8 plus also removes three of the coefficients of the fitted curve, features 10, 11, and 12.

### 4.4.2 Results

As described, experiments were performed with different combinations of features, galaxies, classes, and machine learning methods in order to compare classification

accuracy more effectively. In total, 216 experiments were performed, but only the first 120 are part of the Primary Sets, which are the feature sets that will be discussed in the majority of the following tables.

Below, results comparing the machine learning methods are shown. In order to obtain the following results, four averages were calculated using only the Primary Sets. For each machine learning method, the mean was taken for the 30 experiments done with that method and feature sets 1-5. The experiments varied the number of galaxies, classes, and subsets of features but are comparable across machine learning methods. For example, Experiments 1-4 included both galaxies, all 28 features, and only the first two classes. The only thing that changed between Experiments 1-4 were the machine learning methods.

| *Machine Learning Method* | *Accuracy Percent* |
|---|---|
| *Multilayer Perceptron* | 95.2205 |
| *Naïve Bayes* | 87.3121 |
| *J48 Decision Trees* | 90.6842 |
| *Random Forest* | 91.2358 |

*Table 4.4.2a: Comparison of Machine Learning Methods and Classification Accuracy*

As can be seen in the table above, the Multilayer Perceptron performed better on average than the other machine learning methods. Naïve Bayes performed the worst. Both J48 and Random Forest are in between and very similar to each other. But it is also important to note that the Multilayer Perceptron performed better than the other machine learning methods in every experiment set, not just on average. Thus, for most of the remainder of the tables in this chapter, only the Multilayer Perceptron machine learning technique will be used in order to simplify the tables of results. As for the other methods, there was no consistent trend throughout the experiment sets for which is second best or third or worst.

Next are the results of star classes on classification accuracy. Two groups of experiments were performed. Group 1 includes only fundamental-mode and first-overtone Cepheids. Group 2 includes all three classes, so second-overtone Cepheids are also included. The averages of the experiments for the Multilayer Perceptron are summarized in the table below and only include the results of the Primary Sets.

| *Classes Included* | *Accuracy Percent* |
|---|---|
| *Fundamental-mode and First-overtone* | 96.4203 |
| *Fundamental-mode, First-overtone, and Second-overtone* | 94.0207 |

*Table 4.4.2b: Comparison of Star Classes and Classification Accuracy*

As can be seen in the table above, the accuracy average is somewhat lower when the second-overtone stars are included. This is actually consistent across the data. For every Multilayer Perceptron experiment set, the experiment with all three classes has a lower accuracy percent than the one with only two classes, frequently by 2-3%. This consistency is also demonstrated in J48 and Random Forest experiments. Naïve Bayes has a little inconsistency though with 3 experiment sets out of 15 being reversed, although the difference is less than 1% in all those cases.

Another aspect tested in the experiments is the galaxy's effect on accuracy. Three groups of experiments were used, one with both galaxies included, one with only the LMC, and the last with only the SMC. The average results for the Multilayer Perceptron are shown in the table below and include only the results of the Primary Sets.

| *Galaxies Included* | *Accuracy Percent* |
|---|---|
| *LMC and SMC* | 95.9013 |
| *LMC* | 93.9115 |
| *SMC* | 95.5193 |

*Table 4.4.2c: Comparison of Galaxies and Classification Accuracy*

As can be seen in the table above, the LMC is associated with a lower accuracy than the SMC, but this did not hurt the average for both galaxies. On the contrary, performance was best when both galaxies were included, which is interesting. This trend can also be seen in the data apart from the averages, although it is not entirely consistent. Further experiments, perhaps with a different dataset, would be needed in order to confirm whether this result is the norm or if it is chance.

Lastly, feature subset experiments were performed. As discussed previously, there are 5 sets included in the Primary Sets and 4 sets included in the Secondary Sets. Only Set 1 includes all 28 features. The creation of these feature sets is based on the results of Information Gain and Gain Ratio in Section 4.2. The base set were the features with the worst performance and, thus, were the safest to remove. The Secondary Sets were only included to demonstrate that the machine learning methods do not perform as well if too many features are removed. The results are shown in the following table.

| *Set Number* | *Accuracy Percent for All Machine Learning Methods* | *Accuracy Percent for Multilayer Perceptron* |
|---|---|---|
| *Set 1* | 90.7684 | 94.1430 |
| *Set 2* | 91.1622 | 94.9511 |
| *Set 3* | 91.2515 | 95.4666 |
| *Set 4* | 91.1858 | 95.6203 |
| *Set 5* | 91.1979 | 95.9216 |
| *Set 6* | 88.4026 | 87.8505 |
| *Set 7* | 88.9274 | 88.3175 |
| *Set 8* | 89.1560 | 89.0293 |
| *Set 9* | 89.1003 | 88.7642 |

*Table 4.4.2d: Comparison of Feature Sets and Classification Accuracy*

As shown in the table above, it is fairly consistent for the accuracy percent to increase as poorly performing features are removed, but the accuracy quickly declines if too many features are taken out. The Multilayer Perceptron accuracy consistently increases in sets 1-5, which are the sets that have poorly performing features removed. It

is especially clear when looking at the Multilayer Perceptron that too many features were removed for the Secondary Sets, sets 5-9, as the accuracy decreases about 8% between Set 5 and Set 6.

## 4.5 Revised Experiments

### 4.5.1 Explanation

In order to do further testing, a revised set of experiments was developed using all 28 features. The larger training set was able to be used for these revised experiments because only fundamental-mode and first-overtone Cepheids of the LMC were included, much like in the initial experiments. The main difference between these revised experiments and the initial ones is the number of features. First, all 28 features were able to be used. Second, subsets of those 28 features were tested in order to compare classification accuracy when poorly performing features were removed. Also, stars that are missing data were included, unlike in the initial experiments. All four machine learning methods were run in these revised experiments.

### 4.5.2 Results

As previously explained, these experiments were similar to the initial experiments already discussed in this chapter but include all 28 features. The results appear in the chart below in addition to the “Custom” set of data. The Custom set is the result from Section 4.4 experiments just discussed, included here for comparison only. As previously explained, the Custom set has a training set with 7 of each type of star, so about 42 stars total, which is a much smaller number of stars than the other training and testing methods. It was expected that the Custom training data would build a classifier with somewhat lower accuracy due to training on such a small number of stars and including all three classes.

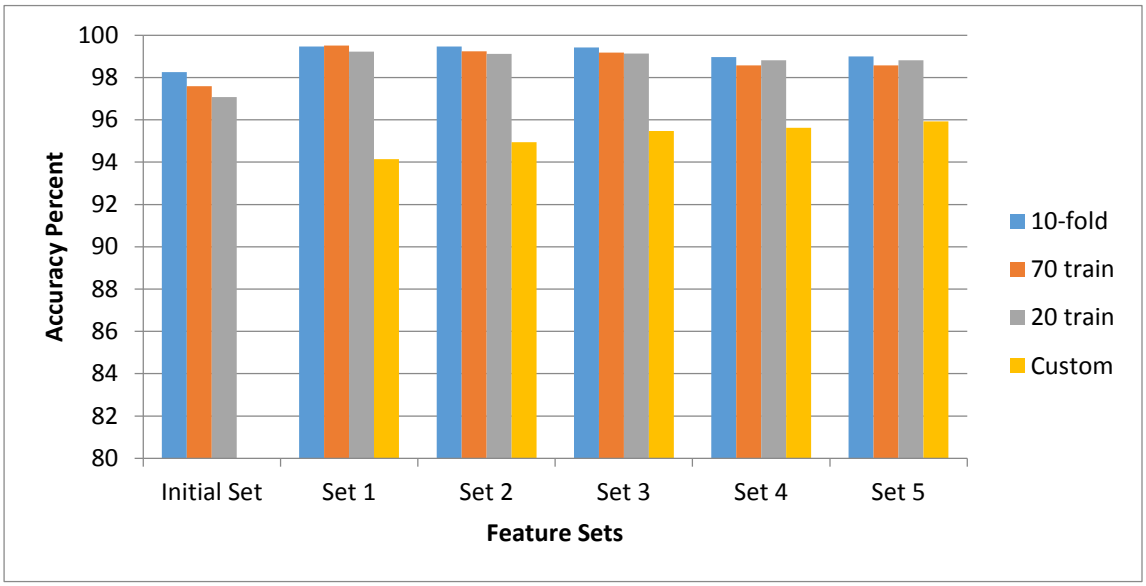


*Figure 4.5.2: Comparison of Feature Sets in Revised Experiments*

As can be seen in the chart above, accuracy increases slightly between the Initial Set and Set 1, which means that removing poorly performing features was somewhat beneficial. This is true for all of the training-testing combinations. But an interesting difference occurs between the Custom data and the other training-test combinations. The Custom data accuracy percent increases a little as more poorly performing features are removed, but the other training-testing combinations have a slight decrease in accuracy. One possibility for this is that having a larger amount of training information overcomes some of the shortcomings of features that do not perform as well.

# Chapter 5:

# Discussion

## 5.1 General Discussion

In the experiments in the last chapter, it was determined that the Multilayer Perceptron performed the best out of the four methods. There may be a number of reasons for this, such as how neural networks and many other methods perform well without requiring prior knowledge. But it may be helpful to incorporate contextual information and prior information to improve the classification accuracy of the Naïve Bayes approach. Also, for Random Forest, it may be helpful to further refine the utility of the features and develop new features with high classification accuracy. It is also important to note that, according to [17], "classification trees are simple, yet powerful, non-parametric classifiers" but even small feature changes can change the tree structure significantly.

It is also worth reviewing the fact that including both galaxies, rather than just one galaxy at a time, yielded higher classification accuracy. This is especially interesting because the galaxies do have different populations of Cepheids, which have slightly different characteristics. But it seems that neural networks are capable of working with those differences between Cepheid populations. It is possible that feeding more data into the machine learning methods is an effective way to achieve high accuracy percentages, even if that information contains variations.

Similarly, higher classification accuracies were achieved when more training information was used when the second-overtone Cepheids were left out, even when poorly performing features were retained. This is a positive thing because we are

gathering huge amounts of data on astronomical objects, so it is becoming more and more possible to train on larger amounts of data.

Overall, the results are very hopeful, and the interdisciplinary field of computer science within astronomy appears to have a bright future. It is important to focus on automatically collecting new data, assessing feature quality, and performing experiments in order to test approaches.

## 5.2 Misclassified Stars

In order to further analyze the classification capability of the developed features and in order to work on the development of new features, misclassified stars were reviewed. In order to do this, all 28 features, both galaxies, and all three star classes were included with the small training set of 42 stars described in the previous chapter. The Multilayer Perceptron was used. Several types of misclassifications will be discussed below with supporting star plots included.

One of the characteristics of fundamental-mode Cepheids, as discussed previously, is that their light curves are rather asymmetrical, unlike first-overtone Cepheids, which are fairly symmetrical. They are also unlike second-overtone Cepheids, which have nearly sinusoidal light curves.

But in the following example (and many other instances not shown here), the fundamental-mode Cepheid OGLE-LMC-CEP-0333 looks very symmetrical. Here, this LMC fundamental-mode Cepheid was classified as a first-overtone Cepheid. Because this goes against symmetry expectations, which ranks high in Information Gain and Gain Ratio rankings previously, it is not surprising that the machine learning methods would misclassify the star. Probably the best way to overcome this fault is to have other strong features that do not revolve around symmetry.

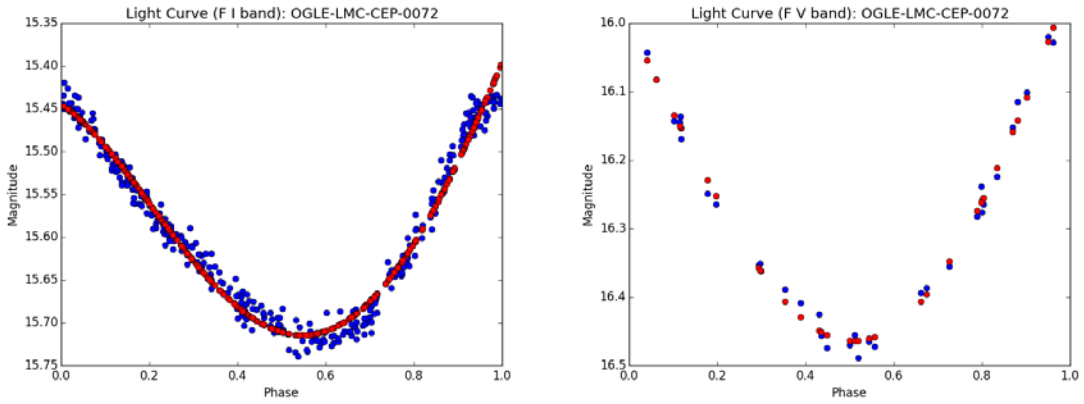


*Figure 5.2a: Symmetrical Fundamental-Mode Example #1*

As can be seen above, the fundamental-mode LMC Cepheid's light curve does look rather symmetrical, much like a first-overtone Cepheid. There are anomalies within each star class, such as this one, whose classes are difficult to distinguish.

Similarly, in the following example, an SMC fundamental-mode Cepheid, OGLE-SMC-CEP-0982, is misclassified as a second-overtone Cepheid.

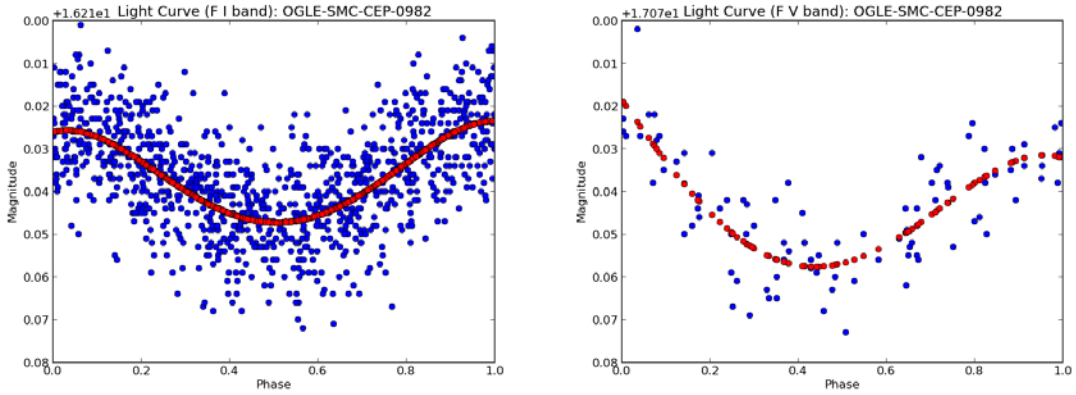


*Figure 5.2b: Symmetrical Fundamental-Mode Example #2*

In the above example, especially in the I-band, the light curve is surprisingly symmetrical for a fundamental-mode Cepheid. Also, the amplitude is very small. It looks very much like a second-overtone star, so it is no wonder that the machine learning methods misclassified it. This may be an anomaly for this type of star, or perhaps a

labeling error has occurred. Further research into this may provide deeper answers that could help with future feature development and star classification.

Another type of problem that arose that seemed to affect classification accuracy negatively is that the magnitude of the light curves is not always represented in the same way. In most of the stars that were classified correctly, the minimum magnitudes occur at minimum and maximum phase values. But in a number of misclassifications, the magnitudes begin to rise again near the minimum and maximum phase values, as shown below in the example star OGLE-LMC-CEP-0223. The following LMC fundamental-mode Cepheid was classified as a first-overtone Cepheid.

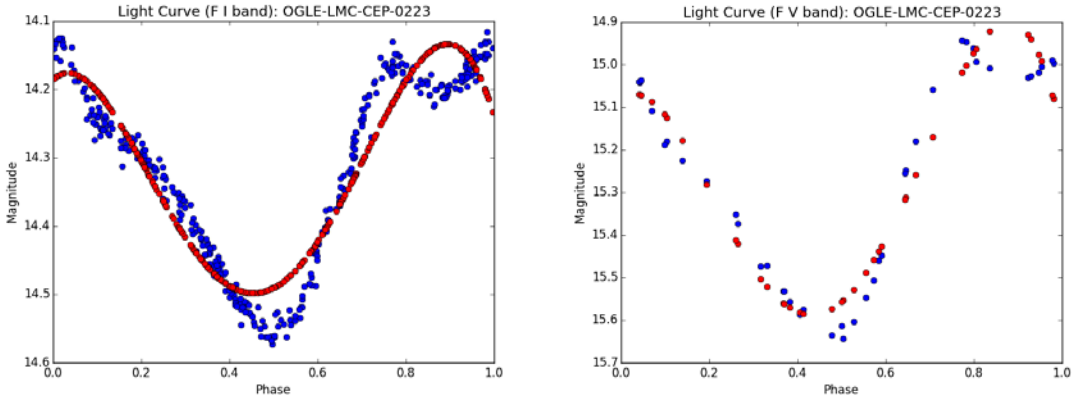


*Figure 5.2c: Magnitude Example #1*

In the above star plots, there are unexpected variations in the magnitude where the phase is a higher value. Normally the magnitude has a gradual decline from approximately the middle phase value to the highest phase value. But, here, the magnitude decreases somewhat rapidly and then increases again before decreasing to its minimum. Also, the fitted curve (in red) clearly had some problems with fitting the data properly in this case. But increasing the degree of the fitted curve did not completely resolve this problem.

Below is an example of a SMC fundamental-mode Cepheid being misclassified as a first-overtone Cepheid, possibly due to the magnitude issues like in the previous example.

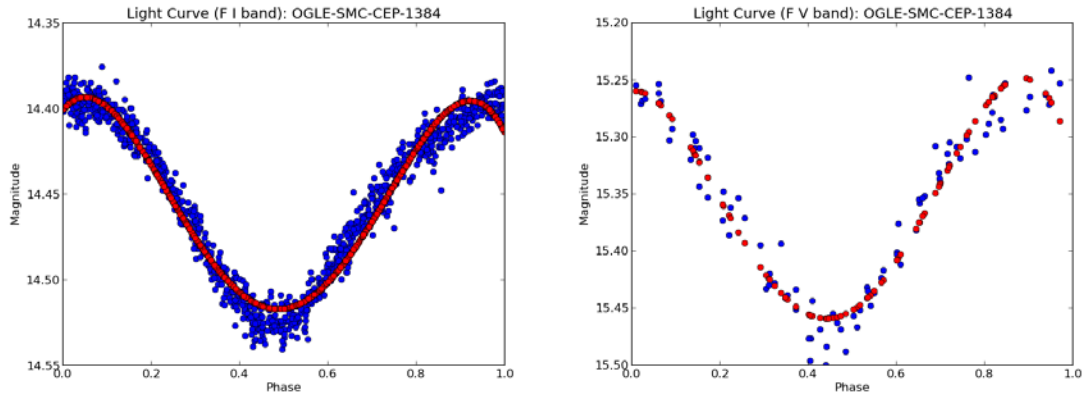


*Figure 5.2d: Magnitude Example #2*

In the above example, the magnitude increases again at both the minimum and the maximum phase values. This may have played a part in the misclassification of this star, even though it is less pronounced than the previous example.

Also, a frequent problem that occurred was that the V-band was misclassified even though the I-band for that star was classified correctly. Fundamental-mode Cepheid OGLE-LMC-CEP-0218 is shown below as an example of this.

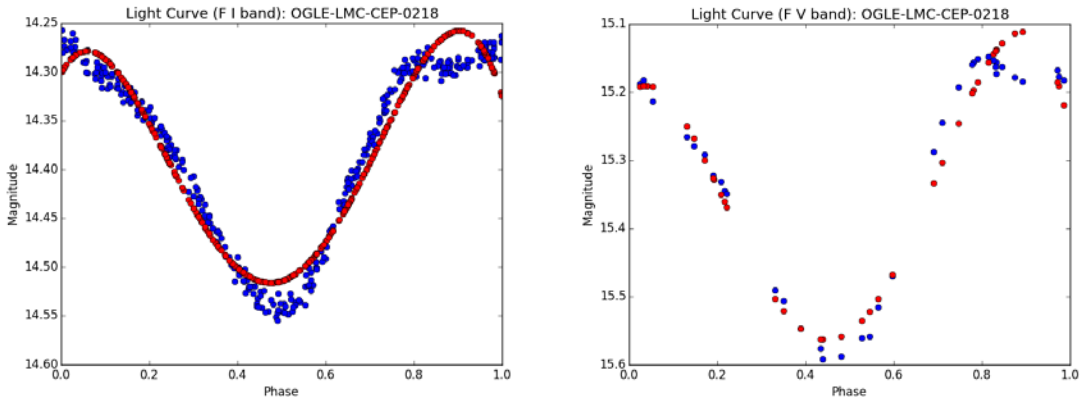


*Figure 5.2e: Misclassified V-Band Example #1*

In the above star plots, it is apparent that the V-band information is much sparser, which may be a large part of why the V-band was misclassified but the I-band was not. It is also interesting to note here that this example also demonstrates two other issues previously mentioned. The light curve is not quite a U-shape, as the magnitude has some unexpected variation at the minimum and maximum phase. Also, this star is more symmetrical than many of the fundamental-mode Cepheids. Perhaps these problems plus the sparser data for the V-band account for the partial misclassification of this star. Another issue that could be playing a part in the misclassification of the V-band instance is that the I- and V-bands have slightly different characteristics. But these characteristics were not taken into account for running the machine learning methods. A more complex approach to this problem may be taken in the future.

Below is another star misclassified only in the V-band, the LMC fundamental-mode Cepheid OGLE-LMC-CEP-0139.

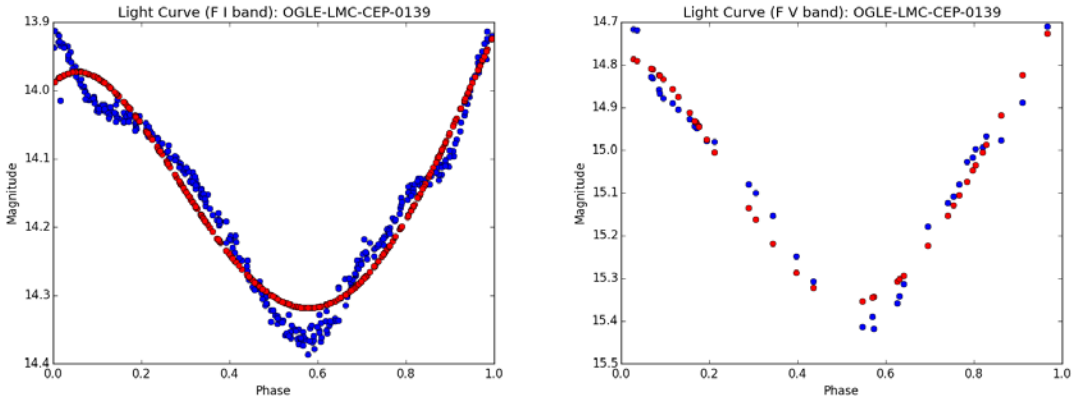


*Figure 5.2f: Misclassified V-Band Example #2*

As in the previous example, the V-band is sparser, and there are some problems with the fitted curve conforming to the data properly. At the smaller phase values in the I-band light curve, the magnitude is especially misinterpreted by the fitted curve. The actual magnitude decreases, but the fitted curve's magnitude increases at the minimum

phase. This occurs in the V-band, as well, although it is not quite as clear or as pronounced. This is interesting because the I-band was correctly classified, despite the problems with its fitted curve. But perhaps because the information is much sparser for the V-band, it does not take as much unexpected variation to cause misclassification, and there may have been less obvious issues with the other features.

The issues discussed may be able to be solved with further research and feature development. Having a variety of features with high classification accuracy that target the differences between each class is a beneficial approach. More about the limitations of this research and potential for future work will be discussed in the next chapter.

# Chapter 6

# Future Work and Conclusion

## 6.1 Limitations and Future Work

Although a number of experiments were performed in the context of this thesis, there are a variety of improvements that could be made and a number of future paths to explore. For example, it would be best if this work were expanded to include other galaxies and other datasets in order to further test the classifying potential of these features. The more variety that is incorporated with new test sets, the more sure we can be of how these features rank in terms of true classification value.

Also, if more datasets are incorporated, it would become more viable to distinguish between a larger number of classes rather than just fundamental-mode, first-overtone, and second-overtone, which would be useful in the future. It would also be interesting to see how these features would work if more diverse classes were included, such as supernovae. Furthermore, being able to distinguish between more classes would make the feature utility clearer.

Another direction that could be taken in the future is to explore sparse data more thoroughly. It would be beneficial to determine just how sparse the data could be but still have high classification accuracy with these features. In this thesis, both the I-band and the sparser V-band data were used for Cepheids. There were numerous examples where the V-band was misclassified, even though the I-band was classified correctly. In the future, experiments could be done to remove a percentage of the available data points and compare classification accuracy. This could be very beneficial for assessing features,

as well, because stars with only sparse data available may be more prone to misclassification. Having features that are sensitive to these conditions could yield higher classification accuracy overall, even when available information is not sparse. Another way to deal with the I-band and V-band issue is to treat them separately and incorporate prior knowledge about each band.

It would be ideal to build on this work in order to be able to classify stars in real-time and with huge amounts of data. It would also be ideal if contextual information could be integrated with the features. It could be possible to use contextual information in order to rule out noise and identify measurement and labeling errors. It would also be beneficial to use different types of machine learning methods, such as Support Vector Machines, because they may have many advantages not explored in this thesis.

## 6.2 Conclusion

The overall goal of this research is to reduce the number of hours humans must spend in order to further astronomy and to increase the amount of automation throughout the system in order to make discoveries as efficiently as possible. This work just scratches the surface of astronomy and machine learning. There is a multitude of ways that this work could be incorporated into something larger. Machine learning is becoming an increasingly crucial part of astronomy. With the rapid growth of data collection, it will only continue to be more so. Developing efficient, scalable, useful, and even real-time methods for star classification and related purposes is essential for making use of the new data. Developing features for classifying the different types of classical Cepheids shows a lot of promise, as these experimental features yield good classification results. There are still many issues to tackle, and it would be good to expand to different kinds of Cepheids and stars and other datasets to make sure that the features are robust.

In this thesis, we have reviewed 28 developed features and numerous experiments and results for LMC and SMC fundamental-mode, first-overtone, and second-overtone Cepheid variable stars using the following machine learning methods implemented in Weka: Multilayer Perceptron, Naïve Bayes, J48 Decision Trees, and Random Forest. Multilayer Perceptron performed the best overall.

# References


[1] Igor Soszyński et al., “Classical Cepheids in the Large Magellanic Cloud,” in *Acta Astronomica, 58, 163*, 2008.

[2] Igor Soszyński & OGLE Team. “OGLE Atlas of Variable Star Light Curves” [Online]. Available: http://ogle.astrouw.edu.pl/atlas/classical_Cepheids.html.

[3] David H. Levy, “Meeting the Family,” in *David Levy's Guide to Variable Stars*, 2nd ed. Cambridge, United Kingdom: Cambridge University Press, 2005, pp. 25-29.

[4] David H. Levy, “Getting Started with Cepheids,” in *David Levy's Guide to Variable Stars*, 2nd ed. Cambridge, United Kingdom: Cambridge University Press, 2005, pp. 30-34.

[5] Gerald North, “Clockwork Pulsators,” in *Observing Variable Stars, Novae, and Supernovae*, Cambridge, United Kingdom: Cambridge University Press, 2004, pp.121-132.

[6] Joseph W. Richards et al., "Active Learning to Overcome Sample Selection Bias: Application to Photometric Variable Star Classification," in *The Astrophysical Journal 744, no. 2*, 2012, *pp. 192-211*.

[7] James P. Long et al., "Optimizing Automated Classification of Variable Stars in New Synoptic Surveys," in *Publications of the Astronomical Society of the Pacific 124, no. 913*, 2012, *pp. 280-295*.

[8] J. S. Bloom et al., "Automating Discovery and Classification of Transients and Variable Stars in the Synoptic Survey Era," in *Publications of the Astronomical Society of the Pacific 124, no. 921*, 2012, pp. 1175-1196.

[9] Vasily Belokurov et al., "Light-Curve Classification in Massive Variability Surveys–II. Transients towards the Large Magellanic Cloud," in *Monthly Notices of the Royal Astronomical Society 352, no. 1*, 2004, pp. 233-242.

[10] Julian Faraway et al., “Modeling Light Curves for Improved Classification,” in *arXiv preprint arXiv: 1401.3211*, 2014.

[11] S. G. Djorgovski et al., "Flashes in a Star Stream: Automated Classification of Astronomical Transient Events," in *E-Science (e-Science), 2012 IEEE 8th International Conference*, 2012, pp. 1-8.

[12] Ciro Donalek et al. "Feature Selection Strategies for Classifying High Dimensional Astronomical Data Sets," in *Big Data, 2013 IEEE International Conference*, 2013, pp. 35-41.

[13] Machine Learning Group at the University of Waikato. “Weka 3 – Data Mining with Open Source Machine Learning Software in Java” [Online.] Available: http://www.cs.waikato.ac.nz/ml/weka.

[14] Jonas Blomme et al. "Automated Classification of Variable Stars in the Asteroseismology Program of the Kepler Space Mission," in *The Astrophysical Journal Letters, 713.2*, 2010, L204.

[15] A. A. Mahabal, et al. "Discovery, Classification, and Scientific Exploration of Transient Events from the Catalina Real-Time Transient Survey," in *arXiv preprint arXiv:1111.0313*, 2011.

[16] Igor Soszynski et al., “Classical Cepheids in the Small Magellanic Cloud,” in *Acta Astronomica, 60, 17*, 2010.

[17] Joseph W. Richards, et al. "On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data," in *The Astrophysical Journal 733.1* (2011): 10.

[18] Ricardo Vilalta, Kinjal Dhar Gupta, and Lucas Macri. "Domain Adaptation Under Data Misalignment: An Application to Cepheid Variable Star Classification," in *Pattern Recognition (ICPR)*, 2014 22nd International Conference on. IEEE, 2014.

[19] Yanxia Zhang and Yongheng Zhao. "Automated Clustering Algorithms for Classification of Astronomical Objects," in *Astronomy & Astrophysics 422.3* (2004): 1113-1121.

[20] Gerald North, “Foundations, Federations, and Finder Charts,” in *Observing Variable Stars, Novae, and Supernovae*, Cambridge, United Kingdom: Cambridge University Press, 2004, pp.1-19.