New Astronomy 52 (2017) 35–47

Contents lists available at ScienceDirect

New Astronomy

journal homepage: www.elsevier.com/locate/newast

Generation of a supervised classification algorithm for time-series variable stars with an application to the LINEAR dataset

K.B. Johnston*, H.M. Oluseyi

Florida Institute of Technology, Physics and Space Sciences Dept., Melbourne, Florida, 32901, United States

HIGHLIGHTS

- We present a new supervised classification methodology for the analysis of time series variables.
- We apply this analysis to the LINEAR survey dataset.
- An anomaly detection algorithm is developed, as are improved estimates of performance.

ARTICLE INFO

Article history:
Received 12 January 2016
Revised 5 October 2016
Accepted 17 October 2016
Available online 19 October 2016

Keywords:
Supervised classification
Anomaly detection
Statistical performance
Stellar variability

ABSTRACT

With the advent of digital astronomy, new benefits and new problems have been presented to the modern day astronomer. While data can be captured in a more efficient and accurate manner using digital means, the efficiency of data retrieval has led to an overload of scientific data for processing and storage. This paper will focus on the construction and application of a supervised pattern classification algorithm for the identification of variable stars. Given the reduction of a survey of stars into a standard feature space, the problem of using prior patterns to identify new observed patterns can be reduced to time-tested classification methodologies and algorithms. Such supervised methods, so called because the user trains the algorithms prior to application using patterns with known classes or labels, provide a means to probabilistically determine the estimated class type of new observations. This paper will demonstrate the construction and application of a supervised classification algorithm on variable star data. The classifier is applied to a set of 192,744 LINEAR data points. Of the original samples, 34,451 unique stars were classified with high confidence (high level of probability of being the true class).

1. Introduction

With the advent of digital astronomy, new benefits and new challenges have been presented to the modern day astronomer. While data is captured in a more efficient and accurate manner using digital means, the efficiency of data retrieval has led to an overload of scientific data for processing and storage. This means that more stars, in more detail, are captured per night; but increasing data capture begets exponentially increasing data processing. Database management, digital signal processing, automated image reduction, and statistical analysis of data have all made their way to the forefront of tools for the modern astronomer. Astro-statistics and astro-informatics are fields which focus on the application and development of these tools to help aid in the processing of large scale astronomical data resources.

This paper will focus on one facet of this budding area, the construction and application of a supervised pattern classification algorithm for the identification of variable stars. Given the reduction of a survey of stars into a standard feature space, the problem of using prior patterns to identify new observed patterns can be reduced to time-tested classification methodologies and algorithms. Such supervised methods, so called because the user trains the algorithms prior to application using patterns with known classes or labels, provides a means to probabilistically determine the estimated class type of new observations. These methods have two large advantages over hand-classification procedures: the rate at which new data is processed is dependent only on the computational processing power available, and the performance of a supervised classification algorithm is quantifiable and consistent. Thus the algorithm produces rapid, efficient, and consistent results.

This paper will be structured as follows. First, the data and feature space to be implemented for training will be reviewed. Second, we will discuss the class labels to be used and the meaning behind them. Third, a set of classifiers (multi-layer perceptron, random forest, k-nearest neighbor, and support vector

* Corresponding author. Fax: +(321) 674 8000.
E-mail addresses: kyjohnston2000@my.fit.edu (K.B. Johnston), holuseyi@fit.edu (H.M. Oluseyi).

http://dx.doi.org/10.1016/j.newast.2016.10.004
1384-1076/© 2016 Elsevier B.V. All rights reserved.

machine) will be trained and tested on the extracted feature space. Fourth, performance statistics will be generated for each classifier and a comparing and contrasting of the methods will be discussed with a "champion" classification method being selected. Fifth, the champion classification method will be applied to the new observations to be classified. Sixth, an anomaly detection algorithm will be generated using the so called one-class support vector machine and will be applied to the new observations. Lastly, based on the anomaly detection algorithm, and the supervised training algorithm a set of populations per class type will be generated. The result will be a highly reliable set of new populations per class type derived from the LINEAR survey.

1.1. Related work

The idea of constructing a supervised classification algorithm for stellar classification is not unique to this paper (see Dubath et al. 2011 for a review), nor is the construction of a classifier for time variable stars. Methods pursued include the construction of a detector to determine variability (two-class classifier Barclay et al. 2011), the design of random forests for the detection of photometric redshifts in spectra Carliles et al. (2010), the detection of transient events Djorgovski et al. (2012), and the development of machine-assisted discovery of astronomical parameter relationships Graham et al. (2013). Debosscher (2009) explored several classification techniques for the supervised classification of variable stars, quantitatively comparing the performed in terms of computational speed and performance which they took to mean accuracy. Likewise, other efforts have focused on comparing speed and robustness of various methods (e.g. Blomme et al. 2011; Pichara et al. 2012; Pichara and Protopapas 2013). These methods span both different classifiers and different spectral regimes, including IR surveys (Angeloni et al. 2014 and Masci et al. 2014), RF surveys (Rebbapragada et al., 2011), and optical (Richards et al., 2012).

2. Data

The procedure outlined in this paper will follow the standard philosophy for the generation of a supervised pattern classification algorithm as professed in Duda et al. 2012 and Hastie et al. (2004), i.e. exploratory data analysis, training and testing of supervised classifier, comparison of classifiers in terms of performance, application of classifier. Our training data is derived from a set of three well known variable star surveys: the ASAS survey (Pojmanski et al., 2005), the Hipparcos survey (Perryman et al., 1997), and the OGLE dataset (Udalski et al., 2002). Data used for this study must meet a number of criteria:

1. Each star shall have differential photometric data in the u-g-r-i-z system
2. Each star shall have variability in the optical channel (band) that exceeds some fixed threshold with respect to the error in amplitude measurement
3. Each star shall have a consistent class label, should multiple surveys address the same star

2.1. Sample representation

These requirements reduce the total training set down to 2,054 datasets with 32 unique class labels. The features extracted are based on Fourier frequency domain coefficients (Deb and Singh, 2009), statistics associated with the time domain space, and differential photometric metrics; for more information see Richards et al. (2012) for a table of all 68 features with descriptions. The 32 unique class labels can be further generalized into four main groups: eruptive, multi-star, pulsating, and "other" (Debosscher, 2009), the breakdown of characterizations for the star classes follows the following classifications:

Table 1
Broad classification of variable types in the training and testing dataset.

Type	Count	% Dist
Multi-star	514	0.25
Other	135	0.07
Pulsating	1179	0.57
Erupting	226	0.11

Table 2
Unique classification of variable types in the training and testing dataset.

Class type	% Dist	Class type	% Dist
a. Mira	8.0%	m. Slowly Puls. B	1.5%
b1. Semireg PV	4.9%	n. Gamma Doradus	1.4%
b2. SARG A	0.7%	o. Pulsating Be	2.4%
b3. SARG B	1.4%	p. Per. Var. SG	2.7%
b4. LSP	2.6%	q. Chem. Peculiar	3.7%
c. RV Tauri	1.2%	r. Wolf-Rayet	2.0%
d. Classical Cepheid	9.9%	r1. RCB	0.6%
e. Pop. II Cepheid	1.3%	s1. Class. T Tauri	0.6%
f. Multi. Mode Cepheid	4.8%	s2. Weak-line T Tauri	1.0%
g. RR Lyrae FM	7.2%	s3. RS CVn	0.8%
h. RR Lyrae FO	1.9%	t. Herbig AE/BE	1.1%
i. RR Lyrae DM	2.9%	u. S Doradus	0.3%
j. Delta Scuti	6.5%	v. Ellipsoidal	0.6%
j1. SX Phe	0.3%	w. Beta Persei	8.7%
k. Lambda Bootis	0.6%	x. Beta Lyrae	9.8%
l. Beta Cephei	2.7%	y. W Ursae Maj.	5.9%

- Pulsating
  - Giants: Mira, Semireg RV, Pop. II Cepheid, Multi. Mode Cepheid
  - RR Lyrae: FO, FM, and DM
  - "Others": Delta Scuti, Lambda Bootis, Beta Cephei, Slowly Pulsating B, Gamma Doradus, SX Phe, Pulsating Be
- Erupting: Wolf-Rayet, Chemically Peculiar, Per. Var. SG, Herbig AE/BE, S Doradus, RCB and Classical T-Tauri
- Multi-Star: Ellipsoidal, Beta Persei, Beta Lyrae, W Ursae Maj.
- Other: Weak-Line T-Tauri, SARG B, SARG A, LSP, RS Cvn

The *a priori* distribution of stellar classes is given in Table 1 for the broad classes and in Table 2 for the unique classes:

It has been shown (Rifkin and Klautau, 2004) that how the classification of a multi-class problem is handled can affect the performance of the classifier; i.e. if the classifier is constructed to process all 32 unique classes as the same time, or if 32 different classifiers (detectors) are trained individually and the results are combined after application, or if a staged approach is best where a classifier is trained on the four broad classes first, then a secondary classifier is trained on the unique class labels in each broad class (Debosscher, 2009). The *a priori* distribution of classes, the number of features to use, and the number of samples in the training set are key factors in determining which classification procedure to use. This dependence is often best generalized as the "curse of dimensionality" (Bellman et al., 1961), a set of problems that arise in machine learning that are tied to attempting to quantify a signature pattern for a given class, when the combination of a low number of training samples and high feature dimensionality results in a sparsity of data. Increasing sparsity results in a number of performance problems with the classifier, most of which amount to decrease generality (over-trained classifier) and decreased performance (low precision or high false alarm rate). Various procedures have been developed to address the curse of dimensionality, most