New Astronomy 50 (2017) 1–11

Contents lists available at ScienceDirect

New Astronomy

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

Variable Star Signature Classification using Slotted Symbolic Markov Modeling

K.B. Johnston[a,*], A.M. Peter[b]

[a] Florida Institute of Technology, Physics and Space Sciences Department, Melbourne, Florida, 32901, USA
[b] Florida Institute of Technology, Systems Engineering Department, Melbourne, Florida, 32901, USA

HIGHLIGHTS

- We present a new feature space for the supervised classification of stellar variables.
- Two surveys are used: data from the UCR database and data from the LINEAR survey.
- Improved linear separation is generated using the new feature space.

ARTICLE INFO

Article history:
Received 18 June 2015
Revised 14 March 2016
Accepted 6 June 2016
Available online 23 June 2016

Keywords:
Stellar variability
Supervised classification
Markov modeling
Time-domain analysis

ABSTRACT

With the advent of digital astronomy, new benefits and new challenges have been presented to the modern day astronomer. No longer can the astronomer rely on manual processing, instead the profession as a whole has begun to adopt more advanced computational means. This paper focuses on the construction and application of a novel time-domain signature extraction methodology and the development of a supporting supervised pattern classification algorithm for the identification of variable stars. A methodology for the reduction of stellar variable observations (time-domain data) into a novel feature space representation is introduced. The methodology presented will be referred to as Slotted Symbolic Markov Modeling (SSMM) and has a number of advantages which will be demonstrated to be beneficial; specifically to the supervised classification of stellar variables. It will be shown that the methodology outperformed a baseline standard methodology on a standardized set of stellar light curve data. The performance on a set of data derived from the LINEAR dataset will also be shown.

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.

A methodology for the reduction of stellar variable observations (time-domain data) into a novel feature space representation is introduced. The proposed methodology, referred to as Slotted Symbolic Markov Modeling (SSMM), has a number of advantages over other classification approaches for stellar variables. SSMM can be applied to both folded and unfolded data. Also, it does not need time-warping for alignment of the waveforms. Given the reduction of a survey of stars into this new feature space, the problem of using prior patterns to identify new observed patterns can be addressed via classification algorithms. These methods have two large advantages over manual-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.

The remainder of this paper is structured as follows. First, the data, prior efforts, and challenges uniquely associated to classification of stars via stellar variability is reviewed. Second, the novel methodology, SSMM, is outlined including the feature space and signal conditioning methods used to extract the unique time-domain signatures. Third, a set of classifiers (random forest/bagged

* Corresponding author. Fax: +(321) 674 8000.
E-mail addresses: kyjohnst2000@my.fit.edu (K.B. Johnston), apeter@fit.edu (A.M. Peter).

http://dx.doi.org/10.1016/j.newast.2016.06.001
1384-1076/

decisions tree, k-nearest neighbor, and Parzen window classifier) is trained and tested on the extracted feature space using both a standardized stellar variability dataset and the LINEAR dataset. Fourth, performance statistics are generated for each classifier and a comparing and contrasting of the methods is discussed. Lastly, an anomaly detection algorithm is generated using the so called one-class Parzen Window Classifier and the LINEAR dataset. The result will be the demonstration of the SSMM methodology as being a competitive feature space reduction technique, for usage in supervised classification algorithms.

1.1. Related work

The idea of constructing a supervised classification algorithm for stellar classification is not unique to this paper (Dubath et al., 2011). Methods pursued include the construction of a detector to determine variability (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., 2013a). Debosscher (2009) explored several classification techniques for the supervised classification of variable stars, quantitatively comparing the performance in terms of computational speed and performance. Likewise, other efforts have focused on comparing speed and robustness of various methods (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; Masci et al., 2014), RF surveys (Rebbapragada et al., 2011) and optical (Richards et al., 2012). Methods for automated supervised classification include procedures such as: direct parametric analysis (Udalski et al., 1999), fully automated neural networking (Pojmanski, 2000; 2002) and Bayesian classification (Eyer and Blake, 2005).

The majority of these studies rely on periodicity domain feature space reductions. Debosscher (2009) and Templeton (2004) review a number of feature spaces and a number of efforts to reduce the time domain data, most of which implement Fourier techniques, primarily the Lomb-Scargle (L-S) Method (Lomb, 1976; Scargle, 1982), to estimate the primary periodicity (Eyer and Blake, 2005; Park and Cho, 2013; Richards et al., 2012; Ngeow et al., 2013; Deb and Singh, 2009). Lomb-Scargle is favored because of the flexibility it provides with respect to observed datasets; when sample rates are irregular and drop outs are common in the data being observed. Long et al. (2014) advance L-S even further, introducing multi-band (multidimensional) generalized L-S, allowing the algorithm to take advantage of information across filters, in cases where multi-channel time-domain data is available. There have also been efforts to estimate frequency using techniques other than L-S such as the Correntropy Kernelized Periodogram, (Huijse et al., 2011) or MUlti SIgnal Classificator (Tagliaferri et al., 2003).

The assumption of the light curve being periodic, or even that the functionality of the signal being represented in the limited Fourier space that Lomb-Scargle uses, has been shown (Palaversa et al., 2013; Barclay et al., 2011) to result in biases and other challenges when used for signature identification purposes. Supervised classification algorithms implementing these frequency estimation algorithms do so to generate an estimate of primary frequency used to fold all observations resulting in a plot of magnitude vs. phase, something Deb and Singh (2009) refer to as "reconstruction" . After some interpolation to place the magnitude vs. phase plots on similar regularly sampled scales, the new folded time series can be directly compared (1-to-1) with known folded time series. Comparisons can be performed via distance metric (Tagliaferri et al., 2003), correlation (Protopapas et al., 2006), further feature space reduction (Debosscher, 2009) or more novel methods (Huijse et al., 2012). It should be noted that the family of stars with the label "stellar variable" is a large and diverse population: eclipsing binaries, irregularly pulsating variables, nova (stars in outburst), multi-model variables, and many others are frequently processed using the described methods despite the underlying stellar variability functionality not naturally lending itself to Fourier decomposition and the associated assumptions that accompany the said decomposition. Indeed this is why Szatmary et al. (1994); Barclay et al. (2011); Palaversa et al. (2013) and others suggest using other decomposition methods such as discrete wavelet transformations, which have been shown to be powerful in the effort to decompose a time series into the time-frequency (phase) space for analysis (Torrence and Compo, 1998; Bolós and Benítez, 2014; Rioul and Vetterli, 1991). It is noted that the digital signal processing possibilities beyond Fourier domain analysis time series comparison and wavelet transformation are too numerous to outline here; however the near complete review by Fulcher et al. (2013) is highly recommended.

2. Slotted Symbolic Markov Modeling

The discussion of the Slotted Symbolic Markov Modeling (SSMM) algorithm encompasses the analysis, reduction and classification of data. Since the *a priori* distribution of class labels are roughly evenly distributed for both experimental studies, the approach uses a multi-class classifier. Should the class labels with additional data become unbalanced, other approaches are possible (Rifkin and Klautau, 2004). Data specific challenges, associated with astronomical time series observations, have been identified as needing to be addressed as part of the algorithm design.

2.1. Algorithm design

Stellar variable time series data can roughly be described as passively observed time series snippets, extracted from what is a contiguous signal (star shine) over multiple nights or sets of observations. The time series signatures have the potential to change over time, and new observations allow for the increased opportunity for an unstable signature over the long term. Astronomical time series data is also frequently irregular, i.e., there is often no associated fixed Δt over the whole of the data that is consistent with the observation. Even when there is a consistent observation rate, this rate is often broken up because of observational constraints. The stellar variable moniker covers a wide variety of variable types: stationary (consistently repeating identical patterns), non-stationary (patterns that increase/decrease in frequency over time), non-regular variances (variances that change over the course of time, shape changes), as well as both Fourier and non-Fourier sequences/patterns. Pure time-domain signals do not lend themselves to signature identification and pattern matching, as their domain is infinite in terms of potential discrete data (dimensionality). Not only must a feature space representation be found, but the dimensionality should not increase with increasing data.

Based on these outlined data/domain specific challenges (continuous time series, irregular sampling, and varied signature representations) this paper will attempt to develop a feature space extraction methodology that will construct an analysis of stellar variables and characterize the shape of the periodic stellar variable signature. A number of methods have been demonstrated that fit this profile (Grabocka et al., 2012; Fu, 2011; Fulcher et al., 2013), however many of these methods focus on identifying a specific time series shape sequence in a long(er) continuous time series,