A Dynamic, Modular Intelligent-Agent framework for Astronomical Light Curve Analysis and Classification

Paul R. McWhirter[12], Sean Wright[1], Iain A. Steele[2], Dhiya Al-Jumeily[1], Abir Hussain[1] and Paul Fergus[1]

[1]Liverpool John Moores University, Applied Computing Research Group, Faculty of Engineering and Technology, Byrom Street, Liverpool, L3 3AF, UK.
P.R.McWhirter@2014.ljmu.ac.uk
S.Wright@2012.ljmu.ac.uk
{D.Aljumeily, A.Hussain, P.Fergus}@ljmu.ac.uk

[2]Liverpool John Moores University, Astrophysics Research Institute, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool, L3 5RF, UK.
I.A.Steele@ljmu.ac.uk

Abstract. Modern time-domain astronomy is capable of collecting a staggeringly large amount of data on millions of objects in real time. This makes it almost impossible for objects to be identified manually. Therefore the production of methods and systems for the automated classification of time-domain astronomical objects is of great importance. The Liverpool Telescope has a number of wide-field image gathering instruments mounted upon its structure. These instruments have been in operation since March 2009 gathering data of multi-degree sized areas of sky around the current field of view of the main telescope. Utilizing a Structured Query Language database established by a pre-processing operation upon the resultant images, which has identified millions of candidate variable stars with multiple time-varying magnitude observations, we applied a method designed to extract time-translation invariant features from the time-series light curves of each object for future input into a classification system. These efforts were met with limited success due to noise and uneven sampling within the time-series data. Additionally, finely surveying these light curves is a processing intensive task. Fortunately, these algorithms are capable of multi-threaded implementations based on available resources. Therefore we propose a new system designed to utilize multiple intelligent agents that distribute the data analysis across multiple machines whilst simultaneously a powerful intelligence service operates to constrain the light curves and eliminate false signals due to noise and local alias periods. This system will be highly scalable, capable of operating on a wide range of hardware whilst maintaining the production of accurate features based on the fitting of harmonic models to the light curves within the initial Structural Query Language database.

Keywords: Data analysis methods, Big Data Mining, Machine Learning, Uneven Time-Series analysis, Light Curve analysis, Variable Stars, Binary Stars, Harmonic Regression, Harmonic Feature Extraction, Multi-agent systems, Period Detection.

1 Introduction

Astronomy is entering a period of unprecedented data gathering capability. Advances in observational, storage and data processing technologies have allowed for extended sky surveys such as the Sloan Digital Sky Survey (SDSS) to be conducted and exploited [1]. Within the next decade a number of even larger surveys are planned such as the Large Synoptic Survey Telescope (LSST) [2]. Technology is now at a point where it has become possible to gather data on wide regions of the sky repeatedly over variable time periods [3]. This data can be analyzed through the use of periodograms to identify periodic structure. This periodic structure can then be used to create physical models by fitting weighted regression learning algorithms. These methods can provide us with valuable knowledge about the presence and classification of astronomical objects that are periodically changing in time as well as identifying transient phenomena [4].

Time domain astronomy is a research area characterized by the large datasets generated by sky surveys containing time-series data [5]. Time-series data contains information on the temporal component of measurements and the whole time-series contains observations across multiple epochs. In many environments this time-series data can be gathered regularly and often. This simplifies the statistical analysis by supplying a large number of observations with consistent time intervals between individual observations. However, in Time-Domain Astronomy, it is common for these observations to have a significantly uneven distribution in time with inconsistent intervals between observations [6]. A major cause of this is weather limitations that can prevent telescope operation for uncertain periods of time. As a result, astronomy maintains a demand for data processing techniques capable of the automated processing of time-series data on individual objects that can contain observations over the space of days followed by no additional observations for a period of months.

In this paper we propose a new theoretical platform for the analysis of this vast quantity of time-domain astronomy data by introducing intelligent-agents. A typical agent is a type of computer system that is embedded in a type of environment that is capable of conducting an autonomous action within that environment in order to meet its objectives. An Intelligent Agent on the other hand is an extension of this approach with the ability to make decisions and adapt to its changing environment.

The rest of this paper is structured as follows. In Section 2, the background of time-domain astronomy is discussed with reference made to the numerous classifiable objects. Section 3 introduces the Small Telescopes Installed at the Liverpool Telescope (STILT) instruments, wide field imaging devices and the pre-processing pipeline used to construct an Structured Query Language (SQL) time-series database from the raw images. In Section 4, the feature extraction is discussed through using light curve model fitting resulting in the extraction of important, magnitude and phase independent, features. In Section 5, a system utilizing intelligent-agents is proposed for

the successful processing and classification of numerous light curves. The final conclusions and proposals of future work are provided within Section 6.

2 Background

Astronomical time-series data is generated through the production of wide-angle images of the sky. The intensity of the image pixels is determined by the activation of the Charge-Coupled Device (CCD) cameras pixels by incoming light from multiple astronomical objects with some background noise and detection bias from the camera [3]. As a result, each image contains important information about the brightness of the detected objects. By identifying objects in multiple images with different observation times, information on the change of the brightness of these objects can be determined. The resulting brightness-over-time data for each individual object is called the objects light curve [3, 7].

Many astronomical objects exhibit brightness variability due to a large number of differing physical processes that uniquely influence an object's light curve. Therefore, the light curve can be used in the classification of variable objects based on the signature of these physical processes and the detection of unknown candidate objects or even unknown variability phenomena [8]. The first major type of variable astronomical phenomenon is Variable Stars [9]. Variable stars are unstable stars and undergo periods of pulsation where they grow and contract in size [10]. These size oscillations produce changes to the stars temperature and brightness resulting in a measurable change upon the light curves [3, 7]. Two common types of pulsating stars are Cepheid and RR Lyrae variables [10]. These different types of variability produce their own signature light curve profile. The light curves of these pulsating stars can be used to produce descriptive features. Models produced from these features can then be used to identify the class of candidate variable objects and the period of their oscillations. For certain classes of object which also exhibit a specific period-brightness relation, it can also be used to determine the object's mass the distance it is located from Earth [7].

A second important type of variable object is the eclipsing binary [11]. In these systems, two or more stars are in close proximity to each other and execute orbits around a common gravitational center-point. The close proximity of the stars often means that they cannot be distinguished on an image and appear as a single source of light. Variations in these objects are caused by the plane of the orbit aligning with the view from Earth. As a result, one star periodically passes in front of another resulting in a change in the brightness of the source of light in the astronomical images. Analysis of the light curves of eclipsing binary candidate light sources can be used to determine the number and types of star present within the system as well as their mass and orbital period. This process can also be caused by exoplanetary transits, another important research area in modern astronomy where one of the objects is a planet orbiting a parent star.

Finally, there are also transient events that result in harder-to-predict phenomena [4]. Flare Stars are stars that can undergo occasional outbursts due to magnetic and plasma processes within their atmospheres. These events can be repetitive but not usually with the degree of periodicity of variable stars. For purely transient events, two of the most studied examples are Novae and Supernovae caused by the cataclysmic eruption of stellar material, producing some of the brightness objects in the known universe as the victim star is destroyed or badly disrupted during the event. Other common transient phenomena are Gamma Ray Bursts caused by a number of physical events including the collapse of supermassive stars into black holes and Gravitational Microlensing events when the light from a background object is brightened by lensing due to a dark, massive object moving in the foreground.

3 The STILT Dataset

The Small Telescopes Installed at the Liverpool Telescope (STILT) dataset is a wide field object SQL database. It contains 1.24 billion separate object observations of 27.74 million independent stellar objects. It was generated through the pre-processing of observational images gathered by the STILT instruments [3]. The Liverpool Telescope is located at the Observatorio del Roque de los Muchachos on La Palma [12]. The STILT instruments consist of three cameras with varying field of views mounted directly to the body of the main Liverpool Telescope aimed co-parallel with the main telescope's field of view.

The first instrument, SkycamA is capable of imaging the entire sky from La Palma. It is primarily used for monitoring the status of weather but it can be of use in the detection of bright transient objects. This camera does not contribute any observations to the STILT database. The next camera is named SkycamT and is responsible for most of the observations. It is a single CCD camera capable of detecting light across the visible spectrum with a wide-angle lens with a field of view of 21 by 21 degrees and a magnitude limit of +12. Finally, the remainder of the database is constructed from observations by the SkycamZ instrument. This instrument contains a CCD camera which is also capable of detecting light from across the visible spectrum attached to a small telescope with a field of view of only one by one degree but with a greatly increased magnitude limit of +18. The database contains time-series data on the magnitude of detected objects over a period of time from March 2009 to March 2012 [3]. It makes use of the SExtractor software to identify potential objects and the astrometry.net software to correctly assign celestial coordinates to each object [13, 14].

As the Skycam images are centered on the view of the main Liverpool Telescope, observations of specific objects are only recorded when they are within the field of view of the camera as the telescope is focused within the vicinity of the objects. This results in time-series with uneven length gaps between observations, greatly increasing the difficulty of identifying variations in the magnitude of the observations.

4 Feature Extraction

The astronomical time-series gathered by the Skycam instruments can contain a number of periodic features buried within measurement and sampling noise. In order to identify variable objects, these features must be extracted from the data and expressed as a fixed set of parameters that describe the state of the specific object. By collecting sufficient features it may be possible for a classification algorithm to successfully discriminate between many different types of variable object based on the physics and timings that govern them [15].

We begin our analysis using a methodology proposed by Debosscher et al. in 2007 and improved upon by Richards et al. in 2011 [16, 4]. The goal is to describe the time-series data for each object as a set of harmonic features that are invariant to the objects mean magnitude and time-translation phase allowing the features to be directly compared to other objects of differing classes. The whole SkycamT database used in this investigation is 180 GB in size with 20 GB of indexes for faster query response times. For each object, a set methodology is applied to generate an associated feature vector. The database is queried for all observations of a specific object. The returned table has its magnitude, modified Julian date and magnitude error columns retrieved. The identification of the dominant periodic oscillation within the object's time-series is then required. There are a number of possible algorithms that can be deployed on uneven time-series.

Phase dispersion minimization [17] and the String Length Lafler-Kinman (SLLK) statistic can identify how well-aligned data points are placed in phase-space across a sample range of periods [18]. This is accomplished through computing the distance between each data point in phase space. This calculation is performed across a frequency spectrum of candidate frequencies. Upon the alignment of the data points at a frequency close to (or a multiple of) the true frequency, this statistic is minimized [18]. An extension to this idea of aligning data within phase space is a recently proposed periodogram based on the Blum-Kiefer-Rosenblatt (BKR) statistical independence test [19]. Instead of utilizing the alignment of the data such as in the string-length methods, a rank correlation test is performed on the phase-folded data within each candidate period phase space. As the phase folded light curve aligns at a strong period, the correlation between the magnitude and phase of each data point rises. The peak of this correlation is a good candidate frequency and can be determined from time-series of limited size [19]. There also exist Information Theoretic approaches such as slotted correntropy and the improved Correntropy Kernelized Periodogram (CKP) [20] that have proven to be very effective and are a focus for future initiatives on the Skycam database [7]. However, for this initial investigation, the methodology proposed by Debosscher et al. in 2007 is utilized [16]. Therefore, a Lomb-Scargle Periodogram (LSP) is utilized to identify the primary periodic signal within the data. The Lomb-Scargle Periodogram algorithm utilized in this method was produced by Thomas Ruf originally for biological periodic pattern recognition [21].

The Lomb-Scargle Periodogram uses a least-squares spectral analysis. It is a method of estimating the frequency spectrum of time-series data by the fitting of multiple sinusoids to the data using least-squares regression [6, 4]. Like the phase dispersion minimization this method is performed over a frequency range resulting in the statistic normalized power that has a larger value if the fitted sinusoid has a lower chi-squared error with a candidate frequency. It operates over a frequency range with a finite set of candidate frequencies separated by intervals. As the objects in the database can have low and high period variations, this interval is set as constant to produce a uniform sample across the full frequency range. The lowest frequency is the longest period that can be expected to be detected by the periodogram. It is defined as the reciprocal of the difference between the maximum and minimum modified Julian Date of the object's observations named the total observation time t_{tot}. The maximum frequency is an interesting discussion area and is related to the minimum periods that can be found from an object's data. Hypothetically scanning down to the minimum possible periods for variable stars is recommended. However, for some pulsating white dwarf stars, this can be as low as 1-2 minutes [16]. The Lomb-Scargle Periodogram is not limited to a Nyquist frequency [4]. However, at very high frequencies many noisy peaks can be generated as any data can be fitted well to a model with such high frequencies. In previous methods, a Pseudo-Nyquist frequency was proposed for the determining of a maximum frequency for unevenly sampled data which approximates the Nyquist frequency by taking the mean of the individual time intervals between the observations of an object. This equation is shown in equation 1.

$$f_{nyq} = 0.5 \ \langle \frac{1}{\Delta T} \rangle \qquad (1)$$

Where f_{nyq} is the Pseudo-Nyquist frequency and ΔT is the time intervals between observations of an object. In the Richards et al. methodology, the frequencies could rise beyond this value with the periodogram normalized power subjected to a soft penalty term that weakens the peaks beyond the Pseudo-Nyquist frequency based on the relative difference between this frequency and the candidate frequency [4]. So far this mechanism has yet to be implemented in this investigation of the STILT database. Currently the maximum frequency is simply limited to the Pseudo-Nyquist frequency as utilized in the Debosscher et al. methodology [16].

This frequency is determined by the mean intervals between observations in an uneven time-series. Gaps in observations are not considered uneven observations and instead are just considered times with a lack of observations and do not contribute to the calculation of the Pseudo-Nyquist frequency. Despite this distinction, a globally accepted definition of a 'gap' and an 'uneven sample' was not identified. Therefore we attempt to approach this problem to provide a good solution. Theoretically, the time intervals between observations could all be considered gaps. This is not really possible as the start and stop times of Skycam observations are unlikely to be an integer number of minutes (the interval between exposures). Ignoring this, theoretically the sampling rate could be the Nyquist frequency of the evenly sampled exposure intervals which is half of the reciprocal of a minute, 720 cycles per day, equivalent to

a period of 2 minutes. For the example with no gaps, every interval is included in the calculation of the Pseudo-Nyquist frequency. This results in a frequency that may result in the loss of low period signals from the periodogram.

Evidence suggests that neither of these options is ideal. For example, the dominant period of the star RR Lyrae has been determined as 0.5668 days [22]. From the STILT data for this star, if none of the time intervals are considered gaps, the Pseudo-Nyquist frequency results in a minimum detectable period of approximately 0.7 days. No candidate frequencies would be evaluated by the periodogram near the 0.5668 day period. By manually allowing for larger frequency range, this period is detected as the dominant period. Conversely, if the 720 cycles per day frequency is used as the Nyquist-frequency as described above, the frequency spectrum is dense enough to result in extreme processing load from the periodogram.

As previously discussed, an adequate definition of what time intervals are to be considered gaps and which are to be considered unevenly sampled intervals would be advantageous. Currently we have defined a gap as an interval greater than two standard deviations for this method. This is unlikely to be an acceptable final answer as the standard deviation has limited meaning on non-normal distributions. Future experiments will investigate this question in order to provide a more concrete definition.

Finally, the frequency step between candidate frequencies must be determined to produce a frequency spectrum with a finite number of frequencies. Both Debosscher et al. and Richards et al. make use of a frequency step of 0.1 divided by the total observation time as defined above [16, 4]. In this method the frequency step is defined as shown in equation 2.

$$f_{step} = \frac{1}{ovsm \times t_{tot}} \qquad (2)$$

Where f_{step} is the frequency step, t_{tot} is the total observation time as defined above and $ovsm$ is the oversampling factor. When the oversampling factor is set to 10 the frequency step equates to that used in the previous methods [16, 4]. Our early experiments suggest this might be too fine a frequency grid for the STILT data as noisy peaks seem to be produced with this oversampling factor for some objects such as the previously mentioned RR Lyrae. The best results for this star appear to be generated with an oversampling factor of between 2 and 6. Interestingly, for the light source Algol, an eclipsing binary system, an oversampling factor of 10 is required to identify the correct period. This remains a challenge in the development of this method. When the Lomb-Scargle Periodogram is applied to the time-series observational data, the frequency associated with the maximum power is recorded as the primary frequency.

An additional operation to eliminate periodicities caused by the sampling times is also performed. This is accomplished by creating a randomized dataset based on the object magnitude but with the same observation times. Any periodicities found by the periodogram in this randomized dataset are purely due to the sampling times. There-

fore, this power spectrum is subtracted from the power spectrum produced by the true time-series to eliminate any peak values not associated with actual variability in the objects magnitude values. Additionally, local synodic periods caused by light variations near Earth such as 0.5 days (solar half-day cycle), one day (solar day cycle) and 29.5 days (lunar synodic period) and associated alias periods are removed.

Upon the determination of the candidate period, a harmonic model with a linear trend is then fitted with this period across the time-series data. In order for the model to have the degree of freedom required to accurately fit the data, artificial data points are bound into the time-series. These artificial data points have a uniform distribution in time and a magnitude equal to the mean magnitude of the time-series. The artificial data points must be assigned zero weight as to not contribute to the model optimization. As the weights were defined as the reciprocal of the magnitude error, the artificial data points are assigned a magnitude error of positive infinity. Weighted linear regression is then performed on this new time-series to fit a four-harmonic sinusoid model using the period detected by the periodogram. This model has ten coefficients and is demonstrated by Equation 3. The symbols within the equation are explained under Equation 4.

$$y(t) = ct + \sum_{j=1}^{4}\{a_j \sin(2\pi j f_1 t) + b_j \cos(2\pi j f_1 t)\} + b_0 \quad (3)$$

This model is then subtracted from the time-series in a process called pre-whitening. This is done as to eliminate any periodic activity within the time-series based on the dominant period detected by the periodogram. This pre-whitened time-series is then used to identify a second period independent of the first dominant period. A harmonic model is then fit for this period and subtracted off in a second pre-whitening phase. Finally, a third period is identified independent to the first two periods. The time-series is then restored to the original time-series archived prior to these operations. A harmonic best-fit is again computed by weighted linear regression using a model with twenty six coefficients is utilized as shown in Equation 4.

$$y(t) = ct + \sum_{i=1}^{3}\sum_{j=1}^{4}\{a_{ij} \sin(2\pi j f_i t) + b_{ij} \cos(2\pi j f_i t)\} + b_0 \quad (4)$$

Where the b_0 parameter is the mean magnitude of the light curve and the c parameter is the linear trend of the time-series. By calculating this linear trend in the same regression operation as the sinusoidal models, a time-series with a non-integer number of wavelengths within the sampling period (with a corresponding trend) will not interfere with the linear trend caused by a gradual brightening or dimming of the object, an important feature for classification. The frequencies f_i and the coefficients a_{ij} and b_{ij} are retained and provide a good description of the light curve as long as it is periodic and well-described by a sum of sinusoids. These coefficients are not yet time-translation invariant and must be transformed into better descriptors of the light curve. This is accomplished by transforming the Fourier coefficients into a set of amplitudes A_{ij} and phases PH_{ij}. The amplitudes are computed by equation 5 and the phases by equation 6 derived from trigonometric identities [16].

$$A_{ij} = \sqrt{{a_{ij}}^2 + {b_{ij}}^2} \quad (5)$$

$$PH_{ij} = arctan(b_{ij},\ a_{ij}) \quad (6)$$

The phases are not yet time-translation invariant and are defined relative to PH_{11}, the phase of the first harmonic of the dominant period using equation 7.

$$PH_{ij}{}' = arctan(b_{ij},\ a_{ij}) - \left(\frac{i\,f_i}{f_1}\right) arctan(b_{11},\ a_{11}) \quad (7)$$

The phases are then constrained between $-\pi$ to $+\pi$ by the transformation in equation 8. For simplicity the double dash is dropped through the rest of the paper.

$$PH_{ij}{}'' = arctan\left(sin(PH_{ij}{}'),\ cos(PH_{ij}{}')\right) \quad (8)$$

As long as the light curves are well described as monoperiodic this results in the production of twenty eight features that are time-translation invariant. Monoperiodic light curves are those that oscillate with one dominant period primarily due to radial pulsations [23]. This assumption does not hold for all potential variable stars but as the primary period is usually highly dominant, this assumption is a good approximation [16]. These features include the slope of the linear trend, the three frequencies used in the final harmonic model, the twelve amplitude coefficients and eleven phase coefficients (as PH_{11} is always zero it is discarded) and the ratio of data variance (called variance ratio) between the variance before the pre-whitening of the harmonic model of the primary period and after. This statistic is a strong indicator of the importance of the primary period to the light curve relative to the other periods. These features are assembled into a feature vector and recorded as a row within a table of feature vectors where the columns are features and the rows are each object found within the original SQL query criteria.

These features have been successfully implemented into classifiers by Debosscher et al in 2007 and Richards et al. in 2011 [16, 4]. These studies made use of well sampled datasets collected by orbital space telescopes and large collaborations. The SkycamT database has very sparse and noisy time-series data. Additionally, the Lomb-Scargle Periodogram is known to strongly identify periodicities for variable stars that are highly sinusoidal such as Mira-class variables. But it can also struggle with less sinusoidal light curves such as eclipsing binaries occasionally missing the period of offering a multiple of the correct period instead of the true period.

Figure 1 demonstrates the model produced by the described method for the SkycamT data collected on the star Mira, the prototype of the Mira class variables showing a clear sinusoidal oscillation. The period of Mira has been widely reported as 332 days, verified by surveys such as HIPPARCOS [24]. Despite the discussed weaknesses, if the periodogram returns a result similar to the stars correct period, the linear regression can produce an accurate model despite the prevalence of noise within the time-series data. The model is not a perfect fit but should be sufficient to generate

features within the ranges expected of Mira class variables.

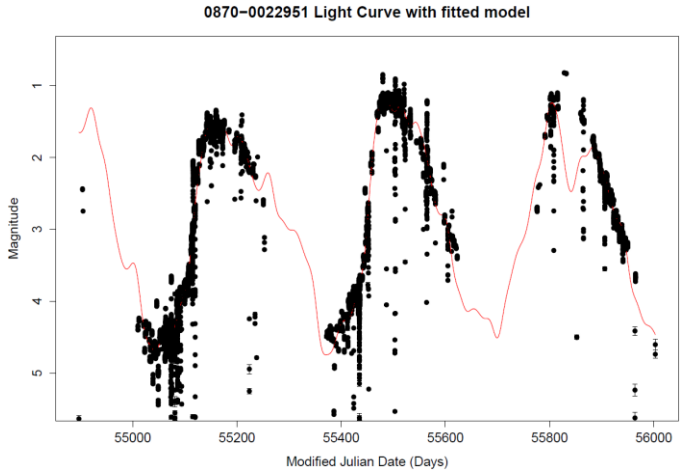

Fig. 1.
The light curve of the star Mira with a harmonic fit with a primary period of 316 days compared to the accepted value of 332 days. Despite this two week discrepancy, the harmonic model is a close match to the data producing features suitable for future classification algorithms.

A good solution for improving the accuracy of the period search is to incorporate a more powerful period finding algorithm such as the correntropy kernelized periodogram with information potential. This method does not assume the variations are sinusoidal and has verifiably improved success [7]. The artificial data points have allowed the harmonic fits the degree of freedom required to fit the unevenly sampled magnitude data resulting in powerful fits such as that demonstrated by figure 2.

Unfortunately this has also resulted in situations where the model deviates drastically in the non-sampled regions due to ‘noisy fringes’ in the data. The linear regression is resilient to noise and can evaluate the signals within very poor data. However, coupled with the uneven sampling rate, the noise can result in the linear regression entering a region devoid of data whilst fitting for a steeper or shallower gradient due to the noise of these last data points. As a result the amplitude of the sinusoids can peak beyond a physically realistic state. A new feature has been implemented which is assigned the value of one when the model peaks outside of the mean plus or minus two times the standard deviation of the object’s time-series and zero otherwise. This will allow for any future method to identify unrealistic fits due to this phenomenon.

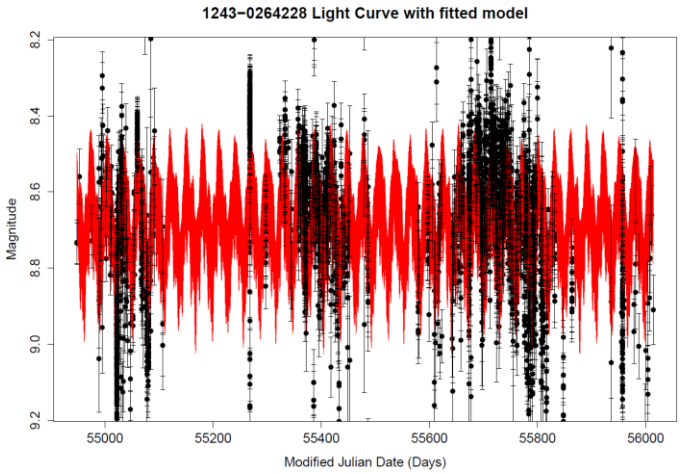

Fig. 2.

A harmonic fit on the star 1243-0264228. By allowing the model to vary within the non-sampled spaces, a superior harmonic fit can be constructed. However, the primary 29.6 day period seen here is likely due to variations in ambient light due to the lunar synodic period.

Finally, as the object database exhibits 27.74 million independent stellar objects, the performance of the analysis system is of great concern. Fortunately, the weighted linear regression is implemented using a very efficient normal equation method. Therefore the Lomb-Scargle Periodogram which is $O(N^2)$ in processing complexity is the primary processing component of this analysis method. This order is a result of needing to run every observation of an object over a high resolution frequency spectrum to extract the dominate periods otherwise important harmonic variations may be missed. The observations cannot be subjected to stochastic or mini-batch operations as each observation refined the range of the harmonic least-squares fitting. They are all interdependent for this operation. On the other hand, the frequency spectrum can contain tens of thousands of candidate periods to be evaluated. Each period, regardless of proximity to each other, is completely independent. Therefore, this loop can be performed stochastically in parallel as the result is the same whether each candidate period is run on the same CPU core or across thousands of separate CPU cores. This allows the Lomb-Scargle Periodogram to be heavily multithreaded if the hardware exists for considerable performance improvement.

5 Proposed Solution

In order to significantly improve performance, we propose a system that makes use of multiple intelligent agents to subdivide the processing tasks between multiple clusters whilst simultaneously a new ‘intelligence service’ will constantly monitor the models being generated by the individual agents. This intelligence system will learn stochastically as models are continuously generated for the light curves of different objects.

We use intelligent agents for this task as they are capable of continuously making decisions based on the quantity of data being processed and the system resources available [25]. The proposed framework will be dynamic and modular as well as scalable in nature. The modular aspect of the framework will allow a user the ability to develop tasks as well as the type of strategies used when replicating an agent. This dynamic aspect of the framework allows it to be scalable by using replication.

Due to the modularity of the framework, a user has the ability to define their tasks. Depending on the tasks and the environment that will be operated; the agent will either know in advance when it is first started or at which time in the future. There are two distinct scenarios in which these agents can operate. In the first scenario, an agent is processing some calculations in a static SQL Database. The agent will know in advance the size of the tables required for the tasks by conducting a count to retrieve the size and it can compute the amount of agents and the tasks that needs to be processed on each agent depending on the user's pre-defined discretion. In the second scenario, an agent is processing in-frequent amounts of data bursts, sometimes small amounts of data and sometimes large amounts of data in a timely manner. The agent will have to monitor its resources and depending on the threshold defined by the user the agent can determine if it can complete the tasks in the required time [25]. When an agent has decided it can either not complete the tasks or has reached a point that it no longer can complete its task due to the load then it will replicate itself. This is the dynamic and scalable aspect of the framework because the agent has to make a decision that it cannot complete its tasks and has to offload some tasks to another agent. This decision is made from what we call 'the dynamic strategy' by looking at the resources of the host machine and comparing them with the threshold defined by the user. This is only used for the replication phase.

Once an agent has requested a replication then that agent will only run the first task and the newly replicated agent will run the remaining tasks. When that agent has decided that it can no longer complete the tasks in a suitable time period then it will replicate itself again. Every time an agent starts to struggle due to the load then it will replicate itself until each task is running on a single agent. Previously replication has mainly been used for fault tolerance [26, 27, 28]. However, we are using replication for performance and scalability increase. Some tasks could possibly be synchronous and some tasks may be asynchronous. This means that tasks could possibly rely on results computed from previous tasks and some tasks may be independent and can run in parallel. We define two types of task, synchronous tasks that rely on the result from a previous task or asynchronous tasks that rely on a type of data either from a previous tasks result or based on the source data.

The system will contain two types of agents, 'The Task Agent' and 'The Resource Management Agent / Resource Controller'. When an agent wants to replicate itself, it will contact all Resource Management agents on the network to see if any resources are available for it to replicate and responses will be send back informing the replicating agent of the amount of usable resources available on each of the machine. If an

appropriate amount of resources are available then the agent will request the resource management agent to replicate the task agent. When replicating, the resource management agent will spawn a container for the replicated agent to run within with a specific amount of resources that the user has defined for the task [29]. We choose to containerize the agent for two reasons. Firstly, containerizing the agent allows us to isolate the agent from other agents on the machine while also having the capability to deploy the agent almost instantly due to their minimal runtime requirements. Secondly, we can configure the container to only use a specific amount of resources, for example a predefined number of CPU Cores and percentage of usage for each CPU, allowing for more control over the resources allocated to the replicated agent [30]. Figure 3 demonstrates the structure of this replication sequence.

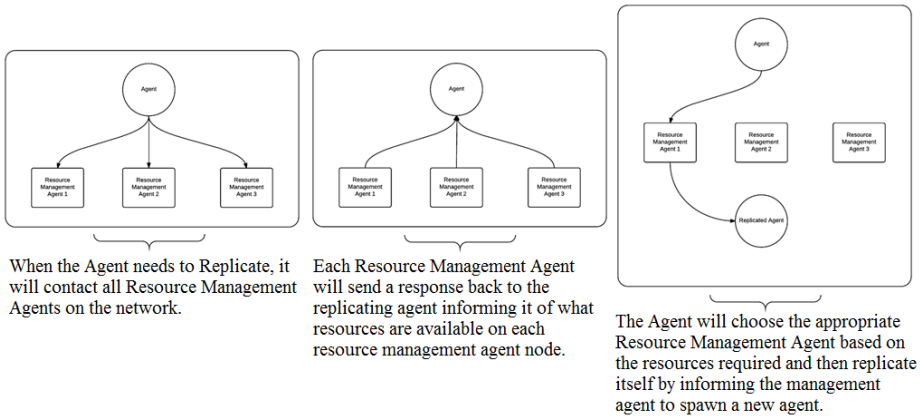

Fig. 3.
The sequence of operations involved in agent replication.

The proposed system for the processing of astronomical light curves will be using the static Skycam database. This means that no new entries are ever added to the database and the data is left standalone. We will use a static based strategy as we know that the database is not going to be updated. This strategy will count the number of objects in the database and then based on the quantity of resources allocated to each agent, a decision that the programmer has made when building their strategy, will either allow a single agent to conduct all the tasks utilizing the total resources (as shown in figure 4) or each machine running an instance of the Resource Management Agent will spawn task agents to run the tasks for multiple objects simultaneously.

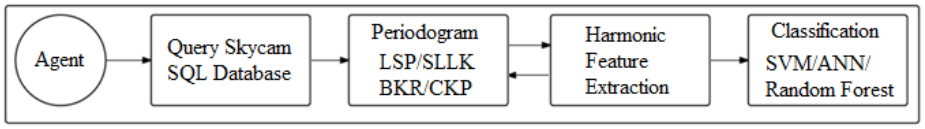

Fig. 4.
A single agent is capable of running every task.

Whilst the multi-agents distribute the processing tasks generating the harmonic best-fit models, a second intelligence service operates independently. This service monitors the outbound models and determines whether they are physically realistic or a result of overfitting on noise or a false period. This service with have a confidence threshold which dictates how well the model has fit the time-series data based on its trained state. If it has a low confidence it attempts to modify the model based on previous patterns that have been discovered within the multiple objects light curves processed previously. The intelligence service is always operating using stochastic, supervised learning trained using the known light curves of well sampled objects of each variability class to continuously improve its prediction based on the models being continuously produced by the agents on the multiple objects currently being processed. This system is envisaged to use a form of neural network containing multiple models based on the different light curve profiles discovered. Both Recurrent and Convolutional neural networks have been shown to be potent at predicting time-series [31, 32]. Additionally, these methods can be extended into deep learning through the addition of more layers if the light curves require the production of more powerful features. In the event of a harmonic fit model receiving a low confidence, the intelligence service will attempt to construct a new harmonic model using neural networks trained on the previous high confidence models through the stochastic learning process.

In order to reduce noise, techniques such as wavelet analysis can also be applied [31]. Wavelet analysis has previously been used successfully in the process of creating abstract images of Gamma Ray Burst transient events that retain both temporal and spectral features for classification [33]. It is very difficult to simply filter the noise from the time-series data as the noise has an amplitude very similar to many of the signals. This amplitude is approximately 0.2 magnitudes. More work is needed in order to refine the design of this intelligence service. However, we are hopeful that the combination of scalable processing and accurate time-series predictions will lead to high performance processing of the STILT database through the generation of robust features by supervised and unsupervised learning for future multi-class classification analysis.

6 Conclusion and Future work

The weighted linear regression harmonic best-fit models can be used to produce time-translation invariant features from uneven time-series. The STILT database contains many objects with sufficient noise and uneven sampling to result in poor or physically unrealistic harmonic models. The Lomb-Scargle Periodogram can produce multiples of the correct period and occasionally it misses the periodic signal completely. This problem is exaggerated by light curves exhibiting highly non-sinusoidal signals such as eclipsing binaries. Replacing the Lomb-Scargle Periodogram with a more powerful and less limiting algorithm such as the correntropy kernelized periodogram might alleviate this problem. The proposed solution seeks to introduce scalability and multi-

threading for the high performance processing of the STILT database. This is accomplished through the use of a multiple intelligent agent platform. This platform is capable of distributing the processing tasks across multiple available machines as required based on the processing workload and available resources. Finally, a new intelligence service using more powerful machine learning algorithms such as Recurrent and Convolutional neural networks can regulate the models generated by the harmonic best-fit producing results that are consistent with astrophysical processes despite the degree of freedom available to the weighted linear regression. Our future work will involve the incorporation of the proposed methods into a newly developed data analytics platform. Following this, the models produced can be evaluated through testing previously classified variable objects in the STILT dataset as well as sourcing external datasets for comparative results. These efforts allow the production of robust light curve features that are well placed for future incorporation into a powerful multi-class classification system to rapidly and intelligently perform automated identification of all variable objects within the STILT database. This methodology can then be extended to all other noisy light curve for a complete survey of all variable objects present within the night sky.

Acknowledgment

This work was funded through a Liverpool John Moores University scholarship in partial fulfilment of the requirements for the degree of Doctor of Philosophy. This paper makes use of the SkycamT database as developed by Neil Mawson as part of his PhD research and Professor Iain Steele of the Astrophysics Research Institute at Liverpool John Moores University. The raw images were gathered by the STILT instruments located on the Liverpool Telescope located on La Palma, Canary Islands and administered by Liverpool John Moores University.

References

[1] D. G. York, J. Adelman, et al., "The Sloan Digital Sky Survey: Technical Summary," The Astronomical Journal, vol. 120, no. 3, pp. 1579-2000, 2000.

[2] Z. e. a. Ivezic, "LSST: from science drivers to reference design and anticipated data products," ArXiv e-prints, 2011.

[3] N. R. Mawson, I. A. Steele and R. J. Smith, "STILT: System design and performance," Astronomische Nachrichten, vol. 334, no. 7, pp. 729-737, 2013.

[4] J. W. Richards, D. L. Starr, et al., "On Machine-Learned Classification of Variable Stars with Sparse and Noisy Time-Series Data," The Astrophysics Journal, vol. 733, no. 1, p. 10, 2011.

[5] S. Vaughan, "Random time series in astronomy," Philosophical Transactions of the Royal Society, vol. 371, no. 20110549, 2011.

[6] J. D. Scargle, "Studies in Astronomical Time Series Analysis. II. Statistical aspects of spectral analysis of unevenly spaced data," The Astrophysical Journal, vol. 263, pp. 835-853, 1982.
[7] P. Huijse, P. A. Estévez, et al., "An Information Theoretic Algorithm for Finding Periodicities in Stellar Light Curves," IEEE Transactions on Signal Processing, vol. 60, no. 10, pp. 5135-5145, 2012.
[8] P. Protopapas, J. M. Giammarco, et al., "Finding outlier light curves in catalogues of periodic variable stars," The Royal Astronomical Society, Monthly Notices, vol. 369, pp. 677-696, 2006.
[9] L. Eyer and N. Mowlavi, "Variable stars across the observational hr diagram," Journal of Physics: Conference Series, vol. 118, no. 1, p. 012010, 2008.
[10] J. Percy, Understanding Variable Stars, Cambridge University Press, 2007.
[11] D. M. LaCourse, K. J. Jek, et al., "Kepler eclipsing binary stars - VI. Identification of eclipsing binaries in the K2 Campaign o data set," Monthly Notices of the Royal Astronomical Society, vol. 452, no. 4, pp. 3561-3592, 2015.
[12] I. A. Steele, R. J. Smith, et al., "The Liverpool Telescope: performance and first results.," in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 2004.
[13] E. Bertin and S. Arnouts, "SExtractor: Software for source extraction," Astronomy & Astrophysics Supplement Series, no. 117, pp. 393-404, 1996.
[14] D. Lang, D. W. Hogg, et al., "Astrometry.net: Blind astrometric calibration of arbitrary astronomical images," The Astronomical Journal, vol. 139, no. 5, p. 1782, 2010.
[15] J. S. Bloom and J. W. Richards, "Data Mining and Machine-Learning in Time-Domain Discovery and Classification," in Advances in Machine Learning and Data Mining for Astronomy, Taylor & Francis Group, 2011.
[16] J. Debosscher, L. M. Sarro, et al., "Automated supervised classification of variable stars I. Methodology," Astronomy and Astrophysics, no. 475, pp. 1159-1183, 2007.
[17] R. F. Stellingwerf, "Period Determination using Phase Dispersion Minimization," The Astrophysical Journal, vol. 224, pp. 953-960, 1978.
[18] D. Clarke, "String/Rope length methods using the Lafler-Kinman statistic," Astronomy and Astrophysics, vol. 2, no. 386, pp. 763-774, 2002.
[19] S. Zucker, "Detection of Periodicity Based on Independence Tests - II. Improved Serial Independence Measure," Monthly Notices Letters of the Royal Astronomical Society, vol. 1, no. 457, pp. 118-121, 2016.
[20] W. Liu, P. P. Pokharel and J. C. Principe, "Correntropy: A Localized Similarity Measure," in Neural Networks, 2006. IJCNN '06. International Joint Conference on, Vancouver, BC, 2006.
[21] T. Ruf, "The Lomb-Scargle Periodogram in Biological Rhythm Research: Analysis of Incomplete and Unequally Spaced Time-Series," Biological Rhythm Research, no. 30, pp. 178-201, 1999.
[22] K. Kolenberg, S. Bryson, et al., "Kepler photometry of the prototypical Blazhko star RR Lyr: An old friend seen in a new light," Monthly Notices of the Royal Astronomical Society, no. 411, p. 878, 2011.

[23] C. Aerts, S. V. Marchenko, et al., "Delta Ceti is not Monoperiodic: Seismic Modeling of a Beta Cephei star from MOST Space-based Photometry," The Astrophysical Journal, vol. 642, pp. 470-477, 2006.
[24] T. R. Bedding and A. A. Zulstra, "HIPPARCOS Period-Luminosity relations for Mira and semiregular variables," The Astrophysical Journal, vol. 506, pp. 47-50, 1998.
[25] L. Padgham and M. Winikoff, Developing Intelligent Agent Systems, A practical guide, Melbourne, Australia: John Wiley & Sons Ltd, 2004.
[26] A. d. L. Almeida, S. Aknine, et al., "Plan-Based Replication for Fault-Tolerant Multi-Agent Systems," in Parallel and Distributed Processing Symposium, 2006, IPDPS 2006. 20th International, Rhodes Island, 2006.
[27] Z. Guessoum, N. Faci and J.-P. Briot, "Adaptive replication of large-scale multi-agent systems: towards a fault-tolerant multi-agent platform," in SELMAS '05 Proceedings of the fourth international workshop on Software engineering for large-scale multi-agent systems, New York, NY, USA, 2005.
[28] D. Sylvain, Z. Guessoum and M. Ziane, "Adaptive Replication in Fault-Tolerant Multi-Agent Systems," in Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on, Lyon, 2011.
[29] Docker, "What is Docker? Understand how Docker works and how you can use it.," 2016. [Online]. Available: https://www.docker.com/what-docker. [Accessed 01 03 2016].
[30] Docker, "Docker Run Reference," 2016. [Online]. Available: https://docs.docker.com/engine/reference/run/. [Accessed 01 03 2016].
[31] M. Langkvist, L. Karlsson and A. Loutfi, "A review of unsupervised feature learning and deep learning for time-series modeling," Pattern Recognition Letters, vol. 1, no. 42, pp. 11-24, 2014.
[32] M. Dalto, "Deep neural networks for time series prediction with applications in ultra-short-term wind forecasting," in Industrial Technology (ICIT), 2015 IEEE International Conference on, Seville, 2015.
[33] T. N. Ukwatta and P. R. Wozniak, "Integrating Temporal and Spectral Features of Astronomical Data Using Wavelet Analysis for Source Classification," in 2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Imaging: Earth and Beyond , Washington DC, 2015.