Southern Horizons in Time-Domain Astronomy
Proceedings IAU Symposium No. 339, 2019
E. Griffin, ed.

doi:10.1017/S1743921318002855

Searching for Pulsating Stars Using Clustering Algorithms†

R. Kgoadi[1], I. Whittingham[1] and C. Engelbrecht[2]

[1]College of Science and Engineering, James Cook University, Townsville, Australia
email: refilwe.kgoadi1@my.jcu.edu.au

[2]Physics Department, Faculty of Science, University of Johannesburg, South Africa

Abstract. Clustering algorithms constitute a multi-disciplinary analytical tool commonly used to summarise large data sets. Astronomical classifications are based on similarity, where celestial objects are assigned to a specific class according to specific physical features. The aim of this project is to obtain relevant information from high-dimensional data (at least three input variables in a data-frame) derived from stellar light-curves using a number of clustering algorithms such as K-means and Expectation Maximisation. In addition to identifying the best performing algorithm, we also identify a subset of features that best define stellar groups. Three methodologies are applied to a sample of *Kepler* time series in the temperature range 6500–19,000 K. In that spectral range, at least four classes of variable stars are expected to be found: δ Scuti, γ Doradus, Slowly Pulsating B (SPB), and (the still equivocal) Maia stars.

Keywords. Astronomical databases: surveys, stars: variables: other, methods: data analysis, methods: statistical

1. Introduction

The classification of variable stars is an initial and vital step of studies in asteroseismology. It is therefore crucial that it be performed with high precision. In the current data revolution, where high dimensionality is common and space-science data are available openly through a number of surveys, Astroinformatics is increasingly being used in astronomy as an analytical tool. Astroinformatics is a sub-discipline that applies machine-learning techniques to large astronomical data-sets in order to gain new insights rapidly and efficiently (Borne 2009). Traditional methods have proven less efficient for performing this significant stage in studies of variable stars. Since most stellar surveys provide data sets that contain physical (surface) properties of stars, the Harvard Classification scheme may still be used as an initial classification to infer the nature of candidate stars. A rapid second stage of asteroseismological classification inference can then be carried out using high-dimensional data-frames with features derived primarily from light-curves through clustering algorithms.

Cluster analysis is a data mining tool in which objects in a data-frame are separated into groups based on their similarities, by minimising objective functions. This is a form of unsupervised learning because object isolations are not dependent on labels or target variables. Clustering algorithms can be either hierarchical or partitioning methodolgies, and the nature of the algorithm determines whether the groupings are hard or probabilistic. The discovery of stellar classes in large stellar data-sets may be achieved through 'hypothetical clustering approaches', whereby the results obtained are used to *confirm* a proposed hypothesis.

† For the full poster, see http://dx.doi.org/10.1017/S1743921318002855

Table 1. Features used in clustering methods. They were engineered with **UPSILON** and colour indices sourced from *Kepler* stellar parameters data

Feature	Description
$g-r$	SDSS Colour Index
$J-K$	2MASS Colour Index
$g-K$	SDSS/2MASS Colour Index
$\log P$	log of the period extracted with Lomb Scargle Analysis
A_1	Amplitude from Fourier Decomposition at period from **upsilon**
R_{21}	2^{nd} to 1^{st} amplitude ratio from Fourier Decomposition
R_{31}	3^{rd} to 1^{st} amplitude ratio from Fourier Decomposition
ϕ_{21}	Relative phase difference of the 2^{nd} and 1^{st} phases from Fourier Decomposition
ϕ_{31}	Relative phase difference of the 3^{rd} and 1^{st} phases from Fourier Decomposition
γ_1	Skewness
γ_2	Kurtosis
Q_{3-1}	Difference between 3^{rd} and 1^{st} quantiles
Ψ^{η}	η (degree of change of trends) of a phased curve
Ψ^{CS}	Range of cumsum of the phased curve

Clustering algorithms are commonly used as a component of the exploration phase of data analyses, as they are able to compress the information contained in high-dimensional data-frames and emphasise some of the features that best describe the structure of the data-sets. This method can therefore be used to explore stellar data-sets prior to in-depth studies. Although clustering methods present themselves as one of the plausible solutions for exploring high-dimensional data of celestial objects, complexities such as computation time and the convergence of algorithms are also important in the evaluation of the performance of those methods. By comparing three commonly-practised methods, we demonstrate that clustering algorithms can be implemented successfully for exploring high-dimensional data for variable stars.

2. Database Generation

2.1. *Data*

A sample of *Kepler* time-series data (Koch *et al.* 2010) from quarters 0 and 17 was transformed to high-dimensional frames (6217×14) prior to assigning them to respective groups through cluster analysis. For predictive purposes, the data also contained 321 target stars from Bradley *et al.* (2015) and Balona *et al.* (2016). They were included in the analysis in order to evaluate the efficiency of the algorithms. This sample of the data-frame was considered to be the *training set*.

2.2. *Feature engineering and selection*

Feature engineering transforms a two-dimensional visual representation of a light-curve into a high-dimensional data set – and thus highlights the complexities of stellar variability. Features generated can be categorised as primary and secondary. Primary features such as colour indices (temperature indicators), period(s) and Fourier parameters are those that are traditionally used to classify variable stars. Colour indices are sourced from survey databases. Period(s) and Fourier parameters are estimated by fitting least-squares spectral analysis models to light-curves. Secondary features tend to describe the distribution of the light-curves, and are hence considered statistical features.

https://doi.org/10.1017/S1743921318002855 Published online by Cambridge University Press

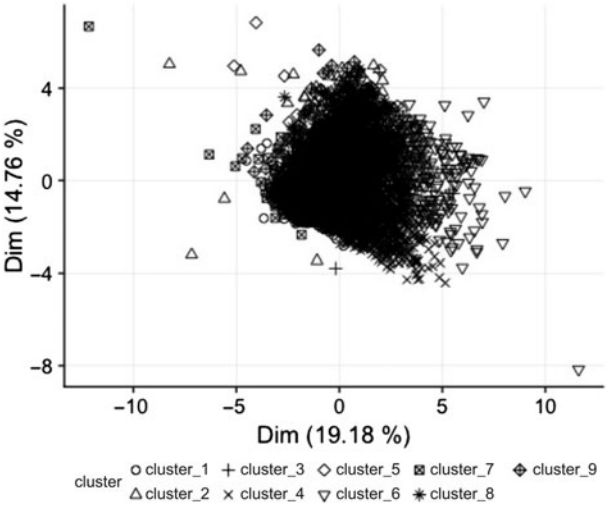

Figure 1. Clustering results from the K-means algorithm suggest that there is an overlap between clusters. Axis labels represent the amount of information retained in the first and second dimension of the data-frame, expressed by the variance of Principal Components (PCs).

Specialised time-domain software packages are being used increasingly to generate features. One such example is that of Kim & Bailer-Jones (2016), which was used in this paper. Their software package is known as A**U**tomated Classification of **P**eriodic Variable **S**tars using Mach**I**ne Lear**N**ing (UPSILON) and the feature generator is accessed as **upsilon.generate_features**. Alternatively, analytical survey web-tools such as the *Caltech Time Series Characterization Service* can be used to generate desired features. A list of features generated in the data-frame is shown in Table 1.

3. Clustering Methods

Three clustering methods were applied to the *Kepler* data-frame: hard (K-means), soft (Expectation Maximization) and K-means applied to features extracted with Principal Component Analysis (PCA) from the original data-frame (K-means via PCA). The K-means approach aims to minimise the distortion, whereas the Expectation Maximization (EM) algorithm assumes data are normally distributed and uses maximum likelihood to assign an object to clusters.

4. Results

Hopkins statistics showed *no* clustering tendency for the *Kepler* data-frame (≈ 0.18), which is lower than the desired threshold of $\mathbf{H} \geqslant 0.5$. Based on *a priori* knowledge from the *training set*, the K-means method was implemented using nine clusters. The EM method resulted in nine optimal clusters in the data-frame (see Fig. 1). Silhouette analysis for the K-means algorithm resulted in coefficients approximately zero (≈ 0.20). That implied that some stars were incorrectly assigned ('misclassified') and that there is an overlap between clusters, suggesting that a probabilistic clustering approach may be the preferred method for stellar studies.

In addition to evaluating the efficiency of the algorithms, the K-means algorithm was computed via PCA with the first seven Principal Components (PCs). Those PCs were chosen so that they contained at least 80% of the data information. This resulted in a 'new reduced' 6217×7 data-frame. Evaluation of the features through PCA showed that the period ($\log P$) of the oscillations contributed the least to the PCs and that

https://doi.org/10.1017/S1743921318002855 Published online by Cambridge University Press

skewness (γ_1), ϕ_{21}, ϕ_{31} and colour index $g-K$ (*gkmag*) contributed significantly to the PCs. Therefore, clustering analysis can arguably be applied in surveys that consist primarily of light-curves. Feature evaluation also showed that colour indices of stars are highly correlated; therefore, in order to reduce the complexity rate, two at most of those features may be included in the data-frame.

When we used the *training set* to infer cluster labels, it became evident that all three algorithms result in one 'empty' cluster, or at least one cluster that had at most two target stars. Rotating variables and δ Scuti stars were both prominent in two clusters in each of the algorithms, which may require modification of the algorithms used such that there is a function that implements cluster merging. Distinct clusters with low misclassification rates were obtained for Maia variables and Eclipsing Binaries. All three algorithms resulted in a cluster that contained at least three types of pulsating stars, the dominant classes being γ Dor, δ Scuti and hybrid γ Dor/δ Scuti stars.

5. Conclusion, and Forthcoming Research

Clustering algorithms can be used as data exploration tools to minimise the time and effort required in the initial procedures of carrying out Asteroseismology. Furthermore, they can aid the discovery of new groups or sub-groups. Given the overlapping characterisation of variable stars, the use of Expectation Maximization or similar probabilistic clustering algorithms may be more beneficial with respect to light-curves of variable stars.

It is evident that there is an imbalance in the *training set* incorporated in the data-frame. That 'imbalance' was due to the way the size and classes were represented. A similar study will therefore be conducted with a more representative training set. In addition to data 'improvements', a large-scale search for pulsating stars in the spectral-type range late B to F will be conducted in various large survey databases by using clustering algorithms. Feature engineering packages and the efficiency of hybrid probabilistic methods such as Modal Expectation Maximization (MEM) will be studied.

Acknowledgements

This project was funded by the James Cook University's Research Training Program Scholarship (RTPS) and the University of Johannesburg.

References

Borne, K. 2009, *BAAS*, 42, 578
Koch, D. G., Borucki, W. J., Basri, G., *et al.* 2010, *ApJ*, 713, L79
Bradley, P. A., Guzik, J. A., Miles, L. F., *et al.* 2015, *AJ*, 149, 68
Balona, L. A., Engelbrecht, C. A., Joshi, Y. C., *et al.* 2016, *MNRAS*, 460, 1318
Kim, D-W., & Bailer-Jones, C. 2016, *A&A*, 587A, 18