Searching for Pulsating stars using Clustering Algorithms

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

[1]College of Science and Engineering, James Cook University, Australia

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

refilwe.kgoadi1@my.jcu.edu.au

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 based on their physical features. This research aims 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 Maximization. In addition to identifying the best performing algorithm, a subset of features that best define stellar groups will be identified. Three methodologies are applied to a sample of *Kepler* spacecraft time series in the temperature range $6500\text{K} - 19000\text{K}$. Given the spectral range, at least four types of stars are expected to be found; $\delta-$Scuti, γ Doradus, Slowly Pulsating B (SPB) and (the still equivocal) Maia stars.

Introduction

Variable star classification is an initial and vital step of asteroseismological studies; therefore, it is crucial that is performed with high precision. In the era of big data where high-dimensionality is common, this stage may be done through Machine Learning (ML) techniques as traditional methods have proven less efficient to perform this significant stage in variable stellar studies. Given that most stellar surveys provide data sets that contain surface physical properties of stars, the Harvard classification method may be used as an initial classification to infer the nature of candidate stars. A rapid second stage of asteroseismological classification inference can be done using high dimensional data-frames with features derived primarily from light curves through clustering algorithms. Cluster analysis is a data mining tool where objects in a data-frame are separated into groups based on their similarity. Clustering algorithms are commonly used as a component of the exploration phase of data analysis 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 datasets.

A sample of *Kepler* Time Series data containing candidates from Bradley *et al.* [2] and Balona *et al.* [1] was transformed to a high dimensional frame (6217×14) prior to assigning them to respective groups through cluster analysis. Labeled candidates from Bradley *et al.* and Balona *et al.* were considered to be a*training set* to infer cluster labels.

Main Objective

The primary objective was to group candidate stars in the late B to F spectral regions based on the similarity of their colour indices and light-curve features such as period, Fourier parameters and skewness, in order to deduce their pulsation class and/or subclass. Another objective was to evaluate the algorithms' efficiencies as applied. A combination of hard and soft hard partitioning methods was used to identify the structure in the generated data-frame consisting of approximately 6,217 candidate stars. Methods used were:

1. ***K-means*:** A hard clustering technique which typically uses the Euclidean method to measure the similarity between objects
2. ***Gaussian Mixed Models as Expectation Maximization (EM)*:** A soft clustering approach where probabilities are used to assign objects to clusters.
3. ***K-means via Principal Component Analysis (PCA)*:** A modified/hybrid K-means method where Principal Components (PCs) derived from Principal Component Analysis are used as a new feature set.

Methodology

The following analytical steps were done to deduce the structure of the *Kepler* data-frame.

1. Generate features that best describe candidate stars in a data-frame. Four types:
   (a) **Colour indices**: Features sourced from *Kepler* survey $g-r$, $J-K$ and $g-K$. These were a combination of SDSS and 2MASS photometric systems.
   (b) **Period searching**: The Lomb Scargle Analysis was used to estimate the best period of the light curves. The corresponding amplitude was also determined.
   (c) **Model fitting through Least Squares Spectral Analysis (LSSA)**: This technique was used to determine Fourier parameters , $R_{j1} = \dfrac{A_j}{A_1}$ and $\phi'_{j1} = \phi_j - j\phi_1$ with $j > 1$ from light curves respectively. R_{j1} is the amplitude ratio and ϕ'_{j1} is the relative phase difference. Parameters were calculated from output of a Fourier analysis model defined by:

$$y(t) = y_0(t) + \sum_{j=1}^{4} A_j \cos(j2\pi f + \phi_j) \tag{1}$$

   where $y_0(t)$ is the average emitted flux, $f = \frac{1}{P}$, P is the period extracted using the Lomb Scargle analysis and ϕ_j is the phase of the waveform.
   (d) **Statistical features**: Based on the overall distribution of the magnitudes in a light/phase curve. These are statistical parameters such as *weighted mean, weighted standard deviation, skewness, kurtosis*.
2. Apply clustering algorithms and infer an efficient algorithm for *Kepler* light curves through validation techniques
3. Find possible labels for groups from a plausible algorithm.

Features listed in 1b, 1c and 1d were generated using a python module **UPSILON** accessed as **upsilon.generate_features_all()** [3]. Periods extracted were validated using **gatspy.LombScargle()** where three significant periods were extracted from light curves. A subset of features was selected based on natural clustering technique. Features in the data-frame are based on Sarro *et al.* [4]. A list of features is shown in table 1.

Table 1: Features used in clustering methods. These 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 FD at period from **upsilon**
R_{21}	2^{nd} to 1^{st} amplitude ratio from FD
R_{31}	3^{rd} to 1^{st} amplitude ratio from FD
ϕ_{21}	Relative phase difference of the 2^{nd} and 1^{st} phases from FD
ϕ_{31}	Relative phase difference of the 3^{rd} and 1^{st} phases from FD
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 Methods

Two clustering algorithms were applied: K-means and Expectation Maximization (EM). The K-means algorithm shown in figure 1 is a hard partitioning method which aims to minimise the distortion defined by:

$$f_{KM} = \frac{1}{n}\sum_{i=1}^{n} d(x_i, c_{ai}) \tag{2}$$

where n is the number of objects in a data frame and $d(x_i, c_{ai})$ is the Euclidean distance between cluster c_i with centroid a and object i.

Algorithm 1: K-Means Clustering Algorithm

```
Input  : Data set X containing n items, number of clusters k
Output: Cluster models M = {m_j, j = 1...k}, assignments
        A = {a_i, i = ...n}
Initialise k cluster with models randomly chosen x_i ϵ X
repeat
    for i = n to n do
        Let c_j be x_i's closest cluster. Set a_i to j.
        ;                                   // Item Assignment
    end for
    for j = n to k do
        Set cluster model m_j to be mean of {x_i |= j}
        ;                                   // Cluster Update
    end for
until;
A does not change
```

Figure 1: K-means Algorithm

The Expectation Maximization (EM) algorithm shown in figure 2 is a probabilistic approach to clustering and is known as soft/fuzzy. Maximum likelihoods are used to assign an object to clusters. The EM algorithm assumes that features in a data frame are normally distributed and aims to minimise the objective function:

$$\begin{aligned} f_{EM} &= \mathcal{L}(\mu, \sigma|\mathcal{X}) \\ &= \prod_{i}^{n}(x_i|\mu_j, \sigma_j) \end{aligned} \tag{3}$$

where $\prod(x_i|\mu_j, \sigma_j)$ is the product of the probabilities.

Algorithm 1: Expectation Maximization Clustering Algorithm

```
Input  : Data set X containing n items, number of clusters k
Output: Cluster models M = {m_j = (μ_j, σ_j), j = 1...k}, assignments
        A = {a_i, i = ..., n}
Initialise k cluster with models randomly chosen x_i ϵ X (or K-means).
repeat
    Estimate the likelihood L using equation eq:em. ; // Expectation
    for i = 1 to n do
        for j = 1 to k do
            Update cluster memberships p(a_i = j, μ_j, σ_j) using
              equation ...
        end for
    end for
    for j = 1 to k do
        Update parameters μ_j, σ_j of eauqtions ... ; // Maximization
        Update prior cluster probabilities p(a_i = j) using equation ...
    end for
until;
A does not change ;                                  // Convergence
```

Figure 2: Expectation Maximization (EM) Algorithm

Results

Hopkins statistics showed *no* clustering tendency for the *Kepler* data-frame (≈ 0.18), which is lower than the desired threshold; $\mathbf{H} \geq 0.5$. Based on a-priori knowledge from the *training set*, K-means was implemented using nine clusters. The EM method resulted in nine optimal clusters in the data-frame (see figure 4). Silhouette analysis for the K-means algorithm resulted in coefficients approximately zero (≈ 0.20). This implies that some stars were incorrectly assigned ("misclassified") and that there is an overlap between clusters, suggesting a probabilistic clustering approach may be the preferred method for stellar studies.

In addition to evaluating the efficiency of the algorithms, K-means via PCA was computed with the first seven Principal Components (PCs). These 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 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, at most two of these features may be included in the data-frame.

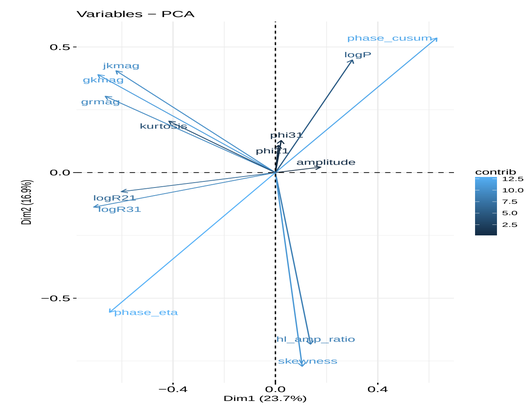

Figure 3: Contribution of variables with respect to K-means via PCA which show that features such as amplitude, ϕ_{21} and ϕ_{31} may be removed from the feature set

Using the *training set* to infer cluster labels, it is evident that all three algorithms resulted 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 with dominant classes being γ-Dor, δ-Scuti and *hybrid* γ-Dor / δ-Scuti stars.

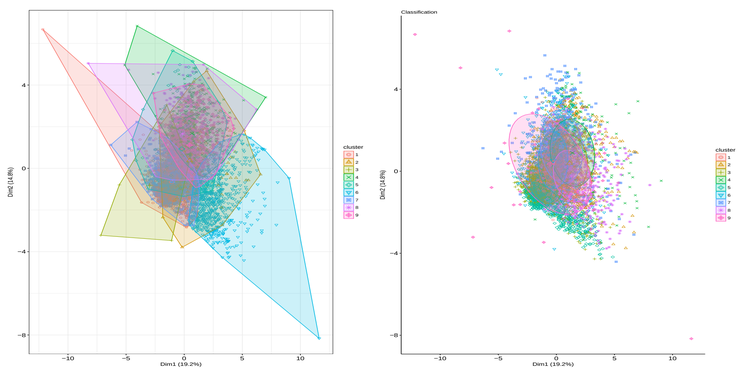

Figure 4: Clustering results from K-means and EM algorithms using *Kepler* Time Series data frame

Conclusions

Clustering algorithms can be the used as a data exploration tool to minimise the time and effort required in the initial procedures of doing Asteroseismology. Furthermore, they can aid the discovery of new groups/sub-groups. Given the overlapping characterisation of variable stars using the EM algorithm, probabilistic clustering algorithms may be more beneficial with respect to variable stellar light curves.

Forthcoming Research

A large scale search for pulsating stars in the late B to F regions stars through clustering algorithms will be conducted in various large survey databases. The efficiency of hybrid probabilistic methods such as Modal Expectation Maximization (MEM) will be studied.

References

[1] L. A. Balona et al. The hot γ Doradus and Maia stars. *Mon. Not. R. Astron. Soc.*, 460(2):1318–1327, 2016.

[2] Paul A. Bradley et al. Results of a Search for gamma Dor and delta Sct Stars with the Kepler Spacecraft. *Astron. J.*, 149(2):1–38, 2015.

[3] Dae-Won Kim and Coryn A. L. Bailer-Jones. A Package for the Automated Classification of Periodic Variable Stars. *Astron. Astrophys.*, 258(2011):1–16, 2015.

[4] L. M. Sarro et al. Comparative clustering analysis of variable stars in the Hipparcos, OGLE Large Magellanic Cloud, and CoRoT exoplanet databases. *Astron. Astrophys.*, 506(1):535–568, 2009.

Acknowledgements

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