Unsupervised classification of eclipsing binary light curves through $k$-medoids clustering

Soumita Modak,$^a$ Tanuka Chattopadhyay$^b$ and Asis Kumar Chattopadhyay$^a$

$^a$Department of Statistics, University of Calcutta, Kolkata, India; $^b$Department of Applied Mathematics, University of Calcutta, Kolkata, India

ABSTRACT

This paper proposes $k$-medoids clustering method to reveal the distinct groups of 1318 variable stars in the Galaxy based on their light curves, where each light curve represents the graph of brightness of the star against time. To overcome the deficiencies of subjective traditional classification, we separate the stars more scientifically according to their geometrical configuration and show that our approach outperforms the existing classification schemes in astronomy. It results in two optimum groups of eclipsing binaries corresponding to bright, massive systems and fainter, less massive systems.

ARTICLE HISTORY
Received 16 August 2018
Accepted 19 June 2019

KEYWORDS
Light curve of variable star; clustering; $k$-medoids method; complexity invariance distance

1. Introduction

Eclipsing binaries (Es) are treated as fundamental probe for studying stellar structure and evolution [1,4,8,16,46]. They can be classified based on their light curves (LCs) to find out the possible sources of homogeneous groups of the stars [22,34,47]. Miller et al. [29] have carried out observations of 1318 new variable stars covering 0.25 square degree region of the Galactic plane centered on Galactic coordinates (latitude, longitude) of (330.94, $-2.28$) deg. The majority of stars in the above region are thought to be associated with the Normal Spiral Arm. They separated the stars subjectively according to the appearance of their LCs into four groups, viz. Algol type (EA), Beta Lyrae (EB), W Ursae Majoris (EW) and un-categorized pulsating stars (PUL). But such subjectivity usually includes degeneracy, i.e. it classifies the stars with different physical properties in the same group. Hence, this traditional scheme is now almost obsolete and can be rather misleading. Most importantly, many of the stars were categorized with uncertainty or ambiguity. These limitations lead us to clustering the stars objectively without assuming any prior information.

In this paper, we carry out an unsupervised classification of the variable stars from [29] by applying the $k$-medoids clustering method [19] to their LCs. Astronomical data collection is often obscured by bad weather conditions. Usually, it is not possible to have repeated observations on astronomical objects due to intermediate celestial objects and...
instrumental restrictions. As a result we have data contaminated with noise, affected by outliers or sparsely distributed \([14,30–32]\), which cannot be analyzed properly by the usual methods. To overcome such problems we apply the nonparametric partitioning based \(k\)-medoids clustering to univariate time series, i.e. LCs, which is robust against noisy or unusual LCs while facilitating any distance measure to extract relevant temporal information properly from the given LCs. Previous works of classification of Es based on LCs \([21,27,28,38,41]\) suffer from many significant drawbacks, e.g. the methods are imposed with a lot of restrictions or model assumptions, and the classification is supervised in the sense that the number of classes is assumed as a priori or the properties of the classes are known. Some performed the classification based on variables derived from the observed or simulated LCs instead of directly using the observed LCs of stars. Their resulting classes overlap significantly with respect to the properties of the stars. However, our method can be used for any new data set where the distribution of the LCs is unknown and assumptions on the number of clusters or the cluster properties are not possible to make before the analysis. This paper uses the \(k\)-medoids method with complexity invariance distance (CID) \([2,36,50]\), dynamic time warping distance (DTW) \([7,15,20,39]\) and Euclidean distance (ED), where CID gives the best clustering in terms of the average silhouette width (ASW) \([40]\) with two resulting groups of Es. We prove the superiority of our approach over DBSCAN method \([13,22]\), \(t\)-SNE technique \([21,22,26]\) and \(k\)-means method \([18,32]\). The success of the proposed method in revealing distinct groups of stars is also confirmed by fuzzy \(k\)-means clustering \([3]\). Simulation study establishes the usefulness of our approach in general as a potential LC-based classifier.

The paper is organized as follows. In Section 2, we describe the data and the transformations on data. Section 3 discusses the clustering method with accuracy measure, distance measures and the simulation study. Result and discussion are presented in Section 4, whereas Section 5 concludes.

2. Data

The present data set is taken from \([29]\) where each of the 1318 variable stars in our Galaxy has a LC file together with R-band magnitude (R), colors (B-R, R-I) and period (P). From these we obtain the variables I as R minus R-I, B-I as B-R plus R-I and B as B-I plus I. According to the subjective classification, the variable stars mainly consist of Es along with some possible pulsating stars (201 uncertain and 118 potential pulsating stars). For each LC, relative flux variation in R-band is given on a continuous time scale in Heliocentric Julian Date (HJD) within the range from 2452450.622 HJD to 2453607.616 HJD. Each LC is unevenly spaced of different length having values at different time points. The length of LCs varies from 130 to 264 (except one LC having length 5) and period ranges from several hours to several weeks.

2.1. Phase computation, interpolation and binning

Comparison of the given LCs is performed in terms of a full cycle over phase interval \([0,1]\), which restores the physical properties of the stars \([10,35,44]\), having observations at \(l\) (say) evenly spaced phase points, obtained as follows.
(i) We obtain phased LCs by transforming the given time points into standard phases (always lie between 0 and 1) using the following equation [10,35]:

\[
\text{decimal portion of } \left[ \frac{(t - t_0)}{P} \right],
\]

where \( t \) denotes the time of measurement of the star (here in HJD), \( t_0 \) is an arbitrary epoch – usually a time of maximum or minimum brightness (here the time of the first observed maximum) and \( P \) represents the period (known and constant) of the star (here in days). Now, we have the \( i \)th LC over the phase interval \([0, \pi_i]\), where \( \pi_i \) (close to 1 but < 1) is maximum of standard phases for the \( i \)th LC, \( i = 1, \ldots, 1318 \) (here \( \pi_i \)s are different for all \( i \)).

(ii) To avoid extrapolation involving larger uncertainty than interpolation, we extend the phase interval of the \( i \)th LC to \([0, \pi_i + 1] \in [0, 2)\), \( i = 1, \ldots, 1318 \), by adding +1 to the standard phases computed in step (i) (since a phase of 0 is the same as a phase of 1, −1 or 2).

(iii) We fit the linear spline [7,37] to the LCs from step (ii) at \( l \) evenly spaced phases over \([0,1]\). Given a tabulated function \( y_i = y(x_i), i = 1, \ldots, l \) with \( x_i < x_{i+1} \) for \( i = 1, \ldots, l - 1 \), the interpolating function joins \( l - 1 \) linear functions of the form \( f_i(x) = a_i y_i + b_i y_{i+1}, x \in [x_i, x_{i+1}] \), where \( a_i, b_i \) are constants satisfying (a) \( f_i(x_i) = y_i \), (b) \( f_i(x_{i+1}) = y_{i+1} \), i.e. \( a_i = \frac{x_{i+1} - x}{x_{i+1} - x_i} \) and \( b_i = 1 - a_i = \frac{x - x_i}{x_{i+1} - x_i}, i = 1, \ldots, l - 1 \).

(iv) The interpolation error in step (iii) decreases with the increase in \( l \), provided \( l \) should not be too large to cause considerable error for under-sampled series. Here the LCs (except one) are well sampled of length at least 130 with most of them of length close to 272 which is our chosen value of \( l \) such that there is no loss of information, significant computational burden or considerable error in approximating the series of lower length. Technically, \( l \) is varied over different plausible values which robustly gives the optimal number of clusters as two in terms of clustering accuracy ASW (see Section 3.3) with \( l = 272 \) corresponding to the best clustering.

3. Classification scheme

Clustering of time series [23,24], either in time domain or frequency domain [5,9], can be performed on series of equal length or unequal length [6,25,45], evenly spaced or unevenly spaced [11,12,33,42]. Here we apply the following unsupervised classification scheme to the evenly spaced LCs of equal length over the phase interval \([0,1]\), which are referred to as ‘objects’ in the cluster analysis.

3.1. \( k \)-medoids: partitioning around medoids (PAM)

\( k \)-medoids is a nonparametric partitioning based clustering method which can be applied to univariate time series like LCs. We use a fast and efficient algorithm ‘PAM’ [19] which is executed using the inbuilt function ‘pam’ in software ‘R’. It is based on the search for \( k \) medoids in the data set, where a medoid is the representative object of the cluster it belongs to. These \( k \) medoids represent various structural aspects of the data set of size \( N \) being partitioned into \( k \) mutually exclusive and exhaustive clusters around \( k \) medoids, where a medoid is that object of the cluster for which the sum of distances to all other objects of
the cluster is minimal. Because of the medoids, this method is robust against noise, outliers or sparsely distributed data [43]. This clustering method allows any distance measure depending upon the nature of the given data.

### 3.2. Competitive methods of unsupervised clustering for comparison

As a competitor of \(k\)-medoids clustering, we consider the very popular distance-based approach DBSCAN method [13,22] with parameters ‘\(\epsilon\)’ and ‘\(\text{MinPts}\)’ which estimates the density around each object by counting the number of objects in a neighborhood (\(\epsilon\)) and applies a threshold (\(\text{MinPts}\)) to identify core, border and noise objects. The core objects form a cluster if there is a chain of them wherein one falls inside the \(\epsilon\)-neighborhood of the next, whereas the border objects are arbitrarily assigned to the clusters. Objects lying outside the \(\epsilon\)-neighborhood of the clustered objects cannot be assigned to a cluster and is considered as noise. We also apply this DBSCAN method to data transformed into a nonlinear form through \(t\)-SNE [21,22,26] which reduces high-dimensional objects in a low-dimensional space, where similar objects are modeled using close transformed data and dissimilar objects are modeled using far transformed data with high probability. Here any appropriate metric can be chosen to compute distance among the objects.

Another classical clustering \(k\)-means [18] is well known to partition the objects into \(k\) clusters in which each object belongs to the cluster with the nearest mean, whereas fuzzy \(k\)-means [3] groups the data set such that each object can be classified into more than one cluster. Here for every object the degree of being member of a cluster is quantified by a membership (\(\in [0, 1]\)), with larger membership value indicating higher probability that the object inherits the properties of that cluster. In fuzzy \(k\)-means, the centroid of a cluster is the weighted mean of all objects where weights are the power function of membership values belonging to the cluster.

### 3.3. Optimal number of clusters

The optimal value of the number of clusters (\(k\)) is chosen from ASW [40], which accounts for the efficacy of the cluster analysis, using the distance measure appropriate for the given objects. For each object \(i\), the silhouette width (SW) \(s(i)\) lies from \(-1\) to 1. Object with a large positive SW is very well clustered, SW around 0 means that the corresponding object lies between two clusters, and object with a small negative SW is probably placed in the wrong cluster. ASW is the average SW over all \(i\), with \(-1 \leq \text{ASW} \leq 1\), which is calculated for \(k = 2, 3, \ldots\), etc. and the value of \(k\) is chosen for which ASW is maximum. For given \(k\), the silhouette plot [40] gives a graphical representation of SW of each object. Robustness of the optimal number of clusters with respect to clustering accuracy measure is verified by another distance-based quantity, the connectivity [17], which indicates the degree of connectedness of the clusters and takes a value between zero and infinity with a minimum corresponding to the best possible clustering.

### 3.4. Distance measure

Time series like the LCs having a large number of peaks, in different quantities, amplitudes or durations are considered as complex time series. So we measure distances between the
In our study, LCs under complexity invariance using CID [2,36,50]. Because Batista et al. [2] empirically showed that CID generally performs best among its possible competitors and is effective in clustering complex time series. Here the LCs show considerable complexity in terms of the complexity estimate defined in Equation (3). CID between two time series \( X \) with values \( x_1, \ldots, x_n \) and \( Y \) with values \( y_1, \ldots, y_n \) corresponding to time points \( t = 1, \ldots, n \) is defined as

\[
CID(X, Y) = ED(X, Y) \times CF(X, Y),
\]

where \( CF \) is a complexity correction factor given by

\[
CF(X, Y) = \frac{\max(CE(X), CE(Y))}{\min(CE(X), CE(Y))},
\]

with the following complexity estimate of time series \( X \)

\[
CE(X) = \sqrt{\sum_{t=1}^{n-1} (x_t - x_{t+1})^2}
\]

and

\[
ED(X, Y) = \sqrt{\sum_{t=1}^{n} (x_t - y_t)^2}.
\]

We consider another popular time series distance DTW [7,15,20,39]. For two time series \( X \) with values \( x_1, \ldots, x_m \) and \( Y \) with values \( y_1, \ldots, y_n \) corresponding to time points \( t = 1, \ldots, m \) and \( t = 1, \ldots, n \) respectively, dynamic time warping finds the warping path \( W = w_1, \ldots, w_L \) of contiguous elements on the local distance matrix having \( (i,j) \)th element \( d(x_i, y_j) = |x_i - y_j| \) \( (i = 1, \ldots, m, j = 1, \ldots, n) \), such that \( w_l = (i_l, j_l) \in \{1, \ldots, m\} \times \{1, \ldots, n\}, l = 1, \ldots, L, \max(m, n) \leq L < m + n - 1 \), satisfy the following conditions. (C1) Boundary conditions: \( w_1 = (1, 1), w_L = (m, n) \), (C2) Continuity: For \( w_{l+1} = (i_{l+1}, j_{l+1}) \) and \( w_l = (i_l, j_l) \), \( i_{l+1} - i_l \leq 1 \) and \( j_{l+1} - j_l \leq 1 \) for all \( l = 1, \ldots, L - 1 \) and (C3) Monotonicity: For \( w_{l+1} = (i_{l+1}, j_{l+1}) \) and \( w_l = (i_l, j_l) \), \( i_{l+1} - i_l \geq 0 \) and \( j_{l+1} - j_l \geq 0 \) for all \( l = 1, \ldots, L - 1 \). Then, DTW which is an optimal path between \( X \) and \( Y \) under the stated restrictions is defined as

\[
DTW(X, Y) = \min \left( \sqrt{\sum_{l=1}^{L} w_l} \right).
\]

Dynamic programming is used to find this path by evaluating the following recursive function [15]:

\[
g[i, j] = \min\left( g[i, j-1] + d(x_i, y_j), g[i-1, j-1] + 2 \cdot d(x_i, y_j), g[i-1, j] + d(x_i, y_j) \right), \quad i = 1, \ldots, m, j = 1, \ldots, n.
\]

In our study, \( m = n \) wherein ED comes as a particular form of DTW with \( w_l = (i_l, j_l), i = j = l \).
Figure 1. (a) Simulated average light curves (with standard error), generated using sine signal with added noise and outliers, of two groups each containing 1000 LCs with respective amplitudes 1 (red) and 1.5 (black), (b) average light curves (with standard error) of two clusters, resulted in $k$-medoids clustering with CID, consisting of 1004 (red) and 996 (black) LCs.

Figure 2. (a) Simulated average light curves (with standard error), generated using cosine signal with added noise and outliers, of two groups each containing 1000 LCs with respective amplitudes 1 (red) and 1.5 (black), (b) average light curves (with standard error) of two clusters, resulted in $k$-medoids clustering with CID, consisting of 1003 (red) and 997 (black) LCs.

3.5. Simulation study

We show the performance of $k$-medoids clustering with CID in exploring inherent groups of astronomical objects based on their LCs through simulation study. Here we generate LCs using a periodic signal contaminated with noise and outliers at 100 evenly spaced time points on $[0,1]$ (see for detail [48,49] and references therein). In the first case, we consider a complete cycle of the sine signal contaminated with signal-to-noise ratio $= 3$, wherein 90% of the noise is related to measurement accuracies and 10% is white noise, as well as 10% outliers added to the measurement accuracies. It simulates 1000 LCs from each of two groups having respective amplitudes 1 and 1.5, whose group-wise average LCs are drawn in Figure 1(a). Second, we generate LCs in the same way using a complete cycle of the
Table 1. The average silhouette width (ASW × 10^2) for different number of clusters (k), from k-medoids clustering with CID, for two simulated data sets with ASW\textsubscript{sin} and ASW\textsubscript{cos} respectively corresponding to the sine and cosine signal generated data sets.

| k  | ASW\textsubscript{sin}  | ASW\textsubscript{cos} |
|----|----------------|----------------|
| 2  | 37.758         | 37.611          |
| 3  | 25.372         | 25.162          |
| 4  | 0.901          | 1.095           |
| 5  | 0.872          | 1.062           |
| 6  | 0.705          | 0.877           |

Figure 3. The average silhouette width (ASW) for different number of clusters (k) corresponding to three different distance measures in combination with k-medoids clustering method. The circles indicate values of ASW corresponding to a value of k.

cosine signal, whose group-wise average LCs are plotted in Figure 2(a). Then, in each case we combine the generated LCs from two groups and perform our method on them which reveals the two existing clusters of the data in terms of ASW (see Table 1). The average LCs of the two resulting clusters are shown in Figure 1(b) for the first case and in Figure 2(b) for the latter. Our approach successfully distinguishes between the closely related inherent groups of noisy and outlier affected data in both the cases with 0.2% and 0.15% misclassification respectively. It is observed that almost identical appearances between Figures 1(a) and 1(b) and between Figures 2(a) and 2(b) validate that the proposed clustering method is significantly efficient in identifying natural clustering under the considered situations with very low misclassification chances.

4. Results and discussion

We compare the k-medoids clustering results obtained through CID, DTW and ED in terms of ASW and show that CID outperforms the other two (see Figure 3) resulting in \(k = 2\) (see Table 2, which also verifies \(k = 2\) by the connectivity measure) with two clusters, denoted by k1 and k2, of respective sizes 838 and 480. The corresponding silhouette plot
Table 2. The average silhouette width (ASW $\times 10^2$) and the connectivity for different number of clusters ($k$) from $k$-medoids clustering with CID.

| $k$ | ASW   | Connectivity |
|-----|-------|--------------|
| 2   | 67.610| 149.690      |
| 3   | 49.309| 246.292      |
| 4   | 37.666| 395.906      |
| 5   | 37.608| 421.292      |
| 6   | 28.659| 513.438      |

Figure 4. The silhouette plot gives a graphical representation of silhouette width of each of the light curves belonging to individual clusters, resulted from the $k$-medoids clustering with CID for $k = 2$. The grey shade indicates the silhouette width of a light curve, arranged in descending order (from top to bottom) for individual clusters. The average silhouette width for two clusters $k_1, k_2$ and the whole data set of respective sizes of 838, 480 and 1318 are computed as 0.77, 0.51 and 0.68, respectively. 

(Figure 4) displays the tightness of individual clusters and the separation between two clusters are significant. ASW for $k_1$, $k_2$ and for the whole data set of size 1318 are computed as 0.77, 0.51 and 0.68, respectively, show that the data is quite well clustered. This classification, irrespective of their subjective one (see Table 3), has template LCs in Figures 5 and 6. Cluster-wise representative LCs (i.e. two sets of observed LCs) in Figures 7 and 8 indicate the similarity with their template LCs, whereas the average properties of two clusters are reported in Table 4.

We compare our method with other existing unsupervised classifiers, where the DBSCAN method applied to the LCs with distance measure CID and parameters $\epsilon = 0.5$, and...
Table 3. Membership of subjective types in two groups k1 and k2.

| Type      | k1 | k2 |
|-----------|----|----|
| EA        | 43 | 44 |
| EB        | 84 | 24 |
| EW        | 234| 260|
| PUL       | 99 | 19 |
| EA:       | 15 | 31 |
| EB:       | 165| 50 |
| EW:       | 7  | 14 |
| PUL:      | 145| 23 |
| CV:       | 1  | 0  |
| EA/EB     | 6  | 4  |
| EW/EA     | 2  | 4  |
| EW/EB     | 8  | 4  |
| EB/PUL    | 11 | 1  |
| DCEP/PUL  | 9  | 0  |
| CV/PUL    | 9  | 2  |
| Total     | 838| 480|

Note: An uncertain type is followed by a colon and an ambiguous type is given with a slash.

Figure 5. Template average light curves, with standard error, of two clusters k1 and k2 obtained from k-medoids clustering with CID.

MinPts = 5 fails to reveal the inherent clustering and results in only one group of size 253 with 1065 LCs assigned as noise objects. Even if we consider the noise objects as a separate group [22], then also it gives a poor discrimination with ASW = 0.24. Again, we transform the LCs into two-dimensional data through t-SNE, where the distance between the LCs is measured by CID. Then DBSCAN (ε = 0.5, MinPts = 5) method, with distance measure ED performed on this transformed data, also fails to identify the inherent groups and gives five scattered clusters of sizes 5 to 7 and 1291 noise objects. However, our method which is robust against noise and outliers successfully exposes the clusters from the LCs with significant accuracy in terms of ASW (see Table 2 and Figure 4). We check superiority and
**Figure 6.** Template medoid light curves, corresponding to the stars with ID V-1221 and V-1138 (see [29]), of two clusters k1 and k2 respectively obtained from $k$-medoids clustering with CID.

**Figure 7.** A pair of representative light curves, which are the observed light curves corresponding to the stars with ID V-94 and V-384 (see [29]), of two clusters k1 (left) and k2 (right) respectively obtained from $k$-medoids clustering with CID.
Figure 8. Another pair of representative light curves, which are the observed light curves corresponding to the stars with ID V-334 and V-817 (see [29]), of two clusters k1 (left) and k2 (right) respectively obtained from \( k \)-medoids clustering with CID.

Table 4. Average values (with standard error) of the variables for two clusters k1 and k2 obtained from \( k \)-medoids clustering with CID.

| Name of cluster | No. of members | P (day) | R (mag) | B (mag) | I (mag) | B-I (mag) | R-I (mag) |
|-----------------|----------------|---------|---------|---------|---------|-----------|-----------|
| k1              | 838            | 2.816±0.045 | 17.783±0.036 | 19.917±0.045 | 16.968±0.037 | 2.948±0.026 | 0.814±0.020 |
| k2              | 480            | 1.400±0.125 | 19.636±0.062 | 22.046±0.075 | 18.657±0.061 | 3.389±0.040 | 0.978±0.028 |

Table 5. The average silhouette width (ASW \( \times 10^2 \)) for different number of clusters \((k)\) from \( k \)-means clustering.

| \( k \) | ASW |
|---------|-----|
| 2       | 52.426 |
| 3       | 47.398 |
| 4       | 35.575 |
| 5       | 31.107 |
| 6       | 23.574 |

robustness of our method by comparing it with \( k \)-means clustering, applied to linear features [32] extracted from the LCs in terms of the first 10 principal components describing more than 80% variation in the LCs, which also hints at two optimal groups of the variable stars (see Table 5).
Table 6. Average values (with standard error) of the variables for two clusters g1 and g2 containing the distinct light curves, obtained from fuzzy k-means clustering.

| Name of cluster | No. of members | P (day)       | R (mag) | B (mag)       | I (mag)       | B-I (mag) | R-I (mag) |
|-----------------|----------------|--------------|---------|---------------|---------------|-----------|-----------|
| g1              | 811            | 2.896±0.135  | 17.955±0.044 | 20.077±0.053  | 17.095±0.044  | 2.981±0.028  | 0.859±0.021  |
| g2              | 427            | 1.375±0.108  | 19.299±0.069 | 21.724±0.081  | 18.394±0.066  | 3.330±0.043  | 0.905±0.031  |

The morphologies of variable stars change continuously, so it may not always be possible to find distinctly separated clusters of the stars from their LCs. Therefore, we verify the distinction between the clusters by fuzzy k-means clustering, wherein the stars with \( k = 2 \) whose LCs share potentially common properties of the groups can be classified into both the clusters. It gives two distinct groups of 811 and 427 LCs, whereas 80 LCs cannot be classified distinctly as their membership values are very close for the two groups. Hence, we eliminate these 80 LCs and obtain two groups, say g1 and g2, which contain only the non-overlapping LCs. Comparison between Tables 4 and 6 and Figures 5 and 9 show the similarity between groups k1 and g1 and groups k2 and g2. It supports the distinction between the two groups obtained by our method. Also, the clustering with groups g1 and g2 has ASW = 0.28, which clearly shows that our method leads to significantly better clustering. Therefore, further astrophysical analyses of the groups are carried out based on the results obtained from k-medoids clustering with CID.

Figures 5–8 and Table 4 indicate that the LCs in k1 have less variation between the two minima and larger average time period compared to those in k2. These suggest k1 systems consist of stars which form a more or less detached or semidetached system. Also the depths of the two minima of LCs for k1 are smaller compared to those for k2. This indicates k1 systems have a less massive secondary, whereas the masses are comparable for k2 systems. The color-magnitude diagram (Figure 10), the color histograms (Figures 11 and 12) and Table 4 show that k1 systems are bluer, i.e. have higher temperature than k2 systems and
Figure 10. Color-magnitude diagram of the stars clustered in two groups k1 and k2 through k-medoids clustering method with CID.

Figure 11. Histograms of B-R color index for the stars clustered in two groups k1 and k2 through k-medoids clustering method with CID.

consist of stars with unequal mass, whereas the systems in k2 are redder and consist of stars with comparable mass. So k1 and k2 systems respectively belong to early and late spectral types.

In this regard, we discuss some of the recent works on classification of Es. Sarro et al. [41] classified 81 Es using Bayesian model-based neural networks and resulted in groups which have a high degree of superposition with respect to properties like mass, period, separation. However, our work is more robust based on a nonparametric classification scheme which does not adopt any model assumptions, and there is a well-defined distinction between the resulting groups both in terms of the LCs (Figures 5–8) and the average properties,
Figure 12. Histograms of R-I color index for the stars clustered in two groups k1 and k2 through k-medoids clustering method with CID.

e.g. the period is almost double in k1 compared to k2 (see Table 4). Sarro et al. [41] also concluded that classification based on the LCs, like we perform, is always better than classification with respect to the variables derived from the LCs. Malkov et al. [27] classified 6330 Es on the basis of their observable variables, but their classification is subjective and restricted by several assumptions unlike our method. Prša et al. [38] classified Es based on the variables, derived from 10,000 synthetic and 50 real LCs, by artificial neural network (ANN) method with presumed five classes, whereas our method is applied to the observed LCs with the number of classes found scientifically. Matijevič et al. [28] used dimension reduction technique based on LLE algorithm and found that the projection onto a two-dimensional space can preserve the local geometry. This is somewhat consistent with our findings of two groups of Es, but finally their groups reduced to a single variable equivalent to ‘detachedness’ of the binaries. Kirk et al. [21] classified about 2,00,000 Es based on their observed LCs by LIE method, but they also assumed the number of classes as prerequisite. In [22,34,47], the classification error is around 10% mainly due to the similarity of LCs originating from different physical systems. In particular, Kochoska et al. [22] have found four groups of which the first two and the last two have similar Kepler polyfit primary depths indicating merely two significant groups of Es which supports our results of two groups.

5. Conclusion

We have classified 1318 variable stars in the Galaxy which lie primarily along the spiral arms. To overcome the deficiencies of the subjective classification method [29], k-medoids clustering with CID is applied to the LCs, which gives rise to two physically interpretable groups of Es and indicates there is no separate group of pulsating stars in the present data set. The accuracy of the resulting clustering is significant in terms of the ASW and having outperformed the established methods in astronomical literature, our approach is a strong competitor for future classification of new variable stars based on their LCs.
Acknowledgements

The authors would like to thank the Editor-in-chief and an anonymous associate editor for encouraging the present work on Astrostatistics and two anonymous referees for their intriguing inquiries which helped the authors to present the results in a more convincing way.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] C. Akerlof, S. Amrose, R. Balsano, J. Bloch, D. Casperson, S. Fletcher, G. Gisler, J. Hills, R. Kehoe, B. Lee, S. Marshall, T. McKay, A. Pawl, J. Schaefer, J. Szymanski and J. Wren, ROTSE all-sky surveys for variable stars. I. Test Fields. Astron. J. 119 (2000), pp. 1901–1913.
[2] G.E.A.P.A. Batista, E.J. Keogh, O.M. Tataw and V.M.A. de Souza, CID: an efficient complexity-invariant distance for time series, Data Mining Knowledge Discovery 28 (2014), pp. 634–669.
[3] J.C. Bezdek, Pattern recognition with fuzzy objective function algorithms, Plenum Press, New York, 1981.
[4] D.H. Bradstreet and D.P. Steelman, Binary maker 3.0 – an interactive graphics-based light curve synthesis program written in java, Am. Astron. Soc. 201st AAS Meeting, id.75.02; Bull. Am. Astron. Soc. 34 (2002), pp. 1224.
[5] J. Caiado and N. Crato, A periodogram-based metric for time series classification, Comput. Statist. Data Anal. 50 (2006), pp. 2668–2684.
[6] J. Caiado and N. Crato, Comparison of times series with unequal length in the frequency domain, Commun. Statist. Simulation Comput. 38 (2009), pp. 527–540.
[7] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata and A. Pulvirenti, Advances in Data Mining Knowledge Discovery and Applications, Chapter 3: Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining, pp. 71–96, Intech, 2012.
[8] T. Chattopadhyay, A. Sinha and A.K. Chattopadhyay, Influence of binary fraction on the fragmentation of young massive clusters – a Monte Carlo simulation, Astrophys. Space Sci. 361 (2016), pp. 120–133.
[9] G.R. Dargahi-Noubary, Discrimination between Gaussian time series based on their spectral differences, Commun. Statist. Theory Methods 21 (1992), pp. 2439–2458.
[10] S. Deb and H.P. Singh, Light curve analysis of variable stars using Fourier decomposition and principal component analysis, Astron. Astrophys. 507 (2009), pp. 1729–1737.
[11] A. Eckner, A Framework for the Analysis of Unevenly Spaced Time Series Data. Working Paper. URL: http://eckner.com/papers/unevenly_spaced_time_series_analysis.pdf, 2014.
[12] A. Eckner, Algorithms for Unevenly-spaced time series: Moving averages and other rolling operators. Working Paper. URL: http://eckner.com/papers/Algorithms%20for%20Unevenly%20Spaced%20Time%20Series.pdf, 2017.
[13] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, pp. 226–231, 1996.
[14] E.D. Feigelson and G.J. Babu, (Eds.) Statistical Challenges in Modern Astronomy V, Lecture Notes in Statistics – Proceedings, vol. 209, Springer Science+Business Media, New York, 2013.
[15] T. Giorgino, Computing and visualizing dynamic time warping alignments in R: The dtw package, J. Statist. Softw. 31 (2009), pp. 1–24.
[16] D. Graczyk, I. Sośnyński, R. Poleski, G. Pietrzyński, A. Udalski, M.K. Szymański, M. Kubiak, Ł. Wyrzykowski and K. Ulaczyk, The optical gravitational lensing experiment. The OGLE-III Catalog of Variable Stars. XII. Eclipsing Binary Stars in the Large Magellanic Cloud. Acta Astronom. 61 (2011), pp. 103–122.
[17] J. Handl, K. Knowles and D. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics 21 (2005), pp. 3201–3212.
[18] J.A. Hartigan and M.A. Wong, A K-means clustering algorithm, Appl. Statist. 28 (1979), pp. 100–108.
[19] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, pp. 68–125, John Wiley and Sons, New Jersey, 2005.
[20] E. Keogh and C.A. Ratanamahatana, Exact indexing of dynamic time warping, Knowledge Inform. Syst. 7 (2005), pp. 358–386.
[21] B. Kirk, K. Conroy, A. Prša, M. Abdul-Masih, A. Kochoska, G. Matijevič, K. Hambleton, T. Barclay, S. Bloemen, T. Boyajian and L.R. Doyle, Kepler eclipsing binary stars. VII. The catalog of eclipsing binaries found in the entire Kepler data set, Astron. J. 151 (2016), pp. 68–88.
[22] A. Kochoska, N. Mowlavi, A. Prša, I. Lecoeur-Taibi, B. Holl, L. Rimoldini, M. Süveges and L. Eyer, Gaia eclipsing binary and multiple systems. A study of detectability and classification of eclipsing binaries with Gaia, Astron. Astrophys. 602 (2017), pp. A110.
[23] T.W. Liao, Clustering of time series data—a survey, Pattern Recognition 38 (2005), pp. 1857–1874.
[24] T.W. Liao, C. Ting and P.-C. Chang, An adaptive genetic clustering method for exploratory minning of feature vector any time series data, Intl. J. Prod. Res. 44 (2006), pp. 2731–2748.
[25] N.R. Lomb, Least-squares frequency analysis of unequally spaced data, Astrophys. Space Sci. 39 (1976), pp. 447–462.
[26] L.V.D. Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learning Res. 15 (2014), pp. 3221–3245.
[27] O.Yu. Malkov, E. Oblak, E.A. Avvakumova and J. Torra, Classification of Eclipsing Binaries. Solar and Stellar Physics Through Eclipses, in ASP conference series. Vol. 370. O. Demircan, S. O. Selam and B. Albayrak, eds., 2007.
[28] G. Matijevič, A. Prša, J.A. Orosz, W.F. Welsh, S. Bloemen and T. Barclay, Kepler Eclipsing binary stars. III. classification of kepler eclipsing binary light curves with locally linear embedding, Astron. J. 143 (2012), pp. 123–128.
[29] V.R. Miller, M.D. Albrw, C. Afonso and Th. Henning, 1318 new variable stars in a 0.25 square degree region of the Galactic plane, Astron. Astrophys. 519 (2010), pp. A12.
[30] S. Modak and U. Bandyopadhyay, A new nonparametric test for two sample multivariate location problem with application to astronomy, J. Statist. Theory Appl. 18 (2019), pp.136–146.
[31] S. Modak, T. Chattopadhyay and A.K. Chattopadhyay, Two phase formation of massive elliptical galaxies: study through cross-correlation including spatial effect, Astrophys. Space Sci. 362 (2017), pp. 206–215.
[32] S. Modak, A.K. Chattopadhyay and T. Chattopadhyay, Clustering of gamma-ray bursts through kernel principal component analysis, Commun. Statist. Simul. Comput. 47 (2018), pp. 1088–1102.
[33] C.S. Moller-Levet, F. Klawonn, K. Cho and O. Wolkenhauer, Fuzzy clustering of short timeseries and unevenly distributed sampling points, Adv. Intell. Data Anal. V Lect. Notes Comput. Sci. 2810 (2003), pp. 330–340.
[34] N. Mowlavi, I. Lecoeur-Taibi, B. Holl, L. Rimoldini, F. Barblan, A. Prsa, A. Kochoska, M. Süveges, L. Eyer, K. Nienartowicz, G. Jevardat, J. Charnas, L. Guy and M. Audard, Gaia eclipsing binary and multiple systems. Two-Gaussian models applied to OGLE-III eclipsing binary light curves in the large magellanic cloud, Astron. Astrophys. 606 (2017), pp. A92.
[35] J.R. Percy, Understanding variable stars, Cambridge University Press, New York, 2007.
[36] R.C. Prati and G.E.A.P.A. Batista, A complexity-invariant measure based on fractal dimension for time series classification, Int. J. Natur. Comput. Res. 3 (2012), pp. 59–73.
[37] W.H. Press, S.A. Teukolsky, W.T. Vetterling and W.T. Flannery, Numerical Recipes in C. The Art of Scientific Computing, 2nd ed., Cambridge University Press, Cambridge, 1992, 105–128.
[38] A. Prša, E.F. Guinan, E.J. Devinney, M. DeGeorge, D.H. Bradstreet, J.M. Giammarco, C.R. Alcock and S.G. Engle, Artificial intelligence approach to the determination of physical properties of eclipsing binaries. I. The EBAI project, Astrophys. J. 687 (2008), pp. 542–565.
[39] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc., Upper Saddle River, NJ, 1993.
[40] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987), pp. 53–65.
[41] L.M. Sarro, C. Sánchez-Fernández and Á. Giménez, *Automatic classification of eclipsing binaries light curves using neural networks*, Astron. Astrophys. 446 (2006), pp. 395–402.

[42] J.D. Scargle, *Studies in astronomical time series analysis. III – Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data*, Astrophys. J. 343 (1989), pp. 874–887.

[43] S.S. Singh and N.C. Chauhan, *K-means v/s K-medoids: A comparative study*. National Conference on Recent Trends in Engineering And Technology, 2011.

[44] I. Soszyński, A. Udalski, M.K. Szymański, Ł. Wyrzykowski, K. Ulaczyk, R. Poleski, P. Pietrukowicz, S. Kozłowski, D.M. Skowron, J. Skowron, P. Mróz and M. Pawlak, *The OGLE collection of variable stars. over 45 000 RR Lyrae stars in the magellanic system*, Acta Astron. 66 (2016), pp. 131–147.

[45] A. Stefan, V. Athitsos and G. Das, *The Move–Split–Merge metric for time series*, IEEE Trans. Knowledge and Data Eng. 25 (2013), pp. 1425–1438.

[46] R.A. Street, D.J. Christian, W.I. Clarkson, A. Collier Cameron, N. Evans, A. Fitzsimmons, C.A. Haswell, C. Hellier, S.T. Hodgkin, K. Horne, S.R. Kane, F.P. Keenan, T.A. Lister, A.J. Norton, D. Pollacco, R. Ryans, I. Skillen, R.G. West and P.J. Wheatley, *Status of superWASP I (La Palma)*, Astron. Nachrichten. 325 (2004), pp. 565–567.

[47] M. Süveges, F. Barblan, I. Lecoeur-Taïbi, A. Prša, B. Holl, L. Eyer, A. Kochoska, N. Mowlavi and L. Rimoldini, *Gaia eclipsing binary and multiple systems. supervised classification and self-organizing maps*, Astron. Astrophys. 603 (2017), pp. A117.

[48] A.M. Thieler, M. Backes, R. Fried and W. Rhode, *Periodicity detection in irregularly sampled light curves by robust regression and outlier detection*, Statist. Anal. Data Mining. 6 (2013), pp. 73–89.

[49] A.M. Thieler, R. Fried and J. Rathjens, *RobPer: An R package to calculate periodograms for light curves based on robust regression*, J. Statist. Softw. 69 (2016), pp. 1–36.

[50] Y. Wei, *Multi-dimensional time warping based on complexity invariance and its application in sports evaluation*. 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), IEEE. pp. 677–680, 2014.