University of Groningen

Globular cluster detection in the GAIA survey

Mohammadi, Mohammad; Petkov, Nicolai; Bunte, Kerstin; Peletier, Reynier; Schleif, F.-M.

Published in:
Neurocomputing

DOI:
[10.1016/j.neucom.2018.10.081](https://doi.org/10.1016/j.neucom.2018.10.081)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Mohammadi, M., Petkov, N., Bunte, K., Peletier, R., & Schleif, F.-M. (2019). Globular cluster detection in the GAIA survey. *Neurocomputing*, 342(SI), 164-171. <https://doi.org/10.1016/j.neucom.2018.10.081>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Globular cluster detection in the GAIA survey

M. Mohammadi^a, N. Petkov^a, K. Bunte^a, R.F. Peletier^b, F.-M. Schleif^{c,*}

^aJohann Bernoulli Institute, University of Groningen, The Netherlands

^bKapteyn Institute, University of Groningen, The Netherlands

^cDepartment of Computer Science, University of Applied Science Würzburg-Schweinfurt, Würzburg, Germany

ARTICLE INFO

Article history:

Received 20 July 2018

Revised 6 October 2018

Accepted 17 October 2018

Available online 14 February 2019

Keywords:

Astroinformatics

Stellar structure analysis

Nearest neighbor retrieval

Novelty detection

ABSTRACT

Many techniques for the detection of interesting stellar structures in astronomical data sets require full phase-space or color information, which is not always available. The first data release of the GAIA satellite, for example, provided highly accurate positions and magnitudes for more than one billion sources. Therefore the question arises if such structures can also be automatically found without waiting for more detailed information in future data releases. In this contribution we propose and compare two conceptually different strategies to find globular clusters in the GAIA DR1 survey. The first approach is a nearest neighbor retrieval and the second an anomaly detection. Both techniques are able to find most of the known globular clusters within our observation frames consistently, as well as potential candidates for further investigation. Furthermore we address approximation approaches to scale the strategy to larger data.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

A major interest in astronomy is finding characteristic stellar structures in the sky that are present in large modern astronomical databases, but hidden by structures in front, like dust. With more powerful sensor technology and data surveys ever increasing in size the demand for automated techniques to find and investigate such astronomical structures has grown tremendously. The GAIA mission, for example, launched a satellite at the end of 2013 to survey the whole sky for the creation of an extraordinarily precise three-dimensional map for stars in our Galaxy (the Milky Way) and beyond. Data produced by this mission provides a massive amount of observational data useful to analyze various important problems related to the origin, structure and evolutionary history of our Galaxy.

Good candidates for the study of stellar evolution are structures known as globular clusters (GCs). A globular cluster is a stellar population which assumes that the involved stars were formed at the same time from a single cloud of interstellar gas and dust and therefore being of similar age and chemical compositions, which makes GCs effective tracers of Galaxy formation and evolution [1]. Moreover, the dynamical interactions of GCs with the galactic po-

tential inferred from effects on stellar density and velocity profiles, can be used to study the mass distribution of the Galaxies [2]. However, GCs are comparably small and not necessarily greatly compact stellar structures and therefore not trivial to find.

The development in observing capabilities enable us to collect different types of observations. In the presence of diversity in measurements, we need different techniques to detect stellar structures. A few algorithms like ROCKSTAR [3] and OPTICS [4] identify substructures, such as GCs, from full phase-space information. If only position, magnitudes and color information of the stars is available matched filter methods are typically applied [5]. Also, particular designed clustering algorithms have been proposed along this line [6,7] using color information to detect cluster of galaxies. Contrary, in this contribution we are looking for much smaller structures in data sets not providing color information.

In this paper we focus on the detection of GCs using the data release 1 (DR1) catalog collected by the GAIA satellite. The first release contains highly accurate positions, parallaxes, and proper motions for more than one billion stars brighter than magnitude 20.7 in the white-light photometric band G [8]. However, neither the full phase-space information nor the color information of the stars is available in the GAIA DR1 catalog, and therefore above mentioned methods cannot be used. A very recent contribution proposed the use of the magnitude distribution to detect compact stellar structures such as GCs and dwarf galaxies in GAIA DR1 [5], which inspired this work. They introduce a histogram to approximate magnitude distribution. However, using a one-dimensional

* Corresponding author.

E-mail addresses: m.mohammadi@rug.nl (M. Mohammadi), n.petkov@rug.nl (N. Petkov), k.bunte@rug.nl (K. Bunte), peletier@astro.rug.nl (R.F. Peletier), frank-michael.schleif@thws.de, fschleif@techfak.uni-bielefeld.de (F.-M. Schleif).

histogram has two disadvantages: a histogram is sensitive to the placement of bins and the sole use of the magnitude distribution disregards the information induced by position.

To date, Computer Science has yielded an impressive amount of machine learning algorithms to collect, store, retrieve, visualize and analyze complex interdisciplinary data, which have repeatedly proven useful in decision support and discovery in many fields [9]. In this contribution, we investigate the automatic detection using machine learning when only limited information is available, such that the first releases of forthcoming catalogs can be efficiently scanned for potentially interesting structures right away. For the detection we follow two strategies: a retrieval technique using known GCs to detect similar structures and a novelty detection technique using the fact that most of the sky is empty and does not contain any GCs. The main reason to select these strategies is that both methods are suitable in situations where the availability of examples of the interesting pattern is limited. The contribution is an extension of the conference paper [10]: instead of a one dimensional Kernel Density Estimation (KDE) we extend the strategy using three dimensional KDE and provided more detailed theoretical discussion of the efficiency of the methodology. Furthermore, we extend the demonstration of stellar structure detection to three selected regions of the GAIA DR1 survey. In summary the paper provides the following three major contributions:

- A globular cluster detection technique following two independent strategies, namely similarity retrieval and novelty detection.
- Discussion of the computational costs and approximation schemes to scale the approach to larger data sets.
- An extended set of experimental results showing the efficiency of the approach and a list of potential interesting stellar structures found.

The contribution is structured as follows: in Section 2, we explain the main properties of globular clusters and why they are important for investigation. The data and preprocessing is addressed in Section 3. In Section 4 we explain the retrieval and novelty detection techniques followed by a complexity analysis in Section 5. The effectiveness of the strategies is demonstrated in detail in the experimental Section 6, followed by the results in Section 7 and the contribution is finalized with conclusions and future work.

2. Laboratory for stellar evolution

A globular cluster (GC) is a spherically shaped collection of stars bounded by gravity with high stellar densities toward its center. A star is a luminous spheroid of gas, mostly hydrogen and helium, held together by its own gravity. To measure its brightness, the apparent magnitude m is computed by

$$m = -2.5 \log_{10} F + c, \quad (1)$$

where F denotes the flux of the star received on Earth and c represents a photometric constant. If the magnitude values of stars in two bands are available, we can calculate their color which are closely related to their surface temperature. The Hertzsprung–Russell (HR) diagram uses the relationship between the apparent magnitude (or luminosity) and color to illustrate the evolution of stars.

A star in the main sequence fuses hydrogen into helium in the core region. When its core hydrogen is exhausted, the (low or intermediate mass) star becomes a red giant (the hydrogen in its shell starts burning and larger and lower temperature are observed on the main sequence). While the star is becoming a red giant, the temperature in its core becomes high enough to fuse the helium into carbon. The conversion of helium to carbon takes place very quickly (called helium flash). After that, the core starts fusing of

Fig. 1. The HR diagram of the globular cluster NGC 5024. Important groups are marked by RGB: Red Giant Branch, HB: Horizontal Branch, MS: Main Sequence.

helium to carbon steadily. In this stage, the star reaches the horizontal branch. Since the outer layers of the star is far from the center, the star loses some mass and the sub-sequence evolution depends on the amount of mass that is lost [11].

Fig. 1 depicts an example HR diagram containing the magnitude and color information of stars within the globular cluster NGC 5024. Within a GC we can find stars in different stages of their life which makes them good candidates to study stellar evolution. Furthermore, GCs can help to have a better understanding of the size of our galaxy. Depending on the distance of the cluster, the position of horizontal branch will shift in the apparent magnitude axis which can be described by the distance modulus formula:

$$m - M = 5 \log_{10} d - 5, \quad (2)$$

where m , M and d are apparent magnitude, absolute magnitude and the distance (in parsec), respectively. The absolute magnitude of a star is its intrinsic brightness (in magnitude) which is expected to be $M = -0.5$ for a star in HB [12]. Hence, finding the distances of GCs takes us one step closer to the goal of the GAIA mission of creating a 3-dimensional map of the Milky Way.

3. Data and preprocessing

The GAIA DR1 catalog captures stars of magnitudes up to 20.7 magnitude (mag). We use the magnitude distributions of the stars contained in rectangular windows to detect globular clusters in the GAIA database. As proposed in [5] and [10] we consider windows of 0.5 degree length in both right ascension (RA) and declination (Dec). The left panel in Fig. 2 displays the magnitude distribution as estimated via one dimensional Kernel Density Estimation (KDE) based on the magnitudes extracted from a window centered at position (198.23 RA, 18.17 Dec) of globular cluster NGC5024 in GAIA. In contrast to random windows (Fig. 2 right panel) the distribution of a window containing a globular cluster will exhibit a bump and then a sharp increase which is expected if we collapse the HR diagram in the color dimension (when color information is not available). By comparing Figs. 1 and 2, we see a clear bump at a characteristic magnitude 17, which comes from the horizontal branch (HB). For higher magnitudes, we see a sharp increase towards faint stars, up to the magnitude limit of 20.7 mag in the GAIA survey.

3.1. Preprocessing

The preprocessing procedure we perform throughout the experiments is summarized in Fig. 3. First we divide the sky into smaller

Fig. 2. Left: Magnitude distribution of a window of 0.5 degree centering at globular cluster NGC5024. Right: Magnitude distribution of a random window.

Fig. 3. Summary of preprocessing: (1) divide the sky into smaller windows, (2) extract sources centered around the region of largest density and (3) compute 3D-KDE of the extracted data.

windows of size 0.5×0.5 degree² to ensure each region is taken into account for the analysis. For each bright source, we have three numbers which represent its position (right ascension and declination) and its apparent magnitude. Since the center of GCs are the most dense place in comparison to stars far from the center, we approximate the position of the center finding the most dense place in each window. Then we extract the information (position and apparent magnitude values) of sources contained in the bins of size 0.17×0.17 degree² around the center.

Suppose $(x_1, \dots, x_n)$ is the extracted magnitude values from a window, we can approximate the magnitude distribution by one dimensional KDE as follows:

$$f_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (3)$$

with bandwidth h and the Epanechnikov function as kernel K [13]. Here, we extend the original proposal in [10] to use the full position information. This can be achieved by using three dimensional KDE [14]. Since different coordinates have different units, we need to allow different amount of smoothing in each of the coordinates. Hence, we use 3D KDE which relies upon the product kernels with different bandwidth for each coordinates, as follows:

$$f_H(x) = \frac{1}{nh_1 h_2 h_3} \sum_{i=1}^n K\left(\frac{ra - ra_i}{h_1}\right) K\left(\frac{dec - dec_i}{h_2}\right) K\left(\frac{x - x_i}{h_3}\right) \quad (4)$$

The variables (ra_i, dec_i) denote the position of i th star inside the window and $H = (h_1, h_2, h_3)$ depicts the bandwidths along right ascension, declination and magnitude axes, respectively.

4. Methods

We compare the magnitude densities extracted from a window regarding each magnitude distribution as a sequence (replacing time by magnitude values) and make use of Dynamic Time Warping (DTW) [15,16]. Consider the two D-dimensional sequences $X = x_1, x_2, \dots, x_m$ and $Y = y_1, y_2, \dots, y_n$ where $x_i, y_j \in \mathbb{R}^D$. DTW aligns them non-linearly (see right panel in Fig. 4) in such a way to minimize the distance between them. The alignment can be represented by a sequence of tuples:

$$F = a_1, a_2, \dots, a_K \quad (5)$$

where $a_k = (i_k, j_k)$ contains the indices from X and Y with $i_k \in \{1, 2, \dots, m\}$, $j_k \in \{1, 2, \dots, n\}$. Hence, it assigns an element of X to an element of Y . The sequence F is called warping path, and it should fulfill the following conditions:

- Boundary condition: alignment of the first and last elements of sequence, i.e. $a_1 = (1, 1)$ and $a_K = (m, n)$.
- Monotonicity condition: the assignment can not be backward in time, i.e. if we have $k < l$, then $i_k \leq i_l$ and $j_k \leq j_l$ should hold.
- Step size condition: no element of X and Y can be omitted, and all tuples of F should be distinct.

Given a warping path, we can measure the dissimilarity between them as follow:

$$d(X, Y) = \sum_{k=1}^K \|x_{i_k} - y_{j_k}\| \quad (6)$$

where $\|\cdot\|$ is the euclidean distance. The optimal warping path with the lowest distance d can be achieved by warping the time axes, which can be efficiently obtained using dynamic programming [15]. However, the resulting dissimilarity does not fulfill all metric properties, but it has the potential to detect similar structures in the densities, such as bumps, despite their shifting location on the magnitude axis.

4.1. Nearest neighbor retrieval

Nearest Neighbor Search (NNS) finds similar objects of interest in a data base. For a given example, this technique returns a desired number of most similar objects from the data base according to the distance measure used. By using this strategy we aim to find interesting structures in the GAIA survey by retrieving similar densities to those extracted from the known globular clusters using DTW as dissimilarity measure. To compare the extracted distribution with a k nearest neighbors of the known structures, the number of common GCs for each window is counted. The higher the obtained count the more likely it is that there is something interesting in it. This technique can be easily extended to deal with even larger data sets by using approximate nearest neighbor search approaches, as for example locality-sensitive hashing [17], best bin first [18] and balanced box-decomposition tree based search [19].

Fig. 4. Comparison of densities using Euclidean distance and Dynamic Time Warping.

4.2. Novelty detection

Since the sky is mostly empty we can regard the search for interesting structures as novelty or outlier detection. We use an approach which was inspired by the Support Vector Classifier called Support Vector Data Description (SVDD) [20]. Assume given data $\vec{x}_i \in \mathbb{R}^n$ and a fixed chosen kernel $K: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ which is associated to the feature map $\Phi: K(\vec{x}, \vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle$.

The motivation for such an embedding comes with the hope that the non-linear transformation of input data into a higher dimensional space allows for using linear techniques. Kernelized methods process the embedded data points in a feature space utilizing only the inner products $\langle \cdot, \cdot \rangle$ (kernel trick) [21], without the need to explicitly calculate Φ . The kernel function can be very generic. Most prominent are the linear kernel with $K(\vec{x}, \vec{y}) = \langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle$ where $\langle \Phi(\vec{x}), \Phi(\vec{y}) \rangle$ is the Euclidean inner product and Φ identity mapping, or the RBF kernel $K(\vec{x}, \vec{y}) = \exp\left(-\frac{\|\vec{x}-\vec{y}\|^2}{2\sigma^2}\right)$, with $\sigma > 0$ as a free scale parameter. In any case, it is always assumed that the kernel function $K(\vec{x}, \vec{y})$ is positive semi-definite (psd).

The goal of SVDD is to find a generalized linear mapping $\vec{x} \mapsto \text{sgn}(f(\vec{x})) = \text{sgn}(\vec{w}^T \Phi(\vec{x}) - \rho)$, which defines a separation of the given data to outliers by means of its sign. Where the separation boundary corresponds to a linear separation in the feature space induced by Φ . The problem to find suitable parameters $\vec{w}$ and ρ for a given data set of typical points $\vec{x}_i$ can be formalized as an optimization problem which aims at a separation of the given data from the origin with maximum margin. This leads to the following primal optimization problem SVDD (primal): $\min_{\vec{w}, \rho, \xi_i} \frac{1}{2} \cdot \|\vec{w}\|^2 - \rho + \frac{C}{2} \cdot \sum_i \xi_i^2$, such that $\vec{w}^T \Phi(\vec{x}_i) \geq \rho - \xi_i \quad \forall i$, where $C > 0$ is a fixed constant, and the parameters ξ_i refer to the slack variables to allow for some error tolerance. Using the Karush-Kuhn-Tucker conditions, the Lagrange dual problem becomes

SVDD(dual)

$$\begin{aligned} \max_{\alpha_i} \quad & -\frac{1}{2} \cdot \sum_{ij} K(\vec{x}_i, \vec{x}_j) \alpha_i \alpha_j + \frac{1}{C} \cdot \sum_i \alpha_i^2 \\ \text{such that} \quad & \alpha_i \geq 0 \quad \forall i \\ & \sum_i \alpha_i = 1 \end{aligned}$$

This dual problem can be directly optimized relying on linearly constraint convex quadratic optimization. The solution $\vec{w}$ and ρ of the primal problem can then be recovered from the dual variables α_i . These are non-vanishing for support vectors only and hence result in a sparse description. For large data sets it is possible to use optimized SVDD solvers or to apply the equivalent core vector data description technique (CVDD), which effectively scales to millions of points [22].

The tightness of the data description can be controlled by setting an error tolerance or using outlier examples one desires to detect. However, for this approach we require a positive definite kernel also known as Mercer kernel [23], which we cannot guarantee if the DTW dissimilarity is used due to missing metric prop-

erties. To obtain a kernel (similarities) from a dissimilarity measure (like DTW) it is possible to use a technique called double-centering [23]:

$$S = -DJJ/2 \text{ with } J = (I - \mathbf{1}\mathbf{1}^T/n) \quad (7)$$

with D being a (squared) dissimilarity matrix as obtained from a given dissimilarity function and identity matrix I and vector of ones $\mathbf{1}$. However in case of DTW the obtained kernel will be non-psd.

Non-metric dissimilarity measures and non-psd, or indefinite kernels have recently attracted a lot of research activities, with a summary given in [16]. These measures have often found to be very effective and more flexible than standard metric measures. Most kernel methods are restricted to psd (mercator) kernels and a variety of transformation approaches have been suggested to convert non-psd kernels into psd-kernels [16].

A natural way to address this problem and to obtain a psd similarity matrix is to correct the eigenspectrum of the original similarity matrix S . Popular strategies include *flipping*, *clipping*, *shift correction*. The non-psd similarity matrix S is decomposed as

$$S = U\Lambda U^T, \quad (8)$$

where U contains the eigenvectors of S and Λ contains the corresponding eigenvalues. The following approaches are commonly used:

- clipping - the negative eigenvalues in Λ are fixed to 0
- flipping - the sign of the negative eigenvalues in Λ is changed to positive
- square - all eigenvalues are squared
- shift - the eigenspectrum (from Λ) is shifted by the most negative eigenvalue value, such that all entries in Λ are non-negative.

Spectrum shift enhances all the self-similarities by the amount of ν and does not change the similarity between any two different data points. Detailed equations, further transformation techniques as well as an in depth analysis of the various methods, is provided in [16]. We approximate a positive definite kernel (psd) based on the DTW dissimilarity following the strategy discussed in [16] using the *square approach*, which was found to be of good performance in a number of tasks. We follow two steps to convert the DTW pairwise dissimilarity matrix D between magnitude distributions into an appropriate kernel as input to the SVDD:

1. Constructing the dissimilarity matrix by computing DTW for all pairs of magnitude distributions
2. computing similarity matrix S from D via double centering and
3. converting S to a psd kernel by $K = S \cdot S^T$.

For large scale data this approach can be optimized by using additional methods detailed in [24].

Fig. 5. Number distribution of all the stars in GAIA DR1 with $10 < G < 20.7$ [5]. Boxes marked three selected regions with the positions: (1) $0^\circ < \alpha < 80^\circ$ and $-90^\circ < \delta < 0^\circ$, (2) $120^\circ < \alpha < 246^\circ$ and $-2^\circ < \delta < 60^\circ$, and (3) $310^\circ < \alpha < 360^\circ$ and $-90^\circ < \delta < 0^\circ$.

5. Complexity analysis

The alignment distances can be efficiently calculated using dynamic programming with a worst-case runtime of $\mathcal{O}(L_1 \cdot L_2)$, with L_1 and L_2 being the number of spectrum elements in the first and the second input spectrum, respectively. The conversion from dissimilarities to positive-semi-definite similarities can be achieved with linear cost in $\mathcal{O}(n)$ using an approximation strategy (including an effective out-of sample extension) suggested in [24].

The application of the nearest neighbor classifier has linear complexity in $\mathcal{O}(N)$, but can also be reduced to potentially log-linear costs using e.g. k-d-trees [25]. The SVDD algorithm has roughly $\mathcal{O}(n^2)$ complexity but using the CVDD approach the complexity can be reduced to $\mathcal{O}(n)$. All together using specific optimization concepts and data structures, the NN approach has $\mathcal{O}(\log n)$ runtime and $\mathcal{O}(n)$ memory complexity in the test phase. With NN a training is not needed. Using the SVDD / CVDD technique we get $\mathcal{O}(n)$ runtime and memory complexity during training and due to sparsity and approximation concepts sub-linear memory and runtime complexity during testing. In the presented results we have used exact approaches only.

6. Experiments

We apply the proposed strategies on three regions around the disk of Milky Way as depicted in Fig. 5. As explained in Section 3.1, the sky is divided into smaller windows of 0.5 degree bin width. Then we find the most dense place in each window and extract information of sources in the neighborhood centered at it with the size of 0.17 degree. We use the 3D-KDE with bandwidth $H = (0.1, 0.1, 0.15)$. And we extract the KDE values in a regular grid of size $3 \times 3 \times 500$: 3 along RA and Dec axes and 500 along magnitude axis, respectively. The resulting grid is reshaped as a sequence of vectors $(v_i \in \mathbb{R}^9, i = 1, \dots, 500)$.

There are 26 known globular clusters in the selected regions. Since we have a limited magnitude of stars of maximal 20.7 in the GAIA survey, we do not include the more distant known GCs, which cannot be expected to be captured in the data. A total of 17 known GCs remains for the analysis in our experiments (see Table 1).

6.1. Nearest neighbor retrieval

For the nearest neighbor retrieval we use the magnitude distributions of the 17 known GCs as example structures. All these examples are then compared with DTW to windows of equal size extracted by sliding with 0.5 degree through the parts of the sky under observation. Using the counting strategy discussed above the

Table 1

Names and positions of the known GCs in the selected regions [26].

Name	Position		Region
	Right ascension	Declination	
NGC 1851	78.53	−40.05	1
47 Tuc	6.02	−72.08	1
NGC 288	13.19	−26.58	1
NGC 362	15.81	−70.85	1
NGC 1261	48.07	−55.22	1
NGC4147	182.53	18.54	2
NGC 5024	198.23	18.17	2
NGC5053	199.11	17.7	2
NGC 5272	205.55	28.38	2
NGC5466	211.36	28.53	2
NGC5904	229.02	−0.11	2
Palomar5	229.64	2.08	2
M72	313.37	−12.54	3
M2	323.36	−0.82	3
M30	325.09	−23.18	3
Palomar 12	326.66	−21.25	3
NGC 7492	347.11	−15.61	3

Table 2

Mean hit number (and std) for different hyper-parameter settings k and E in retrieval and outlier detection respectively.

HP	Candidates		GCs found	
k 10	79.25	(11.18)	8.75	(1.26)
k 20	135.75	(19.00)	11.00	(1.41)
k 50	309.25	(31.06)	12.50	(0.58)
k 100	540.00	(51.20)	14.00	(0.00)
k 250	1174.00	(101.78)	16.00	(0.00)
E 1%	90.20	(52.82)	14.40	(0.55)
E 5%	243.80	(37.93)	15.60	(0.89)
E 10%	305.80	(20.05)	16.80	(0.45)
E 20%	318.80	(27.81)	17.00	(0.00)

promising candidates are identified. To evaluate this technique we use 4-fold cross validation. Due to the relatively small number of GCs (seventeen), we divide them into four non-overlapping folds, such that the method is tested on structures not used for training. Then DTW is used to find the most similar distributions in the test set to queries in the training set.

6.2. Novelty detection

For the outlier detection we build 5 psd kernels for cross validation based on the DTW dissimilarities as mentioned before. As a proof of concept we build kernels of 6000 samples (2000 per region) each containing:

- a training subset of 4200 random windows taken from GAIA (1400 per region) resembling empty sky¹ and
- a test set containing the windows extracted around the known GCs, potential candidates from the retrieval and 600 random windows per region.

Therefore, we train 5 SVDD models on different random subsets of empty sky to get models depicting typical uninteresting windows. To evaluate our method, we use the testing set (b) and report the detection rate of known GCs as well as overlap with the retrieval method. In our experiments we investigate the threshold ρ for the decision boundary and therefore the tightness of the data description by an error tolerance E .

¹ The probability of accidentally picking a window containing an interesting structure is very low and the SVDD has a hyper-parameter to allow outliers.

Fig. 6. Workflow of the process: at first, the preprocessing step is done (Section 3.1), then we use two different approaches: Nearest neighbor retrieval (Section 4.1) and Novelty detection (Section 4.2). At the end, DoG (Section 6.3) is used for the candidates obtained by two methods.

Fig. 7. The average votes for the candidate using 40 NN retrieval and the circles mark interesting regions with the color indicating the similarity to known GCs.

6.3. Difference of Gaussian filters

A difference of Gaussian (DoG) filter [27] is constructed by subtraction of a Gaussian filter from a tighter one as follows:

$$DoG_{\sigma}(x, y) = \frac{\exp\left(-\frac{x^2 + y^2}{2(0.5\sigma)^2}\right)}{2\pi(0.5\sigma)^2} - \frac{\exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)}{2\pi\sigma^2}, \quad (9)$$

where σ is standard deviation. This filter gives high response to a circular shape. To determine the exact position of a GC and to reduce false positives we use DoG filters to judge the candidates. The work-flow of the process is depicted in Fig. 6.

7. Results

Both methods retrieve an increasing number of potential candidates with increasing their corresponding hyper-parameter. The summary of the results is given in Table 2. Since the retrieval technique uses examples of the GCs, it is more specific in comparison to the outlier detection strategy. To evaluate the retrieval technique, we use 4-fold cross validation to see whether it can detect the known GCs (see the first 5 rows of Table 2). For this technique, 47 Tuc, NGC 362 and Palomar 5 are more difficult. This is due to the fact that the radii of 47 Tuc and NGC 362 are big (even bigger than 0.25 deg), making them difficult to detect using smaller windows. However, we still can detect them if we increase the hyper-parameters or the size of the windows. The average vote for $k = 40$ nearest neighbors and corresponding candidates together

Fig. 8. Two candidates found at positions (40, −34.5) and (330.78, −51.3). Left: the distribution of stars and gray levels representing the magnitude values of sources from GAIA DR1. Right: stacked responses of spatial Difference of Gaussian filter at different magnitudes.

with known GCs in the investigated regions are shown in Fig. 7. Windows with more than one third of the maximum vote are clustered based on Ra and Dec and hits closer than 0.75° denote the same candidate.

Using the outlier detection strategy models are trained solely on random windows. Therefore the models are not provided with any example of the known structures, but still detect almost all 17 known GCs used for testing when allowed an error tolerance $E \geq 10\%$. Furthermore it is important to note that the outliers found often overlap with candidates detected using the retrieval method.

To reduce false positives, specially for the outlier detection, we use DoG filters [27] in a post processing step. The right panels of Fig. 8 shows filter responses to two potential interesting candidates found by our strategies at different magnitudes stacked on top of each other. On the left side we show original magnitudes of sources at the corresponding RA and Dec position taken from the GAIA DR1 survey. We observe that the filter starts giving high responses in the center of the structure (i.e. a circular shaped group of stars) from some magnitude up to limit magnitude of GAIA, but a weak response (i.e. uniformly distributed stars) at lower magnitude, which can be inferred that there are fewer bright stars and more faint stars in these structure.

8. Conclusion and outlook

Techniques to detect interesting stellar structures often require full-phase space or color information, which is not always available especially in early releases of astronomical data. In this contribution two independent strategies are presented to find interesting stellar structures targeted at astronomical surveys which provide only limited information such as 2D positions and brightness. Previous work [5] uses histograms as features and manual inspection to retrieve possible candidates. In comparison we use three dimensional KDEs as features and automatic retrieval and novelty detection techniques to find possible candidates. The methods are demonstrated for the detection of globular clusters in the GAIA

DR1 survey. Using a nearest neighbor retrieval method we compare examples extracted from 17 known GCs to other parts of the sky. Furthermore, assuming the sky is mainly empty we treat the problem of finding interesting structures as outlier detection problem. Therefore we used some random windows of the sky to train a support vector data description, a kernel-based anomaly detection technique. Both techniques find most of the 17 known GCs used for testing and also depict overlapping agreement on potential candidates found. Since these methods could give false positive candidate which is the case when their hyper-parameters are big, we use Difference of Gaussian filters on the candidate. However, since the GAIA DR1 does not provide the color information, it is still possible to get false positive. For example, it is possible to find some other types of structures, such as dwarf galaxies because some of them have similar magnitude distributions and also spherical shapes. Recently, the new version of GAIA survey (GAIA DR2) has been released which includes color information and proper motions. Since the size of GAIA DR2 is much bigger than GAIA DR1, we can use our result found here (instead of starting from zero) to detect and classify structures. Future work includes the detailed investigation of the promising candidates (using GAIA DR2) and efficient out-of-sample extension.

Acknowledgment

This work was supported by the European H2020-MSCA-ITN Survey Network for Deep Imaging Analysis and Learning (SUNDIAL), project ID 721463. We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster.

References

- [1] J.E. Barnes, Evolution of compact groups and the formation of elliptical galaxies, *Nature* 338 (1989) 123–126.
- [2] C. Rockosi, M. Odenkirchen, E. Grebel, W. Dehnen, K. Cudworth, J. Gunn, D. York, J. Brinkmann, G. Hennessy, Z. Ivezić, A matched-filter analysis of the tidal tails of the globular cluster palomar 5, *Astronom. J.* 124 (1) (2002) 349.

- [3] P. Behroozi, R. Wechsler, H.-Y. Wu, The rockstar phase-space temporal halo finder and the velocity offsets of cluster cores, *Astrophys. J.* 762 (2) (2012) 109.
- [4] S.S. Fuentes, J. De Ridder, J. Debosscher, Stellar halo hierarchical density structure identification using (f) optics, *Astron. Astrophys.* 599 (2017) A143.
- [5] H. Tian, Hide and seek in the halo of the MW, University of Groningen, 2017 Ph.D. thesis.
- [6] C. Miller, R. Nichol, D. Reichart, R. Wechsler, A. Evrard, J. Annis, T. McKay, N. Bahcall, M. Bernardi, H. Boehringer, A. Connolly, T. Goto, A. Kniazev, D. Lamb, M. Postman, D. Schneider, R. Sheth, W. Voges, The C4 clustering algorithm: clusters of galaxies in the sloan digital sky survey, *Astron. J.* 130 (3) (2005) 968–1001. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-25844455623&doi=10.1086%2f431357&partnerID=40&md5=56450620136028ee521627a5df8ba92>.
- [7] D. Murphy, J. Geach, R. Bower, ORCA: the overdense red-sequence cluster algorithm, *Mon. Not. R. Astron. Soc.* 420 (3) (2012) 1861–1881.
- [8] A. Brown, A. Vallenari, T. Prusti, J. De Bruijne, F. Mignard, R. Drimmel, C. Babusaix, C. Bailer-Jones, U. Bastian, et al., GAIA data release 1—summary of the astrometric, photometric, and survey properties, *Astron. Astrophys.* 595 (2016) A2.
- [9] C.M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.
- [10] M. Mohammadi, R.F. Peletier, F.M. Schleif, N. Petkov, K. Bunte, *Globular cluster detection in the GAIA survey*, in: *Proceedings of the ESANN*, 2018.
- [11] M.L. Kutner, *Astronomy: A Physical Perspective*, Cambridge University Press, 2003.
- [12] M. Riello, S. Cassisi, G. Piotto, A. Recio-Blanco, F. De Angeli, M. Salaris, A. Pietrinferni, G. Bono, M. Zoccali, The red giant branch luminosity function bump, *Astron. Astrophys.* 410 (2) (2003) 553–563.
- [13] C.-Y. Chu, D.J. Henderson, C.F. Parmeter, On discrete epanechnikov kernel functions, *Comput. Stat. Data Anal.* 116 (2017) 79–105, doi:10.1016/j.csda.2017.07.003. <http://www.sciencedirect.com/science/article/pii/S0167947317301512>.
- [14] J.S. Simonoff, *Smoothing Methods in Statistics*, Springer Science & Business Media, 2012.
- [15] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust. Speech Signal Process.* 26 (1) (1978) 43–49.
- [16] F. Schleif, P. Tiño, Indefinite proximity learning: a review, *Neural Comput.* 27 (10) (2015) 2039–2096, doi:10.1162/NECO_a.00770.
- [17] J. Leskovec, A. Rajaraman, J. Ullman, *Mining of Massive Datasets*, Cambridge University Press, 2014.
- [18] J. Beis, D. Lowe, Shape indexing using approximate nearest-neighbour search in high-dimensional spaces, in: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, 1997, pp. 1000–1006.
- [19] S. Arya, D. Mount, N. Netanyahu, R. Silverman, A. Wu, An optimal algorithm for approximate nearest neighbor searching fixed dimensions, *J. ACM (JACM)* 45 (6) (1998) 891–923.
- [20] D. Tax, R. Duin, Support vector domain description, *Pattern Recognit. Lett.* 20 (11) (1999) 1191–1199, doi:10.1016/S0167-8655(99)00087-2. <http://www.sciencedirect.com/science/article/pii/S0167865599000872>.
- [21] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis and Discovery*, Cambridge University Press, 2004.
- [22] I.W. Tsang, J.T. Kwok, J.M. Zurada, Generalized core vector machines, *IEEE Trans. Neural Netw.* 17 (5) (2006) 1126–1140.
- [23] E. Pekalska, R. Duin, *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*, World Scientific, 2005.
- [24] A. Gisbrecht, F. Schleif, Metric and non-metric proximity transformations at linear costs, *Neurocomputing* 167 (2015) 643–657.
- [25] S.S. Skiena, *The Algorithm Design Manual*, Springer-Verlag, Berlin, Heidelberg, 1998.
- [26] W. Harris, A catalog of parameters for globular clusters in the milky way, *Astronom. J.* 112 (1996) 1487.
- [27] B. Gecer, G. Azzopardi, N. Petkov, Color-blob-based cosfire filters for object recognition, *Image Vis. Comput.* 57 (2017) 165–174.

Mohammad Mohammadi has two master degrees: Applied Mathematics (Numerical Analysis) from Shahid Beheshti University, Tehran, Iran and Applied Mathematics in Digital Media from Mittweida University of Applied Science, Mittweida, Germany. He received his B.Sc. degree in Applied Mathematics from University of Kashan. He is a Ph.D. candidate at the Bernoulli Institute and Engineering for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen, the Netherlands since July 2017. His research interests are in machine learning, in particular, dimensionality reduction and visualization techniques.

Nicolai Petkov received the Dr.sc.techn. degree in Computer Engineering (Informationstechnik) from Dresden University of Technology, Dresden, Germany. He is professor of computer science and head of the Intelligent Systems group of the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen, the Netherlands. He is the author of two monographs and coauthor of another book on parallel computing, holds four patents and has authored over 100 scientific papers. His current research is in image processing, computer vision and pattern recognition, and includes computer simulations of the visual system of the brain, computer applications in health care and life sciences and creating computer programs for artistic expression. He is a member of the editorial boards of several journals.

Kerstin Bunte graduated at the Faculty of Technology at the University of Bielefeld, Germany. She is an assistant professor and Rosalind Franklin Fellow at the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence of the University of Groningen, the Netherlands since July 2016. He received her Ph.D. in Computer Science in 2011 from the University of Groningen. She had research visits at the University of Rochester (USA) and postdoctoral positions have taken her to Bielefeld University (Germany), Aalto University (Finland), and Université catholique de Louvain (Belgium). In 2015 she got a European Marie Skłodowska-Curie Fellowship (Project ID: 659104) at the University of Birmingham and is partner in the European ITN: SUNDIAL, Project ID: 721463. Her recent work has focused on the development of interpretable machine learning techniques for interdisciplinary data analysis and visualization. Further information can be obtained from <http://www.cs.rug.nl/~kbunte/>.

Reynier F. Peletier (M.Sc. University of Leiden. Ph.D. University of Groningen) has been a postdoc at the Harvard-Smithsonian Center for Astrophysics, a postdoctoral fellow at the European Southern Observatory, and a postdoc in Groningen. In 1997 he obtained a PPARC Advanced Fellowship at the University of Durham, two years later he was appointed as lecturer at the University of Durham, and in 2003 he moved back to Groningen. He is currently a Full Professor at the Kapteyn Astronomical Institute at the University of Groningen. He is interested in the evolution of galaxies, their dynamics and stellar populations, but also in instrumentation for large astronomical telescopes. He has edited several books and written more than 200 papers in refereed journals. At present, he is leading an EU International Training Network, named SUNDIAL, integrating computer scientists and astronomers to develop novel data analysis techniques for use in astronomy.

Frank-Michael Schleif (Dipl.-Inf., University of Leipzig, Ph.D., TU-Clausthal, Germany) was a Marie Curie Senior Research Fellow at the University of Birmingham, Birmingham, UK and a Post-Doctoral Fellow in the group of Theoretical Computer Science (TCS) at the University of Bielefeld, Bielefeld, Germany, where he also received a *venia legendi* in applied computer science in 2013. He was also a software developer and consultant for the Bruker Corp. Since 2016 he is with the University of Applied Sciences, Wuerzburg, Germany, where he is a Professor for Database Management and Business Intelligence. His current research interests include data management, computational intelligence techniques and machine learning for non-metric models and large scale problems. Several research stays have taken him to UK, the Netherlands, Japan and the USA. He is a member of the German chapter of the European Neural Network Society (GNNS), the GI and the IEEE-CIS. He is editor of the Machine Learning Reports and member of the editorial board of the Neural Processing Letters.