Exp Astron
DOI 10.1007/s10686-016-9499-9

ORIGINAL ARTICLE

Open cluster membership probability based on K-means clustering algorithm

Mohamed Abd El Aziz[1,2] · **I. M. Selim**[3,4] · **A. Essam**[3]

Received: 10 November 2015 / Accepted: 3 May 2016

Abstract In the field of galaxies images, the relative coordinate positions of each star with respect to all the other stars are adapted. Therefore the membership of star cluster will be adapted by two basic criterions, one for geometric membership and other for physical (photometric) membership. So in this paper, we presented a new method for the determination of open cluster membership based on K-means clustering algorithm. This algorithm allows us to efficiently discriminate the cluster membership from the field stars. To validate the method we applied it on NGC 188 and NGC 2266, membership stars in these clusters have been obtained. The color-magnitude diagram of the membership stars is significantly clearer and shows a well-defined main sequence and a red giant branch in NGC 188, which allows us to better constrain the cluster members and estimate their physical parameters. The membership probabilities have been calculated and compared to those obtained by the other methods. The results show that the K-means clustering algorithm can effectively select probable member stars in space without any assumption about the spatial distribution of stars in cluster or field. The similarity of our results is in a good agreement with results derived by previous works.

Keywords Galaxy · Open clusters and associations · Individual · k-means clustering algorithm · NGC 2266 · NGC 188 · Color-magnitude diagram (CMD)

✉ I. M. Selim
i_selim@yahoo.com

[1] Department of Mathematics, Faculty of Science, Zagazig University, Zagazig 44519, Egypt

[2] Faculty of Computer Science, Nahda University, Beni Suef, Egypt

[3] National Research Institute of Astronomy and Geophysics Astronomy Dept., Cairo, Egypt

[4] Higher Technological Institute (HTI), Cairo, Egypt

Published online: 13 May 2016

1 Introduction

Open star clusters are the excellent tool to tracers the properties of star formation, the formation and the evolution of the disk of our galaxy and structural points of view, [1]. Study of star clusters provides important clues to the formation and enrichment history of the Milky Way. Distance, age, chemical content, and foreground extinction, they provide stronger constraints on models of stellar evolution than field stars. Determination of open cluster membership probability is a big astronomical problem. So membership determination is an essential task to study an open cluster, which can directly influence estimation of physical parameters and accurate photometric study for star clusters, Selim et al [2]. The first step to determine the physical parameters of an open cluster is determination probable members in the vicinity of the cluster, [3]. Disentangling cluster members from contaminating foreground and background field stars is an important task to study open clusters, which can directly influence on the estimation of open clusters physical parameters. Various methods based on the analysis of positions, proper motions, radial velocities, magnitudes and their combinations have been proposed to determine the membership of open clusters e.g. [4–13]). Few clusters have a set of radial velocity and proper-motion data. When kinematic data are available, it is most valuable criterion for establishing recognized that the membership probabilities obtained from the analysis of proper motions or radial velocities are more reliable.

In the last few years, some authors have developed algorithms with a view to optimization and automatize the analyses process and structure of open clusters e.g. [14–17, 12].

In this, work we present our results of member stars for the open cluster NGC 188 and NGC 2255 using the K-means algorithm.

The rest of this paper is organized as follows: in section 2, the new algorithm based on K-means is introduced. In section 3 the data and result. The conclusion and summary presented in section 4.

2 The methodology based on K-means

In this section we illustrate that the new algorithm solves the open cluster membership based on geometry and photometric properties. Our algorithm consists of two stages; the first one is to clustering stars based on K-means (geometry) and the second clustering stars based physical properties (photometric). Where K-means is one of the simplest clustering algorithm, [18, 19], and it has more attention in the recently years, since it has many applications such as image analysis, bioinformatics [20] and other applications. The K-means algorithm is an iterative algorithm that minimizes the objective function (in general is the Euclidean distance) that used to determine the distance between data points and its cluster center [21].

In the other word, K-means aims to clustering a set of n points for $\{x_1, x_2, \ldots, x_n\}$, where each point is a d-dimensional vector, into K groups, $G = \{G_1, G_2, \ldots, G_k\}$, by

minimizing the squared Euclidean distance:

$$f(C,G)=\sum_{k=1}^{K}\sum_{xi\in G_k}\|x_i-c_k\|^2,\ c_k=1/|G_k|\sum_{x_i\in G_k}x_i \tag{1}$$

Where c_k is the center of cluster G_k. In generally K-means consists of two steps. The first step is the assignment that aims to find the nearest cluster for each point by finding the smallest distance between the points and each cluster center. When all the points belonging to a given data set are included in some clusters, then the first step is completed. The second step is update the centers of K clusters, repeated the first and second step until no more changes in centers.

However, K-means clustering algorithm suffers from some shortcomings, such as its requiring a user to give out the number of clusters at first, and to avoid this drawback the PCA/SVD is used to determine the K principle vectors (clusters). The final K-means is illustrated in Algorithm 1.

Determining the cluster center by maximum spatial density value is requiring reasonable initial values for the center and the radius. The radius of the open star cluster is usually at the maximum value of the radial density profile (the variation of the density of stars per unit area with the distance from the cluster's central coordinates).

Algorithm 1: The k-means clustering algorithm Input:

Input $D=\{d_1,d_2,\ldots,d_n\}$ //set of n data items.
K // Number of desired clusters

Output:

A set of K clusters, c_k:The center of each cluster, ind: index of each pixel and $Wsum$: within-cluster sum of point-to-point centroid.

Steps:

1. K =PCA(D); // estimate the initial K clusters
2. Repeat
   - Calculate the distance between each data point and cluster centers.
   - Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
   - Recalculate the new cluster center using:

$$c_k=(1/|G_k|)\sum_{j=1}^{n_i}x_j$$

Until convergence criteria is met.

The second stage of our method is clustering the results from first stage based on physical (photometric) properties. Then we compute the difference in magnitude

$$Color\,index:\ JH=Jmag-Hmag,\ st-2\le color\ index\le 2 \tag{2}$$

Where Jmag. Is magnitude *Hmag*. is magnitude in any filter, (we implemented on UBV data). Then the *Jmag* that corresponding to *JH* is divided to number of

intervals where the Scott's algorithm [22] is used to compute the optimal number of intervals (bins) $\Delta Jmag_j$, again we determine the elements of $JH(\Delta JH\_E)$ that corresponding to each $\Delta Jmag_j$. The change in color $\Delta color = \Delta(JH)$ is computed from equation:

$$\Delta color_j = B \pm \varepsilon_j, \quad \forall B \in \Delta JH\_E_j, j \in N \tag{3}$$

$$\varepsilon_j = \sqrt{\sum_{i=1}^{n} \left(B_i - \overline{\Delta JH\_E_j}\right)^2} \times \sqrt{\frac{1}{N-1}} \tag{4}$$

Where n the element inside each interval ΔJH_E_j, $\overline{\Delta JH\_E_j}$ is the mean of element in this interval and N is the total number of intervals. Based on (3) select the elements that satisfy the following condition:

$$\overline{\Delta JH\_E_j} - \varepsilon_j \leq JH\_E \leq \overline{\Delta JH\_E_j} + \varepsilon_j \tag{5}$$

The final step in our algorithm is to compute the density of elements before and after applied physical properties (photometric) where a stable measure of central tendency called weighted average distance is used [23]. Each value in the cluster C_H can be weighted with respect to its position, and the weighted values can be averaged. Central points in a cluster will then give the highest weights, while other points will get lower weighting coefficients. The weighted mean is defined as:

$$WM = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \tag{6}$$

A possible problem might be determining weighting coefficients for values from the points in cluster. Indeed, any weighting coefficients w_i could be substituted into the above formula such as variance, such as in [24] but it is affected by outliers. The weighting coefficient for x_i is computed as the inverse mean distance between and the other data points:

$$w_i = \frac{(n-1)}{\sum_{j=1}^{n} |x_i - x_j|}$$

The Proposed algorithm is illustrated in Algorithm 2.

Algorithm 2: Proposed algorithm

Input: the coordinates of pixels $P(x, y)$,$Jmag, Hmag$
Output: $Jmag, Hmag, JH\_E, \Delta color_j$
C_H Element in the cluster that satisfy the geometry and physical properties
// The First stage geometry clustering

1. Using K-mean to compute
   $[C_k, ind, Wsum]$=K-means (P).
2. Determine the center (c_x, c_y) of cluster with largest number of elements (high density)
   $C_H = \arg\max_K Wsum$
3. Compute the radius of C_H, by $R = \max(d), d = \sqrt{(x_i - c_x)^2 + (y_i - c_y)^2}$
4. If P $\notin C_H$ *and* d $\leq R$
   $P \in C_H$
5. End

// The Second stage physical clustering

1. Determine the $Jmag$ and $Hmag$ that corresponding $\forall P \in C_H$.
2. Compute $JH = Jmag - Hmag, \text{st} - 2 \leq JH \leq 2$
3. Choose $Jmag$ and $Hmag$ that corresponding to JH
4. Determine the number of bins (*nbins*) of $Jmag$, and the center of each interval in $Jmag$ (c_{Jmag}) and $B \in \Delta JH\_E_j$

for j=1:nbins
Find the mean of elements in ΔJH_E_j
Compute ε_j using (4),$\Delta color_j$ using (3).
Determine JH_ER that satisfy (5).
Compute density for both JH_E and JH_ER based on (6) and (7)
End

3 Data

K-means algorithm needs high-quality data (star position in the field of cluster direction and magnitude of all stars in the field), we have selected the particular two open stars clusters, the first open star cluster NGC 188 and the second open star cluster NGC 2266. We select open cluster NGC 188 (α=00 47 28, δ=+85 15 18) & (l=122.9 ◦ , b=+22.4 ◦) as a target because it is one of the oldest open clusters known in our Galaxy [25, 26], as well as it is very well known and its membership probabilities, for field stars have been measurements based on redial velocity and proper motions. This particular open cluster is one of the oldest open clusters, have been studded and described by many authors e. g (van den Berg [27], Sandage [28]. Astrometry data of NGC 188 have been obtained using a collection of positions, magnitude and color for 2355 objects in the UBV, the source of data is “WEBDA, [29, 30].

open star cluster NGC 2266 located at (α2000.0 = 06h 43m 19s, δ2000.0 = 26o 58′, l = 187.8o, b = 10.3o) has been selected because it is old open clusters, as well as its membership probabilities for field stars have been measurements using new CCD BVRI observation and JHKs 2MASS data [2]. All-stars in each open cluster field will be used for membership determination based on the K-means algorithm.

4 Results and discussion

4.1 NGC 188

Determination of a cluster's central coordinates is very importance for given the direct impact of its value and thus in its radius estimation.

From the first stage of our algorithm for NGC 188, the geometric clustering is performed then cluster center (c_x, c_y) of a cluster with the largest number of elements (high density) is determined. The value of (c_x, c_y) is (α = 00h 47m 18s, δ = +85d 06m 40s) this value agree with the center in WABDA and the adopted center by (Bonatto et al [25]). The NGC 188 radius is about 9.979 arcmin see Fig. 1.

From the second stage of our algorithm, (physical clustering or photometric clustering), we estimate the total number of probable cluster members based on photometric criteria as described in algorithm above. 545 member stars are selected from 2240 stars in the direction of open cluster NGC 188, see CMD of cluster for selected member and nonmember as shown in Fig. 2. [25] conclude that the large projected area of NGC 188, containing ~ 620 member stars, can be used to search for spatial variations in the stellar content, with statistically significant results [13]

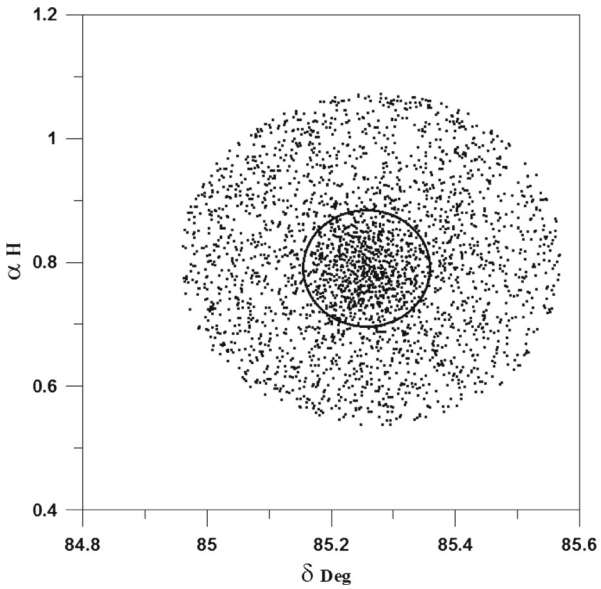

Fig. 1 NGC 188 region with the radius 9.979 arcmin R.A in H And Dec. in Deg

applying a new method (DBSCAN) clustering algorithm to determined membership of NGC 188 and found 472 member stars are selected as a membership in the direction of cluster. Membership probabilities for stars in the NGC 188 field have been measurements based on redial velocity and proper motions (Cannon, Upgren et al. [31]; Xin-Hua [12]). (Bonatto et al. [25]) found that NGC 188, contains a couple of hundred member stars. Its field is not heavily contaminated by background stars. All this makes NGC 188 a perfect target for testing new method for determination of cluster membership.

The final estimated number of cluster members obtained, which means that they are dependent on the completeness level cluster region and physical properties (photometric properties) of cluster. The CMD of the 545 member stars shows a well-defined main sequence and red giant branch Fig. 2, which indicates that our membership determination is very effective.

Finally we compute the density of elements (weighted average distance) after Appling second stage physical properties (photometric) of algorithm where the probability members of stars in NGC 188 is shown in Fig. 3.

4.2 NGC 2266

For NGC 2266 cluster center (c_x, c_y) have been determined with largest number of elements (high density) in the value of (c_x, c_y), it is ($\alpha = 06h\ 43m\ 03s$, $\delta = +26d\ 57m\ 50s$), this value is agree with the center determined by [2] within the cluster radius 7.00 arcmin as shown in Fig. 4.

Appling the second stage of our algorithm on NGC 2266, 908 member stars are selected from 1780 stars in the direction of open cluster, see CMD of cluster Fig. 5. Selim et al [2] conclude that NGC 2266, containing ~ 942 membership stars.

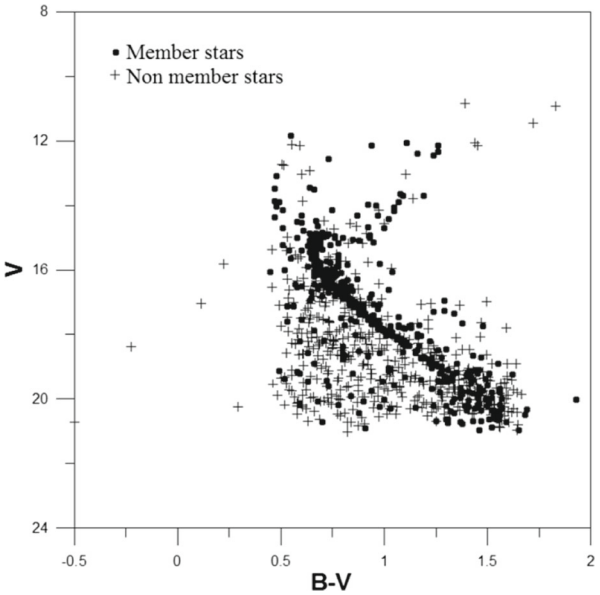

Fig. 2 The CMD of the (•) member stars and (+) of nonmember stars for NGC 188

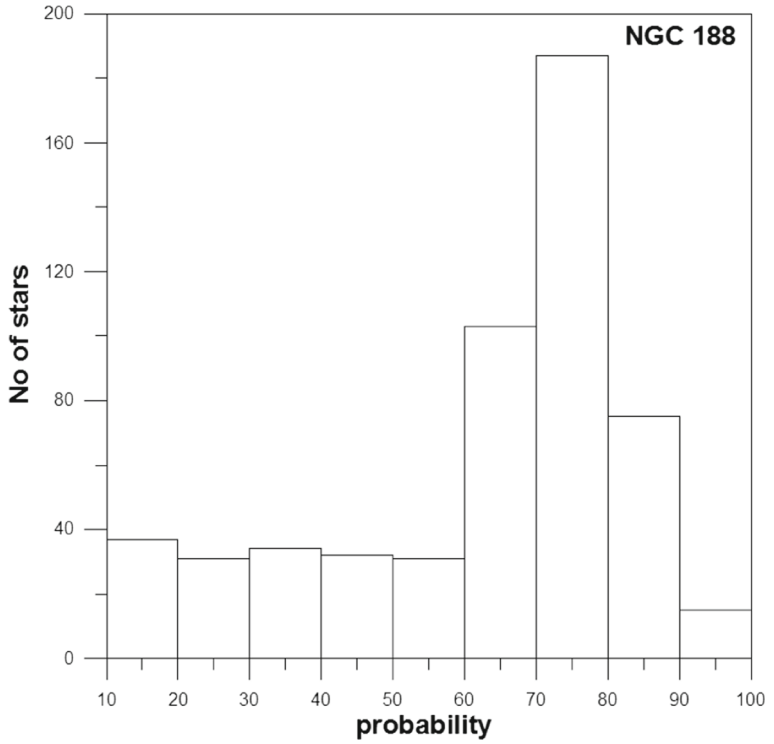

Fig. 3 histogram of membership probabilities stars in the region of NGC 188

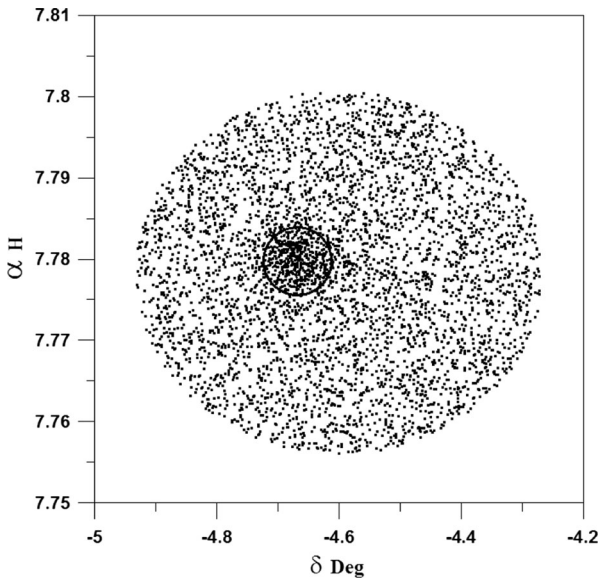

Fig. 4 NGC 2266 region with the radius 7.00 arcmin R.A in H and Dec. in Deg

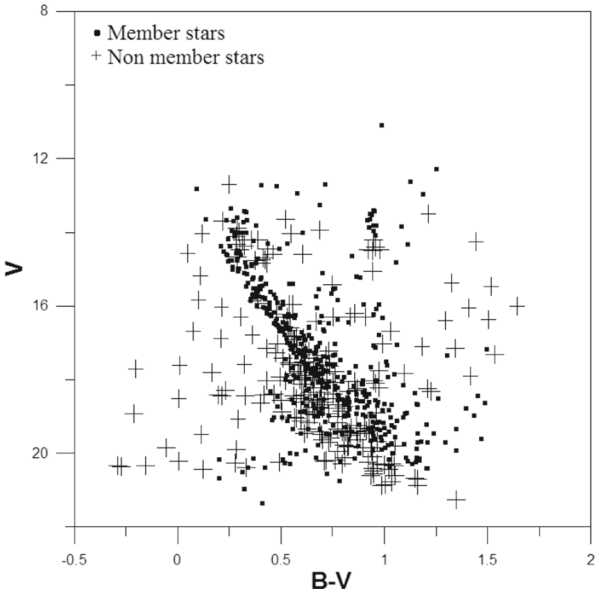

Fig. 5 The CMD of the (•) member stars and (+) of nonmember stars for NGC 2266

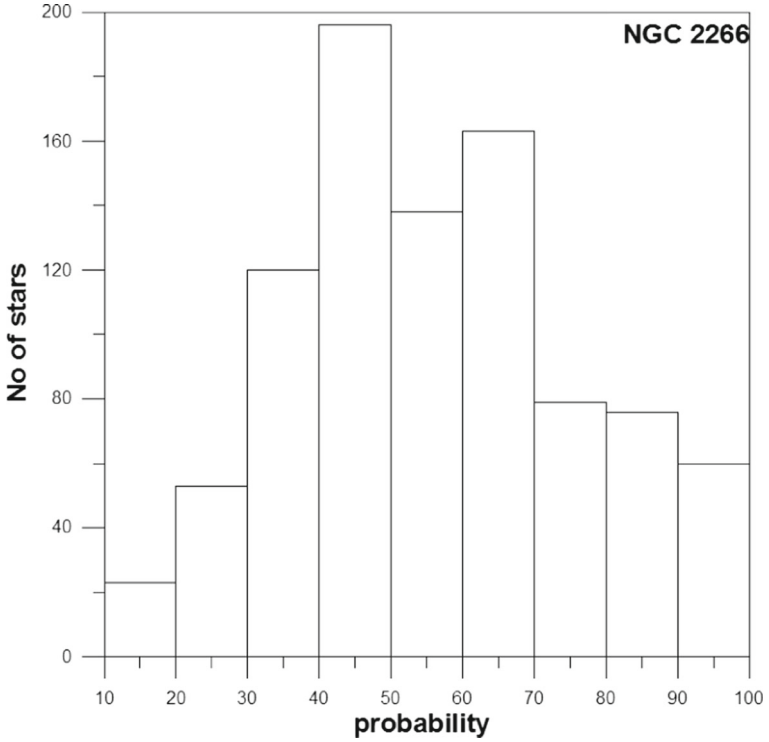

Fig. 6 Histogram of membership probabilities stars in the region of NGC 2266

Color Magnitude Diagram (CMD) of NGC 2266 has been plotted for the field stars (nonmember stars) and membership stars, the 908 member stars shows a well-defined main sequence see Fig. 5.

The difference between our algorithm results and the pervious methods is that, our algorithm does not require initial values of central coordinates of NGC 188, NGC 2266. Approximately all-stars member must be inside this circular of cluster region (area). The probability member of stars in NGC 2266 is shown in Fig. 6.

5 Conclusions

In this paper we presented a new method for open star cluster membership determinations solution. This method allows us to find membership probability based on K-means algorithm. The method and model proposed have proved to be efficient to discriminate between the two populations (cluster and field stars) in the regions of cluster. The method includes functions to perform structure of open star cluster (center and radius), and integrated color range estimations statistically for star membership. To demonstrate the method quality and test how well it handles real clusters, we applied it on two open star clusters NGC 188 and NGC 2266 based on CCD observation and data collected from WEBDA. This method is able to handle large databases both objectively and automatically. We obtained the cluster center and cluster radius, and some parameter have been calculated e.g. the range of the color and membership using available UBV observations, and 2MASS data. The membership probabilities of individual stars in the clusters region have been calculate, shown as a histogram of membership probabilities of stars shown in Figs. 3 and 6. It can be seen that our membership determination is quite effective for determination of the membership probabilities of clusters, found underlines the suitability of the method. Determination the membership probability of the clusters stars involves a careful assessment of the distributions of space positions. We found the cluster members approximately follow a Gaussian profile this results is agreement with the result of Wen et al. [32]. The resulting values were compared with studies using the photometric system or other method when available as well as the resulting cluster center, cluster readies and membership estimates showed very good compared with other studies. The final estimated number of cluster members obtained, which means that they are dependent on the completeness level cluster region and physical properties of cluster. A good result of k-means algorithm confirms that the convergence rate of the different discipline interesting i.e. for determining the cluster center, cluster radius and membership probabilities of star cluster as a compare with paper ([2]; Bonatto et al. [25], [13]).

The method is the best and a first one in the subject of automatization and standardization of membership probabilities study of open star cluster, it is does not require initial values of central coordinates. The method shows that the k-means algorithm can converge the best solution, and it has a high convergence rate and high accuracy. The proposed algorithm has been provided the better and clear way to determine the cluster center, cluster radius and cluster membership. In the future work we apply this method for other open star clusters in different latitude.

References

1. Essam, A., Selim, I. M.: IJAA, **5**, 173–181 (2015)
2. Selim, I.M., Haroon, A.A., Ismail, H.A., Ahmed, N.M., Essam, A., Ali, G.B.: Roman. Astron. J. **24**(2), 159–168 (2014)
3. Wu, Z.-Y., Zhou, X., Ma, J., Jiang, Z.-J., Chen, J.-S.: Publ. Astron. Soc. Pac. **118**, 1104–1111 (2006)
4. Vasilevskis, S., Klemola, A., Preston, G.: A.J. **5**, 173 (1958)
5. Sanders, W.L.: A&A **15**, 368 (1971)
6. Slovak, M.H.: A.J. **82**, 818 (1977)
7. Cabrera-Cano, J., Alfaro, E.J.: A&A **150**, 298 (1985)
8. Mighell, KJ., Rich, R.M., Shara, M., Fall, S.M.: A.J. **111**, 2314 (1996)
9. Kerber, F., Guglielmetti, F., Mignani, R., Roth, M.: A&A **381L**, 9 (2002)
10. Wu, ZY., Wang, JJ., Chen, L.: ChJAA **2**, 216 (2002)
11. Kerber, L.O.,Santiago, B.X.: A&A **435**, 77 (2005)
12. Gao, X-H.: RAA **14**, 159 (2014)
13. Gao, X-H., Chen, L., Hou, Z-j.: ChA&A..38..257G (2014)
14. Javakhishvili, Kukhianidze, V., Todua, M., Inasaridze, R.: A&A **447**, 915–919 (2006)
15. Bonatto, C., Bica, E.: MNRAS **377**, 1301 (2007)
16. Monteiro, H., Dias, W.S., Caetano, T.C.: A&A **516A,** 2 (2010)
17. Maia, L.F., Fernandes, R.F.,; Lobo-Hajdu, G., de Oliveira, L.F.C.: RSPTA **37240**, 200 (2014)
18. Chaturvedi, J.C.A., Green, P.: J. Classification **18**, 35–55 (2001)
19. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y.: IEEE, 24, no. 7 (2002)
20. Amir, B.-D., Shamir, R., Yakini, Z.: J. Comput. Biol. **6**(3/4), 281–297 (1999)
21. Jiawei Han, M.K: An Imprint of Elsevier. (2006)
22. Scott, D.W.: Biometrika **66**, 605–610 (1979)
23. Dodonov, Y.S., Dodonova, Y.A.: Psikhologicheskie Issledovaniya. 5(19) (2011)
24. Alaa, A, H.A., Alsolami, I.Z.: Astrophys Space Sci. 357–21 (2015)
25. Bonatto, C., Bica, E., Santos, J.F.C., Jr.: A&A **433**, 917 (2005)
26. Meibom, Søren; Grundahl, Frank; Clausen, Jens Viggo; Mathieu, Robert D.; Frandsen, Søren; Pigulski, Andrzej; Narwid, Artur; Steslicki, Marek; Lefever, Karolien.: A.J. **137**, 5086 (2009)
27. van den Bergh, S.: Z.A. **46**, 176 (1958)
28. Sandage, A.: ApJ **135**, 349 (1962)
29. Fornal, B., Tucker, D.L., Smith, J.A., Allam, S.S., Rider, C.J., Sung: A. J. **133**, 1409 (2007)
30. Platais, I., Platais, V.K., Mathieu, R.D., Girard, T.M., van Altena, W.F.: A. J. **126**, 2922 (2003)
31. Upgren, A. R., Mesrobian, W.S., Kerridge, S.J.: A.J. **77**, 74 (1972)
32. Wen, W., Zhao, J.-L., Chen, L.: Chin. J. Astron. Astrophys. **30**, 274 (2006)