Scientific Research Publishing

International Journal of Astronomy and Astrophysics, 2020, 10, 1-10
https://www.scirp.org/journal/ijaa
ISSN Online: 2161-4725
ISSN Print: 2161-4717

Virgo Cluster Membership Based on *K*-Means Algorithm

Ibrahim Mohamed Selim[1,2], Passent Elkafrawy[3], Walid Dabour[3,4], Mohamed Eassa[2]

[1]National Research Institute of Astronomy and Geophysics, Cairo, Egypt
[2]Computer Science Department, Integrated Thebes Institute, Cairo, Egypt
[3]Math and Computer Science Department, Faculty of Science, Menoufia University, Shibin El Kom, Egypt
[4]Computer Science Department, Taibah University, Alula Branch, Medina, KSA
Email: i_selim@yahoo.com

How to cite this paper: Selim, I.M., Elkafrawy, P., Dabour, W. and Eassa, M. (2020) Virgo Cluster Membership Based on *K*-Means Algorithm. *International Journal of Astronomy and Astrophysics*, **10**, 1-10.
https://doi.org/10.4236/ijaa.2020.101001

Received: July 12, 2019
Accepted: February 3, 2020
Published: February 6, 2020

Abstract

The Virgo cluster of galaxies is of great importance to study the development of the universe due to its close distance from the earth as well as being the center of the local super cluster. The problem that faces Virgo cluster studies is that it shares the same right ascension (RA) and Declination (DEC) ranges with large number of background as well as foreground galaxies. This study aims to geometrically and statistically estimate Virgo cluster membership. The study employs Virgo cluster data, prepared by Harvard University. The radial velocity (RV) data of the Virgo cluster were treated and employed in exchange of missing galaxies' third dimension, taking advantage of their proportionality. The data were treated by *K*-means algorithm, using Matlab 2014, and visual and logical exclusion of extremity galaxies to determine the rational center of the Virgo galaxies cluster. Results were presented, compared and discussed. Finally distances of galaxies from the Virgo cluster center were employed along with normal probability distribution characteristics to identify the most probable Virgo cluster members from the range of Virgo cluster of galaxies. The results showed that out of 17,466 objects surveyed in Virgo galaxy range, only few of galaxies were estimated to be genuine Virgo members.

Keywords

Virgo Cluster, *K*-Means, Normal Distribution

1. Introduction

The study of the Virgo cluster is important to explore the development stages of the universe. This cluster is the nearest large high-density cluster to earth, at

about 19 Megaparsec (Mpc) [1]. It is easier to monitor with higher accuracy, which makes it an ideal laboratory for testing hypothesis of structures formation of the universe [2]. The Virgo cluster is somewhat an irregular cluster with concentration of galaxies at the center. In three-dimensional (3D) space, the Virgo cluster constitutes the nucleus of the Local Super Cluster (LSC) of galaxies [3]. Its location in the crowded center of our local super cluster causes it to share its right ascension (RA) and Declination (DEC) ranges with large number of unrelated background as well as foreground galaxies [4]. These galaxies would highly influence investigations of Virgo cluster structure [5].

Optimization techniques relay primarily on arranging states of candidate members based on chosen optimization characteristics, within the search space. Once organized states of all members are reached, selection of the optimal solution is a straight forward process [6]. Different state identification techniques are employed for different optimization problems. For cluster memberships assessment, a number of different optimization techniques were employed with varying degrees of complexity and accuracy, such as Hierarchical Clustering Techniques [7], expectation-maximization (EM) algorithm [8], density based clustering algorithm [9] and Cuckoo Search Algorithm [10]. *K*-means represent one of the most promising optimization techniques, being employed for geometrical galaxy membership identification [11].

K-means is used to group galaxies into a given K number of galaxy clusters as well as identify the clusters' centers. This is done through optimization of Euclidian distances between galaxies and identified clusters' centers [12]. Distances are used to evaluate most probable cluster membership of galaxies.

Another important optimization technique is the normal distribution characteristics [13]. This technique could be employed on Euclidian distances between galaxies' locations and their cluster center. In this technique, the probability of each galaxy membership to the cluster is identified and a threshold of probability is employed to differentiate between the cluster members and field galaxies.

The current study attempts to differentiate between Virgo cluster galaxies and the unrelated field galaxies using 2D *K*-means and 3D *K*-means as well as other optimization techniques. This process was performed using Matlab® 2014 on Windows 10 operating system and results were plotted using Grapher©. The proposed technique was verified with other reviewed studies.

2. Related Work

Recently, computational techniques coupled with greater computational powers developed astronomical capacity leading to better understanding of the universe. Studying different stellar objects and their development process leads to better understanding of the past, present and future of our planet, solar system and galaxy.

In [14], authors presented an automated stellar cluster analysis tool that employs the standard tests on stellar clusters to determine their basic parameters. This tool has a set of functions that are used to obtain precise and objective val-

ues for a given cluster's center, radius, luminosity and integrated color magnitude.

In [15], a data clustering technique is proposed. It is a clustering technique based on descriptive data analysis and it can be employed to uncover the structure of multivariate data sets. It depends on *K*-means clustering technique. *K*-means clustering is a center defining model as each cluster is represented by one vector. It was developed by Macqueen [16], as a tool to give researchers qualitative and quantitative insights into multivariate large data sets. *K*-means clustering proved to be useful in data mining and investigative data analysis and is used to provide unique and definitive means of data grouping. Large data sets are the result of advances in information technology and growth of computational powers. Popularity of *K*-means over other clustering techniques stems from its ease of implementation, low memory requirements and computational efficiency.

In [17], *K*-means clustering was employed to develop a new method for open cluster membership determination. The developed algorithm allows efficient discrimination between cluster members and field stars. The results showed that the developed algorithm has the capacity to evaluate stars membership probability without assumptions regarding stars spatial distribution in cluster or field.

Researchers in [18] investigated membership of Virgo cluster, as the closest and consequently most studied cluster of galaxies. They employed classical methods to evaluate membership probability of Virgo cluster. Virgo has a problem that stems from sharing its field with a large number of non-Virgo galaxies and stars. There are many reasons favoring studies of Virgo such as its nearness, which causes it to be a main candidate for surveys. Also, Virgo exhibits a full range of galaxy luminosities and morphological types. Faint Virgo galaxies proved difficult to separate from overwhelming number of background galaxies. This was specially an issue of older literature. Investigations of galaxies are constantly biased in favor of massive and more luminous galaxies. This trend has many reasons such as under representation of low luminous galaxies in most galaxies catalogs due to the larger volume over which more luminous galaxies can be sampled.

3. The Proposed Method

3.1. *K*-Means

The current research employed 2D *K*-means and 3D *K*-means as the tools to estimate the Virgo cluster center. In this section 2D *K*-means algorithm and 3D *K*-means algorithm of clustering probability are presented. These algorithms deal with set of n members coordinates $\{x_1, x_2, \cdots, x_n\}$, each is a d-dimensional vector. *K*-means process is carried out in two steps [19].

In the first step, the members are divided into, given K groups, $G = \{G_1, G_2, \cdots, G_K\}$, by minimizing the mean squared distance between members and predicted group's center [20] as follows:

$$f(C,G)=\sum_{k=1}^{K}\sum_{x_i\in G_k}\left\|x_i-c_k\right\|^2 \quad (1)$$

In second step, the group center is relocated to be situated at the arithmetic mean of the group's members locations [20] as follows:

$$c_k=1/\left|G_k\right|\sum_{x_i\in G_k}x_i \quad (2)$$

Finally, first step and second step are repeated until no more changes in group members are observed.

The 2D and 3D *K*-means are illustrated in Algorithm 1 and Algorithm 2.

3.1.1. 2D K-Means Algorithm

Algorithm 1.

Input $X=\{x_1,x_2,\cdots,x_n\}$ //set of *n* data items.

K // Number of desired clusters

Output:

A set of *K* clusters, c_k : The center of each cluster and

Steps:

Repeat

- Calculate the distance between each data point and cluster centers using:

$$d=\sqrt{\left(x_i-c_x\right)^2+\left(y_i-c_y\right)^2}$$

- 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:

$$c_k=1/\left|G_k\right|\sum_{x_i\in G_k}x_i$$

Until the centers don't change.

3.1.2. 3D *K*-Means Algorithm

Algorithm 2.

Input $X=\{x_1,x_2,\cdots,x_n\}$ //set of n data items.

K // Number of desired clusters

Output:

A set of *K* clusters, c_k : The center of each cluster and

Steps:

Repeat

- Calculate the distance between each data point and cluster centers using:

$$d=\sqrt{\left(x_i-c_x\right)^2+\left(y_i-c_y\right)^2+\left(z_i-c_z\right)^2}$$

- 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:

$$c_k=1/\left|G_k\right|\sum_{x_i\in G_k}x_i$$

Until the centers don't change.

The proposed method consists of two steps. The first, is to employ 2D *K*-means and 3D *K*-means to evaluate the most probable Virgo cluster center,

treated as one group ($K = 1$), and consequently evaluate the distance of each galaxy to the Virgo cluster center. The second step, is to employ the RA and DEC positions with normal probability distribution to derive the membership probability of Virgo cluster galaxies. Finally a set value of standard deviation is employed to eliminate the most improbable galaxies and identify the most probable Virgo cluster galaxies.

3.2. Normal Probability Distribution

Normal probability distribution characteristics evaluate the probability that an individual data point is a member of a given data set [21]. For this purpose the standard normal random variable Z is evaluated for this data point according to the following equation [22]:

$$Z = \left(x - \overline{x}\right)/\sigma \tag{3}$$

Finally the probability is obtained from the following equation [13]:

$$\text{Probability} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

where:

x is the investigated data point.

$\overline{x}$ is the arithmetic mean of the data set under consideration.

σ is the standard deviation of the data set under consideration.

The data set arithmetic mean is calculated as follows:

$$\overline{x} = \frac{\sum_{i=1}^{N}\left(x_i\right)}{N} \tag{5}$$

Standard deviation is a measure that quantifies the amount of variation of a set of data values. Lower standard deviation indicates that data points are gathered closer to the data set mean (center of Virgo cluster of galaxies), while higher standard deviation indicates that data is spread further away from the data set mean [23]. Standard deviation is calculated as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2}{N-1}} \tag{6}$$

The perfect normally distributed data would have equal mean, median and mode values. It would also be asymptotic and have its probability distribution symmetrical and centered on vertical axis at the data mean value. The values of mean and (mean + σ) would enclose 34.1% of all data elements. Also, the values of (mean + σ), (mean + 2σ) and (mean + 3σ) would enclose 13.6% and 2.1% of all data elements respectively [24].

4. Experimental Result

4.1. Virgo Cluster Data

The raw Virgo clusters of galaxies data have been used in the current study are

from Virgo subset of 2 Micron All-Sky Survey catalogs (2MASS). The data is delivered by Center for Astrophysics (CFA) of Harvard university as reported in year 2007 [25]. The raw Virgo data clusters of galaxies contain all the galaxies within RA from 11.0 to 14.0 hours and DEC from −10 to + 35 degrees. This roughly rectangular region is centered around 12.5 hours and +12 degrees, where Harvard believes that Virgo cluster center is located. 2MASS's RA is measured in hours, minutes and seconds format, running from 0 to 24 hours. DEC is measured in degrees, minutes and seconds format running from −90 to +90 degrees. RV, heliocentric radial velocity, of the different Virgo galaxies is employed in exchange of missing galaxies' third dimension, taking advantage of their proportionality [26].

4.2. Result

The first employed methodology is based on 2D *K*-means. The locations of the 17,466 galaxies of Virgo Cluster raw data, namely, RA and DEC were fed to 2D *K*-means algorithm. The number of groups, *K*, was chosen as one group. The main results indicated that center of the Virgo Cluster raw data is at RA and DEC of 12.4 and 10.4 respectively. According to Binggeli and Huchra [27] the center of Virgo cluster is located at 12.4 RA and 10.4 DEC, while eSky [28] defines the center at 12.4 RA and 12.4 DEC. The data was plotted to visually assess the results, as shown in **Figure 1**. The squared shape of the plotted data indicated and further stressed the Harvard suggestion that the data includes background and foreground objects that do not belong to Virgo cluster galaxies.

To further investigate the location of data center, the data of RA and DEC along with RV values, representing the third dimension, and *K* value of 1 were fed to the 3D *K*-means. To study the distribution of the raw data and their probable belonging to the Virgo cluster, the RA and DEC data with RV were plotted in 3D.

The second step, normal distribution parameters of the developed RA and DEC positions of galaxies were used to drive probabilities of each galaxy membership likelihood to the Virgo cluster. Finally, a threshold of σ that supposed to enclose coefficient of determination, R-squared = 0.0351475 Residual mean square, sigma-hat-squared = 34.2374 is employed the data set. This galaxies data set would represent the most probable Virgo cluster members, **Figure 2**. With the center at RA, DEC and RV of 12.4, 10.4 and 425 respectively show that there are galaxies that are scattered far away from the center of data and accordingly, they are most probably not members of the Virgo cluster.

The running time is 1.5 seconds on a DELL laptop employing 64 Bit core i7-36120QM processor and 8GB RAM. The laptop is running Windows 10 operating system and the code is written in Matlab® 2014.

The plotted data indicated the most probable Virgo cluster galaxies in all 2D data of 2Mass catalog, as shown as **Figure 3**. The circle in **Figure 3** indicates the most probable range of Virgo cluster members.

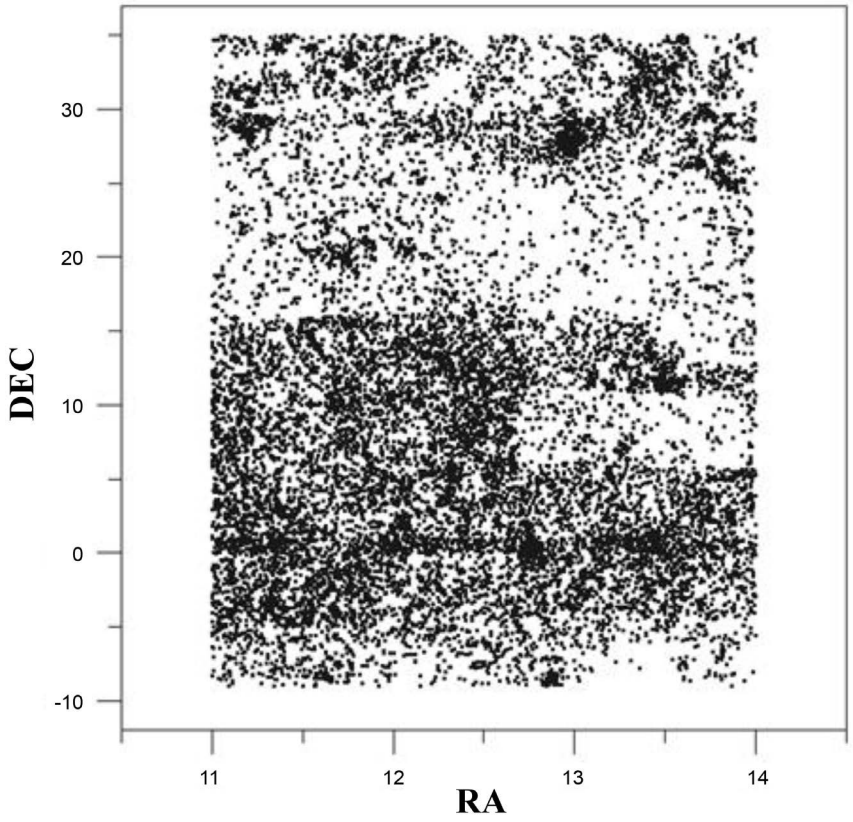

Figure 1. The raw data of Virgo cluster in 2D.

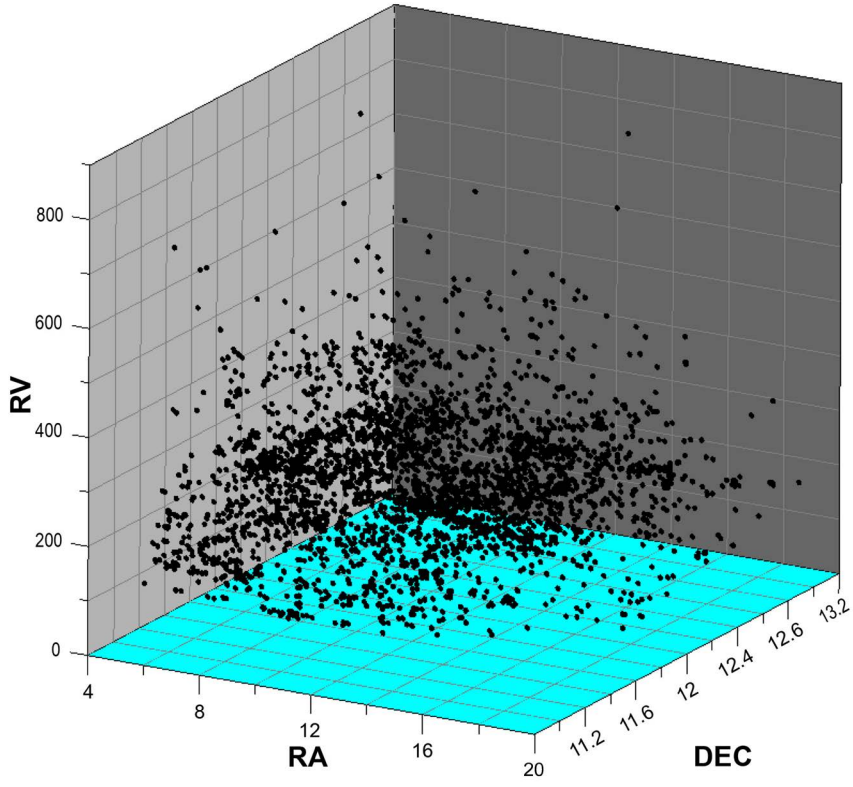

Figure 2. The Proposed Virgo cluster data in 3D.

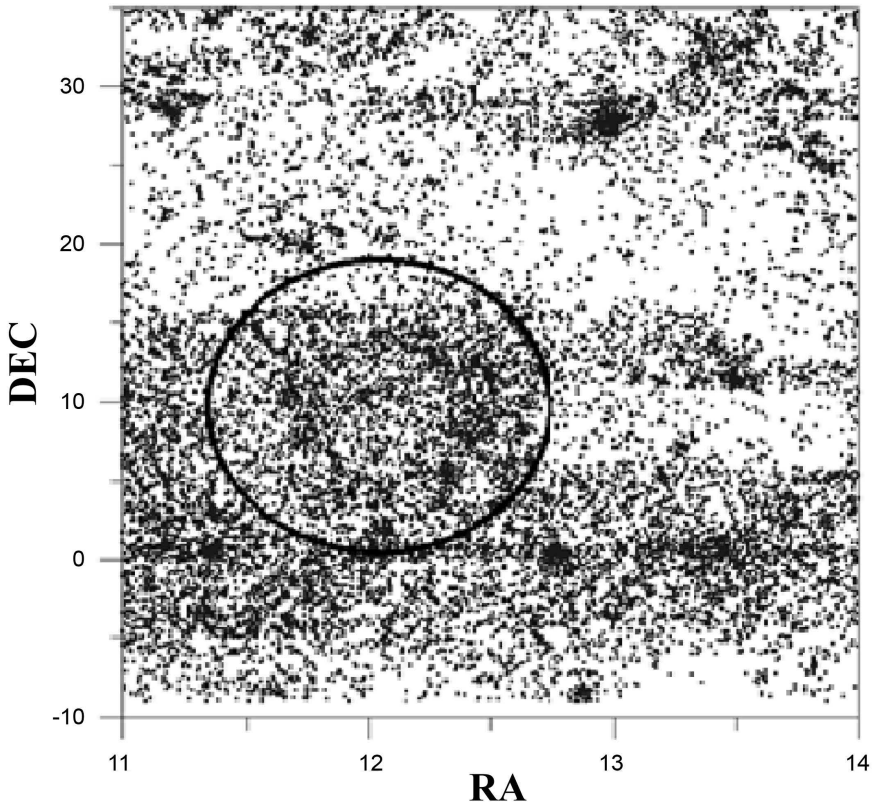

Figure 3. The most probable Virgo cluster range of galaxies member in 2Mass data.

5. Conclusion

In this paper we presented a new method for galaxies cluster membership determinations based on *K*-means algorithm. The application of the 3D *K*-means identification along with normal distribution characteristic was employed to identify the most probable galaxies of Virgo cluster as well as the Virgo cluster center. To demonstrate the method quality and to test how well it handles real clusters the proposed technique was employed to raw Virgo cluster data in 2MASS as supplied by Harvard. Probability results showed that Virgo cluster contains 1300 galaxies in 2MASS data. The above mentioned results support us to use the proposed algorithm to get a better and clearer way to determine the membership of groups. This method has the ability to handle large datasets both objectively and automatically. The method includes functions to identify structure of the cluster. The developed technique proved its ability to successfully perform global searches to solve the set problem.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mei, S., *et al.* (2008) The ACS Virgo Cluster Survey. XIII. SBF Distance Catalog and

the Three-Dimensional Structure of the Virgo Cluster. *The Astrophysical Journal*, **655**, 144. https://doi.org/10.1086/509598

[2] Ouellette, N.N.Q., Courteau, S., Holtzman, J.A., Dalcanton, J.J., McDonald, M. and Zhu, Y. (2014) The Dynamical Properties of Virgo Cluster Disk Galaxies. *Structure and Dynamics of Disk Galaxies*, **480**, 89.

[3] Maccone, C. (2012) SETI among Galaxies by Virtue of Black Holes. *Acta Astronautica*, **78**, 109-120. https://doi.org/10.1016/j.actaastro.2011.10.011

[4] Zhu, X. and Chu, Y. (1995) Association of Quasars and Galaxies in the Field of the Virgo Cluster. *Chinese Astronomy and Astrophysics*, **19**, 129-136. https://doi.org/10.1016/0275-1062(95)00018-N

[5] Andreo, R.B. (2015) Virgo Cluster Galaxies. NASA.

[6] Yang, X.-S. and Deb, S. (2014) Cuckoo Search: Recent Advances and Applications. *Neural Computing and Applications*, **24**, 169-174. https://doi.org/10.1007/s00521-013-1367-1

[7] Kaushik, M. and Mathur, B. (2014) Comparative Study of K-Means and Hierarchical Clustering Techniques. *International Journal of Software & Hardware Research in Engineering*, **2**, 93-98.

[8] Svensson, C.-M., Bondoc, K.G., Pohnert, G. and Figge, M.T. (2017) Segmentation of Clusters by Template Rotation Expectation Maximization. *Computer Vision and Image Understanding*, **154**, 64-72. https://doi.org/10.1016/j.cviu.2016.08.003

[9] Dong, S., Liu, J., Liu, Y., Zeng, L., Xu, C. and Zhou, T. (2018) Clustering Based on Grid and Local Density with Priority-Based Expansion for Multi-Density Data. *Information Sciences*, **468**, 103-116. https://doi.org/10.1016/j.ins.2018.08.018

[10] Abdel-Baset, M., Selim, I.M. and Hezam, I.M. (2015) Cuckoo Search Algorithm for Stellar Population Analysis of Galaxies. *International Journal of Information Technology and Computer Science*, **7**, 29-33. https://doi.org/10.5815/ijitcs.2015.11.04

[11] Tang, R. and Fong, S. (2018) Clustering Big IoT Data by Metaheuristic Optimized Mini-Batch and Parallel Partition-Based DGC in Hadoop. *Future Generation Computer Systems*, **86**, 1395-1412. https://doi.org/10.1016/j.future.2018.03.006

[12] Jaroš, M., *et al.* (2017) Implementation of K-Means Segmentation Algorithm on Intel Xeon Phi and GPU: Application in Medical Imaging. *Advances in Engineering Software*, **103**, 21-28. https://doi.org/10.1016/j.advengsoft.2016.05.008

[13] Nadarajah, S. (2005) A Generalized Normal Distribution. *Journal of Applied Statistics*, **32**, 685-694. https://doi.org/10.1080/02664760500079464

[14] Perren, G.I., Vázquez, R.A. and Piatti, A.E. (2015) ASteCA—Automated Stellar Cluster Analysis. *Astronomy and Astrophysics*, **576**, 29. https://doi.org/10.1051/0004-6361/201424946

[15] Morissette, L. and Chartier, S. (2013) The K-Means Clustering Technique: General Considerations and Implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, **9**, 15-24. https://doi.org/10.20982/tqmp.09.1.p015

[16] Macqueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. Western Management Science Institute Working Paper, 281-297.

[17] Elsayed Abd El Aziz, M., Selim, I. and Essam, A. (2016) Open Cluster Membership Probability Based on K-Means Clustering Algorithm. *Experimental Astronomy*, **42**, 49–59. https://doi.org/10.1007/s10686-016-9499-9

[18] Binggeli, B., Sandage, A. and Tammann, G.A. (1985) Studios of Virgo Cluster. II. A Catalog of 2096 Galaxies in the Virgo Cluster Area. *The Astronomical Journal*, **90**, 1681-1758. https://doi.org/10.1086/113874

[19] Chaturvedi, A., Green, P.E. and Caroll, J.D. (2001) K-Modes Clustering. *Journal of Classification*, **18**, 35-55. https://doi.org/10.1007/s00357-001-0004-3

[20] Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R. and Wu, A.Y. (2002) An Efficient *K*-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**, 881-892. https://doi.org/10.1109/TPAMI.2002.1017616

[21] Wei, Y.-C., *et al.* (2007) The Influences of the Velocity Distribution and the Thickness Effect of Galactic Disk on the Z-Distribution of Normal Pulsars. *Chinese Astronomy and Astrophysics*, **31**, 11-20. https://doi.org/10.1016/j.chinastron.2007.01.008

[22] LaMorte, W.W. (2016) The Standard Normal Distribution. The Role of Probability, School of Public Health, Boston University, Boston, MA.

[23] Shi, B.-Q., Liang, J. and Liu, Q. (2011) Adaptive Simplification of Point Cloud Using *K*-Means Clustering. *Computer-Aided Design*, **43**, 910-922. https://doi.org/10.1016/j.cad.2011.04.001

[24] Taillon, J. (2013) Gaussian Shift (Mean Shift) Clustering and Variance Approximation. Williams College, Williamstown, MA.

[25] Huchra, J.P. (2007) The Virgo Cluster. Center for Astrophysics, Harvard University, Cambridge, MA.

[26] Zhao, J.-L., Pan, R.-S., Huang, S.-N. and He, Y.-P. (1991) The radial velocity membership of the Virgo Cluster. *Chinese Astronomy and Astrophysics*, **15**, 95-102. https://doi.org/10.1016/0275-1062(91)90015-P

[27] Binggeli, B. and Huchra, J. (2006) Virgo Cluster. IOP Publishing Ltd., London.

[28] http://www.glyphweb.com/esky/concepts/virgocluster.html