An intuitive clustering algorithm for spherical data with application to extrasolar planets

Wen-Liang Hung\textsuperscript{a}, Shou-Jen Chang-Chien\textsuperscript{b} & Miin-Shen Yang\textsuperscript{b}

\textsuperscript{a} Department of Applied Mathematics, National Hsinchu University of Education, Hsinchu, Taiwan
\textsuperscript{b} Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li, Taiwan, Taiwan

Published online: 18 Mar 2015.


To link to this article: http://dx.doi.org/10.1080/02664763.2015.1023271
An intuitive clustering algorithm for spherical data with application to extrasolar planets

Wen-Liang Hung\(^a\)*, Shou-Jen Chang-Chien\(^b\) and Miin-Shen Yang\(^b\)

\(^a\)Department of Applied Mathematics, National Hsinchu University of Education, Hsinchu, Taiwan
\(^b\)Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li, Taiwan, Taiwan

(Received 26 June 2014; accepted 23 February 2015)

This paper proposes an intuitive clustering algorithm capable of automatically self-organizing data groups based on the original data structure. Comparisons between the proposed algorithm and EM [1] and spherical \(k\)-means [7] algorithms are given. These numerical results show the effectiveness of the proposed algorithm, using the correct classification rate and the adjusted Rand index as evaluation criteria [5,6]. In 1995, Mayor and Queloz announced the detection of the first extrasolar planet (exoplanet) around a Sun-like star. Since then, observational efforts of astronomers have led to the detection of more than 1000 exoplanets. These discoveries may provide important information for understanding the formation and evolution of planetary systems. The proposed clustering algorithm is therefore used to study the data gathered on exoplanets. Two main implications are also suggested: (1) there are three major clusters, which correspond to the exoplanets in the regimes of disc, ongoing tidal and tidal interactions, respectively, and (2) the stellar metallicity does not play a key role in exoplanet migration.

Keywords: EM algorithm; extrasolar planets; mixtures of von mises distributions; spherical data; spherical \(k\)-means algorithm

1. Introduction

Clustering is a powerful exploratory approach to grouping data and to revealing structural information in data. It is a data-driven procedure that classifies a datum in one of several classes by looking at proximity and homogeneity in a feature space. Conventional approaches to clustering can be categorized broadly as hierarchical or nonhierarchical. The mixture likelihood approach to clustering, which is nonhierarchical, has the potential to handle structured data because it is model based. However, traditional clustering algorithms experience difficulty clustering high-dimensional data. To solve this problem, Dhillon and Modha [7] proposed the spherical \(k\)-means (spkmeans) algorithm that implemented the \(k\)-means algorithm using cosine similarity instead of Euclidean distortion. Recently, Banerjee et al. [1] proposed two variants of EM algorithms

\(*\)Corresponding author. Email: wlhung@mail.nhcue.edu.tw

© 2015 Taylor & Francis
for estimating the mean and concentration parameters of the mixture high-dimensional von Mises–Fisher distribution [14]. Empirical results of high-dimensional text clustering and gene-expression data show that the proposed algorithms perform well in handling high-dimensional data. However, a number of issues remain to be resolved in these approaches. For example, the spkmeans algorithm usually requires an initial partition to start the iterative process, and the number of clusters must be given a priori. In addition, this type of algorithms suffers from the problem of being trapped within local minima (or maxima), which is a result of a poor selection of initial partitions. On the other hand, there are shortcomings connected with the EM algorithm. These include slow convergence, the need for a suitable stopping rule that can detect whether the algorithm has reached the maximum, and the choice of initial values in order to reach the global maximum in fewer iterations.

This paper proposes a simple and intuitive clustering algorithm, which is different from spkmeans and EM algorithms, to improve clustering performance. The proposed algorithm simulates the process of how the data points perform self-clustering based on the data structure. After the convergence of the process, every data point reaches an equilibrium position without further movements. Data points that arrive at the same position are considered to belong to the same cluster. Furthermore, numerical results show that the performance of the proposed algorithm is better than those of spkmeans and EM algorithms, using the correct classification rate (cRate) and the adjusted Rand index (aRan) as evaluation criteria [5,6].

To date, more than 1000 exoplanets have been detected. These discoveries form part of a new page in astronomy, which could eventually answer questions regarding the formation and evolution of planetary systems, including our Solar system. With cluster analysis, which is a data analysis tool for finding clusters of a data set with the most similarity in the same cluster and the most dissimilarity between different clusters, it might also be possible to better understand the formation and evolution of planetary systems. Therefore, the proposed clustering algorithm is used to gain insight into the data of exoplanets.

The remainder of the paper is organized as follows. In Section 2, the proposed algorithm is introduced in full detail. Section 3 presents simulation results based on the indices of cRate and aRan [5,6]. Real data applications are given in Section 4. Finally, conclusions are given in Section 5.

2. Methodology

2.1 The proposed clustering algorithm

Let \( \mathbb{R} \) denote the set of real numbers, \( S^{p-1}(p \geq 2) \) denote the \((p - 1)\)-dimensional sphere embedded in \( \mathbb{R}^p \), and the norm \( || \cdot || \) denote the \( L_2 \) norm. A \( p \)-dimensional unit random vector \( X \) (i.e. \( x \in \mathbb{R}^p \) and \( ||x|| = 1 \), or equivalently \( x \in S^{p-1} \)) is said to have \( p \)-dimensional Langevin distribution if its probability density function (pdf) is defined as

\[
 f(x; a, \kappa) = C_p(\kappa) \exp(\kappa a \cdot x), \quad \kappa \geq 0, \quad ||a|| = 1,
\]

and

\[
 C_p(\kappa) = \frac{\kappa^{p/2-1}}{(2\pi)^{p/2}I_{p/2-1}(\kappa)},
\]

where \( I_p \) denotes the modified Bessel function of the first kind and order \( p \). The two parameters, \( a \) and \( \kappa \), represent the mean direction and the concentration, respectively. \( f(x; a, \kappa) \) is called the von Mises distribution when \( p = 2 \), and it is called the Fisher distribution when \( p = 3 \).
Let \( X = \{x_1, \ldots, x_n\} \) be a finite set of sample unit vectors drawn independently from a mixture of Langevin distribution with the pdf
\[
f(x; a, \kappa) = \sum_{i=1}^{c} \alpha_i f_i(x; a_i, \kappa_i),
\]
where \( a = (a_1, \ldots, a_c), \kappa = (\kappa_1, \ldots, \kappa_c) \) and \( \sum_{i=1}^{c} \alpha_i = 1, \, 0 \leq \alpha_i \leq 1 \). To sample a data point from this mixture pdf, the \( i \)th Langevin distribution is randomly chosen with probability \( \alpha_i \), and then a point (on \( S^{n-1} \)) is sampled following Equation (1). The aim is to group \( X \) into \( c \) clusters. For a given cluster number \( c \), Dhillon and Modha \cite{7} proposed the spkmeans algorithm, which adapts the \( k \)-means algorithm to normalized data by using the cosine similarity for cluster location, and also by re-normalizing the cluster means to unit length.

Banerjee et al. \cite{1} developed two variants of the EM methods to estimate the mean and concentration parameters of Equation (1). However, two main drawbacks of the EM algorithm are its slow convergence and the dependence of the solution on the initial value used. On the other hand, spkmeans also requires a set of initial values to begin the iterative process. To address these problems, this paper proposes another clustering algorithm inspired by Chen and Shiu \cite{4} and Hung et al.'s \cite{10} ideas, called the self-updating (SU) clustering algorithm, which can self-organize local optimal cluster numbers without using cluster validity functions.

The proposed SU algorithm is a simple iterative process that updates each data point to the weighted average of data points in the neighborhood of the target point. Obviously, it is possible to assign weights that smoothly reduce the distance from the target point (see Equation (2)). This location is attained through a weight function or kernel \( K_\lambda(x_0, x_j) \), which gives a weight \( x_j \) based on its distance from the target point \( x_0 \). The kernels \( K_\lambda \) are indexed by a parameter \( \lambda \) that indicates the width of the neighborhood. Thus, the SU clustering algorithm for spherical data can be summarized as follows.

**The SU clustering algorithm for spherical data**

S1. Initialize \( z^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)}) = (x_1, \ldots, x_n) \) and give \( \epsilon > 0 \); Set iteration counter \( t = 0 \);
S2. Every point is updated according to
\[
z_j^{(t+1)} = \frac{\sum_{j=1}^{n} K_\lambda(z_j^{(t)}, z_j^{(t)}) \cdot z_j^{(t)}}{\sum_{j=1}^{n} K_\lambda(z_j^{(t)}, z_j^{(t)})}, \quad j = 1, \ldots, n,
\]
where \( K_\lambda \) is a truncated Gauss kernel
\[
K_\lambda(z_j^{(t)}, z_j^{(t)}) = \begin{cases} 
\exp\left(-\frac{d}{\lambda}\right) & \text{if } d = d(z_j^{(t)}, z_j^{(t)}) \leq r; \\
0 & \text{if } d > r,
\end{cases}
\]
and \( r = (1/(n^2)) \sum_{j \neq j} d(x_j, x_j) \), i.e. \( r \) is the average point scatter, which is a constant given datum.
Increment \( t \); Until max \( d(z_j^{(t+1)}, z_j^{(t)}) < \epsilon \).
S3. Process agglomerative hierarchical clustering (AHC) with the single linkage for the final states of the data points.
S4. Find the optimal cluster number \( c^* \) according to the hierarchical clustering tree.
S5. Identify these \( c^* \) clusters.
In Equation (3), \(d\) is defined as the distance between the two data \(u, v \in \mathbb{S}^{n-1}\):

\[
d \equiv d(u, v) = 1 - u^T v.
\]

Clearly, \(d(u, v)\) satisfies (i) \(d(u, v) = 0 \iff u = v\); (ii) \(d(u, v) = d(v, u)\). However, \(d\) does not satisfy the triangle inequality. So, \(d\) is a semi-metric. This paper chooses \(\lambda = r / 5\) \([2,10]\). The following is a discussion of how the parameter \(\lambda\) determines the number of clusters in the data.

**Theorem 1**  In Equation (3), when \(\lambda \to 0\), each of the data points is its own cluster.

**Proof**  As \(\lambda \to 0\), then

\[
\lim_{\lambda \to 0} K_\lambda(z_j^{(t)}, z_j^{(t)}) = \lim_{r \to 0} K_{r/5}(z_j^{(t)}, z_j^{(t)}) = \begin{cases} 
1 & \text{if } z_j^{(t)} = z_j^{(t)}; \\
0 & \text{otherwise.}
\end{cases}
\]

This implies \(z_j^{(t+1)} = z_j^{(t)}\) in Equation (2). That is, each of the data points is its own cluster as \(\lambda \to 0\). \(\square\)

**Theorem 2**  In Equation (3), when \(\lambda \to \infty\), the data have only one cluster on the sample mean.

**Proof**  As \(\lambda \to \infty\), then

\[
\lim_{\lambda \to \infty} K_\lambda(z_j^{(t)}, z_j^{(t)}) = \lim_{r \to \infty} K_{r/5}(z_j^{(t)}, z_j^{(t)}) = 1.
\]

This implies, in Equation (2),

\[
z_j^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} z_j^{(t)},
\]

which indicates that the data have only one cluster on the sample mean when \(\lambda \to \infty\). \(\square\)

### 2.2 Convergence

At this point, it must be determined if the proposed algorithm converges. The methods of Chen and Shiu \([4]\) and Chen \([3]\) are thus used to study the convergence property of the proposed algorithm. The convergence of the proposed SU algorithm means that all the data points no longer move in Step S2. Furthermore, the convergence of the SU algorithm depends on the function \(K_\lambda\) in Step S2. Chen and Shiu \([4]\) and Chen \([3]\) proved that PDD (positive and decreasing with respect to distance) provides sufficient conditions to guarantee convergence.

**Definition 1**  A function \(f\) is PDD (positive and decreasing with respect to (w.r.t) distance), if

(i) \(0 \leq f(u, v) \leq 1\) and \(f(u, v) = 1\) only when \(u = v\).

(ii) \(f(u, v)\) depends only on \(d(u, v)\), the distance between \(u\) and \(v\).

(iii) \(f(u, v)\) is decreasing w.r.t. \(d(u, v)\).

The PDD conditions ensure that \(f\) has the following properties. Condition (i) describes the non-negativity of \(f\). In practice, \(f(u, v)\) may be negative, and data points may move further apart and may diverge in the updating process. This is why \(f\) is non-negative. In general, the value of \(f(u, v)\) can be any positive number. For simplicity, it is normalized to be 1. Condition (ii) means
that the influence between \( u \) and \( v \) is only determined by the distance \( d(u, v) \), i.e. \( f(u, v) = f(v, u) \) whenever \( d(u, v) = d(v, u) \). Since \( f(u, v) \) represents the influence that data points \( u, v \) receive from each other, a larger value of \( f(u, v) \) is assumed when \( u \) and \( v \) are closer, as described in condition (iii). The following Theorem 3 ensures the convergence of the proposed SU algorithm. The proof of Theorem 3 is essentially the same as in \[3,4\].

**Theorem 3** If the function \( K_\lambda \) in Equation (2) is PDD, there exists \( \{x_1, \ldots, x_n\} \) such that

\[
\lim_{t \to \infty} x_i^{(t)} = x_i \quad \forall i.
\]

Although Theorem 3 ensures the convergence of the proposed SU algorithm when \( K_\lambda \) has PDD conditions, there are some \( K_\lambda \) that produce a trivial clustering result, in which all data points are grouped into one single cluster. The following Corollary 1 identifies such cases.

**Corollary 1** Let \( d_M \) be the maximum pairwise distance between any two data points. If \( K_\lambda \) is PDD with \( K_\lambda(d_M) > 0 \) there exists \( c \) such that

\[
\lim_{t \to \infty} x_i^{(t)} = c \quad \forall i.
\]

In Equation (2), the assumption of Corollary 1 implies that \( d(z_j^{(t)}, z_j^{(t)}) \leq d_M \) for every \( t \). Since \( K_\lambda \) is decreasing w.r.t. distance,

\[ K_\lambda(z_j^{(t)}, z_j^{(t)}) \geq K_\lambda(d_M) > 0 \quad \text{for every } j \text{ and } j'. \tag{4} \]

Equation (4) implies that all data points converge to the same position, say \( c \).

To prevent the trivial clustering result, the function \( K_\lambda \) has to be zero on \((r, \infty)\) for some \( r < d_M \). Furthermore, to solve the initialization problem, all data points are set to be the initial centers (i.e. \( z^{(0)} = (z_1^{(0)}, \ldots, z_n^{(0)}) = (x_1, \ldots, x_n) \)) (see Step S1). The following examples demonstrate that the proposed algorithm first uses all data points as initial cluster centers and then self-organizes the data based on the original data structure.

### 3. Numerical examples and comparisons

In this section, the proposed SU algorithm is compared with soft-EM, hard-EM and spkmeans algorithms on the data generated by mixtures of Langevin distributions to demonstrate their clustering performance according to the criteria of cRate and aRan \[5,6\]. The aRan index is simply used to compare a priori partition with the partition obtained from the clustering algorithm. The value of aRan belongs to the interval \([0, 1]\), where the value 1 indicates perfect agreement between partitions. Larger cRate-values and cRan-values indicate higher cluster quality.

**Example 1** Figure 1 shows a 2-cluster data set consisting of 200 data points generated from a mixture of 3-variate mixture Langevin distributions

\[
0.45L_3\left(\left(\frac{\sqrt{6}}{4}, \frac{\sqrt{3}}{4}, \frac{\sqrt{2}}{2}\right), 12\right) + 0.55L_3\left(\left(\frac{1}{2}, \frac{\sqrt{3}}{2}, 0\right), 14\right). \tag{5}
\]

The proposed algorithm is implemented in this data set. The initial states of these data points are shown in Figure 2(a). The states of the data points after 2 and 4 iterations are shown in Figure 2(b) and 2(c), respectively. After 7 iterations, shown in Figure 2(d), the SU algorithm
Figure 1. Two-cluster data set from a mixture of 3-variate mixture Langevin distributions.

Figure 2. The states (shown as black) of the data points. Figure 2(a) represents the initial states and Figure 2(b)–(d) represents the states of the data points after 2, 4 iterations and convergence, respectively.

converges to the two cluster centers. Figure 2(d) indicates that all data points merge to two points which are cluster centers. This clustering result reveals the original data structure. The single linkage algorithm is also processed with the final convergence states of data points, and it gives insight into the optimal cluster number for the data set. Figure 3(a) shows the hierarchical tree of the final states’ data points. Obviously, the hierarchical tree shows that the optimal cluster number is 2, and Figure 3(b) presents the corresponding clustering result.

To evaluate the clustering performance of the proposed algorithm with other algorithms, 50 data sets are generated from the mixture of Langevin distribution in Equation (5). Then, the average values of cRate and aRan, shown in Table 1, are calculated based on these 50 data sets. From Table 1, the SU algorithm performs better than the soft-EM, hard-EM and spkmeans algorithms.
In the study on the exoplanets, many authors [8,17,18] have indicated that there is a possible correlation between the planetary mass and the orbit period. Therefore, it seems worth examining where groups for expolanets on these two important features are. The data are from the Extrasolar Planets Catalog compiled by Jean Schneider (http://cfa-www.harvard.edu/planets/catalog.html). Data as of 2014 December 14 are used, and incomplete data are excluded. There are 731 planets available for this study, and each of them has the values of projected mass ($M_p$), and orbital period ($P$). Following Jiang et al.'s [9] approach, the proposed SU algorithm is applied to the data ($\ln(P), \ln(M_p)$) on $S$. The SU algorithm converges to four cluster centers, as shown in Figure 7. Figure 8 shows the corresponding clustering result. Table 3 shows the corresponding centers and members. The fourth cluster only contains the exoplanet WASP-19 b. The characteristics of clusters PM1, PM2 and PM3 might associate

**Example 2**  In this example, a 3-cluster data set that contains one large cluster and two small clusters is considered. The 300 data points are generated from a mixture of 3-variate mixture Langevin distributions

$$0.8L_3\left(\left(0, \frac{1}{2}, \frac{\sqrt{3}}{2}\right), 12\right) + 0.1L_3\left(\left(\frac{1}{2}, \frac{\sqrt{3}}{2}, 0\right), 20\right) + 0.1L_3\left(\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}, 0\right), 20\right). \quad (6)$$

The proposed algorithm is applied to this data set, as shown in Figure 4. Figure 5(a)–(c) shows the states of data points after 1, 2 and 4 iterations. After 5 iterations, shown in Figure 5(d), the SU algorithm converges to three cluster centers. The single linkage algorithm is implemented with the final convergence states of data points. Figure 6(a) shows the hierarchical tree of the final states’ data points. Obviously, the hierarchical tree shows that the optimal cluster number is 3, and Figure 6(b) presents the corresponding clustering result.

Furthermore, the average values of cRate and aRan for the SU, soft-EM, hard-EM, and spkmeans algorithms based on 50 data sets generated from Equation (6) are compared. According to the results of Table 2, the proposed SU also performs well. However, the spkmeans algorithm produces a poor clustering quality in terms of the criteria of cRate and aRan.

**Example 3**  In the study on the exoplanets, many authors [8,17,18] have indicated that there is a possible correlation between the planetary mass and the orbit period. Therefore, it seems worth examining where groups for expolanets on these two important features are. The data are from the Extrasolar Planets Catalog compiled by Jean Schneider (http://cfa-www.harvard.edu/planets/catalog.html). Data as of 2014 December 14 are used, and incomplete data are excluded. There are 731 planets available for this study, and each of them has the values of projected mass ($M_p$), and orbital period ($P$). Following Jiang et al.'s [9] approach, the proposed SU algorithm is applied to the data ($\ln(P), \ln(M_p)$) on $S$. The SU algorithm converges to four cluster centers, as shown in Figure 7. Figure 8 shows the corresponding clustering result. Table 3 shows the corresponding centers and members. The fourth cluster only contains the exoplanet WASP-19 b. The characteristics of clusters PM1, PM2 and PM3 might associate

*Figure 3. The clustering results. Figure 3(a) and 3(b) shows the hierarchical tree and identified two clusters, respectively.*
Figure 4. Three-cluster data set from a mixture of 3-variate mixture Langevin distributions.

Figure 5. The states (shown as black) of the data points. Figure 5(a)–(d) represents the states of the data points after 1, 2, 4 iterations and convergence.

with the disc, tidal and ongoing tidal interactions. This result is consistent with that obtained by Jiang et al. [9].

4. Clustering results for extrasolar planets

In 1995, Mayor and Queloz [15] detected the first exoplanet around a Sun-like star. Since then, astronomers have found more than 1000 exoplanets. In this section, the proposed SU algorithm will be used to cluster the data of exoplanets using data from Example 3. There are 731 planets available and each of them has the values of projected mass ($M_p$), semimature axis ($a$), orbital eccentricity ($e$), stellar metallicity ([Fe/H]) and stellar mass ($M_*$) [13]. Marchi [13] applied the
clusters, respectively.

Figure 6. The clustering results. Figure 6(a) and 6(b) shows the hierarchical tree and identified three principal components, say pc1, pc2 and pc3. Furthermore, Marchi [13] applied a hierarchical clustering method for pc1, pc2 and pc3. His clustering result indicated that (i) the optimal cluster number is 5, and that (ii) there exists strong correlation among these 5 variables in some clusters.

According to the above numerical experiments, the proposed SU algorithm produces satisfactory results with the spherical data. Therefore, the SU algorithm will be used to cluster the data \((M_p, a, e, [Fe/H], M_\odot)\) on \(S^4\). The SU algorithm converges to five cluster centers and the corresponding hierarchical clustering tree is shown in Figure 9. The cluster C1 contains 179

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Soft-EM</th>
<th>Hard-EM</th>
<th>spkmeans</th>
<th>SU</th>
</tr>
</thead>
<tbody>
<tr>
<td>cRate</td>
<td>0.9772</td>
<td>0.9747</td>
<td>0.9150</td>
<td>0.9781</td>
</tr>
<tr>
<td>aRan</td>
<td>0.9718</td>
<td>0.9693</td>
<td>0.9099</td>
<td>0.9727</td>
</tr>
</tbody>
</table>

Table 2. The average values of cRate and aRan for different algorithms.

Figure 7. After convergence, there are four cluster centers (shown as blue).

PCA to handle this high-dimensional data \((M_p, a, e, [Fe/H], M_\odot)\). He obtained three important principal components, say pc1, pc2 and pc3. Furthermore, Marchi [13] applied a hierarchical clustering method for pc1, pc2 and pc3. His clustering result indicated that (i) the optimal cluster number is 5, and that (ii) there exists strong correlation among these 5 variables in some clusters.
On the other hand, exoplanets, and the cluster center is presented in Table 4. The significant (with a two-sided probability less than 5%) intracluster correlations are: $M_p - a, M_p - [Fe/H], M_p - M_s, a - [Fe/H]$, and $a - M_s$. The exoplanets’ masses are positively correlated with stellar masses. This confirms that higher $M_s$ implies larger protoplanetary disk surface density, and therefore larger $M_p$ [11]. On the other hand, $a$ anti-correlates with $[Fe/H]$. This implies that either the planetary migration is more pronounced for high $[Fe/H]$ or giant planets of this cluster may be close to stars in high metallicity environments [12]. Compared with Cluster $P_2$ of Jiang et al. [9], this study finds that this cluster might associate with the disc interaction.

Cluster C2 contains 292 exoplanets, and the cluster center is presented in Table 4. The significant intracluster correlations are: $M_p - a, M_p - e, M_p - [Fe/H], M_p - M_s, a - e, a - M_s$, and $[Fe/H] - M_s$. Note that $M_p$ is anticorrelated with $[Fe/H]$. This is because of some very massive planets with negative $[Fe/H]$, such as, 11 Com b, BD20 2457 b, BD20 2457 c, etc. Furthermore, the cluster C2 also contains several very massive close-in planets, CoRoT-27 b ($M_p = 10.39, P = 3.57532$), CoRoT-3 b ($M_p = 21.7, P = 4.256799$), HD 162020 b ($M_p = 14.4, P = 8.428198$), HD 41004 B b ($M_p = 18.4, P = 1.3283$), KELT-1 b ($M_p = 27.38, P = 1.217514$), WASP-18 b ($M_p = 10.43, P = 0.941452$) and XO-3 b ($M_p = 11.79, P = 3.191524$). This implies that these
The significant intracluster correlations are:

Planets will fall into their central stars in the future. Thus, the conjecture is that this cluster experiences ongoing tidal interaction.

Cluster C3 contains 251 exoplanets, and the cluster center is presented in Table 4. The significant intracluster correlations are: \( M_p - e, M_p - [Fe/H], M_p - M_s, a - e, a - M_s, \) and \([Fe/H] - M_s\). Note that \( M_p \) is anticorrelated with \( e \). This result implies that lower mass exoplanets have higher \( e \), thus, the mechanisms for the pumping-up of the eccentricity are more active in low-mass exoplanets, at least for the high semi-semimajor axes, and moderate positive metallicities of this cluster. For example, GJ 3293 d \((M_p = 0.0705, e = 0.37 )\), HD 126614 b \((M_p = 0.38, e = 0.41)\) and HD 137388 b \((M_p = 0.223, e = 0.36)\) belong to this cluster. Furthermore, the cluster center of C3 is within the regime in which the tidal interaction with the central star is important, based on Jiang et al.’s [8] result.

In clusters C2 and C3, the lack of \([Fe/H]\) correlation with \( a \) indicates that the stellar metallicity does not play a key role in exoplanet migration. Finally, cluster C5 only contains the exoplanet GJ 1214 b with a low-mass and low-semimajor axis.

Next, the spkmeans algorithm with \( k = 5 \) is also applied to the data \((M_p, a, e, [Fe/H], M_s)\) on \( S^4 \). Cluster SPC1 contains 110 exoplanets, and the cluster center is presented in Table 5. The significant intracluster correlations are: \( M_p - a, M_p - [Fe/H], M_p - M_s, a - [Fe/H], \) and \( a - M_s\). The characteristic of the cluster SPC1 is the same as that of C1.

The cluster SPC2 contains 183 exoplanets, and the cluster center is presented in Table 5. The significant intracluster correlations are: \( M_p - a, M_p - M_s, a - e, a - M_s, \) and \( e - [Fe/H]\). The cluster SPC3 contains 123 exoplanets, and the cluster center is presented in Table 5. The significant intracluster correlations are: \( M_p - a, M_p - M_s, a - [Fe/H], \) and \( a - M_s\). The cluster SPC4 contains 169 exoplanets, and the cluster center is presented in Table 5. The significant intracluster correlations are: \( M_p - [Fe/H], M_p - M_s, a - e, a - M_s, \) and \([Fe/H] - M_s\).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster center ((M_p, a, e, [Fe/H], M_s))</th>
<th>Members</th>
<th>Pearson corr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>((1.432, 3.839, 0.262, 0.056, 1.035))</td>
<td>179</td>
<td>corr((M_p, a) = 0.765)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, [Fe/H]) = -0.178)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, M_s) = 0.390)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, [Fe/H]) = -0.235)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, M_s) = 0.239)</td>
</tr>
<tr>
<td>C2</td>
<td>((4.863, 1.065, 0.216, 0.045, 1.248))</td>
<td>292</td>
<td>corr((M_p, a) = 0.424)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, e) = 0.199)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, [Fe/H]) = -0.141)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, M_s) = 0.356)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, e) = 0.381)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, M_s) = 0.254)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr([Fe/H], M_s) = -0.114)</td>
</tr>
<tr>
<td>C3</td>
<td>((0.434, 0.139, 0.092, 0.059, 1.026))</td>
<td>251</td>
<td>corr((M_p, e) = -0.198)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, [Fe/H]) = 0.127)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((M_p, M_s) = 0.580)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, e) = 0.341)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr((a, M_s) = 0.357)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>corr([Fe/H], M_s) = 0.192)</td>
</tr>
<tr>
<td>C4</td>
<td>((0.014, 0.231, 0.098, -0.628, 0.318))</td>
<td>8</td>
<td>None</td>
</tr>
<tr>
<td>C5</td>
<td>((0.020, 0.014, 0.270, 0.390, 0.150))</td>
<td>1</td>
<td>None</td>
</tr>
</tbody>
</table>

Table 4. The SU clustering results for exoplanets.
Table 5. The spkmeans clustering results for exoplanets.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster center ((M_p, a, e, [Fe/H], M_\odot))</th>
<th>Members</th>
<th>Pearson corr.</th>
</tr>
</thead>
</table>
| SPC1    | \((1.291, 4.289, 0.253, 0.033, 0.991)\) | 110    | corr\((M_p, a) = 0.749\)  
|         |                                 |        | corr\((M_p, [Fe/H]) = -0.227\)  
|         |                                 |        | corr\((M_p, M_\odot) = 0.419\)  
|         |                                 |        | corr\((a, [Fe/H]) = -0.231\)  
|         |                                 |        | corr\((a, M_\odot) = 0.278\)  
| SPC2    | \((6.974, 1.143, 0.252, 0.011, 1.270)\) | 183    | corr\((M_p, a) = 0.463\)  
|         |                                 |        | corr\((M_p, M_\odot) = 0.434\)  
|         |                                 |        | corr\((a, e) = 0.357\)  
|         |                                 |        | corr\((a, M_\odot) = 0.219\)  
|         |                                 |        | corr\((e, [Fe/H]) = 0.147\)  
| SPC3    | \((1.984, 1.529, 0.260, 0.103, 1.208)\) | 123    | corr\((M_p, a) = 0.757\)  
|         |                                 |        | corr\((M_p, M_\odot) = 0.404\)  
|         |                                 |        | corr\((a, [Fe/H]) = 0.178\)  
|         |                                 |        | corr\((a, M_\odot) = 0.255\)  
| SPC4    | \((0.960, 0.112, 0.050, 0.082, 1.142)\) | 169    | corr\((M_p, a) = 0.329\)  
|         |                                 |        | corr\((M_p, M_\odot) = 0.585\)  
|         |                                 |        | corr\((a, e) = 0.278\)  
|         |                                 |        | corr\((a, [Fe/H]) = -0.160\)  
|         |                                 |        | corr\((a, M_\odot) = 0.567\)  
| SPC5    | \((0.128, 0.162, 0.117, 0.000, 0.908)\) | 146    | corr\((M_p, [Fe/H]) = 0.188\)  
|         |                                 |        | corr\((M_p, M_\odot) = 0.519\)  
|         |                                 |        | corr\((a, e) = 0.208\)  
|         |                                 |        | corr\((a, M_\odot) = 0.347\)  
|         |                                 |        | corr\(([Fe/H], M_\odot) = 0.415\)  

The goal of this section is to determine whether the clustering results provide more important information about the exoplanet formation processes. According to this criterion, the clustering performance of the proposed SU algorithm is far better than that of the spkmeans algorithm.

5. Conclusions

This study developed a method for clustering spherical data based on an intuitive concept. The proposed algorithm uses all data points as initial cluster centers to solve the problem of choosing initial values. It then merges the surrounding data points automatically into the same cluster such that an optimal cluster number can be automatically found based on the structure of the data. The advantages of the proposed algorithm are robust to initial values and cluster numbers. Numerical experiments show that the proposed algorithm is more accurate than EM and spkmeans algorithms.

Finally, the proposed algorithm is used to cluster the data of exoplanets. The clustering results suggest two main implications: (1) there are three major clusters which might associate with the disc, ongoing tidal and tidal interactions, and (2) stellar metallicity does not play a key role in exoplanet migration.

Acknowledgments

The authors would like to thank the anonymous reviewers for their constructive comments to improve the presentation of the paper.
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported in part by the National Science Council, Taiwan, under Wen-Liang Hung’s Grant: NSC 102-2118-M-134-001.

References


