Int. J. Mach. Learn. & Cyber.
DOI 10.1007/s13042-016-0617-9

ORIGINAL ARTICLE

An ordered clustering algorithm based on K-means and the PROMETHEE method

Liuhao Chen[1] · Zeshui Xu[2] · Hai Wang[3] · Shousheng Liu[4]

Received: 31 March 2016 / Accepted: 2 November 2016

Abstract The multi-criteria decision aid (MCDA) has been a fast growing area of operational research and management science during the past two decades. The clustering problem is one of the well-known MCDA problems, in which the K-means clustering algorithm is one of the most popular clustering algorithms. However, the existing versions of the K-means clustering algorithm are only used for partitioning the data into several clusters which don't have priority relations. In this paper, we propose a complete ordered clustering algorithm called the ordered K-means clustering algorithm, which considers the preference degree between any two alternatives. Different from the K-means clustering algorithm, we apply the relative net flow of PROMETHEE to measure the closeness of alternatives. In this case, the ordered K-means clustering algorithm can capture the different importance degrees of criteria. At last, we employ the proposed algorithm to solve a practical ordered clustering problem concerning the human development indexes. Then a comparison analysis with an existing approach is conducted to demonstrate the advantages of the ordered K-means clustering algorithm.

Keywords Ordered cluster · PROMETHEE method · Weight vector · K-means clustering

1 Introduction

Clustering is a fundamental problem in the data analysis, which can be widely applied to machine learning, pattern recognition, information retrieval and data mining [1–5]. The main idea of clustering is to divide the set of data into a certain number of clusters (groups, subsets, or categories) which has high similarity in the same cluster according to the clustering objective.

One of the most well-known clustering algorithms is the K-means algorithm [6], which minimizes the sums of the distances of all the alternatives to the corresponding cluster center. The K-means clustering has become one of the most popular clustering algorithms because it is very fast and simple for implementation. A lot of studies have focused on this category of algorithms. For instance, Melnykov [7] discussed the k-mean clustering algorithm under Mahalanobis distances and proposed a novel approach for initializing covariance matrices. Bezdek [8] proposed the fuzzy c-means algorithm, which is based on the K-means clustering algorithm and fuzzy logic, to deal with nontrivial data and uncertainties encountered in real life [9, 10]. More generally, Xu and Wu [11] extended the fuzzy c-means algorithm to the intuitionistic fuzzy environment [12] and proposed the intuitionistic fuzzy c-means algorithm. Chen et al. [13] investigated the K-means

✉ Zeshui Xu
xuzeshui@263.net

Liuhao Chen
ywchlhao@126.com

Hai Wang
wanghai17@sina.com

Shousheng Liu
ssliunuaa@sina.com

[1] College of Communications Engineering, PLA University of Science and Technology, Nanjing 210007, Jiangsu, China

[2] Business School, Sichuan University, Chengdu 610065, Sichuan, China

[3] School of Economics and Management, Southeast University, Nanjing 211189, Jiangsu, China

[4] College of Sciences, PLA University of Science and Technology, Nanjing 210007, Jiangsu, China

Published online: 10 November 2016

clustering algorithm under hesitant fuzzy environment. In order to produce the nonlinear separating hypersurfaces between clusters, the kernel K-means clustering algorithm and fuzzy kernel K-means clustering algorithm have been developed [2]. In recent years, the K-means algorithms have been employed and developed so that they can benefit big data processing. Deng et al. [14] used a K-means clustering to separate the big dataset into several parts. In Mashayekhy et al. [15], Bolon-Canedo et al. [16] and Duan et al. [17], the K-means algorithm was implemented for real-time big data processing. This algorithm has also been considered for funds classification [18] and electron microscopy [19] in big data setting. A generalized version of K-means algorithm has also been proposed for processing temporal data [20].

However, the existing K-mean clustering algorithms are mainly used to cluster the data into several groups which don't have any relation among them. In multi-criteria decision aid (MCDA), the decision maker (DM) may desire to get "ordered clusters" in which there exist the ordered relations among the clusters. This kind of problems can be referred to as multi-criteria ordered clustering problems. The identification of ordered clusters can provide the priority relations of alternatives for the DMs. Although the ordered clusters can't provide more accurate relations of alternatives than the complete rankings of all the alternatives, the ordered clustering is also necessary in the real life problems. For example, in the ranking of world universities, the DMs may not give the accurate rankings of some universities because they have no obvious differences. Therefore, it is reasonable that we partition the alternatives with no significant differences into the ordered clusters. For multi-criteria ordered clustering problems, De Smet et al. [21] proposed an exact algorithm (we call it De Smet et al.'s method) to find a completely ordered partition based on the valued preference degrees. However, De Smet et al. [21] only used the ordinal properties of the pairwise preference relations to obtain the ordered clusters. They didn't fully exploit the underlying structure of a data set to produce better ordered clustering results.

As we know, the PROMETHEE method developed by Brans [22] is one of the effective outranking methods to solve the multi-criteria decision making (MCDM) problems [23–25]. The net outranking flow of the PROMETHEE II is a key technique to capture the relatively priority degrees of alternatives. Inspired by the PROMETHEE method and the K-means clustering algorithm, we propose an ordered clustering algorithm which is referred to as the ordered K-means clustering algorithm (OKM). Particularly, the OKM fully employs the net outranking flow of the PROMETHEE method to measure the alternatives with respect to the respective ordered cluster. The OKM first randomly generates the ordered cluster center, and constructs an optimization model to minimize the total net outranking flow of all ordered clusters. Meanwhile, the OKM applies the K-means clustering algorithm's idea to produce the better ordered clustering results.

The rest of the paper is organized as follows: Sect. 2 describes the classical PROMETHEE II and the classical K-means clustering algorithm. In Sect. 3, we propose the OKM based on the net outranking flow. In Sect. 4, we employ the human development index problem to demonstrate the applicability and the implementation process of the proposed algorithm. This section also provides a comparative analysis with the De Smet et al.'s method. Section 5 ends the paper with some conclusions.

2 The classical K-means clustering algorithms and the PROMETHEE methods

In the following, we give a review of the classical K-means clustering algorithm and the PROMETHEE method.

2.1 The classical K-means clustering algorithm

The classical K-means clustering algorithm [6] was introduced to deal with the clustering problem. For simplicity, let $A = \{a_1, a_2, \ldots, a_n\} \subseteq \Re^m$ be a sample set to be partitioned into K disjoint clusters $C_1, C_2, \ldots, C_K$. The K-means clustering algorithm obtains the clustering result by minimizing the objective given by

$$F(a_1, a_2, \ldots, a_n) = \sum_{k=1}^{K} \sum_{i=1}^{n} \delta_{ik} \|a_i - c_k\|^2,$$

where δ_{ik} is a clustering indicator variable with $\delta_{ik} = 1$ if $a_i \in C_k$ and 0 otherwise, and $c_k = \sum_{i=1}^{n} \delta_{ik} a_i \Big/ \sum_{i=1}^{n} \delta_{ik}$ is the centroid of the k-cluster, and $\|.\|$ represents the distance norm used by the K-means clustering algorithm. Usually, the Euclidean norm is applied by the K-means clustering algorithm.

For the optimization problem, the K-means clustering algorithm produces a solution by alternating optimization (AO) [26]. The K-means clustering algorithm's procedure is listed as follows:

Step 1 Input the dataset $A = \{a_1, a_2, \ldots, a_n\} \subseteq \Re^m$ and predefine the cluster number K.

Step 2 Randomly select K points of the dataset $A = \{a_1, a_2, \ldots, a_n\}$ as the initial cluster centers $c_k,\ k = 1, 2, \ldots, K$.

Step 3 According to the minimization distance from each point to the cluster center, we assign each point to the respective cluster center:

$$\delta_{ik} = \begin{cases} 1, & \arg\min_{1 \le k \le K} \|a_i - c_k\|^2 \\ 0, & \text{otherwise} \end{cases}. \quad (1)$$

Step 4 Re-compute the cluster center by the following formula:

$$c'_k = \sum_{i=1}^{n} \delta_{ik} a_i \Big/ \sum_{i=1}^{n} \delta_{ik}. \quad (2)$$

Step 5 If the cluster centroid updated by Eq. (2) has only negligible changes in the former cluster center, i.e., $\max_{1 \le k \le K}\left\{\left\|c_k - c'_k\right\|^2\right\} < \varepsilon$, where ε is a predetermined constant (usually let $\varepsilon = 10^{-3}$), then we obtain the cluster results. Otherwise, turn to Step 3.

2.2 The classical PROMETHEE methods

In what follows, we introduce the basic principles and some basic concepts related to the PROMETHEE method (see Brans and Mareschal [22] for more details). We first consider a MCDM problem where $A = \{a_1, a_2, \ldots, a_n\}$ is a set of n alternatives that are evaluated with respect to a set of criteria $G = \{g_1, g_2, \ldots, g_m\}$. Let $f_l(a_i)$ represent the evaluation of the alternative a_i with respect to the criterion g_l. The preference function $P_l(a_i, a_j)$, which should be given by the DM, represents the degree of preference of the alternative a_i to the alternative a_j with respect to the criterion g_l. There are six basic types of preference functions that can be utilized (see Brans and Vincke [27]). Certainly, the DM has to give the value of an indifference threshold q, and the value of a strict preference threshold p for each criterion. We can write the preference function as follows:

$$P_l(a_i, a_j) = F_l(d_l(a_i, a_j)) \quad \forall\, a_i, a_j \in A, \quad (3)$$

where $d_l(a_i, a_j) = f_l(a_i) - f_l(a_j)$, and $F_l(\cdot)$ is a monotonically non-decreasing function varying from 0 to 1.

For any two alternatives $a_i, a_j \in A$, we have

$$\pi(a_i, a_j) = \sum_{l=1}^{m} w_l \cdot P_l(a_i, a_j), \quad (4)$$

where $\pi(a_i, a_j)$ represents the total degree of preference of the alternative a_i to the alternative a_j taking into account all criteria. The weight w_l represents the relative importance of the criterion g_l in the set of all criteria.

In order to get the ranking of all the alternatives, the positive outranking flow, the negative outranking flow and the net outranking flow were introduced by Brans and Vincke [27] as follows:

$$\phi^+(a_i) = \frac{1}{m-1} \sum_{x \in A \setminus \{a_i\}} \pi(a_i, x),$$

$$\phi^-(a_i) = \frac{1}{m-1} \sum_{x \in A \setminus \{a_i\}} \pi(x, a_i),$$

$$\phi(a_i) = \phi^+(a_i) - \phi^-(a_i),$$

where the positive outranking flow $\phi^+(a_i)$ represents how much the alternative a_i prefers to all the other alternatives. The larger $\phi^+(a_i)$, the better the alternative a_i. Similarly, the negative outranking flow $\phi^-(a_i)$ represents how much all the other alternatives prefer to the alternative a_i. The smaller $\phi^-(a_i)$, the better the alternative a_i. Generally, the larger the net outranking flow $\phi(a_i)$, the better the alternative a_i. If $\phi(a_i) = 1$, then the alternative is strictly better than all the other alternative. If $\phi(a_i) = \phi(a_j)$, then the alternative a_i is indifferent to a_j. Therefore, we can get a complete ranking according to the value of net outranking flow for each alternative.

3 The ordered K-mean clustering algorithm

In this paper, we consider a special clustering problem called ordered clustering problem, which was first addressed by Smet and Gilbart [28] for dealing with the country risk evaluation problem. As previously noted, the identification of the ordered clusters can provide a necessary support for the DM to sort the alternatives. Unlike the classical clustering problem, the ordered clustering problem not only partitions the alternatives into the predetermined number of clusters, but also has a completely ranking relationship of these clusters.

Let $A = \{a_1, a_2, \ldots, a_n\} \subseteq \Re^m$ be a sample set whose elements are evaluated under a set of criteria $G = \{g_1, g_2, \ldots, g_m\}$. If a partition satisfies the following conditions, then we say that the partition is an ordered partition:

(i) $A = \bigcup_{i=1,2,\ldots,K} C_i$;

(ii) $\forall i \neq j,\ C_i \cap C_j = \emptyset$;

(iii) $C_1 \succ C_2 \succ \cdots \succ C_K$,

where C_i represents the ith ordered cluster and $\succ$ denotes the priority relation. If $C_i \succ C_j$, then the elements in the ordered cluster C_i is better than the elements in the ordered cluster C_j.

As we know, the classic K-mean clustering algorithm has been widely used for the classical clustering problems. However, the K-mean clustering algorithm can't be used for dealing with the ordered clustering problem. That is because that the K-mean clustering algorithm uses the Euclidean norm to measure the similar degree between alternatives, but the Euclidean norm doesn't consider the relative importance between the criteria for the DM.

Fortunately, the PROMETHEE method not only considers the difference between criteria, but also can get the priority degree between any two alternatives. Inspired by the PROMETHEE method, in the next section, we will propose an ordered K-mean ordered clustering algorithm based on the net outranking flow to identify the best K-ordered cluster.

3.1 The minimum net outranking flow objective

Similar to the K-means clustering algorithm, we define the objective function:

$$\min J(U,V) = \sum_{i=1}^{K} \sum\nolimits_{a_j \in C_i} \left|\phi_{C_i}(a_j)\right|^2, \tag{5}$$

where C_i is the set of the alternatives in the ith cluster and the partial net outranking flow can be obtained by

$$\phi_{C_k}(a_j) = \frac{1}{|C_k|}\left(\sum_{a_i \in C_k} \pi(a_j,a_i) - \sum_{a_i \in C_k} \pi(a_i,a_j)\right). \tag{6}$$

As have been stated before, the net outranking flow of the PROMETHEE method is used to characterize the quality of alternatives. The larger the net outranking flow $\phi(a_i)$, the better the alternative a_i. In order to get an ordered cluster of all the alternatives, we propose the partial net outranking flow to capture the similarity of alternatives and reconstruct the optimization model (5) to cluster the alternatives. Compared with the classical K-means clustering algorithm, our proposed optimization model has the following two advantages: (1) the partial net outranking flow takes the weight of each criterion into account; (2) the partial net outranking flow considers the preferences of all the alternatives in the same cluster.

For the optimization problem, all the alternatives A are divided into K ordered clusters with minimizing the sum of all the alternatives' partial net outranking flows. Using the exhaustive method, we know that there are K^n divisions. The objective function isn't a convex function. As noted above, finding an exact solution to this problem is NP-hard. Inspired by the AO [26], we try to construct two reduced problems:

Problem P_1: Fix $V = \hat{V}$ and solve the reduced problem $J(U,\hat{V})$.

Problem P_2: Fix $U = \hat{U}$ and solve the reduced problem $J(\hat{U},V)$.

The matrix U represents a partition of all the alternatives with $\sum_{j=1}^{K} u_{ij} = 1,\ i = 1,2,\ldots,n;\quad u_{ij} = 0$ or $1,\ i = 1,2,\ldots,n,\ j = 1,2,\ldots,K$. The set $V = \{c_1,c_2,\ldots,c_K\}$ represents the ordered cluster center with c_i is the ith cluster center. In the following two sections, we give the solutions of these two problems respectively.

3.2 Update the cluster

In the classical K-means clustering algorithm, the shortest Euclidean distance is used to update the membership of cluster. The Euclidean distance only considers the actual distance between alternatives. However, this is not suitable for the ordered clustering of MCDM. The net outranking flow can reflect the relative importance of each alternative for the cluster, but it is relative to the alternatives in the cluster. We try to find the relationship between the partial net outranking flow and the corresponding cluster center, and then obtain

$$\begin{aligned}
\phi_{C_k}(a_j) &= \frac{1}{|C_k|}\left(\sum_{a_i \in C_k} \pi(a_j,a_i) - \sum_{a_i \in C_k} \pi(a_i,a_j)\right) = \frac{1}{|C_k|}\left(\sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot P_l(a_j,a_i) - \sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot P_l(a_i,a_j)\right)\\
&= \frac{1}{|C_k|}\left(\sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot F_l(f_l(a_j) - f_l(a_i)) - \sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot F_l(f_l(a_i) - f_l(a_j))\right)\\
&= \frac{1}{|C_k|}\left(\sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot F_l(f_l(a_j) - f_l(c_k) + f_l(c_k) - f_l(a_i)) - \sum_{a_i \in C_k}\sum_{l=1}^{m} w_l \cdot F_l(f_l(a_i) - f_l(a_k) + f_l(a_k) - f_l(a_j))\right),
\end{aligned}$$

where $f_l(a_i)$ represents the evaluation of the alternative a_i with respect to the criterion g_l, and c_k represents the center of the kth cluster C_k.

If the function $F_l(\cdot)$ is a linear function, then

$$\begin{aligned}\phi_{C_k}(a_j) &= \frac{1}{|C_k|}\left(\sum_{a_i\in C_k}\sum_{l=1}^{m} w_l\cdot F_l(f_l(a_j)-f_l(c_k))\right.\\ &\quad\left.-\sum_{a_i\in C_k}\sum_{l=1}^{m} w_l\cdot F_l(f_l(a_k)-f_l(a_j))\right)\\ &\quad+\frac{1}{|C_k|}\left(\sum_{a_i\in C_k}\sum_{l=1}^{m} w_l\cdot F_l(f_l(c_k)-f_l(a_i))\right.\\ &\quad\left.-\sum_{a_i\in C_k}\sum_{l=1}^{m} w_l\cdot F_l(f_l(a_i)-f_l(c_k))\right)\\ &= \phi_{c_k}(a_j)+\phi_{C_k}(c_k),\end{aligned}$$

where c_k represents the center of the kth cluster C_k and $\phi_{c_k}(a_j)=\left(\sum_{a_i\in C_k}\pi(a_j,c_k)-\sum_{a_i\in C_k}\pi(c_k,a_j)\right)$.

Brans and Vincke [27] gave six types of particular preference functions: (1) usual criterion, (2) U-shape criterion, (3) V-shape criterion, (4) level criterion, (5) V-shape with indifference criterion, and (6) Gaussian criterion. These preference functions (except for Gaussian criterion) are linear functions. Therefore, the transformation can be established.

For the large data, we can see $\phi_{C_k}(a_j)\approx\phi_{c_k}(a_j)+\phi_{C_k}(c_k)$ if the linear preference functions are adopted. For simplicity, we use the center c_k of cluster to represent the corresponding cluster and compute the relative distance between the alternative and the center of cluster.

For the problem P_1, the cluster center $V=\{c_1,c_2,\ldots,c_K\}$ is given, and the solution can be given by

$$u_{il}=1,\text{ if }\left|\phi_{c_l}(a_i)\right|\le\left|\phi_{c_j}(a_i)\right|\text{ for }1\le j\le K,$$
$$u_{it}=0\ \text{ if }t\ne l. \tag{7}$$

In order to facilitate the presentation, Eq. (7) can be rewritten as follows:

$$u_{ik}=\begin{cases}1, & k=\arg\min_{1\le l\le K}\left|\phi_{c_l}(x_i)\right|^2\\ 0, & \text{otherwise}\end{cases}. \tag{8}$$

Thus, each alternative is assigned to the ordered cluster which has the smallest relative preference with the alternative.

3.3 Update the centroid of cluster

For the optimization problem P_2, we have to update the centroid of cluster according to the clustering result. In the classical K-mean clustering algorithm, Eqs. (1) and (2) are used to update the centroid of cluster. Equation (2) is the optimal solution of the following optimization problem:

$$c_i=\arg\min_{a\in\Re^m}\frac{1}{2}\sum_{a_j\in C_i}\left\|a-a_j\right\|^2,\ i=1,2,\ldots,K. \tag{9}$$

To capture the ordered feature of cluster, the clustering center of P_2 can be obtained by:

$$c_i=\arg\min_{x\in\Re^m}\frac{1}{2}\left|\phi_{C_i}(x)\right|^2,\ i=1,2,\ldots,K, \tag{10}$$

where $\phi_{C_i}(x)$ represents the partial net outranking flow of the data $x\in\Re^m$ and can be computed by Eq. (6).

The model (9) of the classical K-means clustering algorithm uses the Euclidean distance to identify the center of cluster, while our proposed optimization model (10) utilizes the partial net outranking flow to identify the center of cluster. According to Eq. (6), we can find that the partial net outranking flow is the total preference degree in the kth ordered cluster. Thus, the optimization model (10) can capture the relationship between alternatives in the same cluster. Meanwhile, the partial net outranking flow takes the importance degrees of criteria into account. The model is suitable for the multi-criteria ordered clustering algorithm.

However, the exact result of the optimization model (10) usually cannot be obtained directly, because the optimization objective usually can't be differentiable. Thus, we propose an approximate method to estimate the clustering center:

$$c_i=\arg\min_{a_j\in C_i}\left|\phi_{C_i}(a_j)\right|^2. \tag{11}$$

Note that Eq. (11) is not a K-means-type updating but is a K-medoids-type one. We use it instead of Eq. (10) so that the complexity of the optimization model can be reduced. If the number of the alternatives is sufficiently large, then the value obtained from Eq. (11) will be a good approximation of the cluster. In fact, as will be seen in Sect. 4, Eq. (11) can perform very well. However, if the data set is not densely distributed in each cluster, Eq. (11) may not perform so well. In such a case, we may need to consider Eq. (10) or its other kinds of approximations.

3.4 The proposed algorithm

The procedure of the OKM is developed as follows:

Step 1 Input the dataset $A=\{a_1,a_2,\ldots,a_n\}\subseteq\Re^m$ and predefine the cluster number K.

Step 2 Randomly select K points of the dataset $A=\{a_1,a_2,\ldots,a_n\}$ as the initial cluster centers $c_k(k=1,2,\ldots,K)$, and reorder the cluster centers as $c_{(k)}(k=1,2,\ldots,K)$ with $c_{(1)}\succ c_{(2)}\succ\cdots\succ c_{(K)}$, where the order $\succ$ is defined according to the net outranking flow, i.e., $c_{(i)}\succ c_{(j)}$ if $\phi(c_{(i)})>\phi(c_{(j)})$ for any $i,j=1,2,\ldots,K$.

Step 3 According to the minimization distance from each point to the cluster center, we assign each point to the respective cluster center:

$$u_{ik} = \begin{cases} 1, & k = \arg\min_{1 \le l \le K} \left|\phi_{c_l}(x_i)\right|^2 \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

Step 4 Re-compute the cluster center by Eq. (11).

Step 5 If the cluster centroid updated by Eq. (11) has only negligible changes in the former cluster center, i.e., $\max_{1 \le k \le K}\left\{\left|\phi_{c'_k}(c_k)\right|^2\right\} < \varepsilon$, where ε is a predetermined constant (usually let $\varepsilon = 10^{-3}$) and c'_k is the updated center of the cluster C_k, $k = 1, 2, \ldots, K$, then we obtain the clustering results. Otherwise, turn to Step 3.

Since the objective function of the optimization problem (5) isn't a convex function, then finding an exact solution to this problem is NP-hard. Until now, there is no any polynomial time algorithm to obtain the global optimization solution, so we try to find out a local optimization solution. We find that the OKM has the following result:

Theorem 1 *The OKM converges to a local optimal solution of the problem (5) in a finite number of iterations.*

Proof As pointed out in Sects. 3.2 and 3.3, the solutions obtained by Eqs. (11) and (12) is the local (or approximate) optimal solution of the problems P_1 and P_2. Therefore, the objective value $J(U, V)$ is monotonically decreasing in the process of the OKM. Meanwhile, the objective function has a lower bound as the collection of feasible solutions is finite. Thus, the proof is completed.

4 Application and comparative analysis

In order to illustrate the effectiveness of the OKM, let us consider the human development index (HDI) problem, which is adopted by Ref. [21]. The HDI was created to emphasize that people and their capabilities should be the ultimate criteria for assessing the development of a country, not only economic growth. The United Nations Development Program (UNDP) has proposed the so-called HDI ranking where 179 United Nations countries are evaluated on the basis of three criteria: the life expectancy, the education and the income index. The HDI ranking can be obtained according to the values which are aggregated by three criteria for each country. However, some scholars [29–32] have criticized the method adapted by the UNDP. In this section, we don't care the exact ranking problem of countries, and only try to partition the countries into several ordered clusters.

In the following, we use the OKM to regroup the countries: the complete lists of the countries, as well as their performances on the three criteria for the year 2008, are given in the Appendix of Ref. [21]. In order to apply our method, we first compute the preference degrees between the countries under the above three criteria. As Behzadian et al. [33] pointed out that the PROMETHEE method has lots of advantages compared to other outranking methods (such as the ELECTRE methods [34]). Thus, we use the PROMETHEE outranking method to compute the preference degrees. For each criterion, we consider the following V-shape preference function:

$$F_l(d) = \begin{cases} 0, & d \le q_l \\ (d - q_l)/(p_l - q_l), & q_l < d \le p_l\,, \\ 1, & d \ge p_l \end{cases} \quad l = 1, 2, 3, \tag{13}$$

where the values of two thresholds and the weights of three criteria are listed in Table 1 (for more details about the parameters, please refer to De Smet et al. [21]).

Based on the above parameters of the PROMETHEE and the linear preference function, we can apply Eq. (4) to obtain the preference degree between any two countries. Before applying the OKM, the number of clusters should be specified by the DM. In the following, we first consider the clustering problem where the number of clusters is 4. They can be respectively signed as the very highly developed countries, highly developed countries, medium-developed countries and low developed countries. The ordered clustering results obtained by using the OKM are presented in Fig. 1, where the x-axis represents the label of the countries (from 1 to 179) and the y-axis denotes the number of clusters. For example, we can get that the alternative 30 belongs to the cluster one and the alternative 100 belongs to the cluster two. From Fig. 1, we can easily find that the ordered clustered results are highly consistent with the HDI ranks, which are given by the UNDP. There is only one alternative (the alternative 157) that is inconsistent with the HDI ranks. The first 50 countries belong to the very high developed countries, the countries 51–124 are the high developed countries, the countries 125–159 except the country 157 are the medium developed countries and

Table 1 The indifference, preference thresholds and the weight of each criterion

Parameters	Life expectancy	Adult literacy index	GDP
Strict preference threshold: p_l	0.704	0.719	0.828
Indifference threshold: q_l	0	0	0
Weight of criterion: w_l	0.333	0.333	0.333

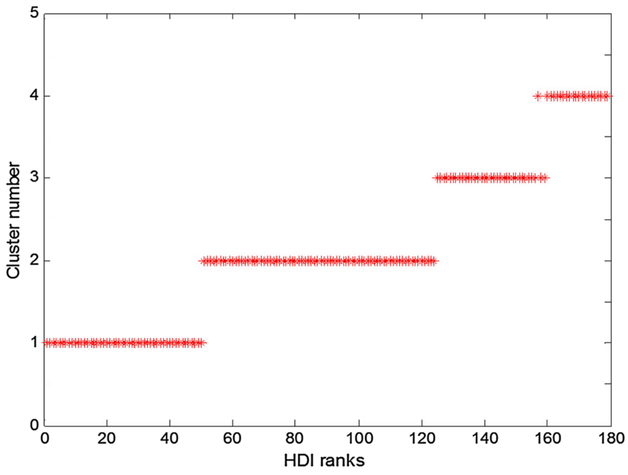

Fig. 1 The results of the ordered partition for four clusters by using the OKM. The *x-axis* represents the label of the countries and the *y-axis* denotes the number of clusters

other countries are the low developed countries. In order to get intuitive feeling of the result, we mark out each country in a three-dimensional map according to the values of three criteria. The results are presented in Fig. 2, where the first axis represents the life expectancy, the second axis denotes the adult literacy index and the third axis indicates the GDP.

To illustrate the advantage of the OKM, we make a comparative analysis with the similar methods. As mentioned previously, the classical K-means clustering algorithm is an effective method for clustering data. For the above problem, we apply the classical K-means clustering algorithm to partition the countries and the results are shown in Fig. 3. The meanings of two axes coincide with Fig. 1. We notice that the partition result is seriously inconsistent with the HDI ranks. The main reason of this phenomenon is that the classical K-means clustering algorithm uses the Euclidean distance to measure the similar degree between any two countries. Since the symmetry of the Euclidean distance, there is no preference relationship between the clusters obtained from the classical K-means clustering algorithm.

Based on the above analysis, the OKM has lots of advantages over the classical K-means clustering algorithm for the ordered clustering problems. However, the OKM is based on the idea of classical K-means clustering algorithm, and thus, there are two inherent deficiencies: the cluster number is predefined before clustering and the result is dependent on the initial values. To reduce the effect, we will run the OKM ten times and select the partition with the least objective function value. If the DM can't fix the cluster number, then we compute all the clustering results for several possible cluster numbers and find out the optimization cluster number.

Figure 4 shows the drift of the total sum of all alternatives' net outranking flow with respect to the cluster numbers. We test the cases when the numbers of cluster increase from 1 to 9. It is obvious that the total sum is decreasing in regards to the cluster numbers. Therefore, the larger the cluster number, the smaller the total sum of all the alternatives' net outranking flow. Consider that more classes will not lead to deeply increase the quality of the partition. Thus, we suggest that the best cluster number should be 5 or 6.

In order to further illustrate the advantage of the OKM, we make a comparative analysis with the similar method

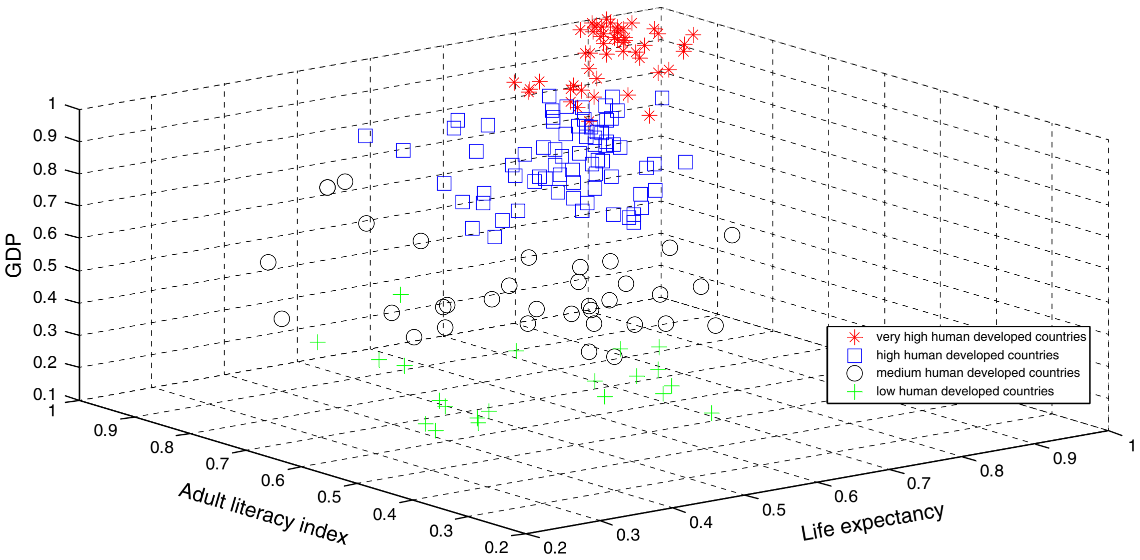

Fig. 2 The ordered clustering results using the OKM. The *three axes* represent the life expectancy, the adult literacy index and the GDP, respectively

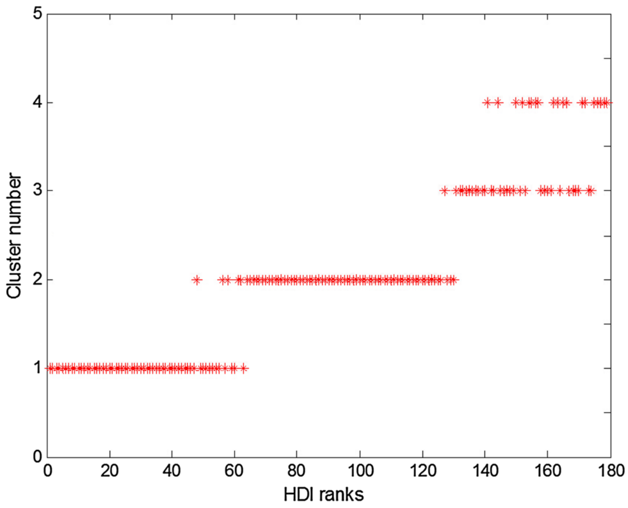

Fig. 3 Cluster results obtained from the classical K-means clustering algorithm (four clusters). The *x-axis* represents the label of the countries and the *y-axis* denotes the number of clusters

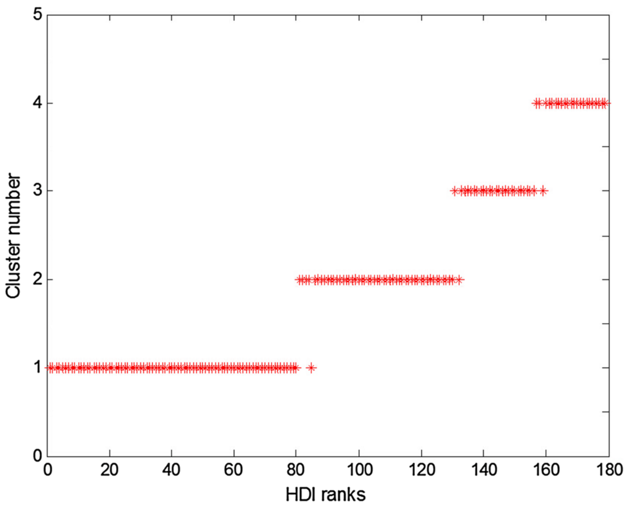

Fig. 5 The results of the ordered partition for four clusters by using the De Smet et al.'s method. The *x-axis* represents the label of the countries and the *y-axis* denotes the number of clusters

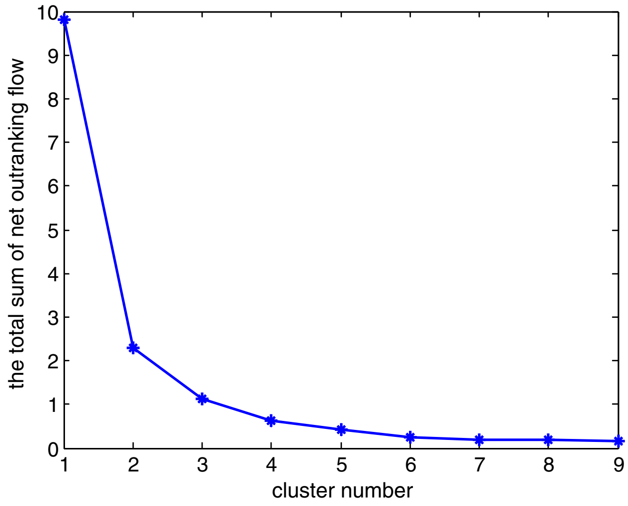

Fig. 4 The sum of the net outranking flows under different cluster numbers

proposed by De Smet et al. [21] as mentioned previously. It is worthwhile to point out that the De Smet et al.'s method is a better clustering algorithm than the classical K-means clustering algorithm for the multi-criteria ordered clustering problems. The main procedure of the De Smet et al.'s method is summarized as follows:

Step 1 Input the preference matrix π of all the alternatives and the predefined cluster number K, let the matrix $M = 0$.

Step 2 Select the largest value π_{ij} of the preference matrix π, if $\pi_{ij} > 0$, go to Step 3; if $\pi_{ij} = 0$, then go to Step 4.

Step 3 Test if putting directed arc from the alternative i to the alternative j creates a cycle or a path longer than $K - 1$ in the graph induced by the new M matrix. If not, then we let $M_{ij} = 1$ and $\pi_{ij} = 0$. If yes, then we let $M_{ij} = 0$ and $\pi_{ij} = 0$. After that, we update the preference matrix π and go to step 2.

Step 4 The K-ordered partition is obtained by the determination of the ranks of the graph induced by the matrix M.

Figure 5 given by De Smet et al. [21] presents the clustering results of the HDI ranking problem when the cluster number is 4. From Fig. 5, we observe that the ordered clustering results are very high consistent with the HDI ranks. By comparing Figs. 1 and 5, we can see that the clustering results of De Smet et al.'s method and the OKM are different. The first 80 countries belong to the first cluster in Fig. 5, while only the first 50 countries belong to the first cluster in Fig. 1. Meanwhile, the number of alternatives in the higher priority cluster is larger than the number of alternatives in the lower priority cluster in Fig. 5.

Since the De Smet et al.'s method is based on the lexicographic comparison of the inconsistent preference matrix, then the largest preference value can be considered first. As De Smet et al. [21] pointed out, if several elements of the preference matrix have the same value, then the order of the elements will affect the results. Therefore, the De Smet et al.'s method is too sensitive to the changes of the elements in the preference matrix.

We have analyzed the proposed approach and the De Smet et al.'s method [21] and the clustering results derived from these two methods about the HDI ranking problems. It is not hard to see that the OKM has some desirable advantages over the De Smet et al.'s method [21]. We summarize them as follows:

1. The OKM can estimate the suitable clustering number according to the total sum of all alternatives' net outranking flows; while the De Smet et al.'s method [21] can only get the ordered clustering result under the predefined cluster number and can't judge which cluster number is suitable.
2. The OKM uses the net outranking flow to measure the similarity of alternatives, and the net outranking flows consider the relative importance of each alternative for all the rest alternatives. Thus, the clustering result obtained from the OKM can't change dramatically. However, the De Smet et al.'s method [21] is too sensitive to the changes of the elements in the preference matrix.
3. The OKM partitions the alternatives into several ordered clusters according to the similarity of alternatives, thus the intra-relationships of alternatives can be captured; while the De Smet et al.'s method [21] considers the order of elements in the preference matrix to identify the best K-ordered partition and doesn't consider the intra-relationships of alternatives.

It is obvious that the proposed OKM can be demonstrated by some other datasets. We adopt the HDI problems for demonstration just for the purpose of comparing the OKM with similar techniques.

In our opinion, the OKM possesses potential effectiveness for practical applications, such as big data processing. In fact, in the era of big data, clustering and classification are crucial techniques for discovering knowledges from data. Priorities and orders exist in clusters and classes due to the nature of specific problems. For instance, when facing the imbalanced data sets, such as medical big data, the patterns in the negative cluster (patterns of illness) own great priorities, comparing with patterns in the positive cluster (normal patterns). Another example can be found in sentimental analysis of social networks. In this case, discovering negative comments is more important than understanding positive comments. Because managers and organizations can realize the weaknesses of their services or products. Therefore, the ordered clustering and classification commonly exist in big data processing. However, it should be noticed that the OKM should be well improved and implemented so that it can meet other requirements of bid data processing.

5 Concluding remarks

In this paper, we have mainly considered the multi-criteria ordered clustering problems. In order to deal with this type of problems, we have proposed an ordered clustering algorithm based on K-means clustering algorithm, which is called ordered K-means clustering algorithm. As the net outranking flow is an effective way to compute the sorting of all alternatives, we have applied the idea of net outranking flow to identify the clustering center of each cluster. Different from the classical k-means clustering algorithm, the sum of all alternatives' net outranking flow can be used as the objective function. There is a complete ordered relationship between the clusters, which is obtained from the ordered K-means clustering algorithm.

The effectiveness of the ordered K-means clustering algorithm has been illustrated by the human development index problem. The ordered clustering results obtained from the ordered K-means clustering algorithm are very highly consistent with the HDI ranks. Meanwhile, the De Smet et al.'s method [21] has been introduced to compare with the ordered K-means clustering algorithm. The comparison analysis with the De Smet et al.'s method [21] has shown that (1) the ordered K-means clustering algorithm has a good robust; (2) the ordered K-means clustering algorithm can select the suitable cluster number according to the total sum of all alternatives' net outranking flows; (3) the clustering results obtained from the ordered K-means clustering algorithm can take full account of the intra-relationships between alternatives.

For future work, we can consider the application of the proposed OKM to other practical problems related to ordered clustering. The deep discussions of using the nonlinear preference function in the OKM are interesting. In addition, the introduction of the proposed OKM to big data clustering is also meaningful.

Acknowledgements The authors thank the anonymous reviewers for their helpful comments and suggestions, which have led to an improved version of this paper. The work was supported by the National Natural Science Foundation of China (Nos. 61273209, 71571123).

References

1. Xu R, Wunsch D (2005) Survey of clustering algorithms. Neural Netw IEEE Trans 16:645–678
2. Filippone M, Camastra F, Masulli F, Rovetta S (2008) A survey of kernel and spectral methods for clustering. Pattern Recogn 41:176–190
3. Berkhin P (2006) A survey of clustering data mining techniques. Grouping multidimensional data. Springer, Berlin, pp 25–71
4. Baraldi A, Blonda P (1999) A survey of fuzzy clustering algorithms for pattern recognition. I. Syst Man Cybern Part B Cybern IEEE Trans 29:778–785
5. Xu ZS (2013) Intuitionistic fuzzy aggregation and clustering. Springer, Berlin
6. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall Inc, USA
7. Melnykov I, Melnykov V (2014) On K-means algorithm with the use of Mahalanobis distances. Stat Probab Lett 84:88–95
8. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum, New York

9. Hai-Yan T, Bao D, Qi WG (2009) Research on traffic mode of elevator applied fuzzy C-mean clustering algorithm based on PSO. In: Measuring technology and mechatronics automation, 2009. ICMTMA'09. International Conference on, IEEE, pp 582–585
10. Havens TC, Bezdek JC, Leckie C, Hall LO, Palaniswami M (2012) Fuzzy c-means algorithms for very large data. IEEE Trans Fuzzy Syst 20:1130–1146
11. Xu ZS, Wu JJ (2010) Intuitionistic fuzzy c-means clustering algorithms. J Syst Eng Electron 21:580–590
12. Yu DJ, Shi SS (2015) Researching the development of Atanassov intuitionistic fuzzy set: using a citation network analysis. Appl Soft Comput 32:189–198
13. Chen N, Xu ZS, Xia MM (2014) Hierarchical hesitant fuzzy K-means clustering algorithm. Appl Math J Chin Univ 29(1):1–17
14. Deng ZY, Zhu XS, Cheng DB, Zong M, Zhang SC (2016) Efficient kNN classification algorithm for big data. Neurocomputing 195:143–148
15. Mashayekhy L, Nejad MM, Grosu D, Zhang Q, Shi WS (2015) Energy-aware scheduling of mapReduce jobs for big data applications. IEEE Trans Parallel Distrib Syst 26:2720–2733
16. Bolon-Canedo V, Fernandez-Francos D, Peteiro-Barral D, Alonso-Betanzos A, Guijarro-Berdinas B, Sanchez-Marono N (2016) A unified pipeline for online feature selection and classification. Expert Syst Appl 55:532–545
17. Duan HC, Peng YB, Min GY, Xiang XK, Zhan WH, Zou H (2015) Distributed in-memory vocabulary tree for real-time retrieval of big data images. Ad Hoc Netw 35:137–148
18. Wang H, Wu JJ, Yuan S, Chen J (2016) On characterizing scale effect of Chinese mutual funds via text mining. Sig Process 124:266–278
19. Jesse S, Chi M, Belianinov A, Beekman C, Kalinin SV, Borisevich AY, Lupini AR (2016) Big data analytics for scanning transmission electron microscopy ptychography. Sci Rep 6:26348
20. Soheily-Khah S, Douzal-Chouakria A, Gaussier E (2016) Generalized k-means-based clustering for temporal data under weighted and kernel time warp. Pattern Recogn Lett 75:63–69
21. De Smet Y, Nemery P, Selvaraj R (2012) An exact algorithm for the multicriteria ordered clustering problem. Omega 40:861–869
22. Brans JP, Mareschal B (2005) PROMETHEE methods. Multiple criteria decision analysis: state of the art surveys. Springer, Berlin, pp 163–186
23. Liao HC, Xu ZS (2014) Multi-criteria decision making with intuitionistic fuzzy PROMETHEE. J Intell Fuzzy Syst 27:1703–1717
24. Chen LH, Xu ZS (2015) A new prioritized multi-criteria outranking method: the prioritied PROMETHEE. J Intell Fuzzy Syst 29:2099–2110
25. Yu XH, Xu ZS, Ma Y (2013) Prioritized multi-criteria decision making based on the Idea of PROMETHEE. Proced Comp Sci 17:449–456
26. Bezdek J, Hathaway R (2003) Convergence of alternating optimization. Nueral Parallel Sci Comput 11:351–368
27. Brans JP, Vincke P (1985) A preference ranking organization method. The PROMETHEE method for multiple criteria decision-making. Manag Sci 31:647–656
28. De Smet Y, Gilbart F (2001) A cluster definition method for country risk problems. Technical Report SMG, IS-MG 2001/13
29. Noorbakhsh F (1998) The human development index: some technical issues and alternative indices. J Int Dev 10:589–605
30. Noorbakhsh F (1996) Some reflections on the UNDP's human development index. Centre for Development Studies, University of Glasgow, Scotland
31. Trabold-Nübler H (1991) The human development index–A new development indicator? Intereconomics 26:236–243
32. Kelley AC (1991) The human development index: "handle with care". Popul Dev Rev 17:315–324
33. Behzadian M, Kazemzadeh RB, Albadvi A, Aghdasi M (2010) PROMETHEE: a comprehensive literature review on methodologies and applications. Eur J Oper Res 200:198–215
34. Figueira JR, Greco S, Roy B, Słowiński R (2013) An overview of ELECTRE methods and their recent extensions. J Multi Criteria Decis Anal 20:61–85