Pattern Analysis and Applications
https://doi.org/10.1007/s10044-019-00783-6

THEORETICAL ADVANCES

# Effect of cluster size distribution on clustering: a comparative study of *k*-means and fuzzy *c*-means clustering

**Kaile Zhou**[1,2,3] · **Shanlin Yang**[1,2]

Received: 26 October 2017 / Accepted: 30 January 2019


**Abstract**
Data distribution has a significant impact on clustering results. This study focuses on the effect of cluster size distribution on clustering, namely the uniform effect of *k*-means and fuzzy *c*-means (FCM) clustering. We first provide some related works of *k*-means and FCM clustering. Then, the structure decomposition analysis of the objective functions of *k*-means and FCM is presented. Afterward, extensive experiments on both synthetic two-dimensional and three-dimensional data sets and real-world data sets from the UCI machine learning repository are conducted. The results demonstrate that FCM has stronger uniform effect than *k*-means clustering. Also, it reveals that the fuzzifier value $m=2$ in FCM, which has been widely adopted in many applications, is not a good choice, particularly for data sets with great variation in cluster sizes. Therefore, for data sets with significant uneven distributions in cluster sizes, a smaller fuzzifier value is preferred for FCM clustering, and *k*-means clustering is a better choice compared with FCM clustering.

**Keywords** Clustering · Data distribution · *k*-means · Fuzzy *c*-means (FCM) · Fuzzifier · Uniform effect

## 1 Introduction

Clustering [1–3] is an unsupervised learning process, which means that the data objects are clustered into several groups according to the similarities/dissimilarities among them, without prior knowledge. It is one of the important tasks of data mining and statistical machine learning. Currently, many different clustering algorithms have been proposed [4–6], which can be divided into different categories. Generally, clustering algorithms can be divided into partition-based [7, 8], hierarchical-based [9–11], density-based [12], grid-based [13, 14] and model-based [15, 16] methods. From the perspective of whether each data object definitely belongs to one cluster of the clustering results, clustering algorithms can also be divided into crisp (hard) clustering [17–19] and fuzzy clustering methods [20–22]. These clustering methods have been successfully used in many areas, such as pattern recognition of gene expression [23], image processing [24], electricity consumption pattern mining [25], anomaly detection [26], protein structures clustering [27] and information retrieval [28].

Though there have been many successful applications of clustering, some inherent deficiencies of traditional clustering algorithms have significantly weakened their performance, which have attracted more and more attention of researchers, for instance the selection of initial cluster centers [29], the determination of optimal number of clusters [30], as well as the influence of outliers [31]. The characteristics of a given data set, such as the dimension, size, noise and outliers, and type of attributes, are also important factors which can strongly affect the unsupervised clustering process [32]. It is important to understand how data distributions can have impact on the results of clustering. Currently, there have some research efforts that focused on the influence of data distribution on the performance of individual clustering algorithm, including *k*-means, hierarchical clustering and fuzzy *c*-means (FCM) [33–35]. However, few are known about the quantitative and comparative effects of data distribution on the performance of different clustering algorithms. This is mainly due to the facts that the machine

✉ Kaile Zhou
zhoukaile@hfut.edu.cn

[1] School of Management, Hefei University of Technology, Hefei 230009, China

[2] Key Laboratory of Process Optimization and Intelligent Decision-Making of Ministry of Education, Hefei University of Technology, Hefei 230009, China

[3] City University of Hong Kong, Kowloon, Hong Kong SAR, China

Published online: 06 March 2019

learning research community usually focuses on the generalized ability of reasoning algorithms and no prior knowledge is available in unsupervised learning. Moreover, there is not one clustering algorithm well capable of processing any distribution of data. Therefore, this study aims at investigating the different uniform effects of *k*-means clustering algorithm and FCM clustering algorithm.

In this study, we present a comparative study of the effect of skewed data distribution on *k*-means and FCM clustering. By means of the structural decomposition analysis of the objective functions, we formally illustrate that *k*-means and FCM have the similar uniform effect, i.e., both *k*-means and FCM tend to produce clusters with relatively uniform sizes, even if the input data cluster sizes are varied. However, the uniform effect of FCM is more complex. The fuzzifier parameter in FCM may have a significant effect on the clustering results. Based on the theoretical analysis, we conduct extensive experiments on both two-dimensional and three-dimensional synthetic data sets and UCI machine learning real-world data sets [36]. The experimental results reveal that, in most cases, FCM has stronger uniform effect than *k*-means, and the uniform effect changes with the change of the fuzzifier value. These findings have great significant in supporting the understanding of the performance and better applications of *k*-means and FCM clustering.

The remainder of this paper is organized as follows: Section 2 introduces the basic theory of clustering as well as the *k*-means and FCM clustering. Then, structure decomposition analyses of *k*-means and FCM clustering are presented in Sect. 3. Section 4 provides the experimental results on both synthetic and real-world data sets. Finally, conclusions are drawn in Sect. 5.

## 2 Related work

### 2.1 Clustering

Clustering is an unsupervised pattern recognition process. Its objective is to partition the data objects within a given data set into several clusters, such that the data objects in the same group are as similar as possible, while those in different groups are dissimilar [25, 37].

For a given data set $X = \{x_1, x_2, \ldots, x_n\}$, the $n$ data objects are divided into $c$ clusters, namely $(V_1, V_2, \ldots, V_c)$, and the partition matrix $U(X)$ is obtained. The partition matrix can be expressed as $U(X) = \left[\mu_{ij}\right]_{c\times n}$ $(i = 1, \ldots, c,\ j = 1, \ldots, n)$, where $\mu_{ij}$ is the membership degree of the data object $x_j$ to group $V_i$. Cluster $V_i$ and its corresponding cluster center $v_i$ are determined by:

$$\begin{cases} V_i = \left\{ x_j \middle| \left\| x_j - v_i \right\| \le \left\| x_j - v_p \right\|,\ x_j \in X \right\}, & p \neq i,\ p = 1, \ldots, c \\ v_i = \sum_{x_j \in V_i} x_j \Big/ |V_i|, & i = 1, \ldots, c \end{cases} \tag{1}$$

where $||\bullet||$ denotes the similarity measure between data objects. $|V_i|$ represents the number of data objects in cluster $V_i$.

The $c$ clusters of the clustering partitions satisfy:

$$\begin{cases} \cup_{i=1}^{c} V_i = X \\ V_i \cap V_j = \emptyset, & i, j = 1, \ldots, c;\ i \neq j \\ V_i \neq \emptyset, & i = 1, \ldots, c \end{cases} \tag{2}$$

Based on the different constraints of membership degree, hard (crisp) and fuzzy clustering can be, respectively, defined as [38]

$$M_{\mathrm{HCM}} = \left\{ U | \mu_{ij} \in \{0, 1\} \quad \forall i, j; \quad 0 < \sum_{j=1}^{n} \mu_{ij} < n \quad \forall i; \quad \sum_{i=1}^{c} \mu_{ij} = 1 \quad \forall j \right\} \tag{3}$$

$$M_{\mathrm{FCM}} = \left\{ U | \mu_{ij} \in [0, 1] \quad \forall i, j; \quad 0 < \sum_{j=1}^{n} \mu_{ij} < n \quad \forall i; \quad \sum_{i=1}^{c} \mu_{ij} = 1 \quad \forall j \right\} \tag{4}$$

For hard (crisp) clustering method, e.g., *k*-means, each data object can only be divided into one cluster with an exact membership degree value of 0 or 1. While for fuzzy clustering, e.g., FCM, each data object has a membership degree between 0 and 1 to all the clusters.

### 2.2 *k*-means clustering

*k*-means is a typical crisp clustering algorithm. It starts by determining a number of clusters $c$, followed by random determination of initial cluster centers. In the next steps, the cluster centers are updated by

$$v_i = \sum_{x_j \in V_i} \left( x_j / n_i \right) \tag{5}$$

where $x_j$ is the $j$th data object. $V_i$ is the $i$th cluster. $n_i$ is the number of data objects partitioned into cluster $V_i$.

The object function is calculated by:

$$J_{\mathrm{KM}}^{(c)} = \mathrm{SSE} = \sum_{i=1}^{c} \sum_{x_j \in V_i} \left\| x_j - v_i \right\|^2 \tag{6}$$

Iteration is terminated when it reaches the maximum number of iterations, or the change of the objective function values for two successive iterations is within a threshold.

### 2.3 Fuzzy *c*-means (FCM) clustering

Compared with *k*-means clustering, the fuzzifier parameter $m$ and the nonzero membership degree $\mu_{ij}$ are introduced in FCM clustering [25]. It also starts by determining an appropriate number of clusters $c$, followed by the random selection of initial cluster centers. Then, the initial membership degrees of each data object $x_j$ to cluster $v_i$ are calculated. The membership degree $\mu_{ij}$ and cluster center $v_i$ are updated, respectively, as follows.

$$\mu_{ij} = 1 \Big/ \sum_{r=1}^{c} [d(x_j, v_i)/d(x_j, v_r)]^{\frac{2}{m-1}} \tag{7}$$

$$v_i = \sum_{j=1}^{n} \mu_{ij}^m x_j \Big/ \sum_{j=1}^{n} \mu_{ij}^m \tag{8}$$

For each iteration, the objective function of FCM is calculated as follows:

$$J_{\text{FCM}}^{(c)} = \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij}^m ||x_j - v_i||^2 \tag{9}$$

where $v_i$ is the cluster center of cluster $V_i$, and $v_i = (1/n_i) \sum_{x_j \in C_i} x_j$. $n_i$ is the number of data objects in cluster $V_i$. $d(x_j, v_i)$ is the Euclidean distance of data object $x_j$ to cluster $V_i$. $m$ is the fuzzifier parameter.

### 2.4 Selection of fuzzifier value in FCM

Fuzzifier parameter $m$, a.k.a fuzziness parameter or weighting exponent, is an important parameter in FCM, which controls the extent of sharing among groups partitioned by FCM. The fuzzifier value can significantly affect the performance of FCM clustering [37]. From the definition of membership degree in Eq. (7), it should be noted that the cluster center closest to the point will be given much more weight than the others when the value of $m$ is close to 1. Also, when the value of $m$ tends to be infinite, each cluster center will be given an approximately equal membership degree since $\lim_{m\to\infty} \mu_{ij} = \lim_{m\to\infty} \left\{ 1/\sum_{r=1}^{c} [d(x_j, v_i)/d(x_j, v_r)]^{\frac{2}{m-1}} \right\} = 1/c$. Therefore, the value of $m$ cannot be too small or too large. Generally, the value of fuzzifier $m$ in FCM is set to 2.0, since this is equivalent to normalize the coefficient linearly to make their sum 1.0.

Currently, there have been some research efforts on the selection of fuzzifier in FCM. A list of research findings and suggestions on fuzzifier selection of FCM is presented in Table 1.

As can be seen from Table 1, there have been many different propositions on the selection of fuzzifier in FCM. The general and most widely used value is $m = 2$ in applications. However, there is still little theoretical guidance and generally accepted criterion to support the fuzzifier parameter selection in FCM [51, 52]. The fuzzifier value is generally selected by users subjectively in different applications, which can seriously affect the clustering results.

## 3 Structural analysis of *k*-means and FCM clustering

### 3.1 Structure decomposition of the objective function of *k*-means

Based on the definition of the objective function of *k*-means in Eq. (6), it has been proposed that *k*-means has the following propositions [33].

**Proposition 1** [33]. *Let* $d(V_p, V_q) = \sum_{x_j \in V_p} \sum_{x_l \in V_q} \left\| x_j - x_l \right\|^2$, *then*

**Table 1** Suggestions on fuzzifier selection of FCM

| Author(s) | Findings and suggestions on fuzzifier value | References |
|---|---|---|
| Shen et al.; Yang and Nataliani; Memon; Memon; Suleman | $m = 2$ | [39–43] |
| Bezdek; Janalipour and Mohammadzadeh | [1.1, 5.0] | [44, 45] |
| Ozkan and Turksen | [1.4, 2.6] | [46] |
| Wu | [1.5, 4.0] | [47] |
| Idri et al. | [1.5, 3.5] | [48, 49] |
| Chan and Cheung | [1.25, 1.75] | [50] |
| Zhou et al. | [2.5, 3.0] | [37] |

$$J_{\mathrm{KM}}^{(c)} = \sum_{i=1}^{c}\left(\frac{1}{2n_i}\sum_{x_j,x_l\in V_i}\left\|x_j - x_l\right\|^2\right) = \frac{1}{2}\sum_{i=1}^{c}\frac{d(V_i,V_i)}{n_i} \tag{10}$$

***Proof*** When we substitute $v_i = \sum_{x_j\in V_i}\left(x_j/n_i\right)$ to the objective function of $k$-means $J_{\mathrm{KM}}^{(c)}$, Eq. (10) can be obtained.

**Proposition 2** [33]. *The sum of all pairwise distances of data objects in c clusters can be expressed as follows:*

$$D_c = \sum_{j=1}^{n}\sum_{l=1}^{n}\left\|x_j - x_l\right\|^2 = \sum_{i=1}^{c}\left(\frac{n}{n_i}d(V_i,V_i)\right) + 2\sum_{1\le i<k\le c}\left(n_i n_k\left\|v_i - v_k\right\|^2\right) \tag{11}$$

***Proof*** Equation (3) can be proved by mathematical induction, which can be found in Ref. [33].

It is true that $D_c$, the sum of pairwise distances of all data objects, is a constant for a given data set, regardless of the number of clusters $c$ and the clustering result.

**Proposition 3** *The objective function of k-means can be written as*

$$J_{\mathrm{KM}}^{(c)} = \frac{D_c - 2\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right]}{2n} \tag{12}$$

***Proof*** If we substitute $J_{\mathrm{KM}}^{(c)}$ in Eq. (10) and $D_c$ in Eq. (11), we have

$$D_c = 2nJ_{KM}^{(c)} + 2\sum_{1\le i<k\le c}\left(n_i n_k\left\|v_i - v_k\right\|^2\right) \tag{13}$$

Then, we can know that Proposition 3 is true.

**Definition 1** If the distribution of cluster sizes generated by a clustering method has less variation than the underlying true distribution of cluster sizes, we say a uniform effect exists for the clustering method.

Based on Propositions 2 and 3, it can be seen that the minimization of the objective of $k$-means $J_{\mathrm{KM}}^{(c)}$ is equivalent to the maximization of $\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right]$, since $D_c$ and $n$ are constants and $\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right] \ge 0$. If the effect of $\left\|v_i - v_k\right\|^2$ is isolated, the maximization of $\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right]$ implies the maximization of $\sum_{1\le i\le k\le c} n_i n_k$, i.e., $n_1 = n_2 = \cdots = n_c$. Therefore, the $k$-means clustering has the uniform effect.

We also note that, in real-world situations, the interaction effect is more complex. We will further illustrate it by extensive experiments in Sect. 4.

### 3.2 Structure decomposition of the objective function of FCM

Based on the definition of the objective function of FCM in Eq. (9), there are the following propositions.

**Proposition 4** [35]. *The objective function of FCM can be rewritten as*

$$J_{\mathrm{FCM}}^{(c)} = \left\{\frac{D_c - 2\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right]}{2n}\right\}\sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m + \sum_{i=1}^{c}\sum_{x_j\notin V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2 \tag{14}$$

***Proof*** According to the original form of the objective function of FCM shown in Eq. (9), we have

$$J_{\mathrm{FCM}}^{(c)} = \sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2 + \sum_{i=1}^{c}\sum_{x_j\notin V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2 \tag{15}$$

Then, if we substitute $J_{\mathrm{KM}}^{(c)}$ in Eq. (6) and Eq. (12) to Eq. (15), we have

$$\begin{aligned} J_{\mathrm{FCM}}^{(c)} &= J_{\mathrm{KM}}^{(c)}\sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m + \sum_{i=1}^{c}\sum_{x_j\notin V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2 \\ &= \left\{\frac{D_c - 2\sum_{1\le i<k\le c}\left[n_i n_k\left\|v_i - v_k\right\|^2\right]}{2n}\right\}\sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m + \sum_{i=1}^{c}\sum_{x_j\notin V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2 \end{aligned} \tag{16}$$

In FCM clustering, $\mu_{ij}\in[0,1]$. For any $x_j\in V_i$, we have $1/c \le \mu_{ij} \le 1$. Then we have $n/c \le \sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m \le n$. Therefore, $1/c \le \sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m \Big/ n \le 1$. While for any $x_j\notin V_i$, $\mu_{ij}\le 1/c$. So the second term $\sum_{i=1}^{c}\sum_{x_j\notin V_i}\mu_{ij}^m\left\|x_j - v_i\right\|^2$ is significantly less than the first term of $J_{FCM}^{(c)}$ in Eq. (14).

To achieve the minimization of $J_{\mathrm{FCM}}^{(c)}$, when the influence of $\sum_{i=1}^{c}\sum_{x_j\in V_i}\mu_{ij}^m \Big/ n$ and $\left\|v_i - v_k\right\|^2$ is controlled, it must satisfy the maximization of $\sum_{1\le i\le k\le c} n_i n_k$, i.e.,

$n_1 = n_2 = \cdots = n_c$. However, compared with the decomposition result of the objective function of $k$-means shown in Eq. (12), the term $\sum_{i=1}^{c} \sum_{x_j \in V_i} \mu_{ij}^m$ is included in the objective function of FCM, i.e., in the left half part of Eq. (16). It is clear that this term is a positive value and it satisfies $n/c \leq \sum_{i=1}^{c} \sum_{x_j \in V_i} \mu_{ij}^m \leq n$. As a result, there is a higher slope in the cost function of the FCM clustering algorithm and its uniform effect occurs with fewer iterations. FCM is much easier to achieve the result of $n_1 = n_2 = \cdots = n_c$, so we can say FCM clustering is more likely to generate partitions with uniform cluster sizes. From this perspective, we can say FCM has stronger uniform effect than $k$-means clustering.

### 3.3 Measurement of dispersion in cluster sizes

In this study, we provide the coefficient of variation (CV) as an indicator to measure the dispersion of cluster sizes in a given data set. The CV indicator [33, 35, 53] is defined as:

$$\mathrm{CV} = \frac{s}{\bar{n}} = \frac{\sqrt{\sum_{i=1}^{c} \left(n_i - \bar{n}\right)/(c-1)}}{\sum_{i=1}^{c} n_i / c} \tag{17}$$

In Eq. (17), $s = \sqrt{\sum_{i=1}^{c} \left(n_i - \bar{n}\right)/(c-1)}$ is the standard deviation of cluster sizes. CV is a dimensionless indicator that allows for a comparison of the variation of populations that have significantly different mean values. In general, the larger the CV value is, the greater the variability is in the cluster size of a data set. Also, the CV indicator is simple in calculation. Due to its advantages, CV has become a popular statistical indicator that has been widely used in measuring the dispersion of a data distribution [33–35, 54, 55].

Based on the change of CV values after a clustering process, we can quantitatively measure the effect of data distribution on the $k$-means and FCM clustering. DCV is the difference of CV value before and after clustering, which is defined as:

$$\mathrm{DCV} = \left|\mathrm{CV}_0 - \mathrm{CV}_1\right| \tag{18}$$

DCV is the absolute value of the difference between $\mathrm{CV}_0$ and $\mathrm{CV}_1$. If $\mathrm{CV}_0 > \mathrm{CV}_1$, it implies that the cluster sizes become more uniform after clustering. Therefore, the larger the DCV value is, the stronger the uniform effect is. We note that "uniform effect" of clustering algorithm is a negative effect, since the clustering partition has more deviations from the real distribution.

In this study, we use data_$d$_$c$_$n$ to name a synthetic data set, in which $d$ indicates the number of attributes of the data set. $c$ is the number of "true" classes. $n$ refers to the total number instances in the data set. Here, we give an illustrative example, and the data set used is data_2_2_880, which is shown in Fig. 1.

As Fig. 1 shows, data_2_2_880 is a two-dimensional data set, which composes two clusters and 880 data objects. The sizes of the two clusters in data_2_2_880 are varied. There are 800 data objects in cluster 1 and 80 data objects in cluster 2.

The clustering results of data_2_2_880 using $k$-means and FCM with fuzzifier $m = 2$ are shown in Fig. 2.

The cluster sizes of original data set and different clustering partitions of $k$-means and FCM are summarized in Table 2.

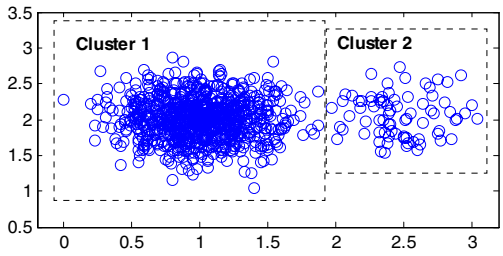
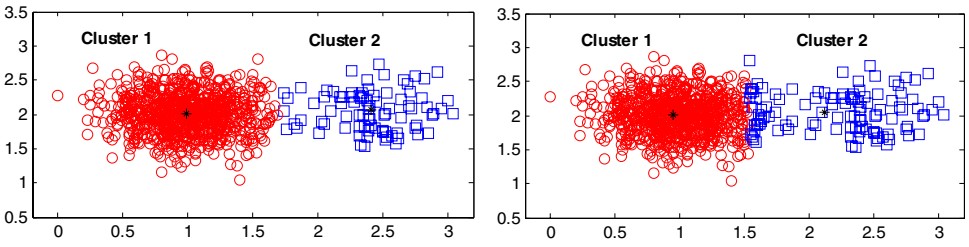



**Fig. 1** Data_2_2_880

**Table 2** Cluster sizes of data_2_2_880 and different clustering partitions

| | Cluster 1 | Cluster 2 | *#inst* misclassified |
|---|---|---|---|
| Original data set | 800 | 80 | 0 |
| $k$-means clustering | 792 | 88 | 8 |
| FCM clustering ($m = 2.0$) | 769 | 111 | 31 |

**Fig. 2** Clustering results of data_2_2_880




**(a)** K-means **(b)** FCM ($m$=2)

As shown in Fig. 2 and Table 2, there are some instances misclassified for both *k*-means and FCM clustering on data_2_2_880. After clustering, the size of the smaller cluster, i.e., cluster 2 in data_2_2_880, tends to be larger, such that the cluster sizes can be more uniform. Also, the results show that FCM has stronger uniform effect than *k*-means on data_2_2_880. Namely, through FCM, more data objects in cluster 2 are misclassified into cluster 1 such that the clustering partitions become more uniform.

In the empirical study section, more extensive experiments will be provided to understand the effect of data distribution on clustering results and the different uniform effects of *k*-means and FCM clustering.

## 4 Experiments

### 4.1 Experimental setup

We conduct experiments over synthetic and real-world data sets to evaluate the effect of data distribution on clustering. All algorithms are implemented in MATLAB R2013a and run in Windows 7 operating system on an Intel(R) Core(TM) i5-4590 CPU @ 3.30 GHz with 4 GB RAM.

Due to the fact that both *k*-means and FCM are partition-based clustering algorithms, they are sensitive to the initial cluster centers. To alleviate the influence of the random selection of initial cluster centers, the clustering process on each of the data sets was implemented 50 times, and the average results are used.

In measuring the dispersion of data distribution with CV indicator, we used $CV_0$ to indicate the distribution of underlying true cluster sizes for a given data set and $CV_1$ to measure the distribution of resultant cluster sizes after clustering.

### 4.2 Data sets

#### 4.2.1 Synthetic data sets

To intuitively observe the clustering result and the influence of data distribution on clustering, we used eight synthetic data sets in the experiments, including both two-dimensional

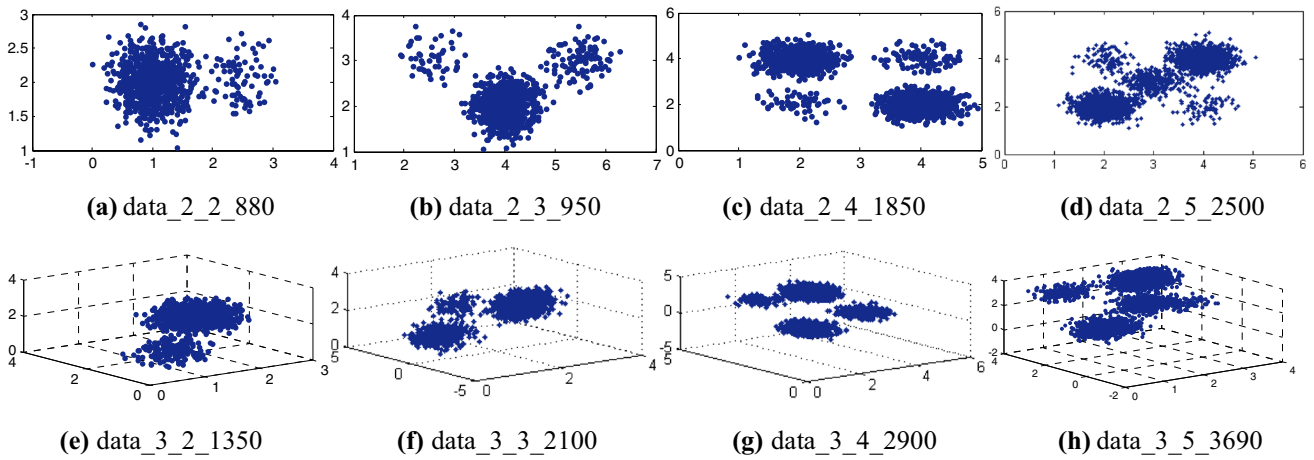



**Fig. 3** Original distribution of eight synthetic data sets

**Table 3** Characteristics of synthetic data sets

| Data set | *#attr* | *#cls* | *#inst* | MinSize | MaxSize | $CV_0$ | Cluster centers |
|---|---|---|---|---|---|---|---|
| data_2_2_880 | 2 | 2 | 880 | 80 | 800 | 1.1571 | (1, 2), (2.5, 2) |
| data_2_3_950 | 2 | 3 | 950 | 50 | 800 | 1.3242 | (4, 2), (2.5, 3), (5.5, 3) |
| data_2_4_1850 | 2 | 4 | 1850 | 60 | 900 | 0.9718 | (2, 2), (4, 4), (2, 4), (4, 2) |
| data_2_5_2500 | 2 | 5 | 2500 | 80 | 1100 | 0.9428 | (2, 2), (2, 4), (4, 2), (4, 4), (3, 3) |
| data_3_2_1350 | 3 | 2 | 1350 | 150 | 1200 | 1.0999 | (1, 1, 1), (2, 2, 2) |
| data_3_3_2100 | 3 | 3 | 2100 | 100 | 1500 | 1.0302 | (1, 1, 1), (3, 1, 2), (2, 3, 2) |
| data_3_4_2900 | 3 | 4 | 2900 | 100 | 1500 | 0.8895 | (1, 1, 1), (1, 3, 3), (3, 1, 2), (3, 3, 3) |
| data_3_5_3690 | 3 | 5 | 3690 | 90 | 1800 | 0.9641 | (1, 1, 1), (3, 3, 3), (1, 3, 3), (3, 1, 2), (3, 3, 1) |

and three-dimensional data sets. Each cluster of the synthetic data sets is randomly generated by the MATLAB function *nngenc*. The nngenc function is defined as *nngenc* (bounds, clusters, points, std_dev), in which bounds parameter refers to the range of generated points in each cluster, clusters parameter is the number of generated clusters, points parameter means the number of points in each cluster and std_dev parameter controls the standard deviation of the generated points to each cluster.

The generated synthetic data sets are shown in Fig. 3.

Table 3 presents the characteristics of the eight synthetic data sets.

All of the eight synthetic data sets have large $CV_0$ values, which imply that the cluster sizes are varied for all of the synthetic data sets.

#### 4.2.2 Real-world data sets

To further explore the uniform effect of *k*-means and FCM, as well as the comparative analysis of the effect of data distribution on *k*-means and FCM clustering, we also used six real-world data sets from the UCI machine learning repository [36] in the experiment. The characteristics of the real-world data sets are provided in Table 4.

Table 4 shows that all of these real-world data sets are high-dimensional data, and their cluster sizes have great variation.

### 4.3 Results and discussion

Under the above experimental setup, and with the above synthetic and real-world data sets, we conduct extensive experiments and provide the experimental results and discussion in this section.

Table 5 provides the basic characteristics of all the experimental data sets before and after *k*-means clustering.

As can be seen from Table 5, after *k*-means clustering, the minimum cluster size becomes large, and the maximum cluster size become small for all the data sets excluding data_3_3_2100, data_3_4_2900, data_3_5_3690, glass identification and page blocks. Therefore, *k*-means clustering has the uniform effect. But the uniform effect of *k*-means is not so strong since the DCV values are not so large.

**Table 4** Characteristics of real-world data sets

| Data set | *#attr* | *#cls* | *#inst* | MinSize | MaxSize | $CV_0$ |
|---|---|---|---|---|---|---|
| Glass identification | 9 | 6 | 214 | 9 | 76 | 0.8339 |
| Yeast | 8 | 10 | 1484 | 5 | 463 | 1.1676 |
| Page blocks | 10 | 5 | 5473 | 28 | 4913 | 1.9528 |
| Breast cancer wisconsin | 10 | 8 | 699 | 17 | 367 | 1.3179 |
| Ecoli | 7 | 8 | 336 | 2 | 143 | 1.1604 |
| Balance scale | 4 | 3 | 625 | 49 | 288 | 0.6623 |

**Table 5** Characteristics of all the experimental data sets before and after *k*-means clustering

| | Before clustering | | | *k*-means clustering | | | DCV |
|---|---|---|---|---|---|---|---|
| | MinSize | MaxSize | $CV_0$ | MinSize | MaxSize | $CV_1$ | |
| *Synthetic data sets* | | | | | | | |
| data_2_2_880 | 80 | 800 | 1.1571 | 85 | 795 | 1.1410 | 0.0161 |
| data_2_3_950 | 50 | 800 | 1.3242 | 50 | 799 | 1.3215 | 0.0027 |
| data_2_4_1850 | 60 | 900 | 0.9718 | 61 | 899 | 0.9705 | 0.0013 |
| data_2_5_2500 | 80 | 1100 | 0.9428 | 82 | 1091 | 0.9331 | 0.0097 |
| data_3_2_1350 | 150 | 1200 | 1.0999 | 156 | 1194 | 1.0874 | 0.0125 |
| data_3_3_2100 | 100 | 1500 | 1.0302 | 100 | 1501 | 1.0311 | 0.0009 |
| data_3_4_2900 | 100 | 1500 | 0.8895 | 100 | 1500 | 0.8895 | 0 |
| data_3_5_3690 | 90 | 1800 | 0.9641 | 90 | 1799 | 0.9633 | 0.0008 |
| *Real-world data sets* | | | | | | | |
| Glass identification | 9 | 76 | 0.8339 | 5 | 124 | 0.8348 | 0.0009 |
| Yeast | 5 | 463 | 1.1676 | 15 | 302 | 0.5486 | 0.6190 |
| Page blocks | 28 | 4913 | 1.9528 | 9 | 4422 | 1.7224 | 0.2304 |
| Breast cancer wisconsin | 17 | 367 | 1.3179 | 28 | 251 | 0.9960 | 0.3219 |
| Ecoli | 2 | 143 | 1.1604 | 10 | 62 | 0.4786 | 0.6818 |
| Balance scale | 49 | 288 | 0.6623 | 175 | 240 | 0.1325 | 0.5298 |

**Table 6** Characteristics of all the experimental data sets before and after FCM clustering ($m=2.0$)

| | Before clustering | | | FCM clustering ($m=2.0$) | | | DCV |
|---|---|---|---|---|---|---|---|
| | MinSize | MaxSize | $CV_0$ | MinSize | MaxSize | $CV_1$ | |
| *Synthetic data sets* | | | | | | | |
| data_2_2_880 | 80 | 800 | 1.1571 | 111 | 769 | 1.0574 | 0.0997 |
| data_2_3_950 | 50 | 800 | 1.3242 | 101 | 565 | 0.7381 | 0.5861 |
| data_2_4_1850 | 60 | 900 | 0.9718 | 384 | 522 | 0.1265 | 0.8453 |
| data_2_5_2500 | 80 | 1100 | 0.9428 | 393 | 582 | 0.1424 | 0.8004 |
| data_3_2_1350 | 150 | 1200 | 1.0999 | 163 | 1187 | 1.0727 | 0.0272 |
| data_3_3_2100 | 100 | 1500 | 1.0302 | 501 | 958 | 0.3345 | 0.6957 |
| data_3_4_2900 | 100 | 1500 | 0.8895 | 300 | 1000 | 0.4328 | 0.4567 |
| data_3_5_3690 | 90 | 1800 | 0.9641 | 90 | 1799 | 0.9633 | 0.0008 |
| *Real-world data sets* | | | | | | | |
| Glass identification | 9 | 76 | 0.8339 | 7 | 66 | 0.6541 | 0.1798 |
| Yeast | 5 | 463 | 1.1676 | 43 | 433 | 0.8437 | 0.3239 |
| Page blocks | 28 | 4913 | 1.9528 | 7 | 4362 | 1.6929 | 0.2599 |
| Breast cancer wisconsin | 17 | 367 | 1.3179 | 25 | 188 | 0.7460 | 0.5719 |
| Ecoli | 2 | 143 | 1.1604 | 27 | 59 | 0.2262 | 0.9342 |
| Balance scale | 49 | 288 | 0.6623 | 165 | 259 | 0.2149 | 0.4474 |

In comparison, the clustering results of FCM with fuzzifier value $m=2.0$ are summarized in Table 6.

Table 6 shows that after FCM clustering, the minimum cluster size becomes large, and the maximum cluster size becomes small for all the data sets excluding data_3_5_3690, glass identification and page blocks. Moreover, most of the positive DCV values of FCM ($m=2.0$) are larger than those of *k*-means. It demonstrates that FCM ($m=2.0$) has stronger uniform effect than *k*-means.

To further explore the influence of fuzzifier parameter on the uniform effect of FCM, we also implemented FCM clustering over the synthetic and real-world data sets with different fuzzifier values, ranging from 1.5 to 5.0.

The clustering partitions of *k*-means and FCM clustering with different fuzzifier values ($m=1.5$, 2.0, 2.5, 3.0, and 3.5) over the two-dimensional and three-dimensional synthetic data sets are shown in Figs. 4 and 5, respectively.

As shown in Figs. 4 and 5, compared with *k*-means, more data objects are misclassified after FCM clustering such that the cluster sizes become more uniform. FCM clustering has stronger uniform effect and it becomes significantly serious with the increase in fuzzifier value.

The changes of DCV values of *k*-means and FCM with different fuzzifier values on the synthetic data sets and real-world data sets are presented in Figs. 6 and 7, respectively.

As illustrated in Fig. 6, the DCV values of FCM with fuzzifier $m=1.5$ are slightly larger than *k*-means clustering over the eight synthetic data sets. However, when *m* is equal to 2.0 or greater, the DCV values of FCM clustering are significantly larger than those of *k*-means. Figure 7 shows that the performance of FCM clustering is more complex, with the change of fuzzifier values. Nevertheless, in most cases, FCM clustering with fuzzifier $m=1.5$ or 2.0 tends to produce more uniform clusters, compared with *k*-means clustering. Therefore, it demonstrated that FCM has stronger uniform effect than *k*-means. Moreover, it also suggested that $m=2.0$ is not a good choice for FCM clustering over data sets with great variations in cluster distributions.

## 5 Conclusions

In this study, we focused on the effect of data distribution on clustering and presented a comparative analysis of *k*-means and FCM clustering. Through the structure decomposition analysis of the objective functions of *k*-means and FCM algorithm, a theoretical comparison of the uniform effect of *k*-means and FCM was carried out. The extensive experiments on both synthetic data sets and real-world data sets demonstrated that FCM has stronger uniform effect than *k*-means. It demonstrated that FCM has a higher slope in the cost function and its uniform effect occurs with fewer iterations. FCM is more likely to generate partitions with uniform cluster sizes. Further, the fuzzifier parameter in FCM is a critical parameter. The value of fuzzifier parameter *m* cannot be too small or too large, since it degrades into hard *k*-means when $m=1.0$, and each cluster center has equivalent membership degree when $m \to \infty$. Due to the fact that some machine learning problems are black-box modeling in which we do not have prior knowledge in advance, and some parameters in unsupervised learning are determined by the data-driven approach. The data-based experiments in many studies have shown that $m=2.0$ is not a good choice for FCM.

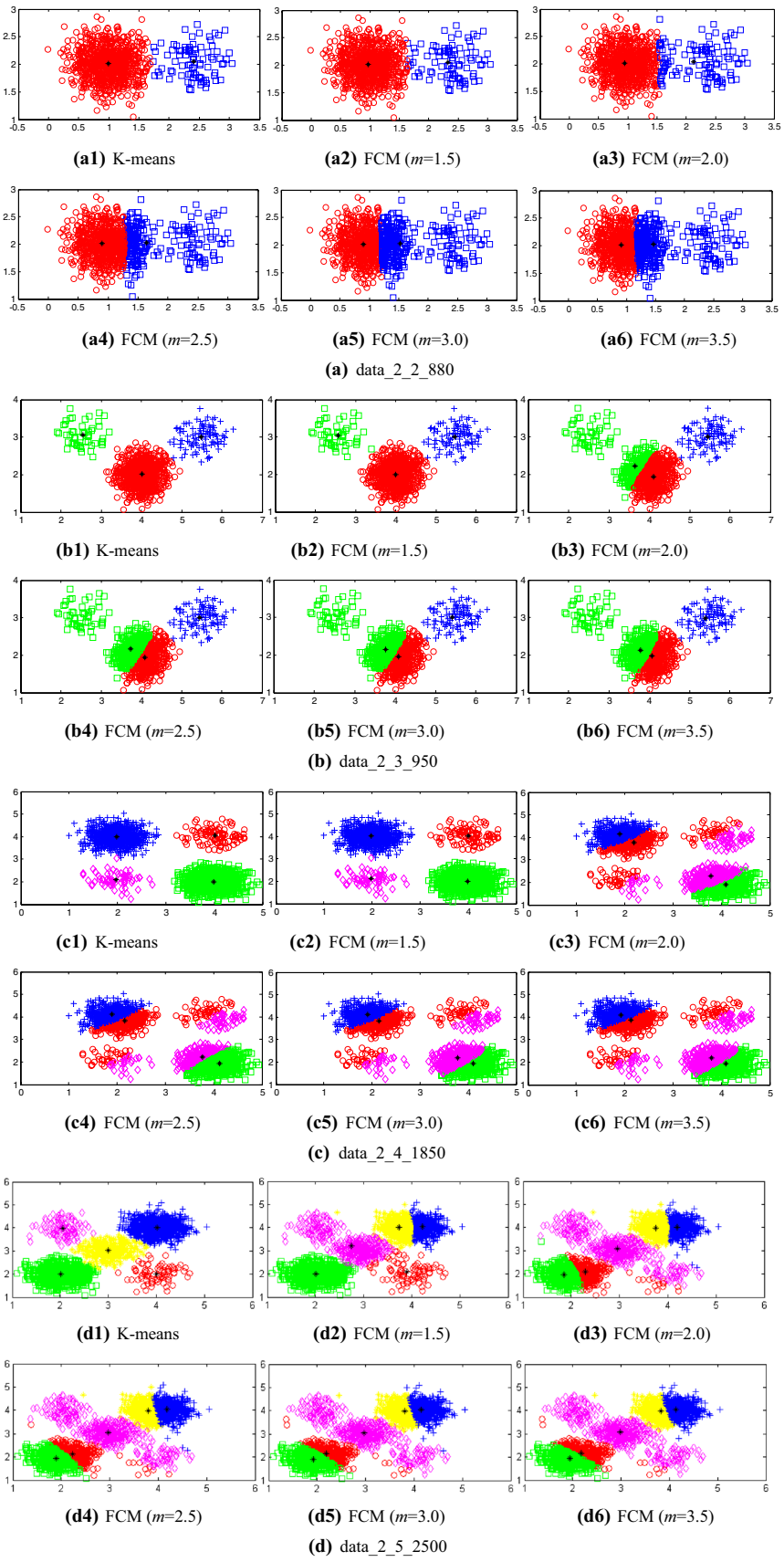
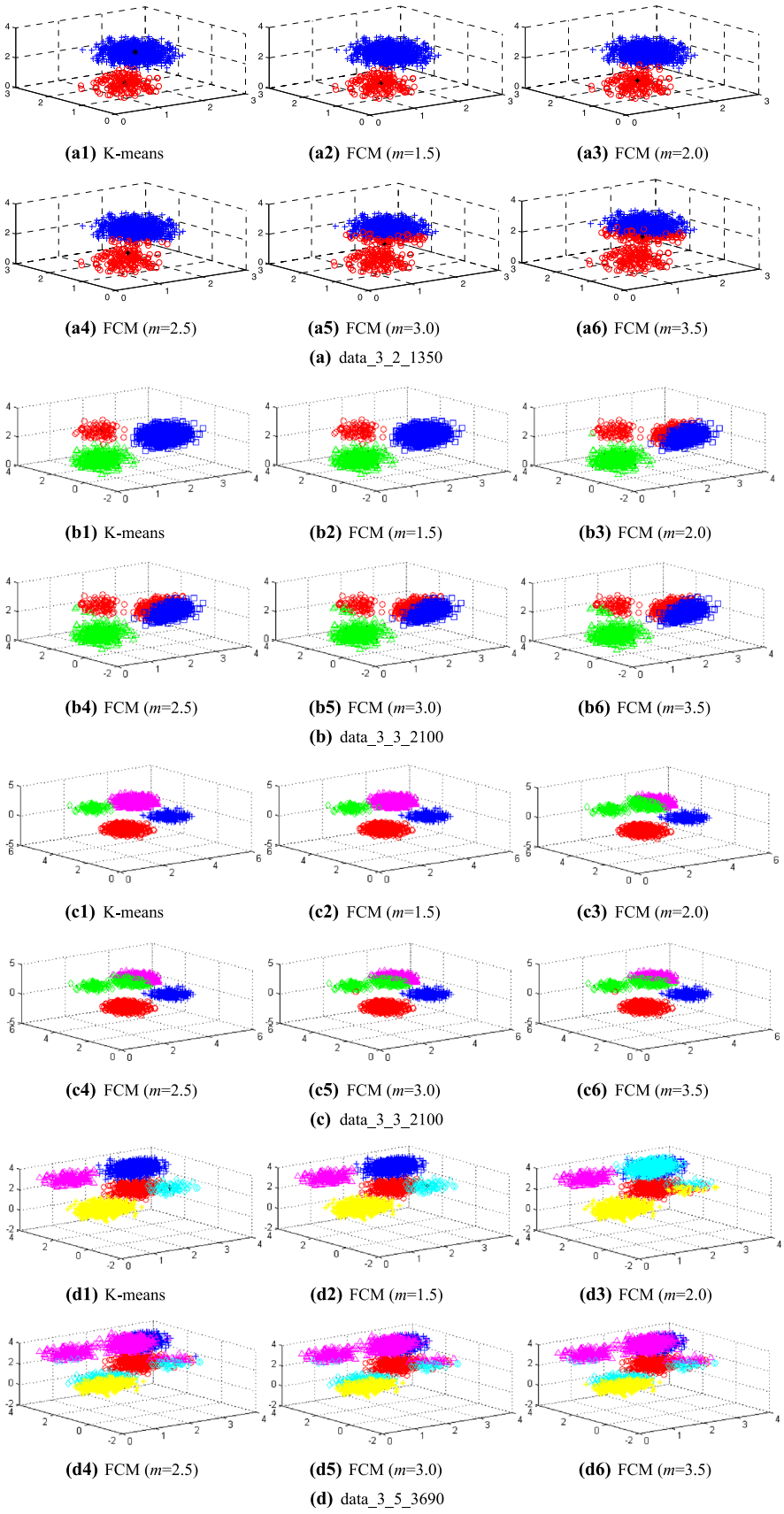



**Fig. 4** Clustering partitions of two-dimensional synthetic data sets



**(a1)** K-means **(a2)** FCM ($m$=1.5) **(a3)** FCM ($m$=2.0)

**(a4)** FCM ($m$=2.5) **(a5)** FCM ($m$=3.0) **(a6)** FCM ($m$=3.5)

**(a)** data_3_2_1350

**(b1)** K-means **(b2)** FCM ($m$=1.5) **(b3)** FCM ($m$=2.0)

**(b4)** FCM ($m$=2.5) **(b5)** FCM ($m$=3.0) **(b6)** FCM ($m$=3.5)

**(b)** data_3_3_2100

**(c1)** K-means **(c2)** FCM ($m$=1.5) **(c3)** FCM ($m$=2.0)

**(c4)** FCM ($m$=2.5) **(c5)** FCM ($m$=3.0) **(c6)** FCM ($m$=3.5)

**(c)** data_3_3_2100

**(d1)** K-means **(d2)** FCM ($m$=1.5) **(d3)** FCM ($m$=2.0)

**(d4)** FCM ($m$=2.5) **(d5)** FCM ($m$=3.0) **(d6)** FCM ($m$=3.5)

**(d)** data_3_5_3690

**Fig. 5** Clustering partitions of three-dimensional synthetic data sets

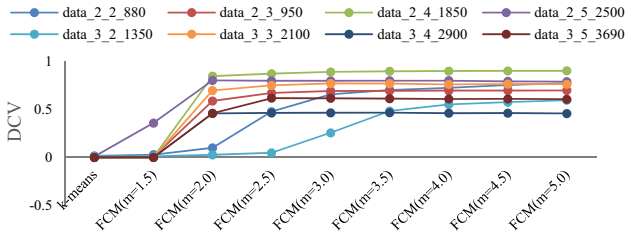



**Fig. 6** DCV values of *k*-means and FCM with different fuzzifier values on synthetic data sets

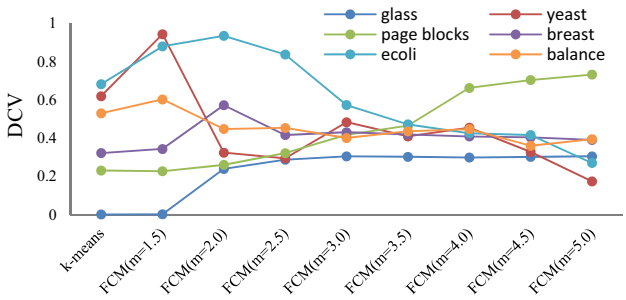



**Fig. 7** DCV values of *k*-means and FCM with different fuzzifier values on real-world data sets

From the data distribution perspective, we also illustrated that $m = 2.0$ for FCM is not a good choice, particularly for data sets with great variation in cluster sizes. Therefore, we suggest that, for data sets with significant uneven distributions, *k*-means clustering is a better choice compared with FCM clustering. When FCM clustering is carried out on this kind of data sets, it is suggested that a smaller fuzzifier value is better.

This study has some limitations. We note that data distribution not only refers to the distribution of cluster sizes. It also means the densities, shapes and other characteristics of a data set. In addition, we understand that clustering is an unsupervised learning process and usually no prior knowledge is available before clustering. In this paper, we mainly focused on investigating the comparison of uniform effect of *k*-means and FCM clustering. Further investigation of the principle of unsupervised clustering has beyond the scope of this study. In our future work, we will conduct more in-depth research on *k*-means and FCM clustering.

**Acknowledgements** The authors would like to thank the anonymous reviewers very much for their valuable comments and suggestions for improving the quality of the paper. This work was supported by the National Natural Science Foundation of China under Grant Nos. 71822104, 71501056 and 71690235, Anhui Science and Technology Major Project under Grant No. 17030901024, China Postdoctoral Science Foundation under Grant No. 2017M612072, and Hong Kong Scholars Program under Grant No. 2017-167.

## References

1. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall Inc., Upper Saddle River
2. Bianchi FM, Livi L, Rizzi A (2015) Two density-based *k*-means initialization algorithms for non-metric data clustering. Pattern Anal Appl 19:1–19
3. Bianchi FM, Livi L, Rizzi A (2016) Two density-based *k*-means initialization algorithms for non-metric data clustering. Pattern Anal Appl 19:745–763
4. Xu R, Wunsch D (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16:645–678
5. Xu YJ, Wu XJ (2016) An affine subspace clustering algorithm based on ridge regression. Pattern Anal Appl 20:557–566
6. Cornuéjols A, Wemmert C, Gançarski P, Bennani Y (2018) Collaborative clustering: why, when, what and how. Inf Fusion 39:81–95
7. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, pp 281–297
8. Gerlhof C, Kemper A, Kilger C, Moerkotte G (1993) Partition-based clustering in object bases: from theory to practice. In: International conference on foundations of data organization and algorithms. Springer, pp 301–316
9. Guha S, Rastogi R, Shim K (2001) CURE: an efficient clustering algorithm for large databases. Inf Syst 26:35–58
10. Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32:241–254
11. Karypis G, Han E-H, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer 32:68–75
12. Hinneburg A, Keim DA (1998) An efficient approach to clustering in large multimedia databases with noise. In: KDD 1998, pp 58–65
13. Sheikholeslami G, Chatterjee S, Zhang A (1998) Wavecluster: a multi-resolution clustering approach for very large spatial databases. In: VLDB 1998. pp 428–439
14. Liao W, Liu Y, Choudhary A (2004) A grid-based clustering algorithm using adaptive mesh refinement. In: 7th workshop on mining scientific and engineering datasets of SIAM international conference on data mining, pp 61–69
15. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B (Methodol) 39:1–38
16. Chen LS, Prentice RL, Wang P (2014) A penalized EM algorithm incorporating missing data mechanism for Gaussian parameter estimation. Biometrics 70:312–322
17. De Carvalho FDA, Lechevallier Y, De Melo FM (2012) Partitioning hard clustering algorithms based on multiple dissimilarity matrices. Pattern Recognit 45:447–464
18. Tîrnăucă C, Gómez-Pérez D, Balcázar JL, Montaña JL (2018) Global optimality in *k*-means clustering. Inf Sci 439–440:79–94
19. Ferreira MRP, de Carvalho FAT, Simões EC (2016) Kernel-based hard clustering methods with kernelization of the metric and automatic weighting of the variables. Pattern Recognit 51:310–321
20. Yang M-S (1993) A survey of fuzzy clustering. Math Comput Model 18:1–16
21. Sert SA, Bagci H, Yazici A (2015) MOFCA: multi-objective fuzzy clustering algorithm for wireless sensor networks. Appl Soft Comput 30:151–165
22. Bonis T, Oudot S (2018) A fuzzy clustering algorithm for the mode-seeking framework. Pattern Recognit Lett 102:37–43

23. Jothi R, Mohanty SK, Ojha A (2017) DK-means: a deterministic *k*-means clustering algorithm for gene expression analysis. Pattern Anal Appl. https://doi.org/10.1007/s10044-017-0673-0
24. Aparajeeta J, Nanda PK, Das N (2016) Modified possibilistic fuzzy *c*-means algorithms for segmentation of magnetic resonance image. Appl Soft Comput 41:104–119
25. Zhou K, Yang S, Shao Z (2017) Household monthly electricity consumption pattern mining: a fuzzy clustering-based model and a case study. J Clean Prod 141:900–908
26. Bigdeli E, Mohammadi M, Raahemi B, Matwin S (2017) A fast and noise resilient cluster-based anomaly detection. Pattern Anal Appl 20:183–199
27. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR et al (2015) Comprehensive assessment of cancer missense mutation clustering in protein structures. Proc Natl Acad Sci 112:E5486–E5495
28. Chifu A-G, Hristea F, Mothe J, Popescu M (2015) Word sense discrimination in information retrieval: a spectral clustering-based approach. Inf Process Manag 51:16–31
29. Kumar KM, Reddy ARM (2017) An efficient *k*-means clustering filtering algorithm using density based initial cluster centers. Inf Sci 418–419:286–301
30. Rodríguez J, Medina-Pérez MA, Gutierrez-Rodríguez AE, Monroy R, Terashima-Marín H (2018) Cluster validation using an ensemble of supervised classifiers. Knowl Based Syst 145:134–144
31. Farcomeni A (2014) Robust constrained clustering in presence of entry-wise outliers. Technometrics 56:102–111
32. Tan PN, Steinbach M, Kumar V (2005) Introduction to data mining. Addison-Wesley, Reading
33. Xiong H, Wu J, Chen J (2009) *k*-means clustering versus validation measures: a data-distribution perspective. IEEE Trans Syst Man Cybern Part B (Cybern) 39:318–331
34. Wu J, Xiong H, Chen J (2009) Towards understanding hierarchical clustering: a data distribution perspective. Neurocomputing 72:2319–2330
35. Zhou K, Yang S (2016) Exploring the uniform effect of FCM clustering: a data distribution perspective. Knowl Based Syst 96:76–83
36. Lichman M (2013) UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml. Accessed July 2017
37. Zhou K, Fu C, Yang S (2014) Fuzziness parameter selection in fuzzy *c*-means: the perspective of cluster validation. Sci China Inf Sci 57:1–8
38. Sledge IJ, Bezdek JC, Havens TC, Keller JM (2010) Relational generalizations of cluster validity indices. IEEE Trans Fuzzy Syst 18:771–786
39. Shen Y, Shi H, Zhang JQ (2000) Improvement and optimization of a fuzzy *c*-means clustering algorithm. Syst Eng Electron 3:1430–1433
40. Yang MS, Nataliani Y (2017) Robust-learning fuzzy *c*-means clustering algorithm with unknown number of clusters. Pattern Recognit 71:45–59
41. Martino FD, Sessa S (2018) Extended fuzzy *c*-means hotspot detection method for large and very large event datasets. Inf Sci 441:198–215
42. Memon KH (2018) A histogram approach for determining fuzzifier values of interval type-2 fuzzy *c*-means. Expert Syst Appl 91:27–35
43. Suleman A (2017) Measuring the congruence of fuzzy partitions in fuzzy *c*-means clustering. Appl Soft Comput 52:1285–1295
44. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
45. Janalipour M, Mohammadzadeh A (2017) Evaluation of effectiveness of three fuzzy systems and three texture extraction methods for building damage detection from post-event LiDAR data. Int J Digit Earth 12:1241–1268
46. Ozkan I, Turksen IB (2007) Upper and lower values for the level of fuzziness in FCM. Inf Sci 177:5143–5152
47. Wu KL (2012) Analysis of parameter selections for fuzzy *c*-means. Pattern Recognit 45:407–415
48. Idri A, Hosni M, Abran A (2016) Improved estimation of software development effort using classical and fuzzy analogy ensembles. Appl Soft Comput 49:990–1019
49. Idri A, Abnane I, Abran A (2017) Evaluating Pred(p) and standardized accuracy criteria in software development effort estimation. J Softw Evol Process 9:9. https://doi.org/10.1002/smr.1925
50. Chan KP, Cheung YS (1992) Clustering of clusters. Pattern Recognit 25:211–217
51. Pal NR, Bezdek JC (1995) On cluster validity for the fuzzy *c*-mean model. IEEE Trans Fuzzy Syst 3:370–379
52. Yu J, Cheng Q, Huang H (2004) Analysis of the weighting exponent in the FCM. IEEE Trans Syst Man Cybern B Cybern 34:634–639
53. Dacunha-Castelle D, Duflo M (1986) Probability and statistics. Springer, New York
54. Wu J, Xiong H, Chen J (2009) Adapting the right measures for *k*-means clustering. In: ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France, June 28–July 2009, pp 877–886
55. Wu J, Xiong H, Wu P, Chen J (2007) Local decomposition for rare class analysis. In: ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, California, USA, Aug 2007, pp 191–220