Pattern Recognition 139 (2023) 109404

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

An Effective and Adaptable K-means Algorithm for Big Data Cluster Analysis

Haize Hu[a], Jianxun Liu[a,*], Xiangping Zhang[a], Mengge Fang[b]

[a] School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, Hunan 411100, China
[b] State Grid Yiyang Power Supply Company, Yiyang, Hunan 413000, China

ARTICLE INFO

Article history:
Received 10 April 2022
Revised 20 January 2023
Accepted 5 February 2023
Available online 12 February 2023

Keywords:
K-means algorithm
Local optimization
Lévy flight
Global search
Clustering centroids

ABSTRACT

Tradition *K*-means clustering algorithm is easy to fall into local optimum, poor clustering effect on large capacity data and uneven distribution of clustering centroids. To solve these problems, a novel *k*-means clustering algorithm based on Lévy flight trajectory (Lk-means) is proposed in the paper. In the iterative process of LK-means algorithm, Lévy flight is used to search new positions to avoid premature convergence in clustering. It is also applied to increase the diversity of the cluster, strengthen the global search ability of *K*-means algorithm, and avoid falling into the local optimal value too early. Nevertheless, the complexity of hybrid algorithm is not increased in the process of Lévy flight optimization. To verify the data clustering effect of LK-means algorithm, experiments are conducted to compare it with the *k*-means algorithm, XK-means algorithm, DDKmeans algorithm and Canopyk-means algorithm on 10 open source data sets. The results show that LK-means algorithm has better search results and more evenly distributed cluster centroids, which greatly improves the global search ability, big data processing ability and uneven distribution centroids of cluster of *K*-means algorithm.

1. Introduction

In the era of big data, hundreds of millions of data is generated every day. How to effectively manage the accurate clustering of the generated data is crucial, and therefore an effective clustering algorithm can be well applied to enhance big data analysis. At present, a variety of optimized clustering algorithms emerge in endlessly, and the application of clustering algorithm has never stopped. For example, clustering algorithm is applied to graph processing [1], pattern recognition [2], spectral density analysis [3], wind farm data integration [4], climate prediction [5], and so on. The main function of clustering analysis is to aggregate the data with high similarity and separate the data with low similarity, which is called data structure splitting. The primary task of data clustering research is to divide the unlabeled $N \times D$ data set (N is the number of samples, D is the data dimension) into k groups according to the similarity, which can be well handled in general algorithms for low dimensional data sets. As a common unsupervised machine learning algorithm, clustering is widely used in data mining and data analysis. However, it is difficult for traditional methods to efficiently classify large capacity data and to implement in high-dimensional data sets. Therefore, many researchers tend to focus on designing a clustering algorithm with higher stability and smaller deviation.

Although clustering algorithm has been developed for decades, *K*-means algorithm is still widely used due to its simple principle, convenience and high efficiency. For example, Ran X et al proposed an artificial constrained *k*-means algorithm for GPS automatic lane detection [6], as well as processing data of virtual driving scene [7]. However, initial *k*-means algorithm has a big problem, i.e., being easy to converge prematurely and then falling into local optimum. To solve the problem, *k*-means algorithm has been continuously optimized by combining it with other algorithm. sunch as, the hybrid algorithm of particle swarm optimization (PSO) and CNAK [8,9], Cuckoo algorithm (CS), Particle Swarm Optimization (PSO) and *K*-means (HCSPSO) [10]. In addition, there are some *k*-means optimization and application studies, for example, Ghadiri et al. [11] proposed a stable deep *k*-means clustering algorithm by combining implicit low-level attribute representations. Song et al. [12] used a bilateral weighted optimization *k*-means clustering algorithm and applied the algorithm to copolymerization and fast

* Corresponding author.
E-mail address: ljx529@gmail.com (J. Liu).

https://doi.org/10.1016/j.patcog.2023.109404
0031-3203/

spectral clustering studies. Zhu et al. [13] used similarity matrix, spectral representation and transformation matrix mind for joint embedding to optimize *k*-means and applied it in spectral analysis. Nie et al. [14] clustered the rows and columns of the data matrix separately and proposed a collaborative clustering heart method FMVBK. Kang et al. [15] optimized from global and local structures and proposed a graph learning based *k*-means optimization method to preserve the integrity of the data. Ma et al. [16] used deep random forest algorithm for *k*-means and proposed a discrete *k*-means clustering algorithm and applied it to the study of price prediction. Zhao et al. [17] used Multi-View Visual Words (MVVW) to fuse feature information in the feature matrix and proposed a joint transfer constrained *k*-means clustering method.

However, compared with the original *K*-means algorithm, optimized algorithm's complexity, space and computation are increased. The others are to enhance the algorithm's search ability, for example, Firefly Algorithm (FA) from the inside out enhanced search and composite enhanced search [18] are adopted, or brain storm optimization algorithm and *K*-means algorithm are combined to enhance the global search ability [19]. Another form is partition iteration, a hybrid approach which can speed up the *k*-means clustering proposed by Borlea et al. [20]. To speed up the clustering and avoid local optimization, the data can be divided into clusters of different sizes, but the complexity and memory cost are greatly increased.

Therefore, a novel clustering algorithm based on the position of cuckoo algorithm and Lévy flight trajectory is proposed in the paper, which is shortly called LK-means. The algorithm was inspired by the XK-means algorithm proposed by Yaying [21]. An exploratory vector was added to the vector of the cluster centroid, so as to jump out of the local optimum in the iterative process. At present, the optimization of *K*-means algorithm mainly focuses on two directions [22], one is the optimization of the initial cluster centroid *k*, the other is the optimization of the updated iterative process. The paper will mainly study the centroid iterative update, based on XK-means clustering algorithm, Lévy flight trajectory was added to the search vector to improve the flight trajectory and optimize the *K*-means clustering algorithm. Levi's flight step size satisfies a heavy tailed Levi's stable distribution, which is a random walk mode that follows Levi's distribution and alternates short distance long distance with occasionally. After several iterations, the distance of random walk tends to be stable. The use of Lévy flight step size enlarges the search range, so that the *k*-means algorithm can jump out of the local optimum and avoid falling into the local optimum. The small step length in the later stage can make the group converge the global optimal solution in a small range. To validate the clustering efficiency of LK-means, multiple benchmark models are used for comparative analysis to evaluate the advantages of the algorithm in terms of both functionality and different data sets.

Our main contributions are as follows:

1. We propose the LK-means clustering algorithm, which has a higher search capability of the clustering algorithm and guarantees the convergence speed of the algorithm.
2. Based on four benchmark algorithms, we applied LK-means to 10 datasets for comparative analysis of clustering effects. The experimental results show that our algorithms show higher accuracy and better clustering centroids.
3. Experimental analysis of the influence factors in the LK-means algorithm is performed to derive the optimal influence factor.

The remainder of this paper is structured as follows. Section 2 introduces the related work. Section 3 presents our approach. Section 4 provides our experimental results, which includes three parts: clustering accuracy, clustering centroid and optimal parameter v value. Finally, Section 5 concludes our work.

2. Preparation

2.1. K-means clustering algorithm

K-means clustering is a vector quantization method derived from signal processing. The goal of *K*-means clustering is to divide *n* data into *k* classes, where each data belongs to the nearest cluster center, as the cluster center. For example, given a set of observations (x_1, x_2, ..., x_n), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* ($<= n$) sets $S = S_1, S_2, ..., S_k$ so as to minimize the Within-Cluster Sum of Squares (WCSS) (i.e., variance). Formally, the objective is to find (as shown in Eq. (1)).

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - u_i||^2 = \arg\min_{S} \sum_{i=1}^{k} |S_i| Var(S_i) \tag{1}$$

where μ_i is the mean of points in S_i. During the clustering calculation, u_i is used as the cluster center for distance calculation and is gradually optimized with iterations. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster (as shown in Eq. (2)).

$$\arg\min_{S} \sum_{i=1}^{k} (\frac{1}{2|S_i|}) \sum_{x,y \in S_i} ||x - y||^2 \tag{2}$$

The equivalence can be deduced from identity (as shown in Eq. (3)).

$$\sum_{x,y \in S_i} ||x - y||^2 = \sum_{x \neq y \in S_i} (x - \mu_i)^T (\mu_i - y) \tag{3}$$

2.2. XK-means clustering algorithm

K-means clustering algorithm has been widely used in engineering because of its fast convergence, but it is easily tend to the local optimal. To solve this problem, XK-means clustering algorithm is proposed, which adds a random exploration vector θ_j to the clustering centroid of *k*-means algorithm(as shown in Eq. (4)).

$$D_j^* = D_j + \theta_j \qquad (0 < j < k) \tag{4}$$

where θ_j is the random search vector iterated in d-dimensional space, which is obtained according to Eq. (5).

$$\theta_j = rand(a_i, b_i) \times randsign(i), i = 1, 2, \ldots, D \tag{5}$$

where $rand()$ is a random function and $sign$(i) is 0 or 1. a_i and b_i are the upper and lower inertia orders of dimension *i*, respectively. a_i and b_i satisfy the Eq. (6).

$$a_i = \beta b_i \tag{6}$$

where β is the given influence factor in the range of [0,1]. After multiple iterations, the cluster centroid tends to be stable, and the interference vector will gradually decrease. Therefore, before a new iteration, the value of b_i changed as Eq. (7).

$$b_i^* = ab_i \tag{7}$$

where α is a fixed constant, the value range is [0,1).

Compared with *k*-means algorithm, XK-means algorithm avoids local optimum to a certain extent. However, according to Eqs. (2)–(5) and Bouyer et al. [23], XK-means algorithm can't completely jump out of local optimum, especially in dealing with low dimensional gene expression clustering problem. The poor results of XK-means algorithm are mainly because the search direction of XK-means algorithm is not diverse enough.

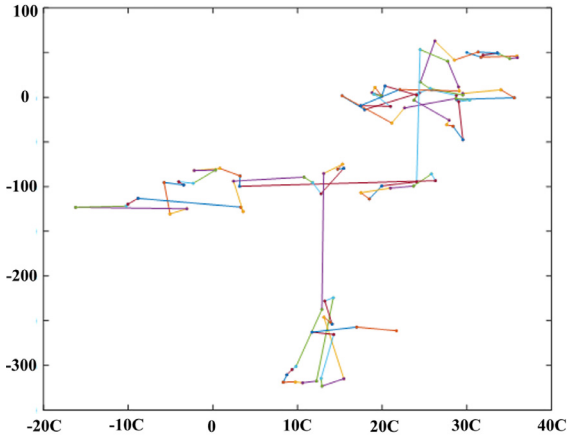


Fig. 1. Lévy flight trajectory.

3. LK-means clustering algorithm

Inspired by the XK-means clustering algorithm, a clustering algorithm based on Lévy flying and *K*-means (LK-means) is proposed to enhance the global exploration of random vectors. The Lévy flight path maximizes the diversity of search agents, which effectively increases the diversity of random search vectors. When the cluster centroid is updated, the Lévy flight trajectory is used to generate the random search vector, which can be expressed by the Eq. (8).

$$\theta_j = rand(a_i, b_i) \times \mu sign\left[rand - \frac{1}{2}\right] \otimes Levy, i = 1, 2, \ldots, D \quad (8)$$

where θ_j, a_i, b_i and $rand()$ are the same as those in Section 2.2, μ is an influence factor that controls the search range, and rand is a random number in the range of [0,1]. $Sign[rand - 1/2]$ represents the direction of the exploration vector, with three values: 0, 1 and -1.

The Eq. (8) essentially is the random flight exploration, which ensures the diversification of random exploration vectors and can optimize the clustering centroid selected by *K*-means to jump out of the local optimum. The algorithm exploration step become wider as the addition of Lévy flight trajectory. Lévy flight provides the Eq. (9). Ghadiri et al. [24]:

$$Levy \sim u = t^{-\lambda}, 1 < \lambda \leq 3 \quad (9)$$

Lévy flight is a power-law behavior with random steps (heavy tail or long tail). It can be seen from Fig. 1 that the Lévy flight always has a large step and a small step in the search process. Mantegna's algorithm [25] imitates the lambda stable distribution by generating a random s step with the same behavior as Lévy flights, as Eq. (10).

$$s = \frac{\rho}{|v|^{\frac{1}{\gamma}}} \quad (10)$$

where s is the random step size of Lévy flight, generally, γ=1.5, where $\lambda = 1 + \gamma$, ρ obeys normal distribution and v obeys normal distribution as Eq. (11).

$$\begin{aligned} \sigma_\rho &= \left[\frac{\Gamma(1+\gamma)+sin(\pi\times\gamma/2)}{\Gamma(((1+\gamma/2))\times\gamma\times2^{(\gamma-1)/2})}\right]^{1/\gamma} \\ \sigma_v &= 1 \end{aligned} \quad (11)$$

The Lévy flight trajectory is generated from the Eqs. (8)–(11), which greatly improves the search ability of XK-means algorithm. Moreover, with the Lévy flight trajectory, the diversity of the algorithm is improved, which ensure that the algorithm jumps out of local optimum. The addition of random search vector is the best objective solution to the error problem of *K*-means clustering algorithm, which ensures the diversity of the algorithm. To verify the advantages of LK-menas clustering algorithm, the evaluation strategy of the algorithm in the next section will be introduced. The clustering process of LK-menas algorithm can be summarized as Fig. 2.

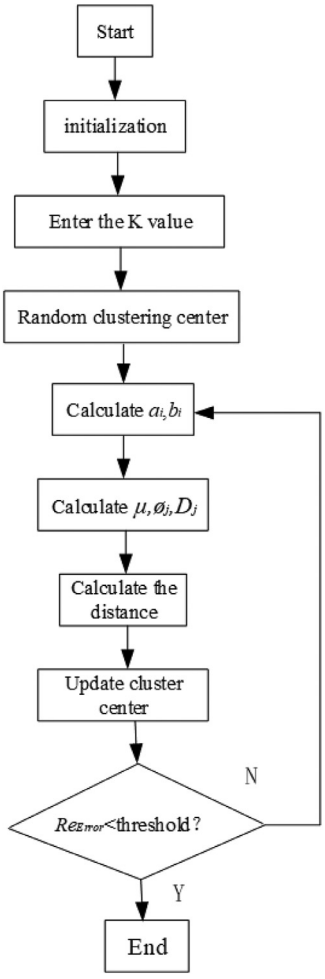


Fig. 2. LK means algorithm flow.

4. Experimental simulation

4.1. Evaluation index

To verify the effectiveness of the proposed algorithm on cluster analysis, four indicators are used to evaluate the algorithm. Mean square error (MSE) [26], Xie Beni index (XB) [27], Davies bouldin index (DB) [28] and Separation index (S) [29] respectively. Specifically as Eqs. (12)–(18).

(1) Mean square error (MSE) (as Eq. (12))

$$MSE = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{N}||x_i - D_j||^2 \quad (12)$$

(2) Xie–Beni index (XB) (as Eq. (13))

$$XB = \frac{MSE}{d_{\min}} \quad (13)$$

where $d_{\min}$ is the smallest distance between cluster centroids. The larger $d_{\min}$, the better clustering effect.

Table 1
Test data sets.

Data sets	No. of vectors N	No. of classes K	No. of attributes D
East cell cycle	6078	256	77
Iris	150	3	4
Ecoli	336	7	8
Wine quality	4898	7	12
Yeast	1484	10	8
Statlog vehicle	946	4	18
Dermatology	366	8	8
Wireless indoor localization	2000	4	7
Page blocks classification	5473	6	10
Image segmentation	2310	7	20

(3) Davies–Bouldin (DB)
As the definition of DB is relatively cumbersome, it is necessary to define the internal separation S_k of class (cluster) first (as Eq. (14)).

$$S_k = (\frac{1}{|D_k|}\sum_{x_i \in D_k} ||x_i - D_k||^2)^{\frac{1}{2}} \tag{14}$$

where D_k represents the samples contained in each class (cluster), and $|D_k|$ represents the number of samples. Then R_k is defined as Eq. (15).

$$R_k = \max_{j,j\neq k}(\frac{S_k + S_j}{||D_k - D_j||}) \tag{15}$$

Then, the Eq. (16) was obtained.

$$DB = \frac{1}{k}\sum_{j}^{k} R_k \tag{16}$$

(4) Separation index (s)

$$S = \frac{1}{\sum_{j=1;i\neq j}^{k} |D_i||D_j|} \sum_{j=1;i\neq j}^{k} |D_i||D_j|||D_i - D_j|| \tag{17}$$

In the numerical iteration process, the end condition is that the relative Re_{Error} is less than the set threshold EPS. The relative Re_{Error} is defined as Eq. (18).

$$Re_{Error} = \left|\frac{MSE_{t-1} - MSE_t}{MSE_t}\right| \tag{18}$$

where MSE_t and MSE_{t-1} represent the MSE values in the current iteration and the last iteration, respectively.

4.2. Data set selection and parameter setting

7 UCI data sets (http://archive.ics.uci.edu/ml/datasets.php), as well as East cell cycle, Ecoli, and Yast [30] gene expression data sets were used to evaluate LK-means algorithm. The east cell cycle dataset was the experimental data of CHO [31], some of the vectors in the east cell cycle were missing. In order to supplement these missing vectors, the method proposed in Gagnon-Bartsch and Speed [32], Hira and Gillies [33], Corso and Cerquitelli [34] was used. KNN algorithm was used to delete, estimate and supplement the data (Gagnon-Bartsch and Speed [32], Hira and Gillies [33] for details, which will not be repeated here). After the supplement, a section of appropriate gene expression data was intercepted (Table 1).

In this paper, core (TM) i5-6200u CPU and 4.00 GB RAM are used in the experiment. Matlab software (matlab r2018a, matlab works, Natick, Ma, USA) is used to cluster data sets. The parameters used in the algorithm are set as follows: error tolerance Etol = 1e−7, $\alpha = 0.50$, $\beta = 0.95$, iterations = 100.

4.3. Result analysis

Evaluation result To intuitively show the advantages of LK-means algorithm, K-means algorithm(Original algorithm), XK-menas algorithm(Source of ideas), DD-kmean algorithm (Iterative optimization algorithm) and Canopy-kmeans algorithm(Iterative optimization algorithm) are used to compare with it through MSE, DB, XB and S in 10 data sets.

From the changes of MSE value shown in Figs. 3–6, it can be seen that in most data sets, the convergence speed of LK-means algorithm is slower than those of k-means algorithm, Xk-menas algorithm, DD-kmean algorithm and Canopy-kmeans algorithm. For example, in the dataset of East Cell Cycle, it can be seen that K-means clustering algorithm converges to a constant value rapidly in the 6th iteration. While MSE-values of XK-means clustering algorithm, LK-means clustering algorithm, DD-kmean algorithm and Canopy-kmeans algorithm continue to converge continuously, and converge in the 10th, 13th, 7th and 13th iteration respectively. However, in all test data sets, the MSE value of LK-means algorithm is significantly lower than those of K-means, XK-means, DD-kmean and Canopy-kmeans algorithm, especially in the East Cell Cycle, Ecoli, East and Image segmentation data sets. Most of these data sets are high-dimensional data, which shows that the proposed algorithm performes better in processing high-dimensional data. In addition, in low dimensional data sets, LK-means algorithm also has the performance of continuing to explore better clustering centroids, which indicates that LK-means algorithm strengthens the global search ability of clustering algorithm and finds the optimal clustering centroid faster.

From the XB curves of 10 data sets shown in Fig. 4, it can be seen that LK-means algorithm has different characteristics in multiple data sets (Due to the length of the manuscript, we only show the results for the East Cell Cycle data.). XB curves of multiple data sets show fluctuation situation because of the global search effect of Lévy flight trajectory, which proves that Lévy flight trajectory avoids the local optimization of the algorithm. Similarly, it was also reflected in the DB curve (Fig. 5) and S curve (Fig. 6). Through the study of simulation diagrams of four indicators (MSE, XB, DB, S), it was found that LK-means algorithm can get better results than K-means algorithm, XK-means algorithm, DD-kmean algorithm and Canopy-kmeans algorithm in 10 data sets. In this part, 10 data sets have been tested for 100 times in five clustering algorithms, and the average values of four indicators (MSE, DB, XB and S) were listed according to the evaluation strategy, as shown in Table 2.

It can be seen from Table 2 that the proposed LK-means algorithm can get the minimum MSE, DB, XB value and the maximum S value, which shows that the overall performance of LK-means algorithm is significantly better than k-means algorithm, Xk-menas algorithm, DD-kmean algorithm and Canopy-kmeans algorithm. The four indexes (MSE, DB, XB, S) values of LK-means algorithm are significantly better than other algorithms in the data sets of East Cell Cycle, Ecoli, Hydrology, and Image segmentation. In addition, for wine quality and East data sets, the MSE value index of LK-means algorithm is the smallest. For wine quality, statlog vehicle and page blocks classification data sets, the XB value index of LK-means algorithm is the smallest. For statlog vehicle and page blocks classification data sets, the DB value index of LK-means algorithm is the smallest. For iris, wine quality, East and statlog classification data sets, the DB value index of LK-means algorithm is the smallest. For vehicle, wireless indoor localization and page blocks classification data sets, the S-value index of LK-means algorithm is the smallest. In total, there are 40 comparative exper-

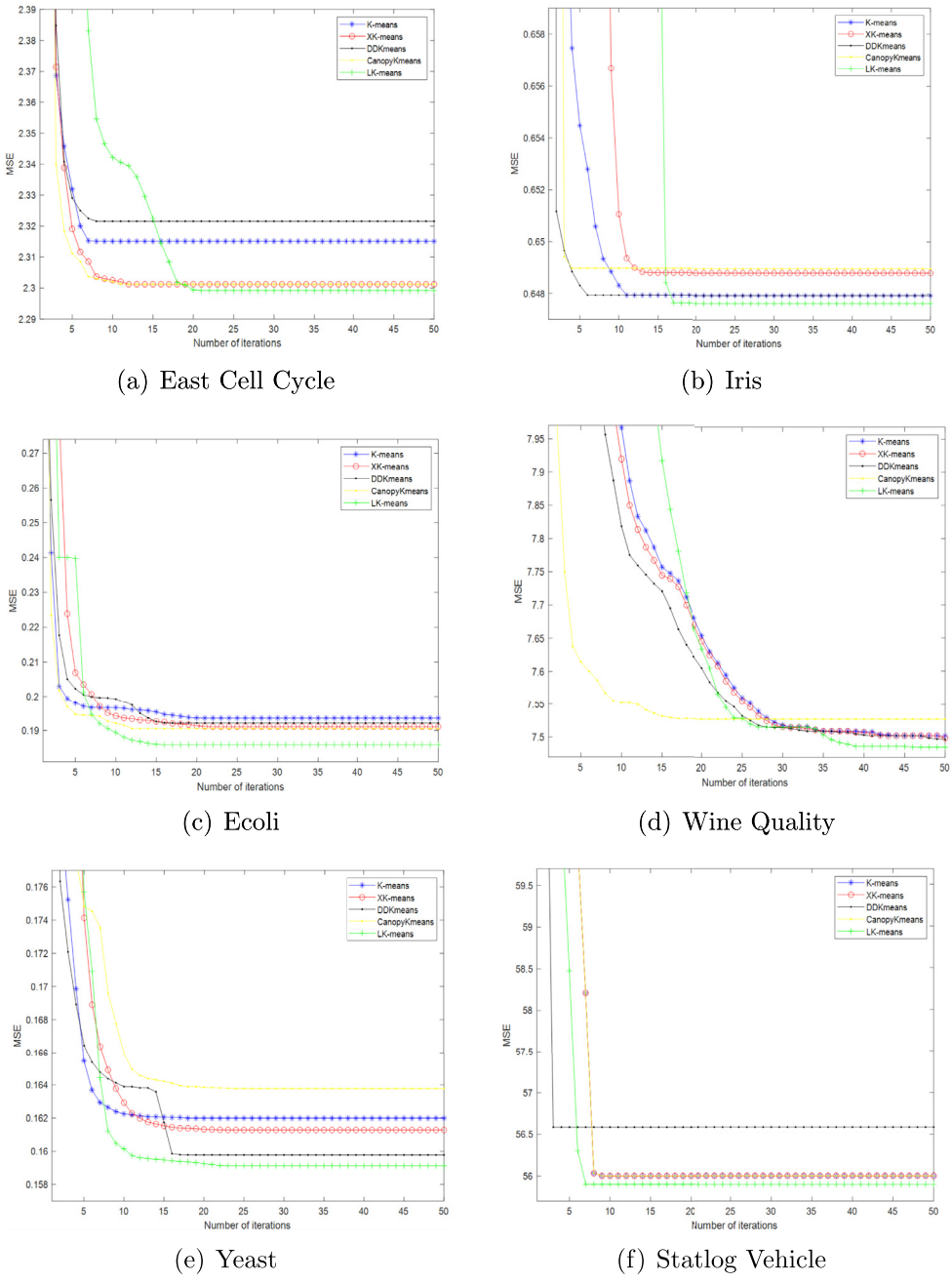
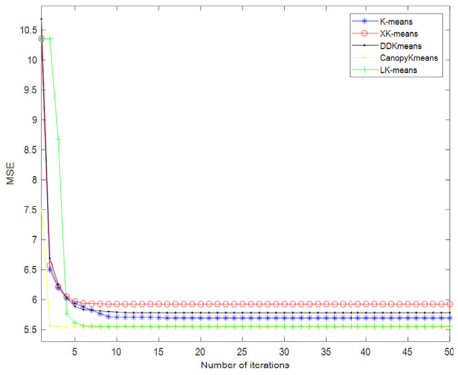


Fig. 3. MSE curve of cluster analysis.

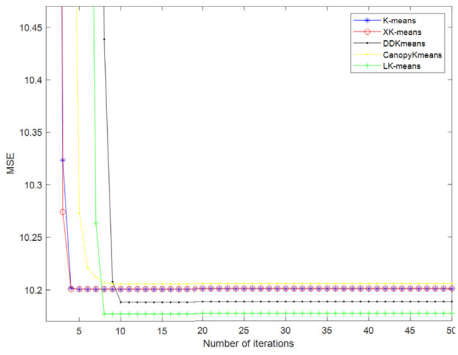
iments in all 10 data sets. It can be seen from the experimental results that the LK-means algorithm has better results in 29 experiments, it shows more than 72.5% of the experimental results prove that the LK-means algorithm was superior to other methods.

Clustering centroids To study the distribution of clustering centroids, the Voronoi diagram is drawn by MPT toolbox in MATLAB to verify the rationality of LK-means algorithm. Based on the first two dimensions of East Cell Cycle dataset, Voronoi diagram is run in *K*-means algorithm, XK-means algorithm, ddkmean algorithm, Canopyk-means algorithm and LK-means algorithm, and the results are shown in Fig. 7. The whole graph evolves in the interval of $-3 < x < 4, -3 < y < 4$. In the two dimensions, the first dimension is the *X*-axis, and the second dimension is the *Y*-axis. 3072 data points are appropriately selected and divided into five clusters. The denser black dots represent data points, where the red dots represent the cluster centroids during the iteration, and the blue dots represent the boundaries between clusters.

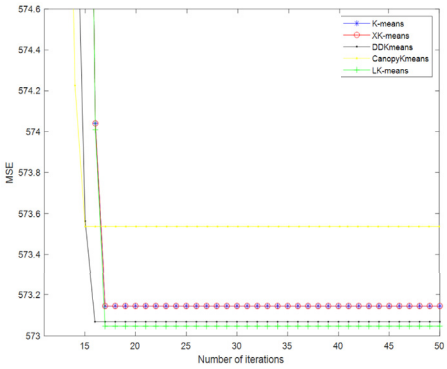
It can be clearly identified from the vonoroi diagram that the number of red dots in the graph c is significantly more than a, b, d and e, which means that the LK-means algorithm has better clustering centroid after iterations. This is mainly because the Lévy flight greatly improves the global search ability of the algorithm, which also proves that the LK-means algorithm effectively solves the problem of excessively fast convergence. Compared with the whole graph, the clustering centroids of *K*-means algorithm and XK-means algorithm are mostly concentrated in high-density areas, but few clusters are used to represent low-density areas, which will produce a high MSE value. The better the MSE value



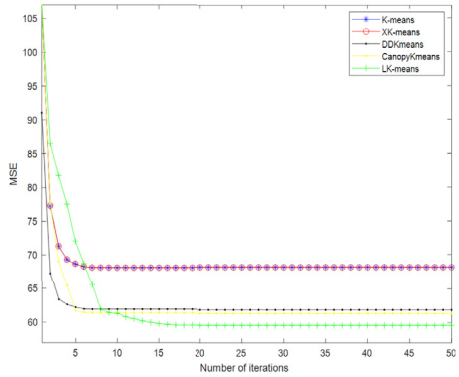
(g) Dermatology



(h) Wireless Indoor Localization



(i) Page Blocks Classification



(j) Image Segmentation

Fig. 3. Continued

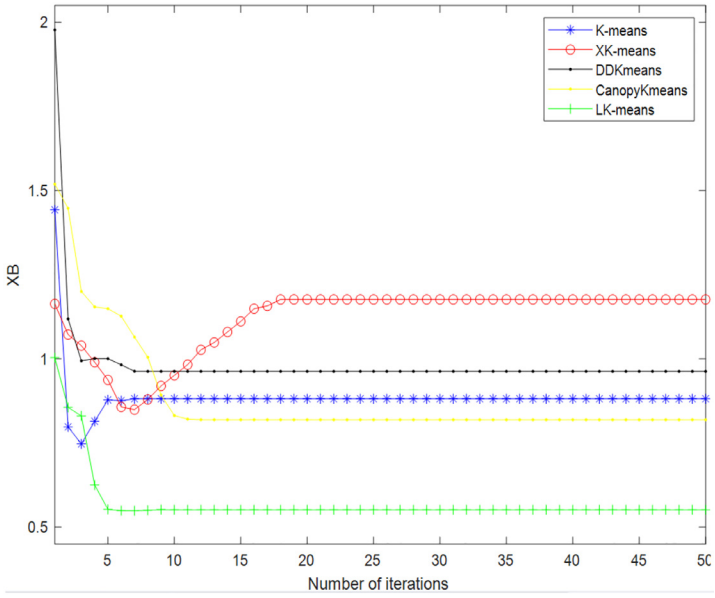


Fig. 4. XB curve of cluster analysis.

(the smaller), the better the compactness within the cluster, and the easier to be classified. Compared with *K*-means algorithm, XK-means algorithm, DDKmeans and Canopy-means, LK-means algorithm provides more uniform clustering centroids and more reasonable distribution in high-density and low-density regions. Similar results can be obtained on other data sets, as shown in Figs. 8–14.

4.4. Selection of influence factor μ of LK-means algorithm

In Section 3, it has been explained that μ is an influence factor to control the search range. In this section, the influence of different μ values on the performance of LK-means algorithm will analyzed. The selection process of parameters α and β have discussed by Lam [22], which will not be repeated here. In all data sets,

Table 2
Average values of MSE, DB, XB and S.

Data sets	Algorithms	MSE	XB	DB	S
East cell cycle	*K*-means	2.3151	0.8815	0.3273	23.2720
	XK-means	2.3120	0.8983	0.3374	23.3407
	DDKmeans	2.3216	0.9629	0.3209	23.2879
	Canopyk-means	2.3110	0.9190	0.9862	23.1733
	LK-means	**2.3098**	**0.8719**	**0.3110**	**23.3732**
Iris	*K*-means	**0.6479**	**0.3623**	**0.0847**	12.2350
	XK-means	0.6521	0.3737	0.0915	12.3010
	DDKmeans	0.6487	0.3634	0.2320	12.2350
	Canopyk-means	0.6490	0.3629	0.2320	12.2350
	LK-means	0.6510	0.3994	0.1146	**12.3936**
Ecoli	*K*-means	0.1936	0.9766	0.1603	0.2937
	XK-means	0.1921	0.9127	0.1595	0.3025
	DDKmeans	0.1921	1.0733	0.2349	0.2971
	Canopyk-means	0.1910	0.8846	0.2694	0.3052
	LK-means	**0.1888**	**0.8837**	**0.1565**	**0.3087**
Yeast	*K*-means	0.1620	**0.9244**	**0.0802**	0.1096
	XK-means	0.1614	0.9697	0.1412	0.1133
	DDKmeans	0.1615	0.9852	0.1663	0.1146
	Canopyk-means	0.1619	0.9420	0.2227	0.1152
	LK-means	**0.1611**	0.9668	0.1391	**0.1157**
Statlog vehicle	*K*-means	56.0015	0.4531	0.2465	9.6163×10^4
	XK-means	56.0013	0.4531	0.2465	9.6163×10^4
	DDKmeans	**56.0011**	0.4526	0.1999	9.6165×10^4
	Canopyk-means	57.0328	0.4513	0.3994	9.6163×10^4
	LK-means	57.0135	**0.4432**	**0.1983**	$\mathbf{1.0076 \times 10^5}$
Dermatology	*K*-means	5.6933	0.8069	0.2325	544.0148
	XK-means	5.8211	0.8565	0.2295	552.2700
	DDKmeans	5.7809	0.9343	0.2147	545.0386
	Canopyk-means	5.7358	0.8029	0.3562	552.3986
	LK-means	**5.6878**	**0.7341**	**0.1992**	**552.5664**
Wireless indoor localization	*K*-means	10.2006	0.4656	**0.1881**	920.6622
	XK-means	10.2012	0.4657	0.1882	920.6734
	DDKmeans	**10.1883**	**0.4651**	0.2207	920.6622
	Canopyk-means	11.0876	0.5469	0.2590	920.4589
	LK-means	10.2584	0.4887	0.2423	**920.8347**
Page blocks classification	*K*-means	573.1447	0.2213	0.0645	1.5563×10^8
	XK-means	573.1312	0.2213	0.0645	1.5563×10^8
	DDKmeans	573.1447	0.2213	0.1044	1.5563×10^8
	Canopyk-means	**573.0469**	0.2214	0.0477	1.5563×10^8
	LK-means	573.2309	**0.2195**	**0.0397**	$\mathbf{1.5765 \times 10^8}$
Image segmentation	*K*-means	68.1357	0.7558	0.0828	3.3453×10^4
	XK-means	67.9043	0.7507	0.0846	3.3596×10^4
	DDKmeans	61.9041	0.7511	0.0613	5.7391×10^4
	Canopyk-means	61.7705	0.7007	0.0382	5.7514×10^4
	LK-means	**61.4808**	**0.6912**	**0.0356**	$\mathbf{5.8025 \times 10^4}$

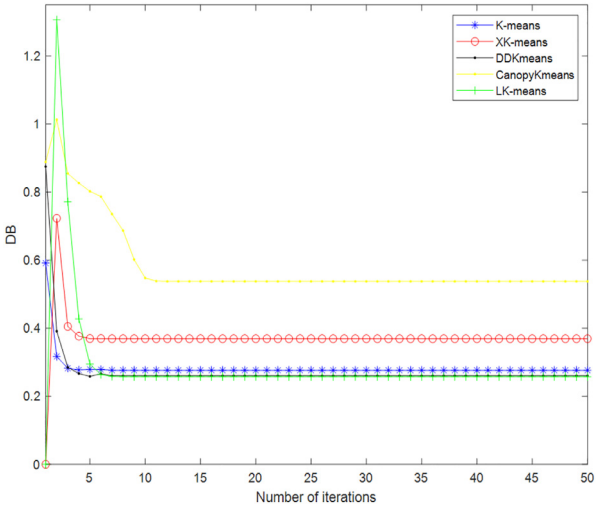


Fig. 5. DB curve of cluster analysis.

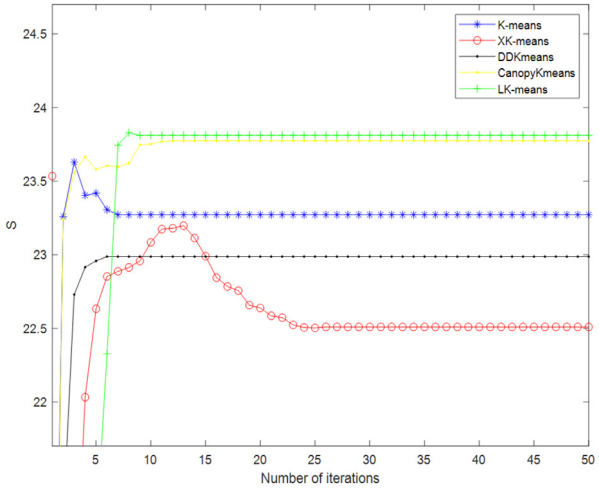


Fig. 6. S curve of cluster analysis.

Table 3
MSE results relative to impact factor μ.

Data sets	μ	MSE	ite
East cell cycle	0.00	2.3151	6
	0.25	2.3084	5
	0.50	2.3080	7
	0.70	2.3078	7
	0.90	2.3068	13
	0.95	2.3074	67
	1.00	2.3087	100
Ecoli	0.00	0.1936	20
	0.25	0.1926	12
	0.50	0.1923	12
	0.70	0.1921	15
	0.90	0.1916	17
	0.95	0.1909	31
	1.00	0.1916	100
Wine quality	0.00	7.5026	30
	0.25	7.5106	29
	0.50	7.5236	29
	0.70	7.5134	34
	0.90	7.4964	34
	0.95	7.4986	46
	1.00	7.5187	100
Yeast	0.00	0.1620	11
	0.25	0.1618	12
	0.50	0.1616	15
	0.70	0.1614	16
	0.90	0.1611	21
	0.95	0.1612	69
	1.00	0.1617	100
Dermatology	0.00	5.6933	9
	0.25	5.7054	3
	0.50	5.7522	6
	0.70	5.7337	6
	0.90	5.6427	7
	0.95	5.6580	29
	1.00	5.6712	100
Image segmentation	0.00	68.1357	6
	0.25	66.9441	5
	0.50	65.5888	5
	0.70	65.1584	6
	0.90	65.1405	10
	0.95	65.4852	37
	1.00	65.8419	100

$\alpha = 0.95$ [22] is the best choice, and the selection of β needs to be choosed by different data sets. To get more accurate results, different β values were selected for different data sets. Through the previous experiments, it was found that LK-means algorithm has a significant effect on the data sets of East Cell Cycle, Ecoli, wine quality, East, Topology and image segmentation. Therefore, the influence of different μ values on algorithm performance is studied for these 6 data sets, with the results shown in Fig. 15.

To further reflect the experimental results of MSE performance of six data sets, MSE samples with μ values in the range of 0 to 1 were taken (MSE was generally selected to carry out μ-parameter experiments in other papers). The experiments of each value were carried 50 times and the average value was calculated. The experimental results are shown in Table 3.

From the experimental results in Table 3 and Fig. 15, it can be seen that the best MSE value can be obtained when μ set to 0.90 for both high-dimensional data set East Cell Cycle and low dimensional data set. Similar to XK-means algorithm, although the best $MSE = 0.1909$ is obtained at the parameter $\mu = 0.95$ in Ecoli dataset, the number of iterations is too large. Therefore, the paper considers both the MSE value and the iteration time, and the parameter $\mu = 0.90$ is selected.

5. Conclusion and next research

5.1. Conclusion

To improve the clustering effect of k-mean algorithm, a new hybrid algorithm LK-means for data clustering is proposed in this paper. LK-means is based on Lévy flight and *k*-means algorithm, and the *k*-means clustering algorithm is optimized using Lévy flight to improve the *k*-means clustering effect. In conclusion, the proposed LK-means algorithm can effectively improve the effect of data clustering and has a better iteration speed.

LK-means can effectively cluster the data, so it can be applied to data management in various industries. Especially, it deals with data with diverse types and feature information. For example, medical data management, due to the diversity of patients, resulting in a variety of conditions, the medical characteristics of the case management to assist doctors to make accurate judgments; Power data management, once the power system is put into operation, it will be in operation for a long time, generating a large amount of power data (including internal system data and external environment data) in real time, and timely and accurate cluster management of power data can effectively assist power workers to maintain power operation and improve power supply reliability. etc.

Advantages of the algorithm proposed in this paper.

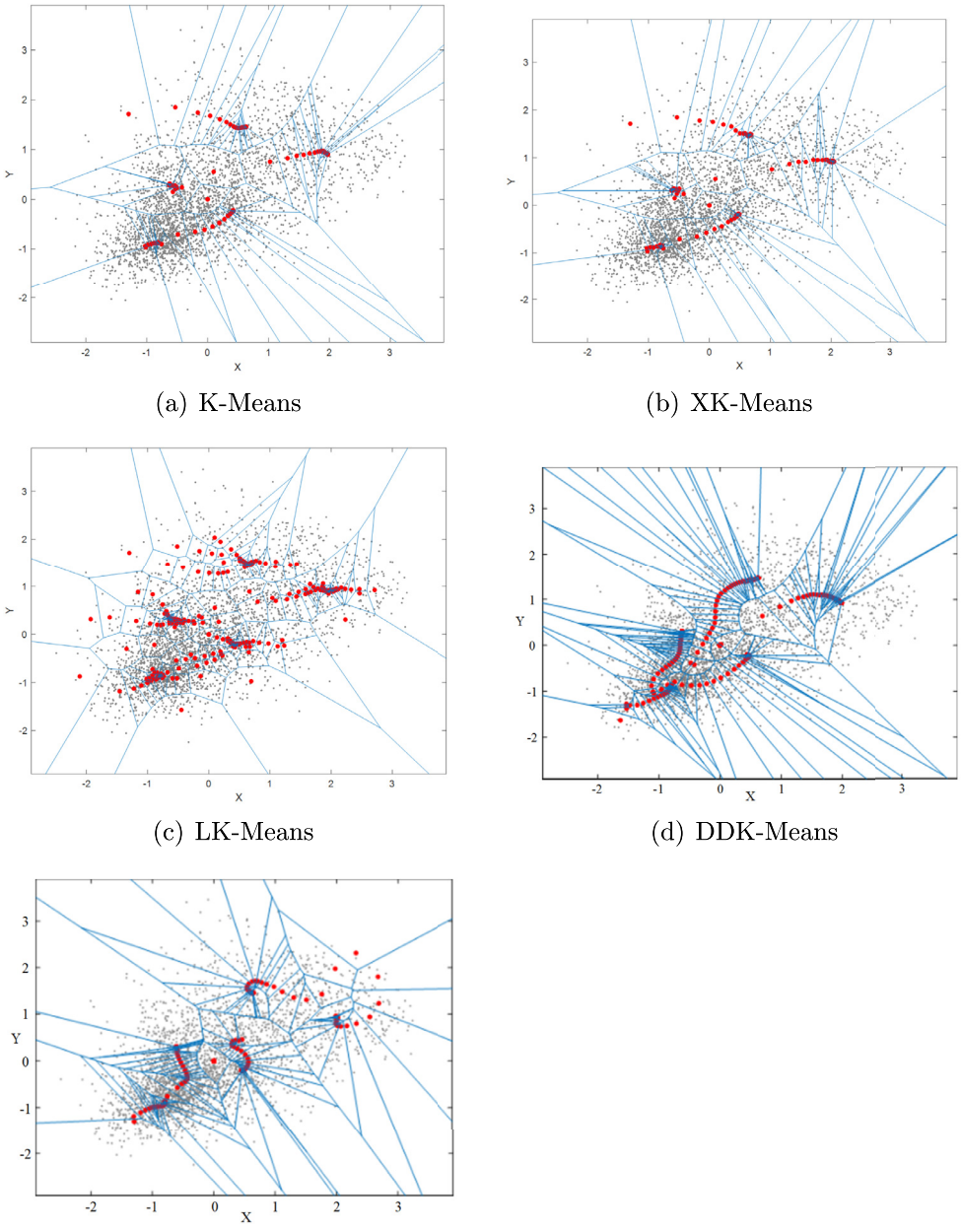
(1) The experiments show that the LK-means algorithm proposed in this paper has higher accuracy in clustering compared to the benchmark model, while the number of iterations under reaching the optimal value is less.
(2) Relative to the benchmark model, the complexity of the LK-means algorithm is not increased, so it has a better efficiency of clustering.
(3) The LK-means algorithm is able to adapt to multiple types of data clustering, such as data sets with size variability and feature type variability.
(4) The LK-means algorithm has fewer parameters, and only the influence factor u needs to be optimized to obtain in the study of the article.

The limited time and resources lead to the proposed algorithm has some limitations.

(1) Limitations of the dataset. This work was validated only on 10 UCIs, and we did not validate the proposed model directly on other datasets because of the variability of the dataset characteristics that prevented direct validation in the proposed model.
(2) Limitations of the initial clustering center study. The initial clustering centers of the LK-means algorithm are not studied, which are generated randomly for the time being, and the study on the selection of the initial clustering centers is lacking.
(3) Limitations of the benchmark model. The article only compares the same type of algorithms for the time being, and has not yet compared other types of algorithms or models, such as machine learning, population algorithms or other iterative algorithms.

5.2. Next research

(1) Optimal selection of clustering centers. This paper mainly focuses on the iterative optimization of *K*-means algorithm, and does not study its clustering center. The selection of clustering center has a great impact on the clustering effect, so in the next study, we will optimize the clustering center of *K*-means algorithm.



(a) K-Means (b) XK-Means

(c) LK-Means (d) DDK-Means

(e) CanopyKmeans

Fig. 7. Voronoi in yeast cell-cycle data.

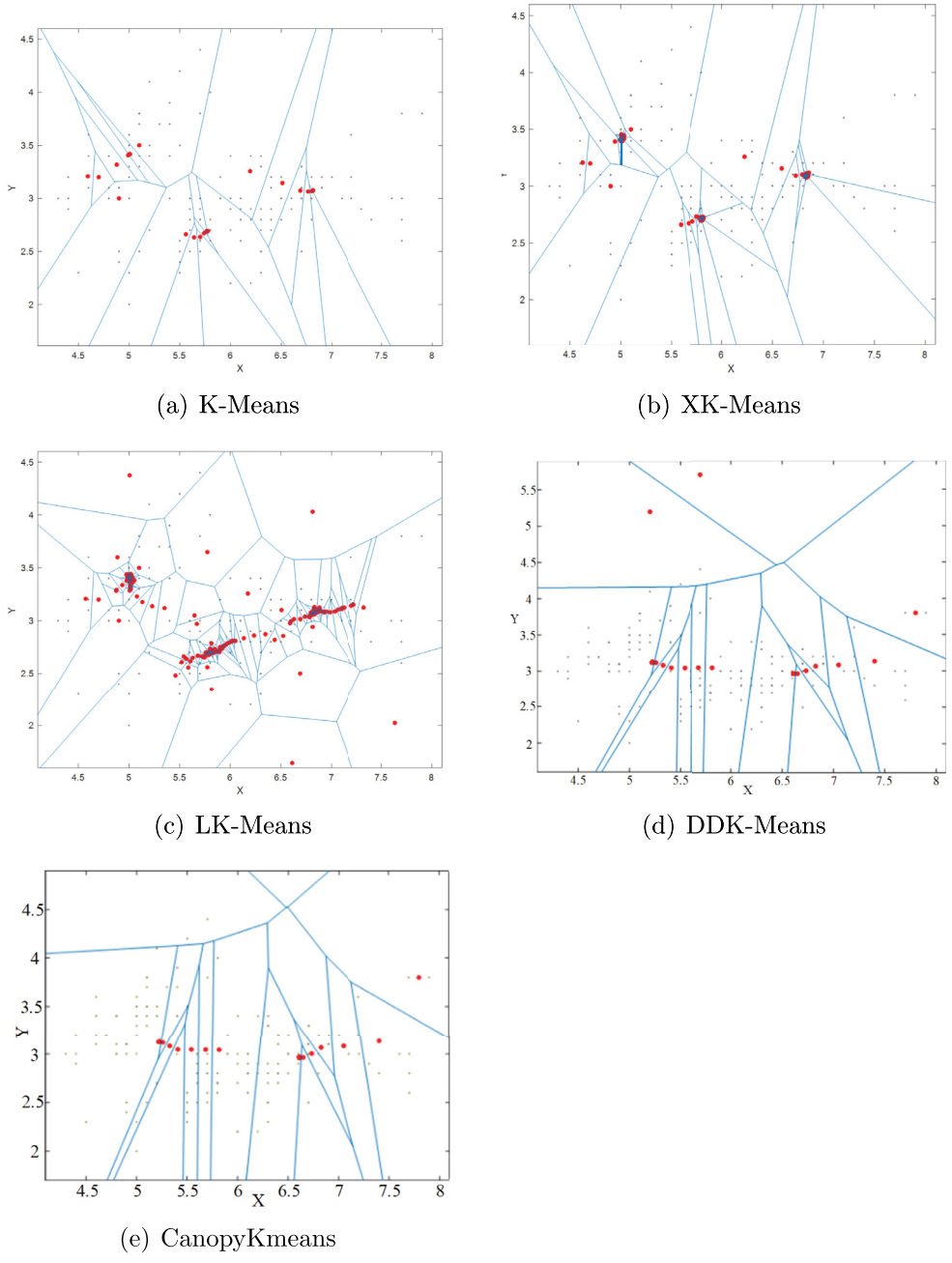


Fig. 8. Voronoi in iris data.

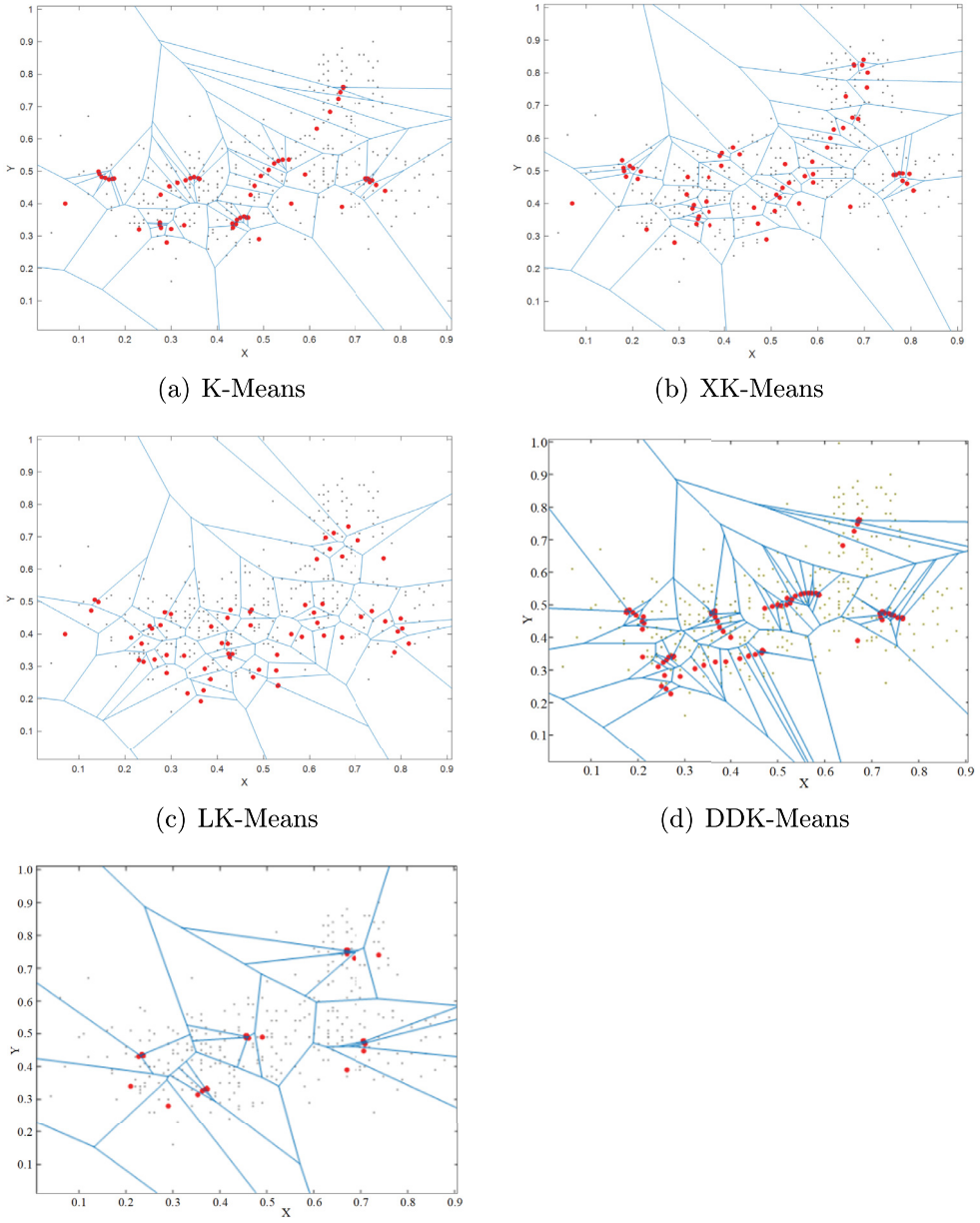


Fig. 9. Voronoi in ecoli data.

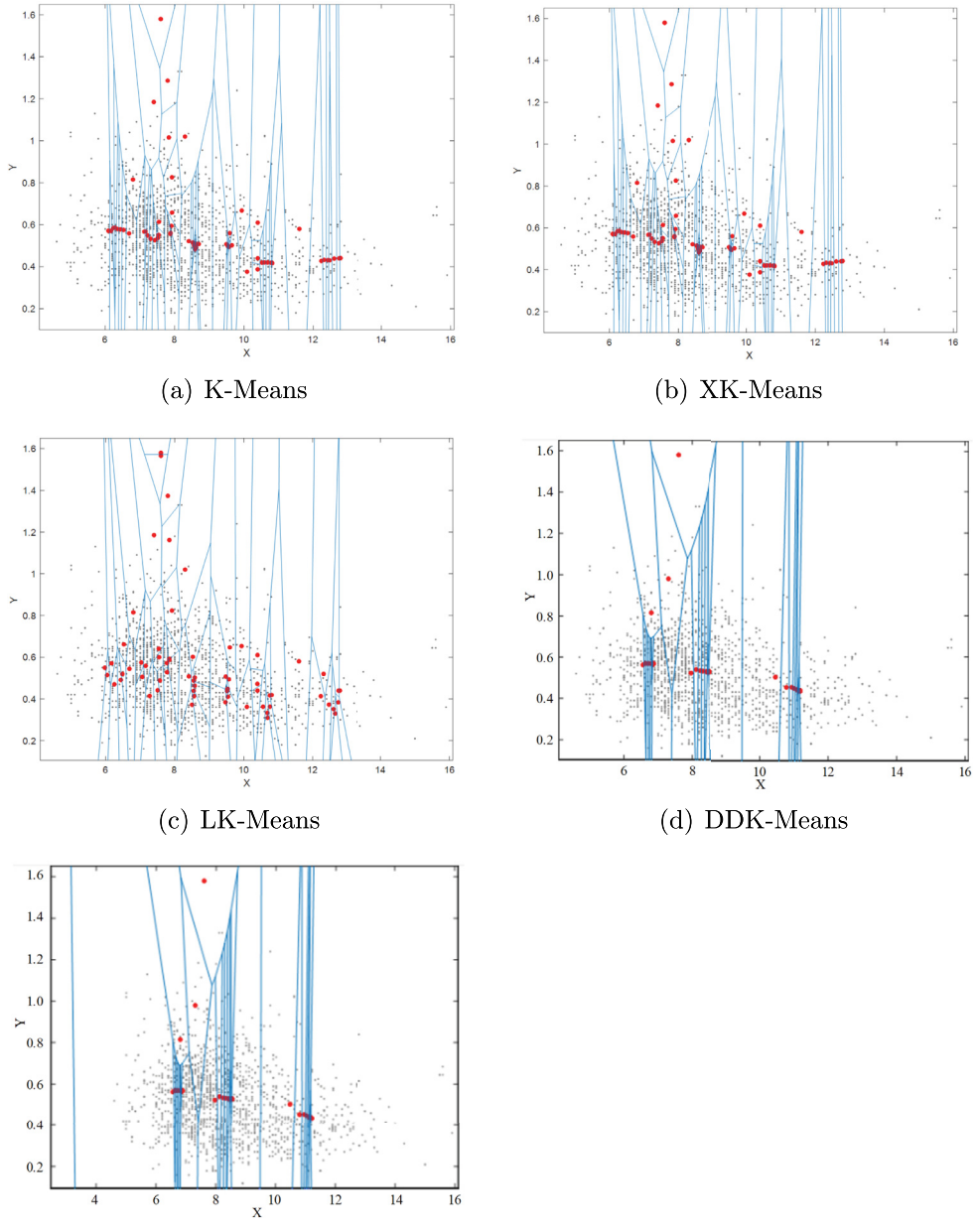
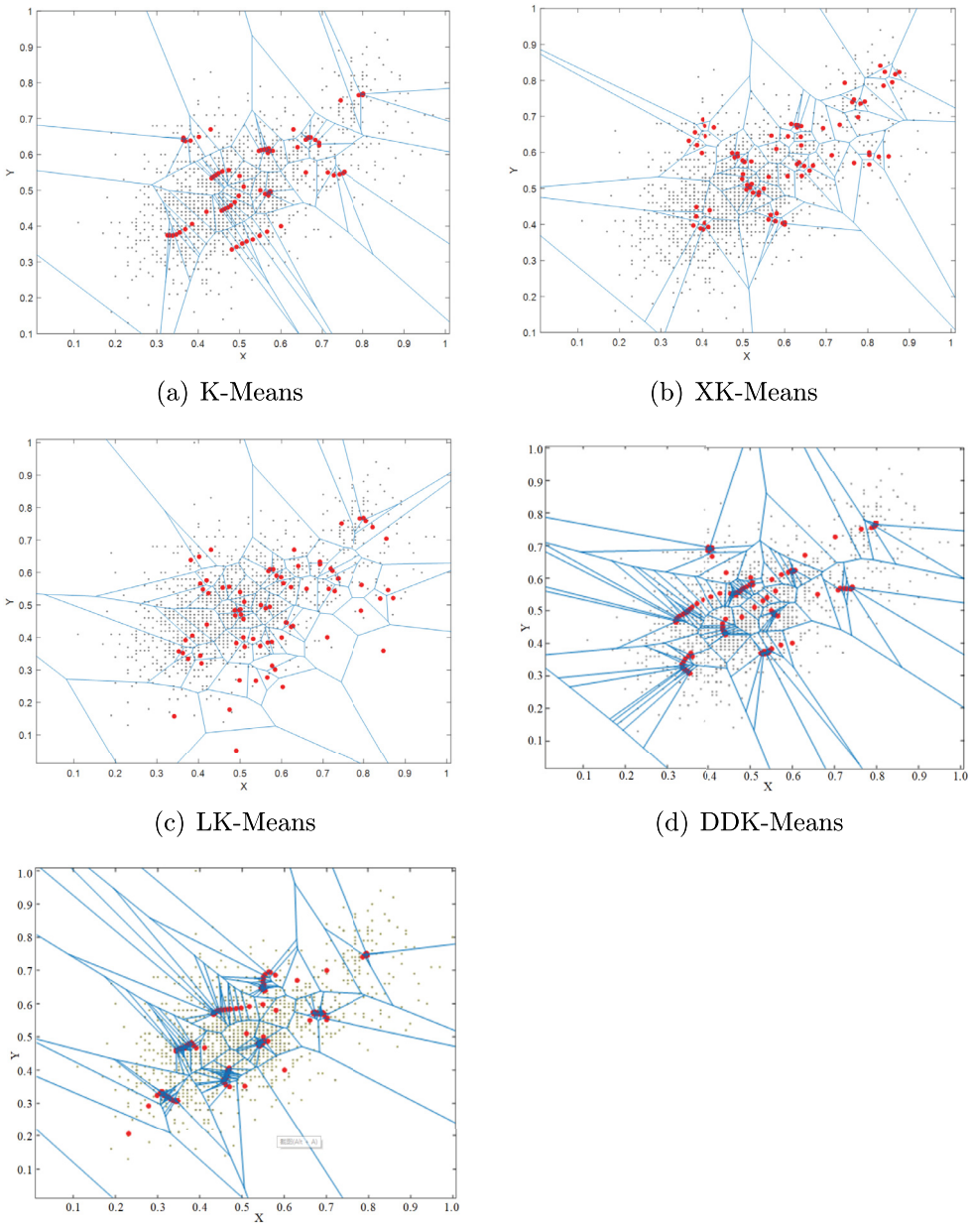


Fig. 10. Voronoi in wine quality data.



(a) K-Means (b) XK-Means

(c) LK-Means (d) DDK-Means

(e) CanopyKmeans

Fig. 11. Voronoi in yeast data.

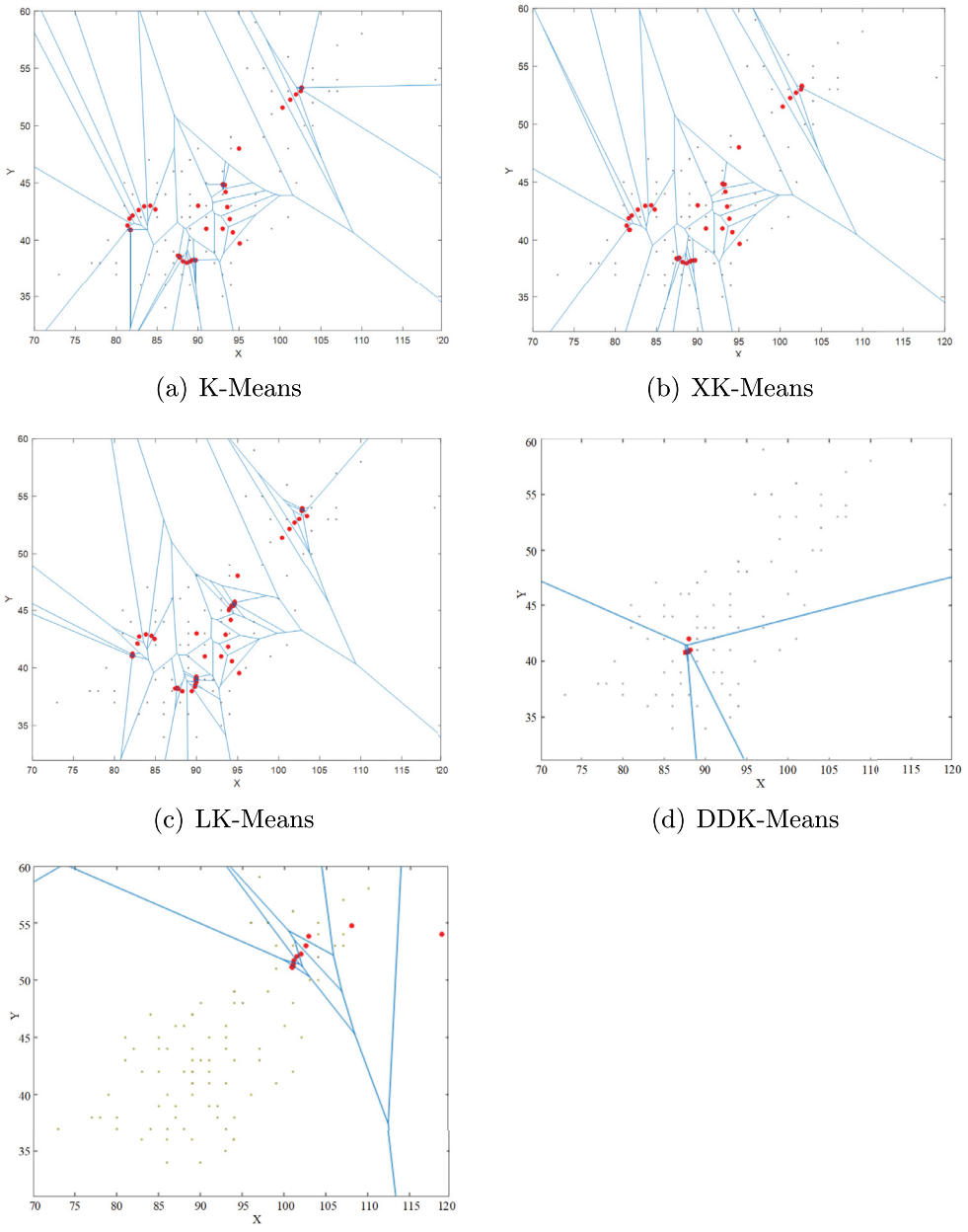


Fig. 12. Voronoi in statlog vehicle data.

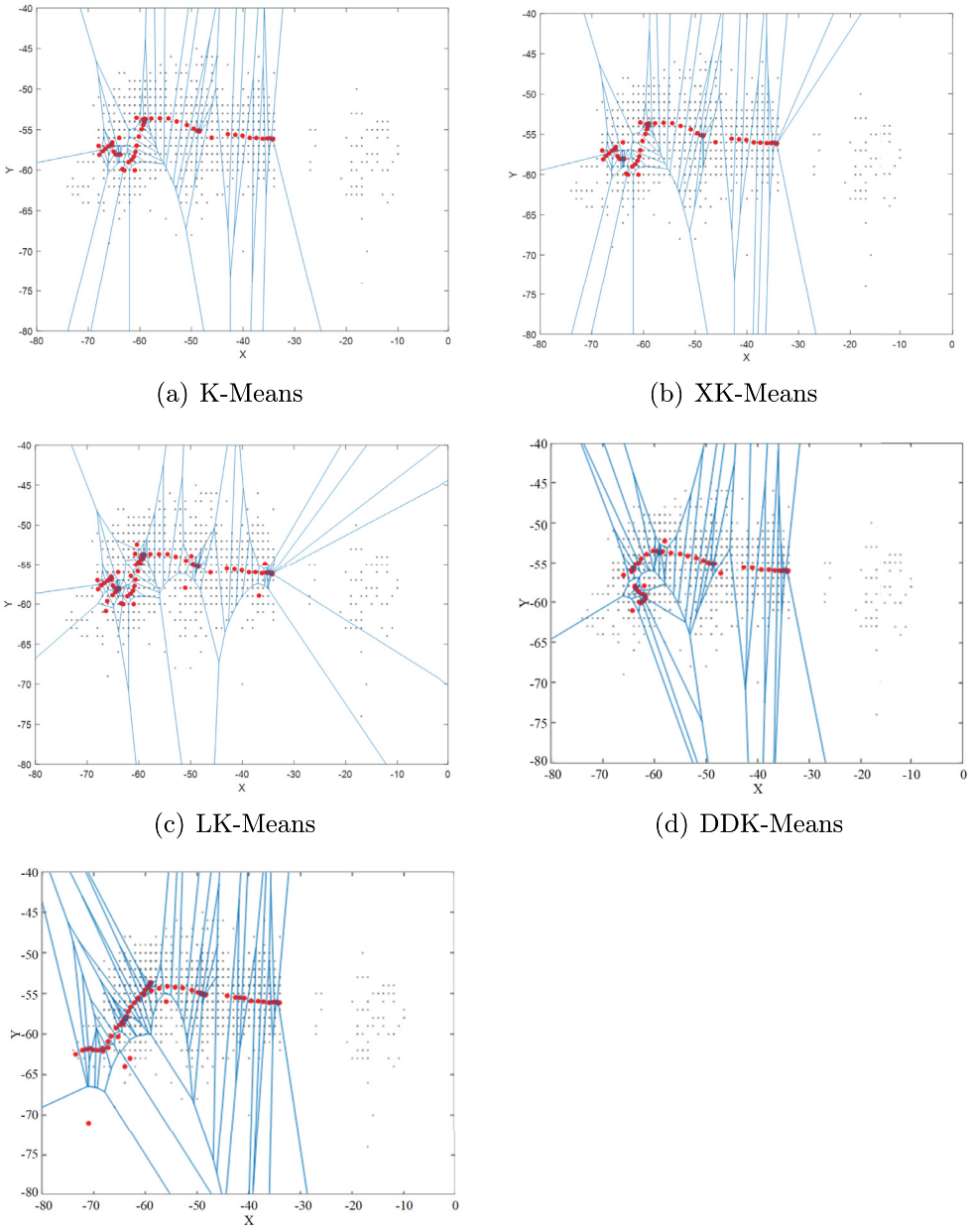


Fig. 13. Voronoi in door localization data.

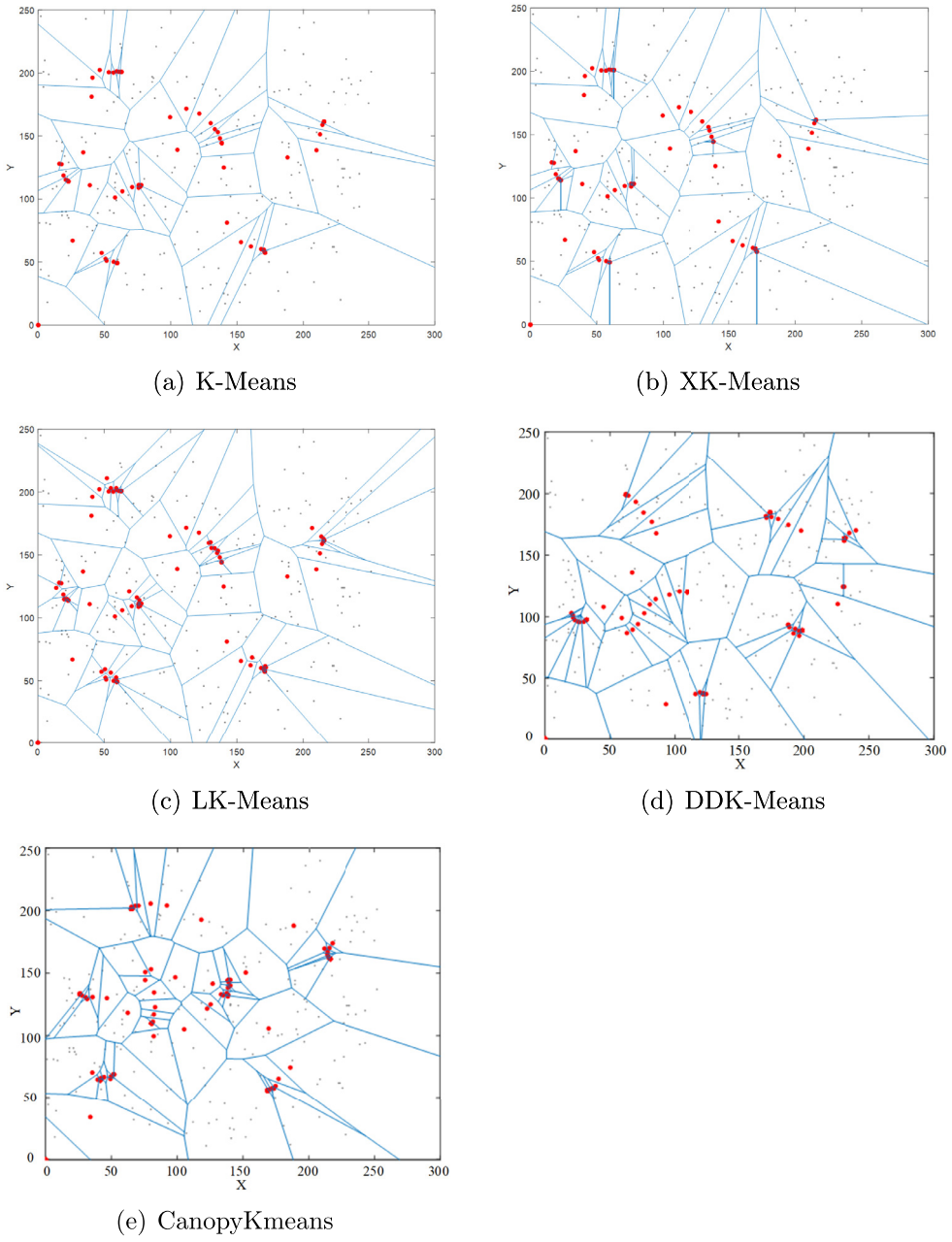
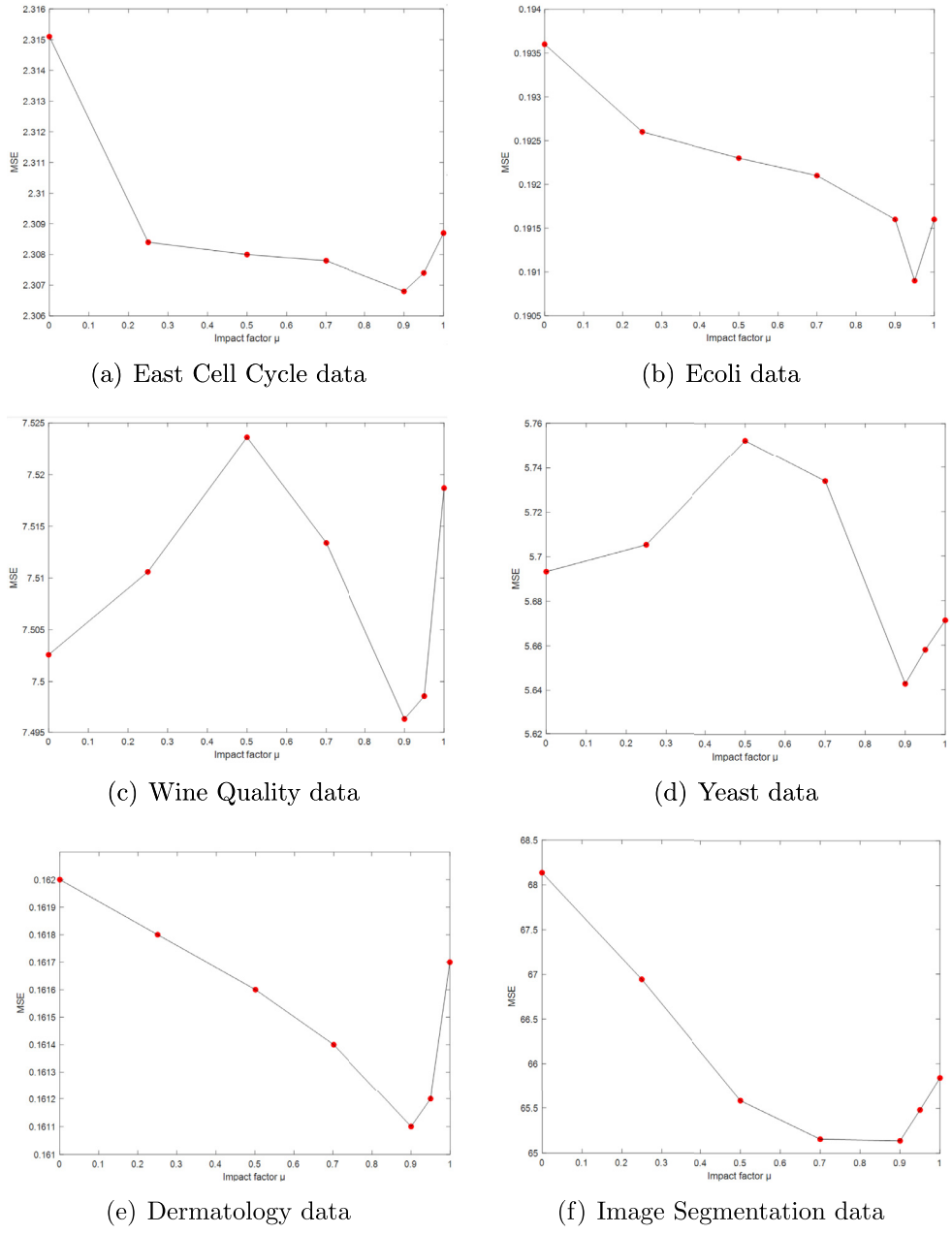


Fig. 14. Voronoi in image segmentation data.

(2) Exploration of datasets. This mainly includes two aspects, on the one hand, the development of large datasets, exploring multi-type and multi-feature datasets to improve the adaptability of the algorithm. On the other hand, we explore the actual data of the industry to optimize the clustering effectiveness of the algorithm with real data sets, and then serve the clustering analysis of big data.

(3) Comparative analysis of algorithms. At present, we have only performed a comparative analysis of *k*-means algorithms of the same type, lacking in other types of algorithms for comparison. For example, particle swarm optimization, Cuckoo Search, Ant Colony optimization and machine learning models. Therefore, the effectiveness of the algorithm will be verified in the next research by continuously optimizing the clustering effect of the algorithm and by exploring multiple types of comparative models.



(a) East Cell Cycle data (b) Ecoli data

(c) Wine Quality data (d) Yeast data

(e) Dermatology data (f) Image Segmentation data

Fig. 15. Plot of sampling inertia weight u values in LK-Means.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Natural Science Foundation of China, under grants 61866013 and 61503152, and the Education Department Key Foundation of Hunan Province in China, under grants 17A173, and the Education Department Foundation of Hunan Province in China, under grants 18C0565.

References

[1] L. Song, Y. Zhou, X. Qian, Graphr: accelerating graph processing using reRAM, in: 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA), vol. 9, 2018, pp. 531–543.
[2] M. Nedyalkova, S. Madurga, V. Simeonov, Combinatorial *k*-means clustering as a machine learning tool applied to diabetes mellitus type 2, Int. J. Environ. Res. Public Health 12 (2021) 1919.
[3] O. Vaulina, E. Sametov, Spectral and structural characteristics for cluster systems of charged Brownian particles, J. Exp. Theor. Phys. 74 (2018) 350–356.
[4] O. Sadeghian, A. Oshnoei, R. Khezri, Data clustering-based approach for optimal capacitor allocation in distribution systems including wind farms, IET Gener., Transm. Distrib. 218 (May 1) (2019) 3397–3408.
[5] N. Salehnia, N. Salehnia, H. Ansari, Climate data clustering effects on arid and semi-arid rainfed wheat yield: a comparison of artificial intelligence and *k*-means approaches, Int. J. Biometeorol. 283 (6) (2019) 861–872.
[6] X. Ran, X. Zhou, M. Lei, et al. A novel *k*-means clustering algorithm with a noise algorithm for capturing urban hotspots[j], Appl. Sci. 2021, 11(23): 11202.

[7] D. Förster, R.B. Inderka, F. Gauterin, Data-driven identification of characteristic real-driving cycles based on *k*-means clustering and mixed-integer optimization, IEEE Trans. Veh. Technol. 3 (4) (2020) 2398–2410.
[8] A. Bouyer, A. Hatamlou, An efficient hybrid clustering method based on improved cuckoo optimization and modified particle swarm optimization algorithms, Appl. Soft Comput. 67 (2018) 172–182.
[9] J. Saha, J. Mukherjee, Cnak: cluster number assisted *k*-means, Pattern Recognit. 110 (110) (2021) 107625.
[10] A. Isazadeh, O. Tarkhaneh, H.J. Khamnei, A new hybrid strategy for data clustering using cuckoo search based on Mantegna Lévy distribution, PSO and *k*-means, Int. J. Comput. Appl. Technol. 91 (2018) 137–143.
[11] M. Ghadiri, S. Samadi, S. Vempala, Socially fair *k*-means clustering, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, vol. 224, 2021, pp. 438–448.
[12] K. Song, X. Yao, F. Nie, Weighted bilateral *k*-means algorithm for fast co-clustering and fast spectral clustering, Pattern Recognit. 109 (99) (2021) 107560.
[13] X. Zhu, Y. Zhu, W. Zheng, Spectral rotation for deep one-step clustering, Pattern Recognit. 105 (2020) 107175.
[14] F. Nie, S. Shi, X. Li, Auto-weighted multi-view co-clustering via fast matrix factorization, Pattern Recognit. 102 (2020) 107207.
[15] Z. Kang, C. Peng, Q. Cheng, Structured graph learning for clustering and semi–supervised classification, Pattern Recognit. 110 (1) (2021) 107627.
[16] C. Ma, Z. Liu, Z. Cao, Cost-sensitive deep forest for price prediction, Pattern Recognit. 107 (9) (2020) 107499.
[17] C. Zhao, X. Wang, W. Zuo, Similarity learning with joint transfer constraints for person re-identification, Pattern Recognit. 156 (2020) 107014.
[18] H. Xie, L. Zhang, C.P. Lim, Improving *k*-means clustering with enhanced firefly algorithms, Appl. Soft Comput. 84 (15) (2019) 105763.
[19] E. Tuba, B. Starnberger I, N, Cooperative clustering algorithm based on brain storm optimization and *k*-means, in: Proceedings of the 2018 ACM Conference on Fairness, Accountability, and Transparency, vol. 9, 2018, pp. 1–5.
[20] I.D. Borlea, R.E. Precup, A.B. Borlea, A unified form of fuzzy c-means and *k*-means algorithms and its partitional implementation, Knowledge-Based Syst. 214 (2021) 106731.
[21] T. Yaying, B. Hazarika, On arithmetic continuity, Bol. Soc. Parana. Mat. 35 (2017) 139–145.
[22] J. Xu, K. Lange, Power *k*-means clustering 22 (4) (2019) 6921–6931.
[23] M. Moshkovitz, S. Dasgupta, C, Rashtchian, et al. Explainable *k*-means and *k*-medians clustering[C]//International conference on machine learning. PMLR, 2020: 7055-7065.
[24] M. Ghadiri, S. Samadi, S. Vempala, Socially fair *k*-means clustering, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, vol. 234, 2021, pp. 438–448.
[25] H. Soneji, R.C. Sanghvi, Towards the improvement of cuckoo search algorithm, in: 2012 World Congress on Information and Communication Technologies, vol. 18, 2012, pp. 3–10.
[26] B.M. Ismail, B.E. Reddy, T.B. Reddy, Cuckoo inspired fast search algorithm for fractal image encoding, J. King Saud University-Computer Inf. Sci. 30 (2018) 462–469.
[27] K. Labed, H. Fizazi, H. Mahi, A comparative study of classical clustering method and cuckoo search approach for satellite image clustering: application to water body extraction, Appl. Artif. Intell. 32 (1–6) (2018) 96–118.
[28] Y.A. Wijaya, D.A. Kurniady, E. Setyanto, Davies Bouldin index algorithm for optimizing clustering case studies mapping school facilities 1099–1103 (2021).
[29] L. KHRISSI, N. ELAKKAD, H. SATORI, Simple and efficient clustering approach based on cuckoo search algorithm, in: 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS), vol. 78, 2020, pp. 1–6.
[30] N. Chumuang, Comparative algorithm for predicting the protein localization sites with yeast dataset, in: 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), vol. 6, 2018, pp. 369–374.
[31] R.J. Cho, M.J. Campbell, E.A. Winzeler, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell 31 (1998) 65–73.
[32] J.A. Gagnon-Bartsch, T.P. Speed, Using control genes to correct for unwanted variation in microarray data, Biostatistics 117 (2012) 539–552.
[33] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinform. 109 (2015) 107560–107571.
[34] D.E. Corso, T. Cerquitelli, METATECH: meteorological data analysis for thermal energy charac-terization by means of self-learning transparent models, Energies 110 (2018) 1336–1342.

Hu Haize obtained his bachelor's degree from Hunan Institute of engineering and master's degree from Changsha University of technology in 2013 and 2016 respectively. After graduation, he has been engaged in teaching in Jishou University for four years and is now studying for a doctor's degree from Hunan University of science and technology.

Jianxun Liu, professor and doctoral advisor, obtained his bachelor's degree, master's degree and doctor's degree from Hunan Institute of engineering, Central South University and Shanghai Jiao Tong University in 1989, 1997 and 2003 respectively.

Xiangping Zhang, received master's degree and bachelor's degree from Hunan University of science and technology in 2016 and 2019 respectively. He is now studying for a doctor's degree in Hunan University of science and technology. His research interests include code representation and code clone detection.

Mengge Fang, received master's degree and bachelor's degree from Changsha University of science and technology in 2020 and 2017 respectively. After graduation, she has been working in State Grid Hunan electric power company.