Hindawi
Mobile Information Systems
Volume 2022, Article ID 2546418, 14 pages
https://doi.org/10.1155/2022/2546418

Research Article

Basketball Big Data and Visual Management System under Metaheuristic Clustering

Hailong Xia[1] **and Long Liu**[2]

[1]*Henan University of Engineering, Zhengzhou 450000, Henan, China*
[2]*Chongqing Preschool Education College, Wanzhou, Chongqing 404100, China*

Correspondence should be addressed to Long Liu; liulong@aynu.edu.cn

Received 12 July 2022; Revised 2 August 2022; Accepted 12 August 2022; Published 21 September 2022

Academic Editor: Yajuan Tang

This study aims to discuss the application value of KMC algorithm optimized by heuristic method in basketball big data analysis and visual management. Because the data in basketball big data is too complicated and incomplete, the extraction of information is not direct and effective enough. Based on the metaheuristic K-Means clustering (KMC) algorithm, the weights and genetic algorithm are introduced to optimize it, and the University of California at Irvine (UCI) data set is applied to analyze the big data clustering performance of the optimized KMC algorithm. The 2018-2019 season National Basketball Association (NBA) shooting guards are selected as the research objects, and the optimized KMC algorithm is used to process the data and analyze the NBA scoring functional factors. It is found that the number of clusters increased from 2 to 16. After optimization, the Between-Within Proportion (BWP) value of the KMC algorithm only drops by 0.35, and the improved BWP (IBWP) value only drops by 0.288, which shows the smallest drop among all the algorithms. When the number of nodes is 4, the running time of the optimized KMC algorithm for processing the COVTYPE data set is 1922 s after optimization, and the running time for processing the IRIS data set is the shortest (113 s). When the number of parallel nodes is 10, the speedup ratio of the optimized KMC algorithm for processing COVTYPE data set is 4.16, and the maximal expansion rate is 0.81. The clustering accuracy of traditional KMC algorithm is 89.33%. After optimization, the clustering accuracy of KMC algorithm is 98.67%. The leader factor, offensive contribution factor, shooting stability factor, and passing ability factor in the core grouping are all at the maximum, which are 0.59, 0.51, 0.47, and 0.43, respectively. The optimized KMC algorithm has been shown to reduce the number of iterations, reduce convergence time, and improve clustering accuracy. The optimized KMC algorithm has been shown to reduce the number of iterations, reduce convergence time, and improve clustering accuracy. The conclusion of this study can provide reference basis for big data clustering and visual management.

1. Introduction

For data research or data application requirements, data visualization is to present specific data in the form of statistical charts and information. Big data analytics refers to the process of extracting potentially valuable information from a large amount of noisy and incidentally incomplete application data [1]. Big data analytics is a poorly multidisciplinary methodology. The main areas are neural networks, pattern recognition, spatial data analysis, image databases, signal processing, artificial intelligence, knowledge base systems, data acquisition, and bioinformatics [2, 3]. Big data analysis has concept description, association analysis, classification and prediction, cluster analysis, external analysis, and evolutionary analysis [4]. Clustering is an unsupervised classification method that automatically divides big data into multiple classes or clusters according to a certain standard. Cluster analysis can preprocess the data by observing the characteristics of each class or concentrating on a certain type of valuable data for further analysis and processing [5]. Cluster analysis is widely used in data analysis, image segmentation, pattern recognition, and other fields [6]. Currently, the common clustering method is the K-Means Clustering (KMC) algorithm based on the

heuristic algorithm. The KMC algorithm is widely used in data statistics, data analysis, and machine analysis due to its short and fast properties [7]. The KMC method based on heuristic algorithm shows significant advantages in small- and medium-scale data analysis. However, when the large-scale data sets are clustered, it is necessary to manually determine the number of clusters, the clustering results are unstable, and the misselecting noise and abnormal points will eventually lead to inefficient data processing and poor clustering quality [8].

The data analysis process is based on a large amount of data, and the ability of human brain to absorb and process information is limited. Visualization technology can transform scientific data into graphic image information that changes with time and space through computer and image processing technology and finally achieve the interactivity, visibility, and multidimensionality of the data [9]. Researchers can analyze the data and its changing trends through graphs and images. Data visualization speeds up data processing and increases the utilization of effective data. Data visualization has been widely used in various fields such as natural sciences, engineering technology, finance, communications, and commerce [10]. Basketball has become a popular sport because of its features such as simplicity, fun, fitness, and education. The depth of basketball is measured by the game. Basketball statistics can make an objective analysis of the data and unearth potential actual combat information. However, there are few studies on applying cluster analysis methods to basketball big data analysis.

In summary, the KMC method based on heuristic algorithm for processing big data has to be further optimized, and there is limited research on applying the clustering data analysis method to basketball data analysis. In this study, the KMC algorithm in the heuristic method is optimized and applied to basketball big data analysis to provide a reference for basketball big data clustering and visual management.

2. Materials and Data

2.1. The Cluster Analysis Methods of Big Data. Big data cluster analysis is the process of grouping a collection of physical or abstract objects into multiple classes composed of similar objects, clustering a collection of data objects in the same cluster. Big data analysis is not a postprocess that obtains effective results after simple analysis of input data. It needs to go through the continuous repetition of a multistep complex process to obtain accurate results. For n vectors in the a-dimensional space Ra, they are assigned to one of the c clusters, so that the distance between each vector and its cluster center is the smallest. Then, the distance between the vectors X_i and X_j can be expressed as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^{a} \left(X_{ik} - X_{jk}\right)^2}, X_i = \{X_{i1}, X_{i2}, \ldots, X_{ia}\}, X_j = \{X_{j1}, X_{j2}, \ldots, X_{ja}\}. \quad (1)$$

Cluster analysis mainly includes two kinds of data matrix and discrepancy matrix [11]. They differ from the matrix diagram method in that they are not filled with symbols on the matrix diagram but filled with data to form a matrix for analyzing the data. The data matrix is a matrix in which d data objects of the entire data set are described with l attributes, and the final data object set is regarded as a $d * l$ matrix. The data matrix can be expressed as follows:

$$\begin{bmatrix} x_{11} \cdots x_{1e} \cdots x_{1p} \\ x_{i1} \cdots x_{ie} \cdots x_{ip} \\ x_{n1} \cdots x_{ne} \cdots x_{np} \end{bmatrix}. \quad (2)$$

The difference matrix refers to the degree of similarity between any two data points in the overall data object set [12], which can be expressed as follows:

$$\begin{bmatrix} 0 & & \\ d(2,1) & 0 & \\ d(n,1) & d(n,2) & 0 \end{bmatrix}, d(i,j) = d(j,i), d(i,j) \geq 0. \quad (3)$$

In (3), n represents the number of data points, and the $d(i, j)$ in the matrix represents the difference degree calculated according to the specified degree of similarity of the data points i and j in the data object collection. The larger the $d(i, j)$ value, the greater the degree of difference between the data objects.

The core of cluster analysis is to obtain the degree of similarity among different data objects [13]. At present, the Minkowski distance calculation method, the Euclidean distance calculation method, and the Chebyshev distance calculation method are commonly used for evaluation [14]. Among them, the data obtained by the Euclidean distance calculation method is not affected by coordinate translation and rotation changes, and it is a commonly used distance similarity measurement method [15]. The calculation method of Euclidean distance is given as follows:

$$d(i,j) = \sqrt{\left|x_{i1} - x_{j1}\right|^2 + \left|x_{i2} - x_{j2}\right|^2 + \cdots + \left|x_{ip} - x_{jp}\right|^2}. \quad (4)$$

In the above equation (4), $d(i, j)$ represents the Euclidean distance between two data points, which satisfies the conditions $d(i,j) \geq 0$, $d(i,j) = d(j,i)$, and $d(i,j) \leq d(i,k) + d(j,k)$.

The similarity factor is mainly used to gauge the similarity among data points [16]. The angle cosine method is a commonly used similarity coefficient calculation method. The value range of the similarity coefficient is [−1, 1]. When

the orthogonal value is 0, it means that the two vectors are completely dissimilar. The calculation method of the similarity coefficient of the angle cosine method is as follows:

$$r_{ij} = \frac{\left|\sum_{k=1}^{p} x_{ik}x_{jk}\right|}{\sqrt{\left(\sum_{k=1}^{p} x_{ik}^2\right)\left(\sum_{k=1}^{p} x_{jk}^2\right)}}. \tag{5}$$

The correlation coefficient method represents the degree of correlation between two data vectors [17], and its value range is [−1, 1]. 0 means that they are not correlated, 1 means that positive correlation is found, and −1 means that negative correlation can be seen. The correlation coefficient method can be expressed as follows:

$$r_{ij} = \frac{\sum_{k=1}^{p}\left(x_{ik} - \overline{x_i}\right)\left(x_{jk} - \overline{x_j}\right)}{\sqrt{\sum_{k=1}^{p}\left(x_{ik} - \overline{x_i}\right)^2\left(x_{jk} - \overline{x_j}\right)^2}}. \tag{6}$$

Appropriate criterion function in cluster analysis can further improve the quality of clustering [18]. The criterion functions commonly used in cluster analysis are as follows: squared margin of error, squared weighted mean value distance sum, and interclass distance sum [19]. The error sum of squares is often used for data analysis with dense samples and little difference between samples [20]. The error sum of square (J_a) can be expressed as follows:

$$J_a = \sum_{i=1}^{n}\sum_{j=1}^{k}\left\|x_i - m_j\right\|^2. \tag{7}$$

In the equation above, m_j is the average value of the class C_k, and $m_j = (1/n_j)\sum_{i=1}^{n_j} x_i$. nj refers to the number of objects in the class C_k.

The interclass distance and criterion (J_b) calculates the distance sum of every clustering epicenter to the global epicenter. The higher the similarity of the research data, the less obvious the clustering result, and the results making J_b the largest result have to be found.

$$J_b = \sum_{j=1}^{k}\left(m_j - m\right)^T\left(m_j - m\right). \tag{8}$$

The weighted average squared distance (J_c) is applicable to data objects with a large disparity in the number of samples, and its calculation method can be expressed as follows:

$$\begin{aligned} J_c &= \sum_{j=1}^{p} p_j s_j^*, \\ s_j^* &= \frac{2}{n_j\left(n_j - 1\right)}\sum_{x\in X_j}\sum_{x\in X_j}\|x - \bar{x}\|^2, p_j = \frac{n_j}{n}. \end{aligned} \tag{9}$$

In the above two equations, s_j^* refers to the average squared distance between samples within a class, and P_j is the prior probability.

2.2. Establishment of Cluster Analysis Method Based on Metaheuristic Algorithm. KMC is the most classic and most widely used clustering method in the metaheuristic algorithm. The kinetic Monte Carlo method (KMC) is simple in principle and highly adaptable, so it is the first choice of researchers in many cases. This method takes Euclidean distance as the correlation measure, and the error sum of squares criterion (J_a) as the criterion function to minimize the evaluation index. The KMC algorithm divides the data set A into the closest classes, and its cluster center is $C_1, C_2, C_3, \ldots, C_k$. The calculation method of each cluster center point is shown in equation (11), in which $i = 1, 2, \ldots, k$ and n_i was the number of data objects in the class C_i.

$$C_i^* = \frac{1}{n_i}\sum_{x_j\in C_i} x_j. \tag{10}$$

The traditional KMC method has a great dependence on the selection of the initial clustering center point, and it is susceptible to the interference of local noise data. The different feature weights assigned to the attributes of each data point can improve the KMC results greatly. The feature weights of variable patterns were assigned to data points, which were named KMC based on density, DK-Mean. The attribute feature weight value of the j-th dimension is assigned to the object data. The calculation method is expressed as follows:

$$w_j = \frac{a_j}{\sum_{j-1}^{m} a_j}, w_j \in [0, 1], \sum_{j-1}^{m} w_j = 1. \tag{11}$$

In (11), a_j is the ratio of the distance between the classes of the attribute and the distance within the classes, and $a_j = d_b/d_i$ (d_b refers to the distance between classes, and $d_b = \sum_{k=1}^{K}(m_{kj} - m_j)^2$; d_i refers to the distance within the class, and $d_i = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ij} - m_{kj})^2$). m_i represents the mean value of the data set on the j-th dimension attribute; K is the number of clusters, and j is the number of attribute bits. Then, the weighted Euclidean distance calculation equation can be written as follows:

$$d(m, n) = \sqrt{\sum_{j=1}^{m} w_j\left(x_{mj} - x_{nj}\right)^2}. \tag{12}$$

The KMC method relies on the cluster center point, which is easy to cause local optimal clustering. Based on the density, the choice of the original aggregation centers is improved accordingly in this study. The clustering criterion function can be denoted as the following equation:

$$J_k = \frac{k_{wi}}{k_{be}}. \tag{13}$$

In (13), k_{wi} represents the distance within the class, and $k_{wi} = \max_{i\in[1,k]}\left\{\min_{j\in[1,C_i]}\left[(1/C_i)\sum_{p=1}^{C_i}\|x_j - x_p\|\right]\right\}$. K_{be} represents the distance between classes, and $k_{be} = \min_{x_p\in C_i, x_q\in C_j, i\neq j}\|x_p - x_q\|$. Then, the density $D(x)$ at sample point X can be expressed as the following equation:

Figure 1: The flow chart of DK-Mean algorithm clustering.

$$D(x) = \left[p \in C \,\middle|\, D_i(x, p) \le r\right]. \tag{14}$$

In (14), D_i represents the weighted Euclidean distance, and r is the specified radius.

Cluster analysis method can solve such problems; cluster analysis method is an exploratory analysis method, which can analyze the inherent characteristics and laws of things and is a commonly used technology in data mining. Genetic algorithm shows good applicability and scalability and can reduce the initialization requirements of traditional clustering algorithms in cluster analysis. The genetic algorithm is introduced further based on the DK-Mean algorithm to increase the accuracy of the clustering algorithm in this study. The genetic algorithm search can minimize the J_k value, and then the fitness feature can be represented as formula (16):

$$f = \frac{1}{J_k} = \frac{k_{be}}{k_{wi}}. \tag{15}$$

The probability of an individual being selected can be expressed as follows:

$$P(x_i) = \frac{f(x_i)}{\sum_{j=1}^{p_s} f(x_j)}. \tag{16}$$

In the above equation, $f(x_i)$ is the fitness value, and $i = 1, 2, \ldots, p_s$.

The crossover operation is performed on two individuals x_1, and x_2, and the new individuals produced by them can be expressed as equations (18)~(19), in which α is the uniform arithmetic crossover parameter.

$$\begin{aligned} \tilde{x_1} &= \alpha x_1 + (1 - \alpha) x_2, \\ \tilde{x_2} &= \alpha x_2 + (1 - \alpha) x_1. \end{aligned} \tag{17}$$

The improved DK-Mean algorithm calculates the data gap matrix and initializes the target eigenvalues to obtain new cluster centers. The specific process of the DK-Mean algorithm is shown in Figure 1.

2.3. Visual Data Analysis and Visualization Based on Clustering Algorithm. The biggest difference between visual analysis and visualization lies in the analysis of this point, the

process of visualizing data for business simulation, correlating multidimensional business data to form a more comprehensive data result, and providing users with auxiliary decision-making process, which is called visual analysis. Parallel coordinate method has the characteristics of mapping high-dimensional data to low-dimensional space and can interact with users at the same time, and it is a commonly used method of visual data analysis at present. In this study, a visual data analysis model based on the KMC algorithm is established based on the optimized KMC algorithm and the parallel coordinate method.

It is assumed that G is a collection of n-dimensional data objects, and $G = \{g_1, g_2, \ldots, g_n\}$, of which gi is an n-dimensional collection $g_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$; the basic coordinate axis $\{x_1, x_2, \ldots, x_n\}$ corresponds to the attribute of the i-th dimension, and each n-dimensional vector can be expressed as $H\{h_1, h_2, \ldots, h_n\}$. The polyline H of the n-dimensional data using linearly independent equations is given as follows:

$$\frac{x_1 - \alpha_1}{\mu_1} = \frac{x_2 - \alpha_2}{\mu_2} = \cdots = \frac{x_n - \alpha_n}{\mu_n}. \tag{18}$$

According to the mapping principle from the midpoint of the coordinate system to the parallel coordinate, the following equation can be obtained:

$$x_{i+1} = m_i x_i + b_i, i = 1, 2, \ldots, n - 1. \tag{19}$$

In the equation, m_i is the slope, and b_i represents the intercept on the axis in parallel coordinates x_{i+1}.

The technology and process of data analysis are applied in the basketball data visualization management system, which can be data processing automation. The data analysis visualization process based on the optimized KMC algorithm is shown in Figure 2. After the sample data is processed through selection operations and cross operations, a new visualization population can be formed.

2.4. The Evaluation Indicators of Cluster Validity. The main indicators in the cluster analysis evaluation are as follows: Accuracy, Precision, Recall, and $F1$ value, four commonly used indicators. The ideal clustering result should reflect the internal structure of the data set as much as possible, so that the sample similarity between classes is the smallest, and the samples within the class are the most similar. In this study, the Between-Within Proportion (BWP) and improved BWP (IBWP) indicators were adopted to analyze the clustering results and performance, where BWP is the ratio of the clustering deviation distance to the clustering distance, and its calculation method is given as follows:

$$\mathrm{BWP}(i, j) = \frac{c(i, j) - w(i, j)}{c(i, j) + w(i, j)}. \tag{20}$$

In the equation, $c(i, j)$ represents the interclass distance of the object i in the j-th class, and $c(i, j)$ represents the intraclass distance of the object i in the j-th class. Among them, $c(i, j) = \min_{a \in [1,k], a \neq j} (1/n_a) \sum_{p=1}^{n_a} \|x_p^a - x_i^j\|^2$, n is the number of data objects, which can be the number of divided clusters, n_a represents the number of elements of the class a, and j represents the class label. The larger the $BWP(i, j)$ value, the more effective the clustering of sample objects.

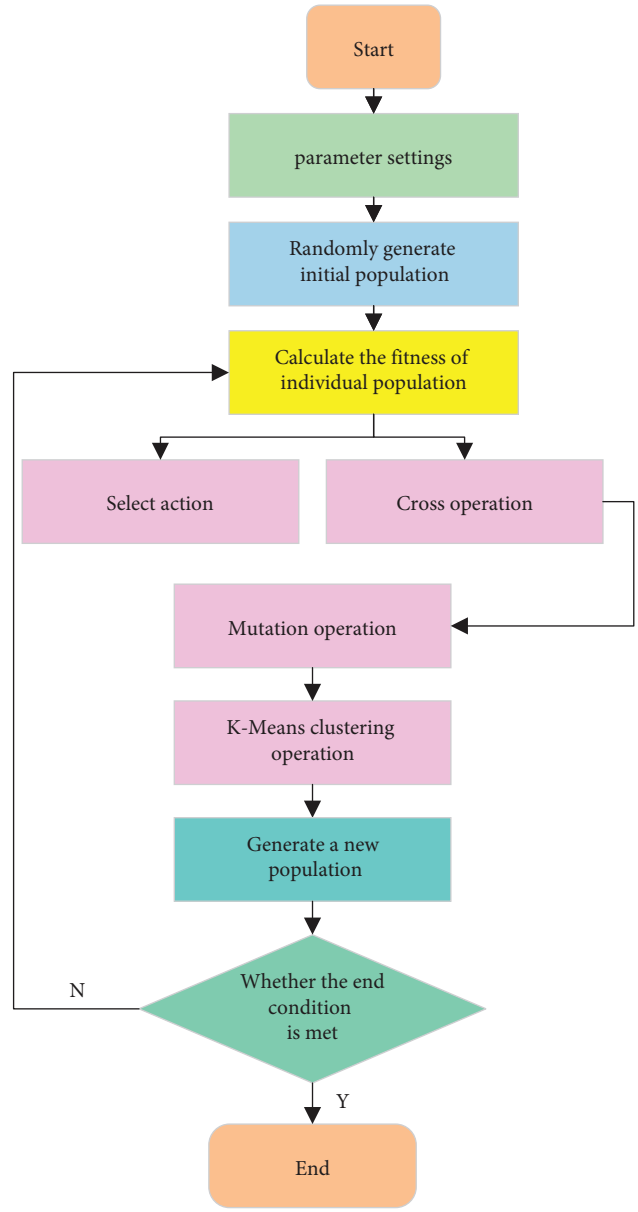

Figure 2: Flow chart of KMC algorithm after optimization.

The IBWP indicators can evaluate the clustering effectiveness of a single data object very well, and its calculation method is shown as follows:

$$\mathrm{IBWP}(i, j) = \frac{ic(i, j) - iw(i, j)}{ic(i, j) + iw(i, j)}. \tag{21}$$

In the (21), $ic(i, j)$ and $iw(i, j)$represent the interclass distance and the intraclass distance of the object i in the j-th class. The larger the value of IBWP index, the more effective the clustering of individual points of the sample.

The speedup ratio and the expansion rate are used to evaluate the parallelization effect and performance of the

Table 1: Data sets under UCI database.

Data sets	IRIS	WINE	SEED	ABALONE	LETTER	COVTYPE
Number of samples	150	178	210	4177	20000	581012
Number of attributions	4	13	8	8	16	54

clustering algorithm in analyzing data. The calculation method of speedup ratio and the expansion rate is given as follows:

$$S = \frac{T_1}{T_m}, \quad S' = \frac{T_n}{T_{mn}}. \tag{22}$$

In the above two equations, T_1 is the data processing time for a single node; T_m is the data processing time for m nodes. N is the size of the processed data; T_n is the time for data to be processed on a child node; and T_{mn} is the time for data of size n to be processed on m nodes.

2.5. Data of Testing Dataset of the Model. In this study, the IRIS, WINE, SEED, ABALONE, LETTER, and COVTYPE data sets in the UCI database are undertaken as the validation sets to verify the algorithm model established. Table 1 shows the total number of samples, dimensions, and categories of the data set. Data analytics is the process of analyzing data sets in order to make decisions about the information they hold, to be used in the business industry to enable organizations to make business decisions.

2.6. Research Objects and Methods of Basketball Big Data. The NBA shooting guards in the 2018-2019 season were selected as the research objects. The relevant indicators included in the study were analyzed statistically using literature data method, logical analysis method, mathematical statistics method, video analysis method, and comparative analysis method. The technical statistical data of the season finals are collected on related websites such as Tencent Sports Video, Hupu NBA, the control video, and official statistics, which were repeatedly confirmed to ensure the authenticity and reliability of the data source. These raw data were adopted for statistical analysis of basketball technical indicators. Data analytics can help businesses better understand their customers, improve their advertising campaigns, personalize their content, and improve their bottom line.

2.7. Analysis on Influencing Factors of Basketball Scoring Based on KMC Algorithm. The 2018-2019 season NBA scoring guards were selected as the research objects. Based on the NBA data query website (https://www.basketball-reference.com/), 17 basic pieces of data such as player scores, rebounds, assists, and steals, as well as advanced data such as passing ability, defensive contribution, and offensive contribution are selected as the original data. The original data removes rebounds, blocks, and fouls and reduces the original data from 36 dimensions to 22 dimensions, including Field goal attempts (FGA), Field goals (FG), Free throws (FT), Free throw attempts (FTA), Assists (AST), Steals (STL), and Points (PTS).

3. Results

3.1. Analysis on Results Based on Metaheuristic Clustering. Using the traditional KMC algorithm to cluster the data in the IRIS data set, the clustering results are divided into 4 clusters, but the clustering results of some data overlap (Figure 3(a)), and the clustering results are significantly different from the real data. The optimized KMC algorithm is used to cluster the data in the IRIS data set. The clustering results are divided into 4 clusters, the data clustering results are of high quality, and there is no crossover phenomenon of different types of data (Figure 3(b)). In addition, there is no difference between the clustering results and the real data.

3.2. Comparison on Classification Indicators of Different Clustering Algorithms. The traditional KMC algorithm is compared with DKMC, the optimized KMC algorithm, self-organizing feature map (SOM) algorithm, quantum evolutionary clustering algorithm (QEAM), and k-medoids algorithm in terms of BWP values (Figure 4(a)). As the number of clusters increases, the BWP values of different algorithms show a downward trend all, and the number of clusters increased from 2 to 16. After optimization, the BWP value of the KMC algorithm only drops by 0.35, which is the smallest drop among all algorithms. The BWP value of an optimized KMC algorithm with the same number of clusters is higher than that of other algorithms. The IBWP values of the different algorithms are compared, and the results are shown in Figure 4(b). As shown in the figure, the IBWP values of the various algorithms show a decreasing trend as the number of clusters increases. The number of clusters increases from 2 to 16. After optimization, the IBWP value of the KMC algorithm only decreases by 0.288, which is the smallest decrease of all algorithms. The IBWP value of an optimized KMC algorithm with the same number of clusters is higher than that of other algorithms.

3.3. Execution Time of Clustering Algorithm. The optimized KMC algorithm was used to perform cluster analysis for six datasets in the UCI database (Figure 5). As the number of nodes increases, the time taken by the KMC algorithm to collect 6 different data sets decreases. If the number of nodes is 4, the COVTYPE dataset output time is 1922 s, and the IRIS dataset output time is 113 s.

3.4. Performance Analysis of Clustering Algorithm in Parallel Data Processing. The speed ratios of the optimized KMC algorithm between the six data sets were analyzed and

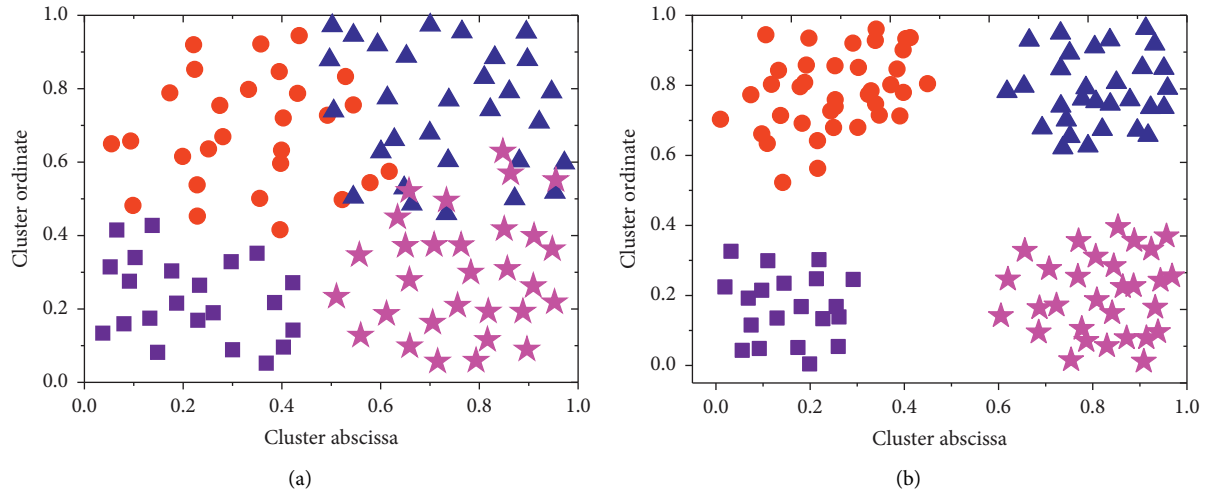

FIGURE 3: Comparison of cluster results of different algorithms. (a) The clustering results of the traditional KMC algorithm; (b) the clustering results of optimized KMC algorithm.

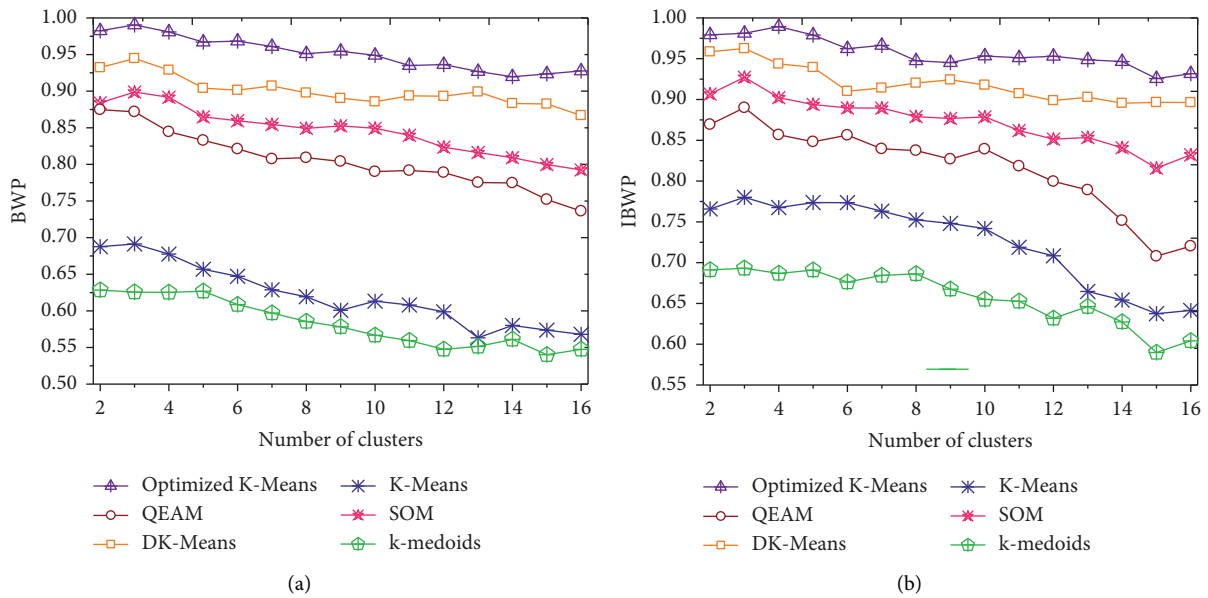

FIGURE 4: Comparison on classification indicators of different clustering algorithms. (a) Comparison on BWP values of different clustering algorithms; (b) comparison on IBWP values of different clustering algorithms.

compared, and the comparison results are shown in Figure 6(a). As the number of parallel nodes increases, the speed ratio of the KMC 6 algorithm when processing the data set shows an increasing trend. If the number of parallel nodes is 10, the maximum speed ratio of the COVTYPE data set is 4.16. The expansion ratios of the KMC-optimized algorithm between the six data sets were analyzed and compared, and the comparison results are shown in Figure 6(b). As the number of parallel nodes increases, the expansion ratio of the KMC algorithm for processing the six data sets appears to decrease. If the number of parallel nodes is 10, the maximum expansion ratio of the COVTYPE data set is 0.81.

3.5. Analysis on Test Performance Rate of Clustering Algorithm. The group accuracy of the different cluster algorithms across the different datasets is compared, and the results are shown in Figure 7(a). There are three different algorithms for the accuracy of data grouping. After optimization, the accuracy of the KMC algorithm when processing the six data sets was clearly higher than the other two algorithms ($P < 0.05$). The convergence times of the different clustering algorithms in the different data sets are compared, and the results are shown in Figure 7(b). The convergence time of the optimized KMC algorithm in the different data sets was shorter than that of the other algorithms, and the difference was statistically significant ($P < 0.05$). The number

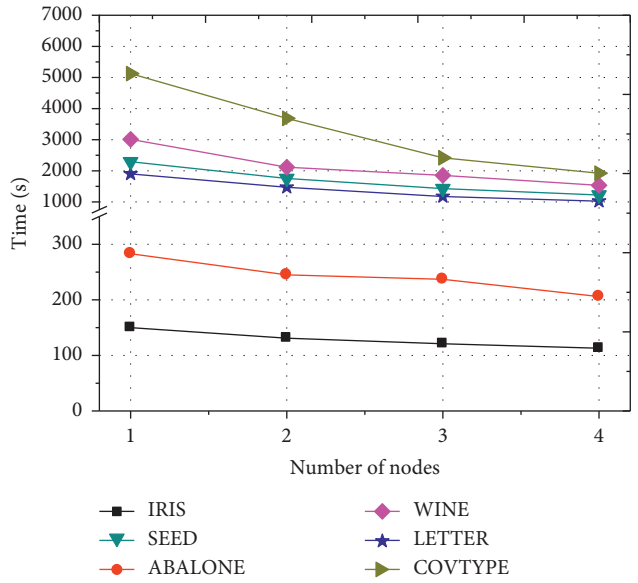

Figure 5: Comparison on execution times of different algorithms under different data sets.

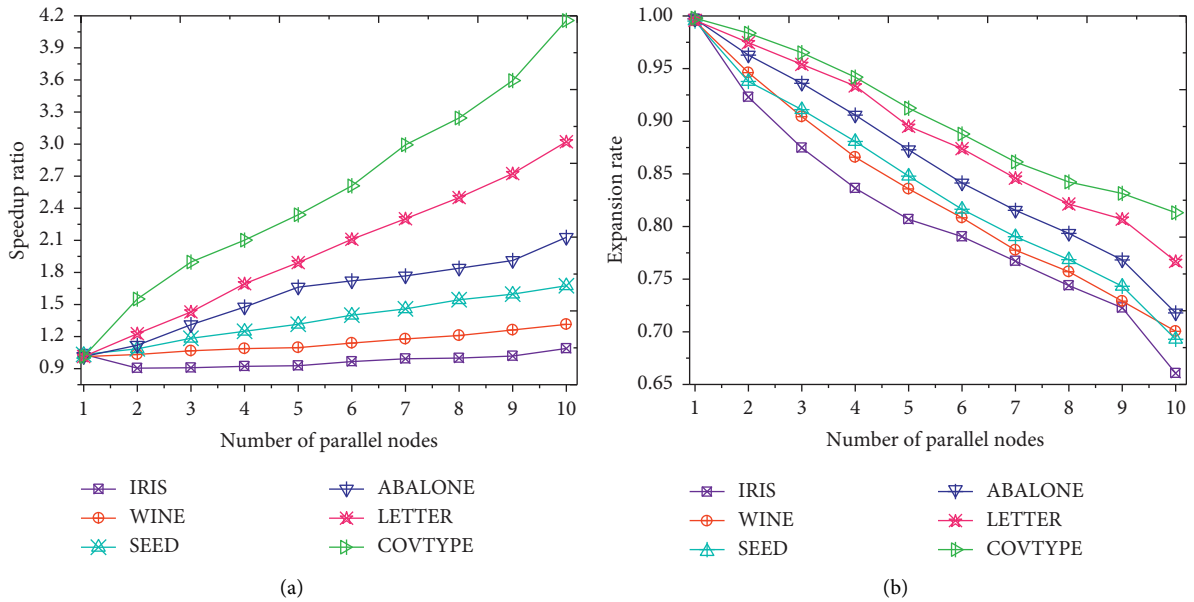

Figure 6: Performance analysis of clustering algorithm in parallel data processing. (a) Comparison on speedup ratio of the optimized algorithm on different data sets; (b) comparison on the expansion rate of the optimized algorithm on different data sets.

of iterations of different clustering algorithms in different datasets is compared, and the results are shown in Figure 7(c). The number of iterations of the optimized KMC algorithm across the different data sets was lower than that of the other algorithms, and the difference was statistically significant ($P < 0.05$).As the convergence time increases, different clustering algorithms are proportional to the data iteration effect.

3.6. Analysis on Cluster Visualization Result. The traditional KMC algorithm and the optimized KMC algorithm are performed with the cluster analysis on the Luanweihua data set in the IRIS data set. The cluster visualization analysis is performed on three types of Luanweihua data in this study. After cluster analysis, all data are divided into 3 clusters, with 50 groups of data in each cluster of original data. After clustering using the traditional KMC algorithm, a total of 16 sets of data have been misclassified, and the clustering accuracy is 89.33%, as illustrated in Figure 8. After clustering using the optimized KMC algorithm, there are two sets of data that are misclassified, and the clustering accuracy is 98.67% (as given in Figure 9). Through the clustering visualization analysis of the traditional KMC algorithm, the

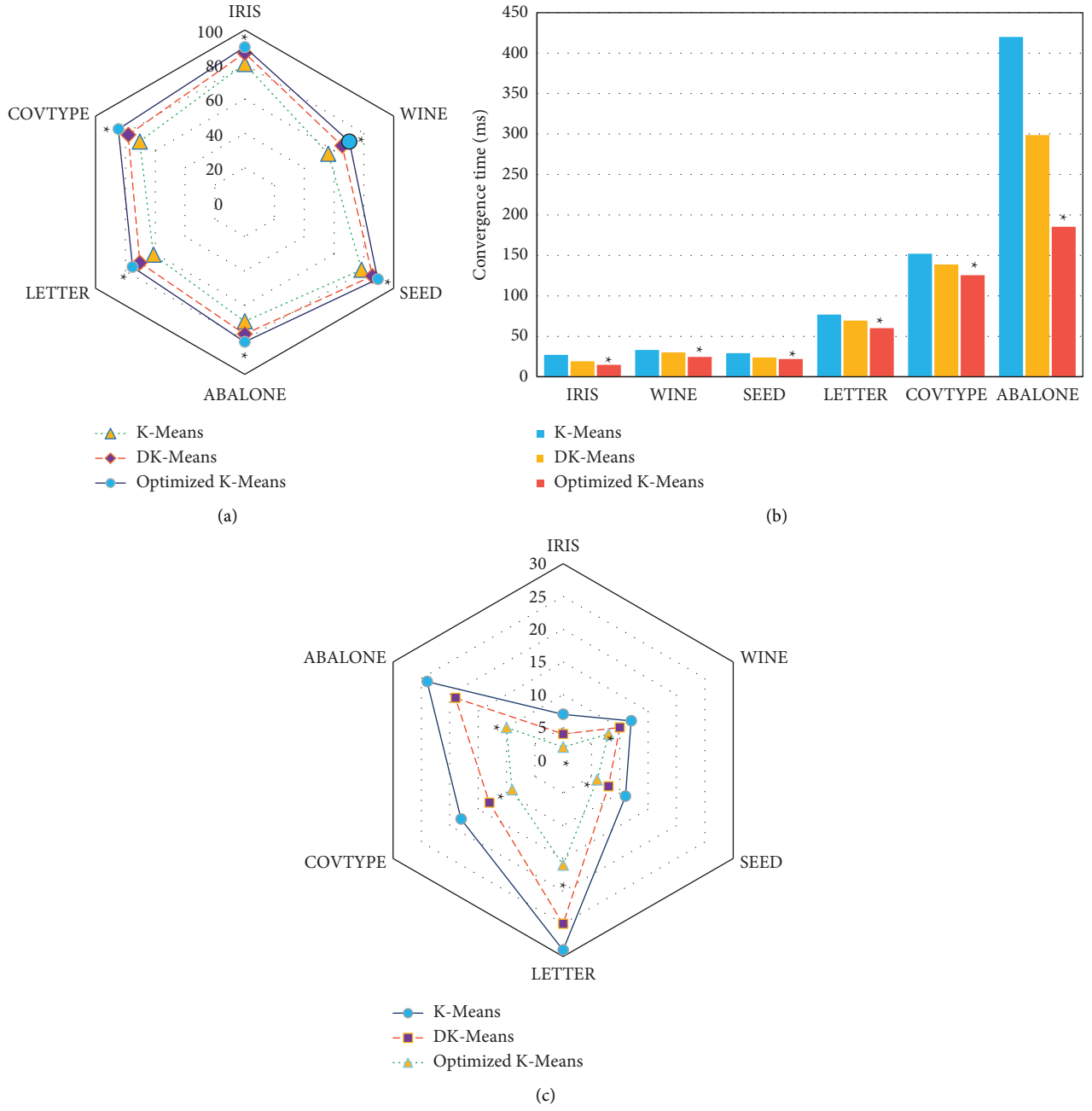

Figure 7: Analysis on test performance rate of clustering algorithm. (a) The clustering accuracy of the clustering algorithm on the test set. (b) The convergence time of the clustering algorithm on the test set. (c) The number of iterations of the clustering algorithm on the test set. Note: $*$ suggested that the difference was statistically great in contrast to the optimized KMC algorithm, $P < 0.05$.

traditional KMC clustering accuracy is about 10% lower than that of the optimized KMC.

3.7. Analysis on the Statistic Results of Basketball Technical Indicators. According to the data from the NBA official website, the technical indicators of the top 10 Eastern teams in the 82 regular seasons in the 2018-2019 season are counted, and the results are shown in Figure 10. The figure illustrates that, except for the significant differences between the lost points and the score items, there is little difference in other indicators. The field goal score of Indiana Pacers is 25.4, which is lower than the first place (Milwaukee Bucks, 38.2 scores), showing a difference of 12.8 between the two. The comparison on the indicators of the first and tenth teams shows that the scores for shots, hits, and rebounds of Milwaukee Bucks are 5.8, 2.8, and 3.4 higher than those of the Miami Heat.

The technical indicators of the top 10 Midwestern teams in 82 regular season games are compared, and the results are given in Figure 11. The free throw of Houston Rockets is as high as 45.4, which is obviously higher than that of other teams. The scores in rebounds and assists of Los Angeles

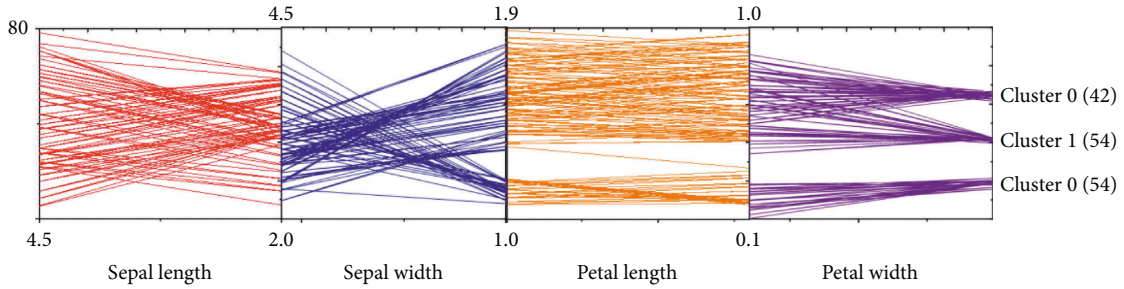

Figure 8: The clustering visualization results of traditional KMC algorithm.

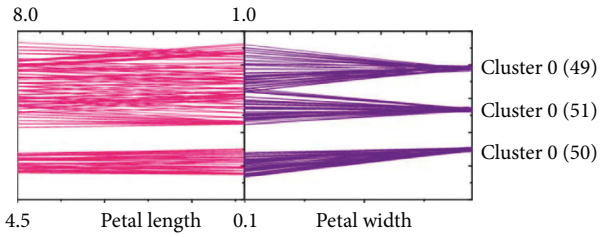

Figure 9: The clustering visualization results of optimized KMC algorithm.

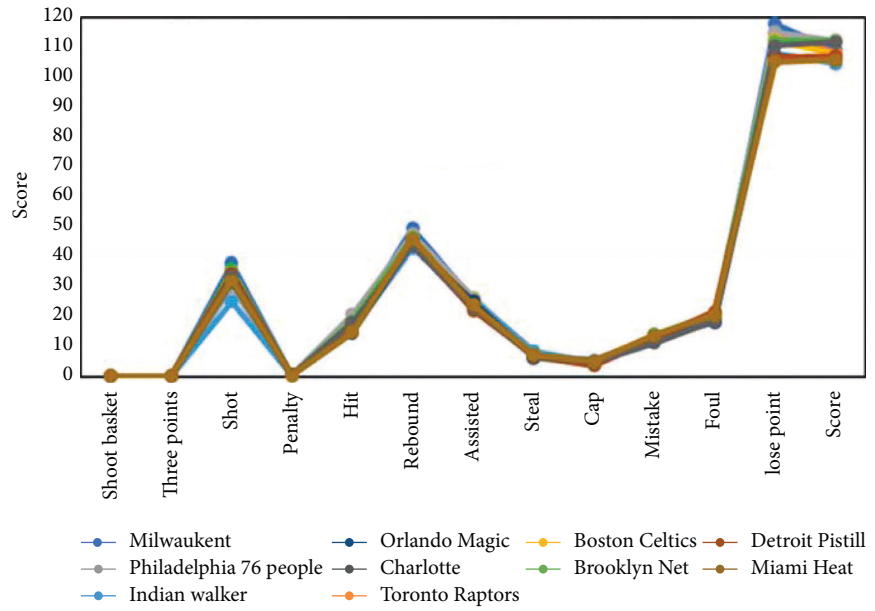

Figure 10: Regular season performance of the top 10 teams in Eastern conference.

Clippers are 22.6 and 28.5, respectively, which are much higher than those of other teams. The shooting percentage of Los Angeles Lakers (the 10th) is 0.7, which is much lower than that of the San Antonio Spurs (0.82) and the Golden State Warriors (0.8).

A cluster analysis is performed in 20 teams in regular games (Figure 12), which shows that 20 teams in the East and West are clustered into 7 categories. Among them, the Golden State Warriors and the Milwaukee Bucks, the first place in the East and West teams, are grouped into the same category. It shows that the top NBA teams have similar characteristics to a certain extent.

The technical indicators of the Golden State Warriors and the Milwaukee Bucks team in the East and West teams are compared, as given in Figure 13. The scores in shooting, attempts, free throws, rebounds, assists, and blocks of the Golden State Warriors and the Milwaukee Bucks are higher than the average scores of all teams. The Milwaukee Bucks and the Golden State Warriors have rebound scores of 49.8 and 46.2, respectively, which are higher 4.6 points and 1 point than the average scores of all teams, respectively, which shows that the rebounding technical indicators of the Milwaukee Bucks have a significant advantage. The scores in assists of the Milwaukee Bucks and the Golden State

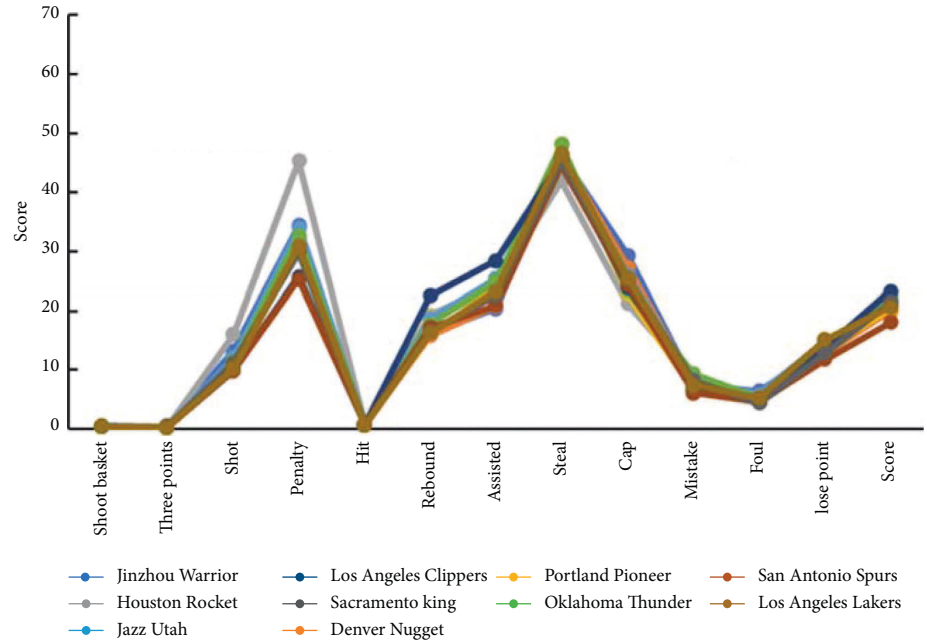

Figure 11: The regular season performance of the top 10 teams in Western.

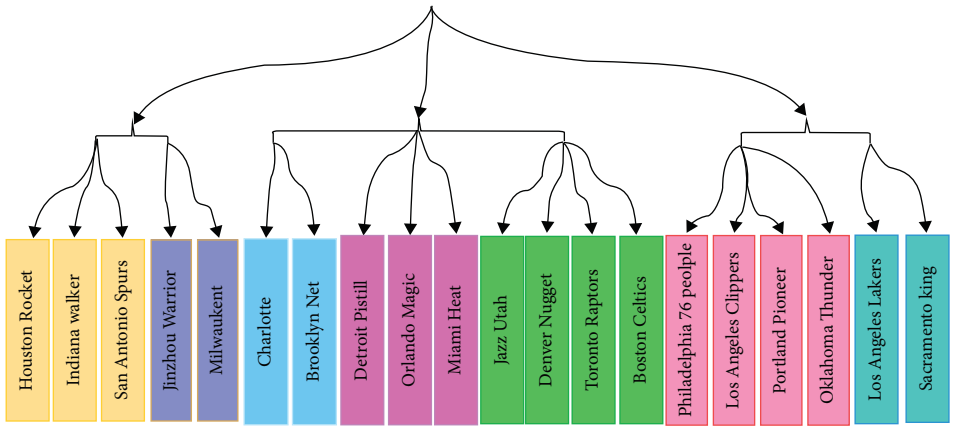

Figure 12: Dendrogram of NBA team clustering results.

Warriors were 26 and 29.4, respectively, showing that the score in assists of the Golden State Warriors has a significant advantage.

3.8. Analysis on the Result of the Influencing Factors of Basketball Scoring. The coefficients of the NBA basketball score scheme in the different groups were compared based on the optimized KMC algorithm. As shown in Figure 14, the cluster boundary factor first increases and then decreases as the number of clusters increases. If the number of clusters is 7, the cluster boundary factor reaches a maximum value of 0.24.

Based on the optimized KMC algorithm, the functional factors of basketball NBA scores are analyzed, and the matrix of different factors after coordinate translation is shown in Table 2. All factor coefficients are close to 0 or 1.

The influencing factors of basketball NBA score are analyzed based on the optimized KMC algorithm, and the distribution of different factor cluster centers is shown in Figure 15. The leader factor, offensive contribution factor, shooting stability factor, and passing ability factor in the absolute core grouping are all the maximum values, which are 0.59, 0.51, 0.47, and 0.43, respectively.

4. Discussion

In this study, the KMC algorithm in the metaheuristic clustering method is optimized, its clustering and visualization performance are analyzed, and it is applied to the analysis of basketball NBA score functional factors. It is found that the clustering results of traditional KMC algorithm have the overlapping of some data clustering results,

Figure 13: Comparison of team technical indicators.

Figure 14: Analysis on changes in cluster profile coefficients.

and there is a big difference with the real data. The optimized KMC algorithm does not have the crossover phenomenon of different types of data, and the clustering results are closer to the real data. Such results prove that the optimized KMC algorithm shows improved quality of clustering results. The clustering centers of the traditional KMC algorithm are randomly selected, which leads to errors in the clustering results, and the clustering analysis of the traditional KMC algorithm requires multiple iterations, so the clustering results are greatly different from the true data distribution. The optimized KMC algorithm has coordinate rotation to select the cluster center, which reduces its randomness, so the initial cluster center can be determined more accurately. The number of clusters increased from 2 to 16. After optimization, the BWP value of the KMC algorithm only drops by 0.35, and the IBWP value only drops by 0.288, which is the smallest drop of all algorithms. Such results suggest that the optimized KMC algorithm shows better clustering results. The BWP and IBWP values of the optimized KMC algorithm are greater than those of other algorithms, indicating that the optimized KMC algorithm shows higher clustering accuracy on the samples. As the number of nodes increases, the time for the KMC algorithm to cluster 6 different data sets shows a downward trend. When the number of nodes is 4, the optimized KMC algorithm can process the COVTYPE data set for a maximum of 1922 s, and the shortest running time for processing the IRIS data set is 113 s. The sample size of the IRIS dataset is observably lower than that of the COVTYPE dataset. Such results indicate that the optimized KMC algorithm takes longer time to process low sample size data. This is because each operation needs to start the Map and Reduce tasks, which takes a certain amount of time, so when the task start time dominates, the small samples are processed. Xu pointed out that as the number of sample nodes increases, the running time of clustering decreases. The larger the data size, the better the acceleration ratio of the algorithm, and the better the algorithm's ability to handle large data. If the number of parallel nodes is 10, the maximum speed ratio of the KMC algorithm optimized for processing the COVTYPE data set is 4.16. He found that the optimized KMC algorithm has several advantages in big data processing. As the number of parallel nodes increases, the level of expansion of the KMC algorithm for processing the six datasets appears to decrease. If the number of parallel nodes is 10, the maximum expansion rate of the COVTYPE data set is 0.81. This is because the amount of communication between each node increases as the number of nodes increases. Some studies have shown that as data size increases, parallel uptime increases, which is similar to the results of this study. This is because the optimized KMC algorithm introduces weights in the calculation of the Euclidean distance, which increases the Euclidean distance between the abnormal point and the cluster center and makes the algorithm iteration result closer to the real data, thereby reducing the number of iterations and

Table 2: Component matrix after coordinate translation.

	Leader factor	Offensive contribution factor	Defensive contribution factor	Three-point ability factor	Shot stability factor	Passing ability factor
FGA	0.942	0.036	−0.058	0.086	−0.053	−0.013
FG%	0.083	0.332	−0.016	−0.22	0.879	−0.059
FT	0.755	0.455	−0.065	−0.259	−0.154	0.074
FTA	0.762	0.414	−0.016	−0.309	−0.146	0.078
AST	0.463	0.127	0.181	−0.164	−0.01	0.78
STL	0.059	0.056	0.91	−0.101	0.029	0.065
PTS	0.906	0.3	−0.098	0.1	0.118	−0.01
AST %	0.563	0.165	0.154	−0.196	−0.018	0.7
STL%	0.042	0.058	0.901	−0.138	0.014	0.049

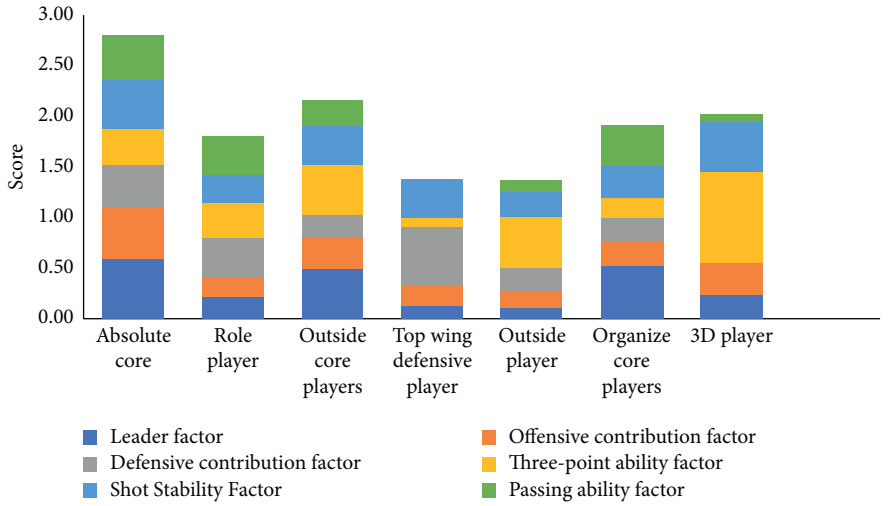

Figure 15: Statistical results for cluster distribution of basketball score influencing factors.

convergence time and improving clustering accuracy. Yin pointed out that the optimized KMC algorithm has improved the clustering efficiency, and its clustering accuracy has not changed greatly. The reason is that the study did not consider the impact of each sample data on the entire clustering result, so the Euclidean distance calculation method was not optimized. The clustering accuracy of the traditional KMC algorithm is 89.33%. After optimization, the clustering accuracy of the KMC algorithm is 98.67%, and the clustering accuracy is improved by 9.34%.

Effective clustering can correctly display the player's status, which is helpful for the rational operation of the team. The research results of this study show that as the number of clusters increases, the cluster contour coefficients first increase and then decrease. When the number of clusters is 7, the cluster contour coefficient reaches the maximum value of 0.24. The leader factor, offensive contribution factor, shooting stability factor, and passing ability factor in the absolute core grouping are all the maximum values, which are 0.59, 0.51, 0.47, and 0.43, respectively. These results show that the absolute core group has an important influence on the team's score. The main influence factors of absolute core are leader factor, offensive contribution factor, shooting stability factor, and passing ability factor.

5. Conclusion

The KMC algorithm in metaheuristic clustering is optimized and applied to the analysis of NBA scoring functional factors in this study. The results of statistical analysis of basketball technical indicators show that there are significant differences in the gain and loss of scores, and other differences are not significant. It turns out that the optimized KMC algorithm reduces the number of iterations and convergence time and improves the clustering accuracy. The leader factor, offensive contribution factor, shooting stability factor, and passing ability factor are functional factors of NBA scoring. However, there are still some shortcomings in this study. Only a preliminary analysis of the functional factors of basketball NBA scores has been carried out, and the clustering results of different players of different teams have not been analyzed and verified. Therefore, it will further increase

the sample size and perform cluster analysis to verify the different players of the team in future. In short, this study provides a reference basis for big data clustering and visual management.

Data Availability

The data used to support the findings of this study can be obtained from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

This work was supported by 2022 Key Scientific Research Projects of Higher Education Institutions in Henan Province, project no. 22A890001.

References

[1] X. Zhang, E. J. Pérez-Stable, P. E. Bourne et al., "Big data science: opportunities and challenges to address minority health and health disparities in the 21st century," *Ethnicity & Disease*, vol. 27, no. 2, pp. 95–106, 2017.

[2] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: a systematic review," *JMIR Medical Informatics*, vol. 4, no. 4, p. e38, 2016.

[3] L. N. Sanchez-Pinto, Y. Luo, and M. M. Churpek, "Big data and data science in critical care," *Chest*, vol. 154, no. 5, pp. 1239–1248, 2018.

[4] A. Madabhushi and G. Lee, "Image analysis and machine learning in digital pathology: challenges and opportunities," *Medical Image Analysis*, vol. 33, pp. 170–175, 2016.

[5] J. S. Beckmann and D. Lew, "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities," *Genome Medicine*, vol. 8, no. 1, p. 134, 2016.

[6] J. Xia, J. Wang, and S. Niu, "Research challenges and opportunities for using big data in global change biology," *Global Change Biology*, vol. 26, no. 11, pp. 6040–6061, 2020.

[7] S. Dirmeier, M. Emmenlauer, C. Dehio, and N. Beerenwinkel, "PyBDA: a command line tool for automated analysis of big biological data sets," *BMC Bioinformatics*, vol. 20, no. 1, p. 564, 2019.

[8] S. U. Park, H. Ahn, D. K. Kim, and W. Y. So, "Big data analysis of sports and physical activities among Korean adolescents," *International Journal of Environmental Research and Public Health*, vol. 17, no. 15, p. 5577, 2020.

[9] H. R. Thornton, J. A. Delaney, G. M. Duthie, and B. J. Dascombe, "Developing athlete monitoring systems in team sports: data analysis and visualization," *International Journal of Sports Physiology and Performance*, vol. 14, no. 6, pp. 698–705, 2019.

[10] A. Alonso-Betanzos and V. Bolón-Canedo, "Big-data analysis, cluster Analysis, and machine-learning approaches," *Advances in Experimental Medicine and Biology*, vol. 1065, pp. 607–626, 2018.

[11] A. M. AbdelAziz, T. Soliman, K. K. A. Ghany, and A. Sewisy, "A hybrid multi-objective whale optimization algorithm for analyzing microarray data based on Apache Spark," *PeerJ Computer Science*, vol. 7, p. e416, 2021.

[12] M. M. Saeed, Z. Al Aghbari, and M. Alsharidah, "Big data clustering techniques based on Spark: a literature review," *PeerJ Computer Science*, vol. 6, p. e321, 2020.

[13] H. Mushtaq, N. Ahmed, and Z. Al-Ars, "SparkGA2: production-quality memory-efficient Apache Spark based genome analysis framework," *PLoS One*, vol. 14, no. 12, Article ID e0224784, 2019.

[14] H. Xia, W. Huang, N. Li, J. Zhou, and D. Zhang, "PARSUC: a parallel subsampling-based method for clustering remote sensing big data," *Sensors*, vol. 19, no. 15, p. 3438, 2019.

[15] V. Ravuri and S. Vasundra, "Moth-flame optimization-bat optimization: map-reduce framework for big data clustering using the moth-flame bat optimization and sparse fuzzy C-means," *Big Data*, vol. 8, no. 3, pp. 203–217, 2020.

[16] V. Mayer-Schönberger and E. Ingelsson, "Big Data and medicine: a big deal?" *Journal of Internal Medicine*, vol. 283, no. 5, pp. 418–429, 2018.

[17] M. A. Levin, J. P. Wanderer, and J. M. Ehrenfeld, "Data, big data, and metadata in anesthesiology," *Anesthesia & Analgesia*, vol. 121, no. 6, pp. 1661–1667, 2015.

[18] B. Karmakar, S. Das, S. Bhattacharya, R. Sarkar, and I. Mukhopadhyay, "Tight clustering for large datasets with an application to gene expression data," *Scientific Reports*, vol. 9, no. 1, p. 3053, 2019.

[19] A. A. Qaffas, R. Hoque, and N. Almazmomi, "The internet of things and big data analytics for chronic disease monitoring in Saudi arabia," *Telemedicine and e-Health*, vol. 27, no. 1, pp. 74–81, 2021.

[20] A. Waschkau, D. Wilfling, and J. Steinhäuser, "Are big data analytics helpful in caring for multimorbid patients in general practice? - a scoping review," *BMC Family Practice*, vol. 20, no. 1, p. 37, 2019.