Journal of Computers Vol. 36 No. 2, April 2025, pp. 105-118
https://doi.org/10.63367/19911599202504360200 8

Academic Early Warning Model Based on Improved Global K-means Algorithm

Jun-Jie Liu, Yong Yang*, and Fu-Heng Qu

College of Computer Science and Technology, Changchun University of Science and Technology
130022, Changchun, Jilin, China
{jjliu, yangy, qufh}@cust.edu.cn

Received 27 February 2024; Revised 21 March 2024; Accepted 21 March 2024

Abstract. In digitizing education, the academic early warning is an important but difficult aspect. By collecting data on students' learning behaviors and analyzing it in real time, it is possible to provide early warning on students' learning status and take timely interventions to optimize the allocation of educational resources. Educational data mining technology provides a reliable method for academic warning. Among them, K-means algorithm, as a reliable data mining technology, has been widely used in the research of students' academic warning. However, the traditional K-means algorithm is very sensitive to the initial center value and easy to fall into the local optimum. In addition, there are a lot of redundant calculations in the iterative process, which have great limitations in practical academic warning applications. In this study, an early academic warning model is proposed to be constructed based on the improved global K-means algorithm. By introducing the incremental center selection method of global K-means algorithm to improve the warning accuracy of the model, and integrating principal component analysis (PCA), incremental center selection of multi sphere splitting and Hamerly algorithm to optimize the computational efficiency of the model. As a result, students' academic performance can be predicted quickly and accurately with fewer input features to achieve the desired early warning effect. The experimental results show that the constructed academic warning model has higher accuracy, precision, recall, and significantly improves the computational efficiency.

Keywords: global k-means algorithm, multi-ball splitting, hamerly algorithm, big data for education, student academic warning

1 Introduction

The swift advancement of digitalization and artificial intelligence is catalyzing significant changes in the educational sector, enhancing teaching methodologies and expanding the use of big data in education [1-5]. The "14th Five-Year Plan for National Informatization," released in December 2021, underscores the imperative for digital transformation in education as a means to drive high-quality educational development and bolster China's overall national strength. The December 2023 "Limitless Possibilities - Digital Development Report of World Higher Education" highlights how digital technologies are refining higher education teaching evaluations, making them more scientific with intelligent, real-time analytics that provide precise problem feedback and targeted instructional guidance [6-9].

Furthermore, personalized learning experiences are being shaped by digital tools that cater to individual learning styles and speeds, supported by algorithms that analyze student data to optimize educational pathways and resources, thus making education more accessible and inclusive [10-12]. The incorporation of virtual and augmented reality into educational settings is revolutionizing traditional classrooms into interactive and immersive learning environments. These technologies not only enhance student engagement but also facilitate a more dynamic exploration of complex concepts [13-21]. Moreover, digital education is enabling global collaboration among educational institutions. Platforms that support cross-border learning and idea exchange allow students from different geographical locations to collaborate on projects, attend virtual lectures by international experts, and access a broader, more diverse set of educational resources [22-27].

Academic early warning, as an important hot branch of educational big data mining research, is an innovative

* Corresponding Author

application of information and artificial intelligence technology in the field of education [28-32]. By mining and analyzing students' academic data, it can help students, parents, and teachers discover problems in learning in a timely manner and take corresponding educational interventions to improve students' academic performance and comprehensive quality [33-35]. Previously, there have been numerous research efforts on the problem of academic warning, but most of them have relied on supervised classification techniques such as Bayesian decision and SVM for academic early warning of students' performance in a particular subject [36-38]. Such techniques require a certain number of labeled historical data features to be acquired in advance. Moreover, the distribution between the historical data of the course and the new data to be classified should have some consistency or correlation, which puts higher demands on the data [39-43].

Compared with supervised classification, the use of unsupervised classification technology such as clustering algorithm for academic early warning has the advantages of less data requirements and high applicability. For example, K-means [44-46], as the most typical clustering algorithm, is increasingly favored by researchers due to its simplicity and ease of implementation. Wang [47] proposed an appropriate and timely warning and pre school K-nearest neighbor algorithm classification model. Based on the ideas of data mining, collecting historical data, and appropriate transformation, statistical analysis techniques were used to analyze the many factors that affect the CET-4 exam, and the CET-4 exam results and their influencing factors were obtained. At the same time, K-weighted K-nearest neighbor algorithm and segmentation algorithm were used in the classification prediction of CET-4 exam scores, and statistical methods were used to study the relevant factors affecting CET-4 exam scores. Screen classification was also performed to predict when to pass the comparative validation. The weight K of input features and adjacent features was weighted. Although the allocation algorithm for adjacent classification performance did not significantly improve, the stability classification was better than the K-nearest neighbor method, which greatly improved the classification efficiency, shortened the classification time, and increased the classification efficiency by 119%. Recent advancements in AI have introduced deep learning techniques that further enhance the predictive capabilities of academic early warning systems. These systems employ neural networks that are capable of identifying complex patterns and correlations in data that traditional models might miss. For example, recurrent neural networks (RNNs) have been applied to sequence prediction problems such as student performance trajectories, allowing for more dynamic and temporally sensitive predictions. Moreover, the integration of sentiment analysis and natural language processing (NLP) techniques into early academic warning systems enabling a more holistic view of student well-being. By analyzing communication patterns, social interactions, and emotional sentiment, educators can gain insights into students' mental health and social dynamics, which are crucial for academic success. These technological advances are being complemented by more sophisticated data collection methods, including real-time data streaming from digital learning platforms. This allows for continuous monitoring and immediate intervention, making academic support more responsive and tailored to individual student needs. Furthermore, the ethical implications and challenges of data privacy in academic early warning systems are being addressed through stricter data protection laws and advanced security measures. Ensuring the privacy and security of student data is paramount, as these systems become more integrated into everyday educational practices.

Based on the outlier data mining technology, Tian et al. [48] proposed to use the K-mean algorithm to analyze students' annual academic performance and establish an academic early warning model, so as to remind students of their academic status in a timely manner and follow up dynamically for early warning research and judgment. Moreover, Li et al. [49] proposed a weighted K-mean algorithm for cluster analysis based on data characterizing students' lifestyle patterns, learning patterns and Internet usage patterns, so as to identify the factors affecting college students' academic performance and to provide decision support to help underperforming students achieve better results.

Nam et al. [50] collected a comprehensive dataset in an online learning environment, which included metrics such as the number of correct responses to classroom questions, the frequency of students' confusion as indicated by interactions with specific slides, and the overall engagement with educational materials such as slides and videos. They applied the K-means clustering algorithm to predict academic outcomes and found it particularly effective at forecasting which students were likely to fail their end-of-semester examinations. This predictive capability forms a crucial foundation for universities to establish academic early warning systems, aiming to identify and support at-risk students proactively.

In a separate study, Wang et al. [51] used the K-means algorithm to analyze the academic records of university students over a four-year period, focusing on compulsory course scores. Their analysis involved deep data mining that uncovered the distribution of student scores across various subjects and assessed the relative importance of each subject. The results from this clustering provided valuable insights that enabled educators to tailor their teaching strategies, offering personalized guidance and "precision education" that could potentially enhance

student performance and engagement. This approach not only helps in optimizing teaching methods but also empowers students to better allocate their time and effort towards their studies.

Additionally, Xiong [52] employed the K-means algorithm to explore learning behavior characteristics among students. The analysis provided early warnings based on three key metrics: grade point prediction, risk of failing courses, and anomalies in behavior patterns. These early warnings are designed to facilitate "early intervention, precise assistance, and timely follow-up," thereby improving educational outcomes and student success rates.

However, despite the benefits demonstrated by these studies, the traditional K-means algorithm has inherent drawbacks. It is highly sensitive to the choice of initial cluster centers, often leads to local optimization rather than global optimum, and involves a significant number of redundant calculations during the iterative process. These issues can result in a low precision of predictions and inefficiencies in computational processes. To overcome these challenges, this study proposes an academic early warning model based on an enhanced version of the global K-means algorithm. This model introduces several innovations: it incorporates a global K-means approach to improve solution accuracy [53], uses principal component analysis (PCA) to reduce data dimensionality and lower the computational demands of processing high-dimensional data [54], integrates BKM and multi-sphere splitting techniques to accelerate the selection of incremental centers [55, 56], and employs the Hamerly algorithm to reduce redundant distance calculations between data points and cluster centers [57], thereby enhancing the overall efficiency and effectiveness of the model.

2 Improve the Global K-means Clustering Algorithm

2.1 Global K-means Clustering Algorithm

The Global K-means clustering algorithm is a distinctive incremental clustering approach designed to address some of the limitations inherent in traditional clustering methods [55]. A primary goal of this algorithm is to minimize the sum-of-squares error, which is the total squared error between each point in a cluster and the cluster's centroid. This makes the algorithm particularly robust in generating clusters that are compact and well-separated. Unlike many clustering techniques, the Global K-means algorithm does not rely on the initial position of any cluster centers and is free from empirically tunable parameters. This characteristic significantly reduces the influence of random initialization, a common issue in traditional K-means clustering algorithms, where the choice of initial centroids can dramatically affect the final clustering results. The absence of dependency on initial cluster positions in the Global K-means algorithm enhances its repeatability and reliability. Traditional K-means often requires multiple runs with different random initializations to obtain a satisfactory outcome, as it can converge to different local optima depending on the starting centroids. In contrast, the Global K-means algorithm systematically adds one cluster center at a time and uses an optimal reassignment step to ensure that each addition improves the overall clustering criterion, thus reducing the chances of falling into suboptimal clustering configurations.

The process begins with the calculation of the clustering solution for one cluster center and incrementally adds one cluster at a time. At each step, the algorithm evaluates all possible locations for the new cluster center (i.e., each point in the dataset) and selects the location that results in the greatest decrease in the sum-of-squares error. This incremental approach ensures that each new center is positioned in a way that optimally enhances the clustering structure. Moreover, the Global K-means algorithm's independence from tunable parameters eliminates the need for manual parameter adjustments, which can be both time-consuming and require a level of expertise that may not be accessible in all application scenarios. This makes the Global K-means a more user-friendly option for practitioners and researchers who may not have extensive backgrounds in data analytics or machine learning. Furthermore, the algorithm's robustness against the initial placement of centers allows it to perform consistently across different datasets and applications. It is particularly useful in applications where data distributions are not well understood beforehand, or where there is a significant amount of noise in the data. The algorithm's ability to methodically explore the data space for optimal cluster centers lends itself well to complex real-world data scenarios where traditional methods might struggle. Additionally, the computational efficiency of the Global K-means algorithm is another significant advantage. While it may appear computationally intensive to evaluate every potential new center, in practice, the algorithm's structured search can be more efficient than the potentially numerous iterations required by traditional K-means to escape poor local optima caused by unfortunate initializations.

In terms of practical applications, the Global K-means clustering algorithm is widely used in various fields

such as market segmentation, image segmentation, and bioinformatics. In market segmentation, for example, it can help identify distinct customer groups based on purchasing behavior or preferences without prior knowledge of the best number of market segments. In bioinformatics, it can be used to classify genes with similar expression patterns, providing insights into functional genomics. The global K-means clustering algorithm is described as follows:

(1) Determine the value of K. Take the average of the data points as the initial first center point, as shown in formula (1), where k=1.

$$c_1 = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

(2) If k-1 cluster center point is known, solve for the next cluster center. Each data point in the data point set $X=\{x_1,x_2,\ldots,x_n\}$ is used as a central candidate for the next cluster.

(3) The candidate points are added to the cluster center set sequentially, and the K-means algorithm is executed for each candidate point until convergence, where the objective function value is calculated for each iteration. Then, the candidate point with the minimum objective function value C_k is selected as the center point of the next cluster and added to the cluster center set $C=\{c_1,c_2,\ldots,c_k\}$.

(4) Termination condition k=k+1. If k=K, the algorithm ends, otherwise go to Step 2.

2.2 Improve the Global K-means Clustering Algorithm

The global K-means clustering algorithm has no initialization problem and is not affected by the location of the initial cluster center. Through deterministic and efficient global search, the sum of squared functions of the clustering errors is effectively minimized and hence the algorithm is very stable. However, since all the K-means values in the global K-means clustering algorithm need to execute the K-means algorithm, the computational tasks are large, and the efficiency and accuracy of the algorithm need to be weighed in practical applications, so this study proposes an improved global K-means clustering algorithm in the following steps:

(1) PCA dimension reduction. The high time complexity of the global K-means algorithm is mainly due to two aspects: distance calculation and exhaustive data point selection for cluster center point. Therefore, in the distance calculation, the higher the data dimension, the higher the time complexity of the algorithm. To solve this problem, this study adopts the PCA dimensionality reduction method [54] to reduce the computation amount in distance calculation. It converts high-dimensional data into low-dimensional data through linear transformation, while retaining the maximum variance in the data, so as to discover the most important features in the data.

(2) Incremental center selection based on multi-sphere splitting. The global K-means algorithm needs to perform k-1 exhaustive operations on N data points when selecting the cluster center point, and run the K-means algorithm for N (k-1) times until convergence. In order to reduce the computation of the exhaustive operations, an incremental center selection method based on multisphere splitting [56] is used for incremental center selection. This method utilizes the features of BKM (Ball K-means, BKM) clustering algorithm [55], which is fast and can record the radius of clusters. By splitting multiple clusters with large radius, the cluster center can be expanded incrementally and quickly.

(3) In the process of optimizing clustering algorithms, particularly when interfacing the global K-means algorithm with traditional K-means, a significant challenge arises due to the excessive number of distance calculations required. This redundancy not only increases the time complexity but also bloats computational overhead. To address this inefficiency, the Hamerly algorithm [57] is employed to streamline distance computations dramatically, thereby enhancing computational efficiency without compromising the accuracy of the global K-means clustering outcomes. The Hamerly algorithm leverages upper and lower distance bounds alongside the triangle inequality to minimize unnecessary distance calculations. This method establishes a more efficient way of evaluating distances between data points and cluster centers by setting a dynamic range within which these distances must fall. For each data point, an upper bound is maintained, always ensuring it is at least as great as the distance from the point to its nearest cluster center. Conversely, the lower bound is set to not exceed the distance to the second closest cluster center. The optimization primarily benefits from two scenarios enabled by these bounds: first, if the upper bound of a data point is less than or equal to the lower bound, then further distance calculations to other cluster centers are unnecessary. Second, if the upper bound is less than or equal to half the distance of the nearest inter-center gap, this also eliminates the need for additional calculations. These criteria allow the Hamerly algorithm to skip over many of the calculations that would otherwise be mandatory, significantly speeding up

the process. This methodological refinement, as detailed in Table 1 of the study, is particularly advantageous in large datasets where the number of clusters is high, making traditional methods computationally prohibitive. By implementing these optimized distance calculations, the Hamerly algorithm not only reduces the computational load but also accelerates the overall clustering process, making it a valuable tool for researchers and practitioners in data-intensive fields.

Table 1. The method used in this paper and its function

Pacing device	Effect
PCA dimensional reduction	Reduce the time complexity of a single distance calculation
Incremental center selection based on multi-sphere splitting	Reduce the exhaustive computation of center selection
Hamerly	Reduce the number of distance calculations

3 Design of Academic Early Warning Model

3.1 Model Implementation Process

In this study, an academic early warning model based on an improved global K-means algorithm is proposed to further improve the accuracy and efficiency of the original method. Firstly, the students' midterm scores were preprocessed and represented by three-dimensional data such as normal scores, midterm exam scores and experimental scores. Firstly, the PCA is used to downsize the data to process the three-dimensional data into one-dimensional data, which is convenient to improve the efficiency of the subsequent algorithms. Then, the processed data was taken to different cluster centers by multi-ball splitting algorithm. Finally, clustering was performed using the Hamerly algorithm. Since the cluster center is three-dimensional, the three score values of the cluster center are added to compare the size of the sum of each cluster center value, and the class of students with the worst score is taken as the warning object. In order to test the accuracy of the model for students' early warning, this study utilizes the mid-term score data of all students for clustering, and takes the class with the worst score as the class that needs to be warned. Then the final scores are clustered to get different classes, and the two clustering results are compared to get the evaluation index of prediction accuracy. The implementation process of academic early warning model is shown in Fig. 1.

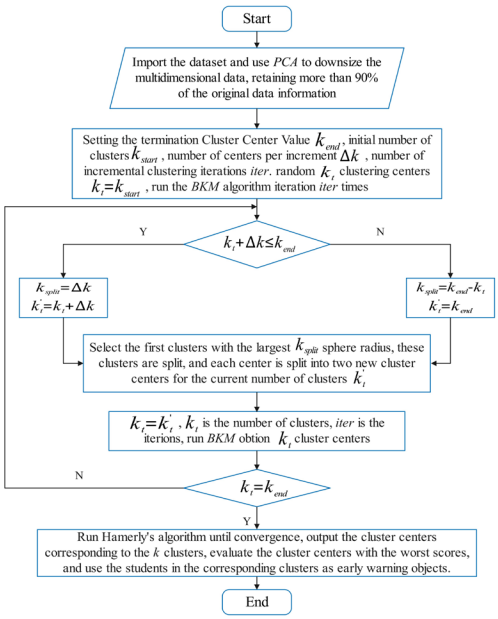

Fig. 1. Academic early warning technology flow chart

3.2 Model Evaluation Index

The evaluation index of academic early warning model includes Accuracy, Precision and F1 value. There are four types of prediction results: TP represents the number of students whose mid-term score is predicted to be early warning and who actually need early warning at the end of the semester; FN indicates the number of students whose mid-term score is predicted to be non-early warning, but who are actually in the lowest grade and need early warning at the end of the semester; FP denotes the number of non-early warning students whose mid-term grades are predicted to be in need of early warning but whose actual final grades are not in the worst tier; TN signifies the number of non-early warning students whose mid-term grades are predicted to be non-early warning but whose actual final grades are not in the worst tier. The confusion matrix contains the actual and predicted classification information. The first row consists of TP and FP, and the second row is composed of FN and TN, which gives a clear view of how correctly and incorrectly the algorithm classified the data results. Accuracy indicates the proportion of correctly predicted warnings and non-warnings to the total number of people.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision represents the percentage of students who are correctly predicted to be warned out of all students that are predicted to be warned.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

F1 value is designed to resolve the conflict between accuracy rate and recall rate in some cases. It is a weighted average of recall rate, accuracy rate and accuracy rate, which can better reflect the comprehensive performance of the model. For problems with unbalanced data distribution, the F1 value can better reflect the performance of the classifier. If the recall rates are used, the performance of the model may be overestimated. Therefore, the F1 value is used in the result analysis to evaluate the comprehensive performance of the prediction model.

$$\frac{2}{F_1} = \frac{1}{Precision} + \frac{1}{Recall} \tag{4}$$

Recall refers to the percentage of students who were correctly predicted to be warned out of all students that have been predicted to be warned.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Sum of Squared Errors (SSE) is one of the commonly used evaluation indexes in K-means clustering. It is the sum of squares of Euclidean distance between elements in each cluster and the center point. The smaller the SSE value, the more similar the elements within the cluster are, and the greater the differences between clusters are. By constantly adjusting the cluster center, the SSE value can be reduced to achieve better clustering results.

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} | x - u_i |^2 \tag{6}$$

4 Experiments and Result Analysis

4.1 Experimental Dataset

The experimental dataset used for early warning accuracy analysis in this study consists of the midterm and final exam scores of 5,476 college students at a university. The midterm scores are composed of three parts: regular scores, midterm exam scores, and lab scores. Final exam scores were compared on a scale of 0-100. The processed data of students' academic performance is shown in Table 2.

For the efficiency analyses, we used randomly generated data on student scores from 0 to 100, generating a total of 10,000, 20,000, 30,000, 40,000, and 50,000 one-dimensional score datasets.

Table 2. Student academic performance data

ID	Regular score	Midterm exam score	Lab score	Final exam score
1	7	25	8	46
2	9	28	10	96
3	8	30	9	98
…	…	…	…	…

K-means *k*=4

Predict / Acutual	warning	non-early warning
warning	144533	60460
non-early warning	68807	0

K-means *k*=5

Predict / Acutual	warning	non-early warning
warning	4438	443
non-early warning	595	0

K-means *k*=6

Predict / Acutual	warning	non-early warning
warning	158128	69657
non-early warning	46015	0

K-means *k*=7

Predict / Acutual	warning	non-early warning
warning	150431	102196
non-early warning	21173	0

Global_K-means *k*=4

Predict / Acutual	warning	non-early warning
warning	4704	251
non-early warning	521	0

Global_K-means *k*=5

Predict / Acutual	warning	non-early warning
warning	155232	65068
non-early warning	53500	0

Global_K-means *k*=6

Predict / Acutual	warning	non-early warning
warning	4506	494
non-early warning	476	0

Global_K-means *k*=7

Predict / Acutual	warning	non-early warning
warning	4303	632
non-early warning	541	0

Fcm *k*=4

Predict / Acutual	warning	non-early warning
warning	157819	50011
non-early warning	65970	0

Fcm *k*=5

Predict / Acutual	warning	non-early warning
warning	188028	42027
non-early warning	43745	0

Fcm *k*=6

Predict / Acutual	warning	non-early warning
warning	193904	39877
non-early warning	36935	0

Fcm *k*=7

Predict / Acutual	warning	non-early warning
warning	196913	43977
non-early warning	32910	0

Hierarchical_K-means *k*=4

Predict / Acutual	warning	non-early warning
warning	3833	814
non-early warning	829	0

Hierarchical_K-means *k*=5

Predict / Acutual	warning	non-early warning
warning	4012	667
non-early warning	797	0

Hierarchical_K-means *k*=6

Predict / Acutual	warning	non-early warning
warning	5,346	22
non-early warning	108	0

Hierarchical_K-means *k*=7

Predict / Acutual	warning	non-early warning
warning	5,346	25
non-early warning	105	0

The proposed algrithm *k*=4

Predict / Acutual	warning	non-early warning
warning	209400	17200
non-early warning	47200	0

The proposed algrithm *k*=5

Predict / Acutual	warning	non-early warning
warning	241700	13950
non-early warning	18150	0

The proposed algrithm *k*=6

Predict / Acutual	warning	non-early warning
warning	256100	5300
non-early warning	12400	0

The proposed algrithm *k*=7

Predict / Acutual	warning	non-early warning
warning	267015	1650
non-early warning	5135	0

Fig. 2. Confusion matrix for various algorithms at k=4, 5, 6, and 7

4.2 Early Warning Accuracy Analysis

For this experimental dataset, the K-means algorithm, global K-means algorithm, fuzzy clustering algorithm (Fcm) [58], a classic hierarchical clustering algorithm, and the algorithm model proposed in this study were used to analyze the data. The final predictions include the number of students to be warned, the final clustering indicators, confusion matrix, and the sum of squared errors (SSE). To reduce the impact of randomness on the experimental results, the algorithms were run 50 times, and the results were averaged. As shown in Fig. 2, and Table 3 and Table 4, since fuzzy clustering and hierarchical clustering do not use formula (5) as an evaluation indicator, this study only evaluates the SSE values of the K-means algorithm, global K-means algorithm and the proposed algorithm in this work.

As shown in Fig. 2, the main diagonal of the confusion matrix represents the number of correctly predicted samples, and the secondary diagonal represents the number of misclassified samples in the prediction process. When k=4, 5, 6, 7, the global K-means/this paper's algorithm correctly predicted the most samples and outperforms the other algorithms, while the Fcm clustering algorithm and hierarchical clustering algorithm predict poorly relative to the other algorithms. The global K-means/this paper's algorithm has the fewest misclassifications for different k value situations, followed by the K-means algorithm, while the Fcm clustering algorithm and hierarchical clustering algorithm are more prone to misclassify non-early warning objects as early warning objects, leading to a higher number of misclassifications and poorer prediction results.

Table 3. Comparison of indicators for various algorithm

		k=4	*k*=5	*k*=6	*k*=7
Accuracy	K-means	0.956410	0.958120	0.963248	0.964530
	Global K-means	0.961538	0.974359	0.974359	0.974359
	Fcm	0.671282	0.805641	0.815513	0.869615
	Hierarchical Clustering	0.820513	0.961538	0.961538	0.961538
	This paper	0.961538	0.974359	0.974359	0.974359
Precision	K-means	0.630952	0.642857	0.666667	0.680952
	Global K-means	0.714286	0.714286	0.714286	0.714286
	Fcm	0.861566	0.769886	0.764706	0.743440
	Hierarchical Clustering	0.333333	0.666667	0.666667	0.666667
	This paper	0.714286	0.714286	0.714286	0.714286
F1 value	K-means	0.595915	0.607493	0.640383	0.639982
	Global K-means	0.625000	0.714286	0.714286	0.714286
	Fcm	0.269516	0.263362	0.265442	0.333988
	Hierarchical Clustering	0.066667	0.400000	0.400000	0.400000
	This paper	0.625000	0.714286	0.714286	0.714286

As shown in Table 3, across different scenarios where k=4, 5, 6, 7, the algorithm presented in this paper demonstrates remarkable performance. Specifically, it attains an accuracy rate exceeding 90%, a precision surpassing 70%, and a comprehensive F1 score superior to other algorithms, aligning closely with the global K-means algorithm. While the Fcm algorithm boasts a higher precision compared to its peers, it unfortunately misclassifies a substantial number of non-early warning samples as warning cases, leading to an elevated count of false predictions and significant divergence from reality. Notably, students flagged by the model as being at risk of academic warning indeed exhibit a high likelihood of receiving warnings during the actual final exams. Consequently, this model serves as a reliable tool for counselors and college

administrators, providing them with a precise list of students at risk of academic warning. Serving as an auxiliary reference, this list enables targeted supervision and assistance to be extended to the students concerned, thereby mitigating their risk of academic warning and underscoring the model's high practical value. Meanwhile, k=4, 5, 6, 7, the algorithm's indicators in this paper rival those of the global K-means algorithm and surpass the metrics of both the standard K-means algorithm and hierarchical clustering. Despite the Fcm algorithm's higher precision, it is plagued by a higher number of misjudgments that undermine its overall effectiveness. Therefore, it can be confidently asserted that the algorithm presented in this paper offers more accurate predictions of actual outcomes and generates fewer misjudgments during the identification process. Such a model is well-suited to address the practical needs of student early warning systems, enhancing the precision and stability of prediction results, better discerning the proportion of positive categories, offering a comprehensive evaluation of the model's strengths and weaknesses, and achieving better outcomes in the context of student academic early warning.

Table 4 compares the SSE values of K-means algorithm, global K-means algorithm and this study's algorithm to evaluate the clustering effect. When k=4, 5, 6, 7, the SSE value of this paper's algorithm is better than that of K-means algorithm, which indicates that this proposed algorithm has a significant advantage over the traditional K-means algorithm.

Table 4. Comparison of SSE values

		k=4	k=5	k=6	k=7
Midterm	K-means	1003.50	1020.01	925.17	904.45
	Global K-means	742.41	567.89	448.12	379.54
	This paper	662.53	605.22	485.45	399.48
Final	K-means	6982.34	5612.54	3180.78	3027.49
	Global K-means	3202.41	2127.81	1330.78	899.25
	This paper	5423.91	3482.41	2325.39	1687.26

Table 5. The total running time (ms) of various algorithms in different dimensions for 10000 pieces of data

Dimension	Global K-means	K-means	Fcm	Hierarchical clustering	This paper
5	4485	23.48	235.06	691.00	7.23
10	7490	40.78	235.08	951.00	7.92
15	9768	47.38	284.44	1983.00	6.98
20	12163	52.48	351.58	2925.00	9.05
25	18977	76.62	420.14	4167.00	7.36

Table 6. The total running time (ms) of various algorithms under different data volumes when the dimension is 10

Data volum	Global K-means	K-means	Fcm	Hierarchical clustering	This paper
10000	7490	40.78	235.08	951.00	7.92
20000	30527	101.58	359.50	10857.00	13.00
30000	83686	198.26	477.08	36125.00	21.77
40000	180646	348.66	692.52	71469.00	32.80
50000	281303	474.86	769.94	114802.00	50.26

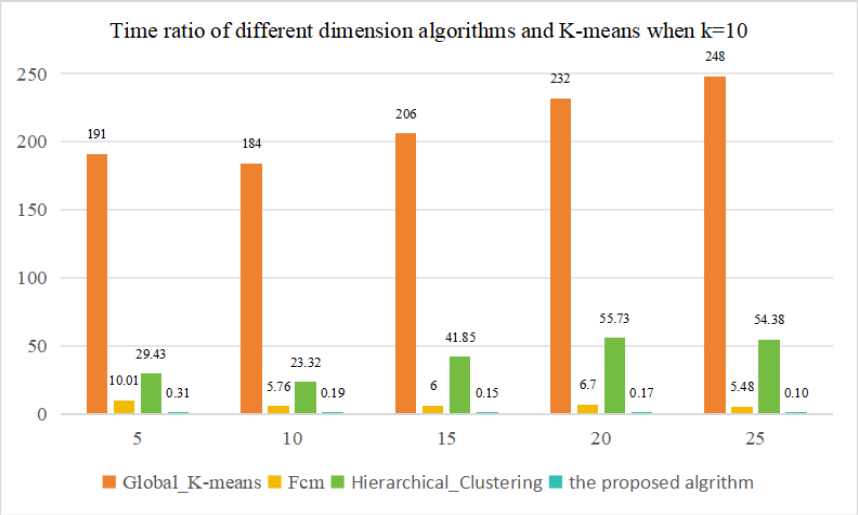

Fig. 3. The ratio of the total running time of the various algorithm and the K-means algorithm in different dimensions for 10000 pieces of data

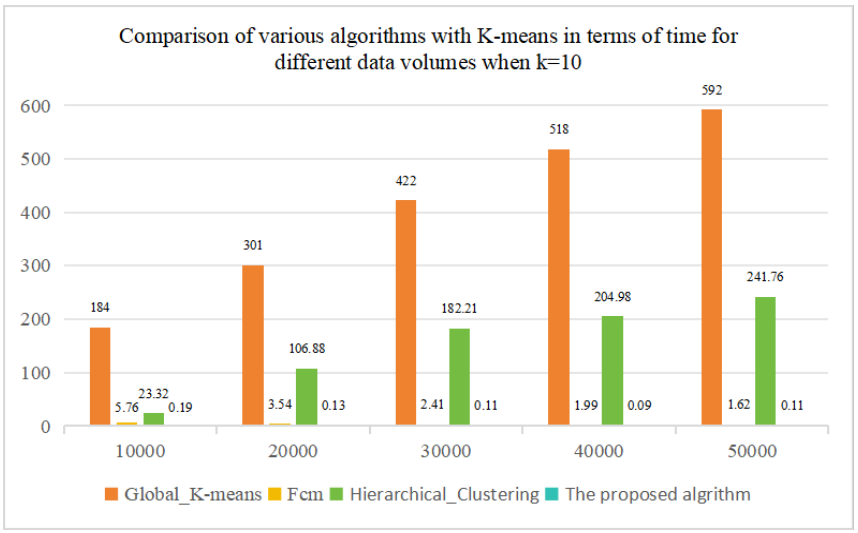

Fig. 4. The ratio of the total running time of the various algorithm and the K-means algorithm under different data volumes when the dimension is 10

4.3 Computational Efficiency Analysis

The purpose of this experiment is to verify the operational efficiency of the K-means algorithm, global K-means algorithm, Fcm algorithm, hierarchical clustering algorithm, and the algorithm introduced in this stduy in operation with different numbers of data points and data dimensions. The data used for the comparative experiments of the five algorithms are mainly normally distributed data in the range of [-1,1] generated by the simulation experiments, with a total of 10,000, 20,000, 30,000, 40,000, and 50,000 data points of different dimensions. The experiment were conducted by fixing the value of k to 10, and Table 5 and Table 6 show the running times of various algorithms for 10,000 data points of different dimensions and different data volumes for a dimension of 10. Fig. 3 and Fig. 4 show the total running time ratios of various algorithms compared to the K-means algorithm.

As can be seen from Table 5 to Table 6 and Fig. 3 to Fig. 4, the efficiency of this algorithm is better than other algorithms in different dimensions or different data volumes, and the advantage is more obvious compared with

the global K-means algorithm. Meanwhile, with the increase of data dimension or data volume, the total running time ratio of this paper's algorithm is further reduced compared with K-means algorithm, and the efficiency advantage is still significant.

5 Conclusion

Addressing the limitations of traditional academic early warning methods, which often require extensive historical data, produce insufficient warning results, and demonstrate low computational efficiency, this study introduces a novel academic early warning model based on an enhanced global K-means algorithm. This model notably improves solution accuracy through the implementation of a global K-means incremental center selection method, and it significantly enhances computational efficiency by optimizing several key areas.

Firstly, the model utilizes principal component analysis (PCA) to reduce the dimensionality of data. This technique is critical in decreasing the computational burden associated with high-dimensional data, thereby expediting the processing time. By transforming the original data into a set of linearly uncorrelated variables known as principal components, PCA enables the model to focus on the most significant features with reduced redundancy and noise. Secondly, the model incorporates the BKM algorithm and multi-sphere splitting concepts to accelerate the selection of incremental centers. The BKM method, an extension of the K-means algorithm, enhances clustering performance by optimizing initial cluster centers more effectively. Multi-sphere splitting, on the other hand, aids in faster convergence by segmenting data into multiple spheres, thus reducing the overall computational complexity and improving the efficiency of the clustering process. Thirdly, the Hamerly algorithm is employed to minimize redundant distance calculations between data points and cluster centers. This algorithm streamlines the clustering process by intelligently determining when it is unnecessary to compute certain distances, which dramatically reduces the number of distance calculations required during each iteration of the clustering process.

The experimental results presented in this study demonstrate that the early warning accuracy of the proposed model aligns closely with that of the global K-means algorithm and surpasses that of the traditional K-means algorithm, as well as other clustering methods. Particularly noteworthy is that the accuracy is significantly higher compared to other clustering algorithms traditionally used in academic settings. Furthermore, in simulation comparison experiments conducted under various dimensional and data count scenarios, the computational efficiency of the proposed algorithm model outperforms the global K-means algorithm, traditional K-means, and other clustering algorithms. This efficiency becomes increasingly pronounced as the dimensionality of the data increases, highlighting the model's robustness in handling large-scale and complex data sets. In addition to these technical improvements, the proposed model also offers practical benefits for educational institutions. By enabling more accurate and timely identification of students at risk of academic failure, schools can implement targeted interventions more effectively, thereby improving educational outcomes and student retention rates. The model's enhanced efficiency also allows for real-time data processing, which is critical in dynamic educational environments where timely data analysis can influence the success of academic interventions.

Overall, the innovative approach of this study to the design of an academic early warning system not only addresses the deficiencies of previous methods but also provides a scalable, efficient solution adaptable to various educational settings. This contributes to a broader understanding and application of data-driven strategies in the educational sector, ultimately fostering an environment where data insights lead to effective student support and academic excellence.

Acknowledgement

We acknowledge the Jilin Provincial Department of Education General Project "Research on Multi view Mining Technology for Education Big Data (JJKH20220777KJ)" and "Key Technology Research on Knowledge Graph Construction for Smart Education (JJKH20230842KJ)".

References

[1] M.-A. Hashim, I. Tlemsani, R. Matthews, Higher education strategy in digital transformation, Education and

Information Technologies 27(3)(2022) 3171-3195. https://doi.org/10.1007/s10639-021-10739-1
[2] G.-D. Sharma, A. Yadav, R. Chopra, Artificial intelligence and effective governance: A review, critique and research agenda, Sustainable Futures 2(2020) 100004. https://doi.org/10.1016/j.sftr.2019.100004
[3] C.A. Bonfield, M. Salter, A. Longmuir, M. Benson, C. Adachi, Transformation or evolution?: Education 4.0, teaching and learning in the digital age, Higher education pedagogies 5(1)(2020) 223-246. https://doi.org/10.1080/23752696.2020.1816847
[4] V. Voronkova, V. Nikitenko, R. Oleksenko, R. Andriukaitiene, J. Kharchenko, E. Kliuienko, Digital technology evolution of the industrial revolution from 4g to 5g in the context of the challenges of digital globalization, TEM Journal 12(2)(2023) 732-742. https://doi.org/10.18421/tem122-17
[5] C. Guan, J. Mou, Z. Jiang, Artificial intelligence innovation in education: A twenty-year data-driven historical analysis, International Journal of Innovation Studies 4(4)(2020) 134-147. https://doi.org/10.1016/j.ijis.2020.09.001
[6] M. Alenezi, Digital learning and digital institution in higher education, Education Sciences 13(1)(2023) 88. https://doi.org/10.3390/educsci13010088
[7] G.-Q. Liang, C.-S. Jiang, Q.-Z. Ping, X.-Y, Jiang, Academic performance prediction associated with synchronous online interactive learning behaviors based on the machine learning approach, Interactive Learning Environments 32(6)(2024) 3092-3107. https://doi.org/10.1080/10494820.2023.2167836
[8] M. Imran, N. Almusharraf, Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature, Contemporary Educational Technology 15(4)(2023) ep464. https://doi.org/10.30935/cedtech/13605
[9] J. Jameson, E-Leadership in higher education: The fifth "age" of educational technology research, British Journal of Educational Technology 44(6)(2013) 889-915. https://doi.org/10.1111/bjet.12103
[10] C. Song, S.-Y. Shin, K.-S. Shin, Implementing the Dynamic Feedback-Driven Learning Optimization Framework: A Machine Learning Approach to Personalize Educational Pathways, Applied Sciences 14(2)(2024) 916. https://doi.org/10.3390/app14020916
[11] M. Li, D. Xu, D.-M Zhang, J. Zou, The seeding algorithms for spherical k -means clustering, Journal of Global Optimization 76(4)(2019) 695-708. https://doi.org/10.1007/s10898-019-00779-w
[12] R.-S. Baker, A. Hawn, Algorithmic bias in education, International Journal of Artificial Intelligence in Education 32(2022) 1052-1092. https://doi.org/10.1007/s40593-021-00285-9
[13] L.-D. Martín, C.-R. Manuel, L.-N. Martín, A.-M. Fernando, Systematic Literature Review of Predictive Analysis Tools in Higher Education, Applied Sciences 9(24)(2019) 5569. https://doi.org/10.3390/app9245569
[14] G. Papanastasiou, A. Drigas, C. Skianis, M. Lytras, E. Papanastasiou, Virtual and augmented reality effects on K-12, higher and tertiary education students' twenty-first century skills, Virtual Reality 23(4)(2019) 425-436. https://doi.org/10.1007/s10055-018-0363-2
[15] J.-N. Bailenson, N. Yee, J. Blascovich, A.-C. Beall, N. Lundblad, M. Jin, The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context, Journal of the learning sciences 17(1)(2008) 102-141. https://doi.org/10.1080/10508400701793141
[16] J. Garzón, An overview of twenty-five years of augmented reality in education, Multimodal Technologies and Interaction 5(7)(2021) 37. https://doi.org/10.3390/mti5070037
[17] A.-C. Dixit, B. Harshavardhan, B. Ashok, M. Sriraj, K. Prakasha, Innovative Pedagogical Approaches for Diverse Learning Styles and Student-Centric Learning, Journal of Engineering Education Transformations 37(2024) 178-188. https://doi.org/10.16920/jeet/2024/v37is2/24039
[18] D. DeSutter, M. Stieff, Teaching students to think spatially through embodied actions: Design principles for learning environments in science, technology, engineering, and mathematics, Cognitive research: principles and implications 2(1)(2017) 22. https://doi.org/10.1186/s41235-016-0039-y
[19] S. Noor, M. Ramly, Bridging Learning Styles and Student Preferences in Construction Technology Education: VARK Model Analysis, International Journal of Academic Research in Progressive Education and Development 12(3)(2023) 2075-2085. http://dx.doi.org/10.6007/IJARPED/v12-i3/19313
[20] C. Dede, The evolution of distance education: Emerging technologies and distributed learning, The American Journal of Distance Education 10(2)(1996) 4-36. https://doi.org/10.1080/08923649609526919
[21] S. Marukatat, Tutorial on PCA and approximate PCA and approximate kernel PCA, Artificial Intelligence Review 56(6)(2023) 5445-5477. https://dx.doi.org/10.1007/s10462-022-10297-z
[22] A. Rof, A. Bikfalvi, P. Marques, Pandemic-accelerated digital transformation of a born digital higher education institution, Educational Technology & Society 25(1)(2022) 124-141. https://www.jstor.org/stable/48647035
[23] C. Nazir, B. Tanya, Placing inclusive education in conversation with digital education, South African Computer Journal 34(2)(2022) 18-34. https://doi.org/10.18489/sacj.v34i2.1084
[24] P. Paudel, Online education: Benefits, challenges and strategies during and after COVID-19 in higher education, International Journal on Studies in Education (IJonSE) 3(2)(2021) 70-85. https://doi.org/10.46328/ijonse.32
[25] I. Langseth, D.-Y. Jacobsen, H. Haugsbakken, The role of support units in digital transformation: How institutional entrepreneurs build capacity for online learning in higher education, Technology, Knowledge and Learning 28(11)(2023) 1745-1782. https://doi.org/10.1007/s10758-022-09620-y
[26] C. Tan, J.Z. Lin, A new QoE-based prediction model for evaluating virtual education systems with COVID-19 side

effects using data mining, Soft Computing 27(3)(2023) 1699-1713. https://doi.org/10.1007/s00500-021-05932-w
[27] A. Glover, Y. Strengers, T. Lewis, The unsustainability of academic aeromobility in Australian universities, Sustainability: Science, Practice and Policy 13(1)(2017) 1-12. https://doi.org/10.1080/15487733.2017.1388620
[28] D. Shi, J. Zhou, D. Wang, X. Wu, Research status, hotspots, and evolutionary trends of intelligent education from the perspective of knowledge graph, Sustainability 14(17)(2022) 10934. https://doi.org/10.3390/su141710934
[29] L. Fu, J. Li, Y. Chen, An innovative decision making method for air quality monitoring based on big data-assisted artificial intelligence technique, Journal of Innovation & Knowledge 8(2)(2023) 100294. https://doi.org/10.1016/j.jik.2022.100294
[30] I.-E. Agbehadji, B.-O. Awuzie, A.-B. Ngowi, R.-C. Millham, Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing, International journal of environmental research and public health 17(15)(2020) 5330. https://doi.org/10.3390/ijerph17155330
[31] C. Wen, J. Yang, L. Gan, Y. Pan, Big data driven Internet of Things for credit evaluation and early warning in finance, Future Generation Computer Systems 124(4)(2021) 295-307. https://doi.org/10.1016/j.future.2021.06.003
[32] Z. Zhang, X. Lin, S. Shan, Big data-assisted urban governance: An intelligent real-time monitoring and early warning system for public opinion in government hotline, Future Generation Computer Systems 144(3)(2023) 90-104. https://doi.org/10.1016/j.future.2023.03.004
[33] R. Asif, A. Merceron, S.-A. Ali, N.-G. Haider, Analyzing undergraduate students' performance using educational data mining, Computers & education 113(5)(2017) 177-194. https://doi.org/10.1016/j.compedu.2017.05.007
[34] K.-L. Ang, F.-L. Ge, K.-P. Seng, Big educational data & analytics: Survey, architecture and challenges, IEEE Access 8(7)(2020) 116392-116414. https://doi.org/10.1109/ACCESS.2020.2994561
[35] N. Tomasevic, N. Gvozdenovic, S. Vranes, An overview and comparison of supervised data mining techniques for student exam performance prediction, Computers & education 143(1)(2020) 103676. https://doi.org/10.1016/j.compedu.2019.103676
[36] E. Howard, M. Meehan, A. Parnell, Contrasting prediction methods for early warning systems at undergraduate level, The Internet and Higher Education 37(2)(2018) 66-75. https://doi.org/10.1016/j.iheduc.2018.02.001
[37] Z. Yang, J. Yang, K. Rice, J.-L. Hung, X. Du, Using convolutional neural network to recognize learning images for early warning of at-risk students, IEEE Transactions on Learning Technologies 13(3)(2020) 617-630. https://doi.org/10.1109/TLT.2020.2988253
[38] X. Bai, F. Zhang, J. Li, T. Guo, A. Aziz, A.Jin, F.Xia, Educational big data: Predictions, applications and challenges, Big Data Research 26(11)(2021) 100270. https://doi.org/10.1016/j.bdr.2021.100270
[39] H. Zeineddine, U. Braendle, A. Farah, Enhancing prediction of student success: Automated machine learning approach, Computers & Electrical Engineering 89(2021) 106903. https://doi.org/10.1016/j.compeleceng.2020.106903
[40] A. Kukkar, R. Mohana, A. Sharma, A. Nayyar, Prediction of student academic performance based on their emotional wellbeing and interaction on various e-learning platforms, Education and Information Technologies 28(8)(2023) 9655-9684. https://doi.org/10.1007/s10639-022-11573-9
[41] T.-M. Alam, M. Mushtaq, K. Shaukat, I.-A. Hameed, M.-U. Sarwar, S. Luo, A novel method for performance measurement of public educational institutions using machine learning models, Applied Sciences 11(19)(2021) 9296. https://doi.org/10.3390/app11199296
[42] A. Dutt, M.-A. Ismail, T. Herawan, A Systematic Review on Educational Data Mining, IEEE Access 5(2017) 15991-16005. https://doi.org/10.1109/access.2017.2654247
[43] B. Albreiki, T. Habuza, N. Zaki, Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models, International Journal of Educational Technology in Higher Education 19(1)(2022) 49. https://doi.org/10.1186/s41239-022-00354-6
[44] Y.-P. Raykov, A. Boukouvalas, F. Baig, M.-A. Little, What to do when K-means clustering fails: a simple yet principled alternative algorithm, PloS one 11(9)(2016) e0162259. https://doi.org/10.1371/journal.pone.0162259
[45] A.-K. Jain, Data clustering: 50 years beyond K-means, Pattern recognition letters 31(8)(2010) 651-666. https://doi.org/10.1016/j.patrec.2009.09.011
[46] C. Chandrashekar, P. Agrawal, P. Chatterjee, D.-S. Pawar, Development of E-rickshaw driving cycle (ERDC) based on micro-trip segments using random selection and K-means clustering techniques, IATSS research 45(4)(2021) 551-560. https://doi.org/10.1016/j.iatssr.2021.07.001
[47] H.-Y. Wang, Analysis and Prediction of CET4 Scores Based on Data Mining Algorithm, Complexity 2021(2021) 5577868. https://doi.org/10.1155/2021/5577868
[48] W.-X. Tian, Research of academic precaution of college students based on outlier data mining technology, Heilongjiang Science 12(07)(2021) 54-56. https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_Gcu8vVv0qKAfWUG5W5es6_sH_5SQ-TXVoEWQoeLbhzRtwHZhj3jgALpWhsY3HUbptuioK7cRbx-CaWr9gwL6n4z9uyhun3JkoKAhT-5PnSxNoJa21UZaVmfWp_XKCCssEjtHohd3e2fNfAa_QKDm4q8JBAU4qVJu9Y5SpFwtZOhfk8ZOlXl8ge_L&uniplatform=NZKPT&language=CHS
[49] X.-L. Li, L. Ma, X.-D. He, H. Xiong, You Are How You Behave – Spatiotemporal Representation Learning for College Student Academic Achievement, Journal of Computer Science and Technology 35(2)(2020) 353-367.

https://doi.org/10.1007/s11390-020-9971-x
[50] S.-J. Nam, P. Samson, Integrating students' behavioral signals and academic profiles in early warning system, in: Proc. 20th International Conference, AIED 2019, 2019. https://doi.org/10.1007/978-3-030-23204-7_29
[51] S.-C. Wang, X.-H. Xu, J.-C. Huang, F.-F. Zhang, Application research of K-means clustering algorithm in college students' performance analysis, Journal of Hubei Normal University (Natural Science) 39(03)(2019) 113-118. https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GfKm4Fi_ZNxe6W-81m_hJKvb8DRcPUAjhO1DMnCphVkBn-N8GkvrCX0Hcvg9d0XlLCPVX4bTnqlMqcBjqOGnvZi48l8U-0WmmGzMmzZuh7Hb8ccEfazkzDszg0biBy9pdn-NeKsG0ANItAwaL8H_DOjkMwxXvd5-iOJGwcJ-QL-1a8xSVFsKgJhrP&uniplatform=NZKPT&language=CHS
[52] D.-L. Xiong, Research on the Academic Early Warning and Assistance for University Students Based on Students' Portraits in Big Data Environment——Taking Xuchang University as an Example, Journal of Xuchang University 42(5)(2023) 104-107. https://kns.cnki.net/kcms2/article/abstract?v=ifIT5_n5_GdiqHkFshEM_XDuuBVf-DmytCLEpn4IbH7EdXU08Co1sJPJfb3Ia4qDotqwGlNzmgFYNbb6I7YykND1--jRg_G4bxmSYjSYTnXpUIE-FE_XNbPgPEJSrD7K6tUqqxoQpoW200eZ1xKdpEQRh2TFByRHSsYFc2hSxLpopNl-hNrLlMiS0oI9Xy_P&uniplatform=NZKPT&language=CHS
[53] G. Vardakas, A. Likas, Global k-means plus plus:an effective relaxation of the global k-means clustering algorithm, Applied Intelligence 54(19)(2024) 8876-8888. https://doi.org/10.1007/s10489-024-05636-2
[54] S. Marukatat, Tutorial on PCA and approximate PCA and approximate kernel PCA, Artificial Intelligence Review 56(6)(2023) 5445-5477. https://doi.org/10.1007/s10462-022-10297-z
[55] S.-Y. Xia, D.-W. Peng, D.-Y. Meng, C.-Q. Zhang, G.-Y. Wang, E. Giem, W.Wei, Z.-Z.Chen, A Fast Adaptive k-means with No Bounds, IEEE transactions on pattern analysis and machine intelligence 44(1)(2020) 87-99. https://doi.org/10.1109/TPAMI.2020.3008694
[56] F.-H. Qu, C.-Y. Qian, Y. Yang, Y. Lu, J.-F. Song, Y.-T. Hu, Incremental k-means clustering algorithm based on multi-sphere splitting, Journal of Jilin University (Engineering and Technology Edition) 52(6)(2022) 1434-1441. https://doi.org/10.13229/j.cnki.jdxbgxb20210098
[57] G. Hamerly, Making k-means even faster, in: Proc. 2010 SIAM International Conference on Data Mining, 2010. https://doi.org/10.1137/1.9781611972801.12
[58] Y. Pang, M.-L. Shi, L.-Y. Zhang, X.-G. Song, W. Sun, PR-FCM: A polynomial regression-based fuzzy C-means algorithm for attribute-associated data, Information Sciences 585(2022) 209-231. https://doi.org/10.1016/j.ins.2021.11.056