Journal of Information Systems Engineering and Management
2025, 10(16s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Improved Student Modeling and Data Clustering for Personalized Online Teaching of International Chinese Teachers: The MKmeans Algorithm Approach

Jing Zhao*, Qian Liu
China International Language and Culture College, Krirk University, Bangkok 10220, Thailand
* Corresponding author: Jing Zhao, huizi8382@sina.com

ARTICLE INFO

Received: 10 Dec 2024

Revised: 28 Jan 2025

Accepted: 12 Feb 2025

ABSTRACT

This study explores personalized online teaching for international Chinese education by enhancing student modeling through advanced data mining. A novel MKmeans clustering algorithm, based on mean shift theory, is proposed to improve clustering stability and noise resilience. Experiments on datasets like Iris and Wine demonstrate MKmeans outperforms classical Kmeans in accuracy and F-measure. The findings enable adaptive teaching strategies, offering significant insights for optimizing online learning systems and advancing personalized education.

Keywords: online teaching; intelligent computing; data clustering algorithms; educational data mining.

1. Introduction

The rapid development of internet technologies has transformed education, making online learning an essential mode of delivering high-quality resources without the constraints of time and space. The “Internet + Education” model enables learners to acquire knowledge efficiently, customize their learning paths, and nurture their interests and abilities. For international Chinese language teachers, online teaching has become an indispensable medium to provide personalized learning experiences to a global audience. Despite its potential, personalized online education still faces challenges, such as insufficient adaptability to diverse learner needs and inadequate utilization of educational data to inform teaching strategies.

Personalized teaching emphasizes "teaching students in accordance with their aptitude," a core principle in education. Online platforms can significantly enhance this by integrating data-driven technologies to tailor learning experiences. However, many existing platforms lack robust mechanisms to account for individual differences in learning ability, background, and preferences. This gap necessitates the development of advanced models and algorithms that can analyze learners' behavioral data to deliver customized instruction effectively.

The advent of educational data mining offers new possibilities to address these challenges. By leveraging data clustering techniques, educators can uncover hidden patterns in learner behavior and categorize students into groups with similar characteristics. Such categorization allows for the design of tailored teaching strategies and the recommendation of optimal learning paths. Among the diverse clustering methods available, the Kmeans algorithm is widely used due to its simplicity and efficiency. However, it has limitations, such as sensitivity to noise, dependence on initial cluster centers, and susceptibility to local optima, which can compromise clustering quality.

To overcome these limitations, this study introduces the MKmeans algorithm, an improved clustering technique based on mean shift theory. The algorithm enhances clustering stability by focusing on high-density data regions during the initialization of cluster centers. This approach ensures more accurate grouping of learners and minimizes the impact of outliers. Experiments using well-known datasets, such as Iris and Wine, demonstrate that MKmeans consistently achieves higher F-measure scores and better clustering performance compared to Kmeans and its classical variants.

This paper also examines the application of MKmeans in online teaching for international Chinese language education. By grouping students based on their cognitive and learning characteristics, the algorithm enables the design of personalized instructional strategies that cater to each group's unique needs. The results reveal that the proposed method significantly improves clustering accuracy and provides actionable insights for educational system optimization.

In summary, this study contributes to personalized online education by proposing an innovative clustering approach that addresses existing shortcomings in student modeling. The findings hold potential to transform online learning platforms, making them more effective, adaptive, and student-centered. This research lays a foundation for future work in educational data mining and personalized teaching for international Chinese educators.

2. Methodology

2.1. Mean shift theory

In a 1975 essay on the gradient function of probability density, Fukunaga established the concept of mean shift, which relates to the mean vector of offset. As mean shift theory develops, mean shift algorithms also appear, albeit initially mean shift is merely a vector. The mean shift theory gained popularity in 1995 thanks to Yizong Cheng[10,11]. In order to establish a link between the offset vector's contribution and the sample's distance from the offset point, he first proposed a kernel function. Secondly, he established a weight coefficient pertaining to the sample point's significance. In the d-dimensional space, given n data sample points x_i, where $i = 1, \ldots, n$. Any point x in the space can have its mean shift vector defined as:

$$M_h(x) = \frac{1}{k} \sum_{x_i \in S_h} (x_i - x) \tag{1}$$

where h is the area of a high-dimensional sphere satisfying the set of points given in Equation (2) with radius h (d may be more than 2). K denotes the zone that k points in n samples fall inside.

$$S_h(x) = \{y: (y - x)^T (y - x) < h^2\} \tag{2}$$

In other words, as seen in Figure 1, each one of n points can be represented as the solid point 1. Using this point as the center and h as the radius, a high-dimensional ball is formed. K points fall into the ball, forming k vectors with the center. Then, as indicated in Equation (1), determine the mean value of the k vectors, which is the mean shift vector, and correlate to the thick arrows in **Figure 1**. Next, create a high-dimensional sphere by ending the vector at a new point. If the earlier steps are repeated, the mean shift vector will converge to the region with a relatively high density.

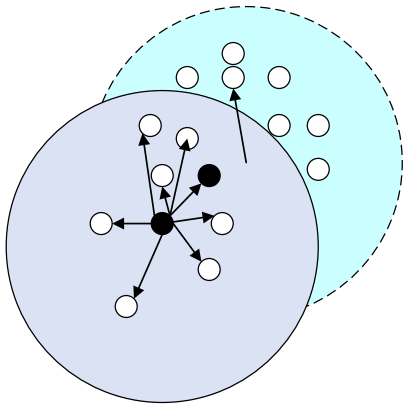

Figure 1. Schematic diagram of mean shift.

Kernel density estimation, sometimes referred to as the Parzen window in pattern recognition technology, is the most popular density estimate method. The following is the definition: Each point in the d-dimensional space, where $i = 1, \ldots, n$, is represented by a column vector, given n data sample points. In the event that x is any of the points, then x's module is.

$$K(x) = k(x^2) \tag{3}$$

In this scenario, $K(x)$ is a piecewise continuous function, if $a < b$, then $K(a) \geq K(b)$, and k is a nonnegative number.

$$\int_0^{\infty} k\,(r)\mathrm{d}r < \infty \tag{4}$$

K(x), as described in Equation (4), is a kernel function if the aforementioned requirements are satisfied. The unit Gaussian kernel function and the unit uniform kernel function are the two unit kernels that are commonly used in mean shift. Two functions are defined as follows: consistent kernel operation within a unit:

$$F(x) = \begin{cases} 1 & \text{if } \|x\| < 1 \\ 0 & \text{if } \|x\| \geq 1 \end{cases} \tag{5}$$

Unit Gaussian kernel function:

$$N(x) = e^{-x^2} \tag{6}$$

Equation (5) shows that point x is treated equally regardless of its distance from the center point as long as it is in a high-dimensional sphere with radius h. There is no weighted value, and the contribution of M x h is the same when calculating. However, we are aware that the closer a point is to the center point, the more significant it is to estimate the statistics surrounding the center point. Yizong Cheng added the kernel function to mean shift in 1995 as a solution to this issue. At the same time, it is also considered that the importance of points $i\ x$ in the dataset is different, so a weight value representing their importance is introduced for each point[12]. Equation (7) thus displays the extended version of mean shift:

$$M(x) = \frac{\sum_{i=1}^{n} G_H\,(x_i - x) w(x_i - x)}{\sum_{i=1}^{n} G_H\,(x_i - x) w(x_i)} \tag{7}$$

Including: $G_H(x_i - x) = |H|^{-1/2} G\left(|H|^{-1/2}(x_i - x)\right)$, $G_H(x)$ is a unit kernel function, and each point's weighted value is determined by $w(x_i)$, and $w(x_i) \geq 0$. H is d × Diagonal matrix of $d\ H = diag[h_1^2, h_2^2, \ldots, h_d^2]$. The simplified form, which necessitates the setting of a coefficient h, is frequently employed in mean shift. Later in this article, this form is used. Equation (8) can therefore be expressed as follows:

$$M_h(x) = \frac{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i) w(x_i - x)}{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i)} \tag{8}$$

If the weight value $w(x_i)$ of all sampling points xi is the same, set it to 1, The fundamental mean shift is obtained by degenerating equation (8). Equation (1) when applying the mean unit kernel function using the unit kernel function 4. Modify Equation (8) to streamline the mean shift algorithm's iteration phases. Equation (8) is transformed into Equation (9) in the manner shown below:

$$M_h(x) = \frac{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i) x_i}{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i)} - x \tag{9}$$

And make the first item as $m_h(x)$, namely:

$$m_h(x) = \frac{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i) x_i}{\sum_{i=1}^{n} G\left(\frac{x_i - x}{h}\right) w(x_i)} \tag{10}$$

Then $mh(x)$ is modified by Equation (11):

$$mh(x) = mh(x) - x \tag{11}$$

2.2. Improved algorithm MKmeans

The aforementioned analysis shows that the initial cluster center chosen at random results in unstable clustering, which greatly affects the clustering quality and is susceptible to noise and easy to reach local extreme points. This work suggests the MKmeans algorithm, an enhanced Kmeans algorithm based on mean shift theory, in light of these

drawbacks and weaknesses. By focusing on high-density regions during the initial cluster centre selection process, this technique improves clustering stability and lowers the likelihood that the algorithm would enter local optimisation[13,14]. There is a coefficient h in the mean shift algorithm that needs to be set. This paper's experiment demonstrates how the degree of variation between dataset properties and the magnitude of h are related. *H* can be set to a lower value when the difference degree is smaller and to a higher value when the difference degree is larger. The modified algorithm's steps are as follows: (1) Determine mh(x), set the coefficient h, and choose a point at random. (2) Assign x to mh(x). (3) Get m points if the cycle has ended; if not, go on to (1). (4) Out of m points, choose the one that has the greatest sum of k distances from other places. (5) Create the initial center by identifying the k points in dataset D that are most similar to these k points. (6) Determine how similar each remaining data sample point is to the k cluster centers in order to sort the sample points into the cluster with the highest similarity. (7) By averaging all of the cluster's data samples, recalculate the centers of each of the k clusters in the new cluster. (8) Using the new center as a guide, reclassify every element in D. Until the number of iterations is reached or the cluster centre of the previous two times does not move, repeat steps (6) through (8).

The *m* points selected by mean shift are all in places with relatively high density, as shown in **Figures 2** .

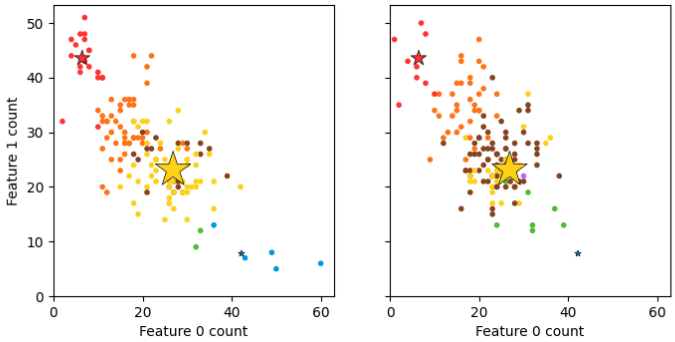

Figure 2. Wine dataset h = 100.

3. Experiments

3.1. Implementation details

All experiments in this paper share three experimental tools MATLAB, SPSS and Weka. Using MATLAB integrated hierarchical clustering, Kmeans clustering, FCM clustering, SOM clustering and other clustering algorithms to do comparative experiments; SPSS was used to analyze the characteristics of the experimental data set; The comparison experiment is made with Weka integrated Xmeans algorithm. The mathematical calculations involved in clustering algorithms are relatively complex and tedious, so professional scientific computing tools should be used. MATLAB is a mathematical software developed by American mathwork for data analysis, algorithm development and numerical calculation[15,16]. Its statistical tools provide several clustering methods: (1) hierarchical clustering, (2) Kmeans clustering, (3) fuzzy C-means clustering FCM: the membership degree of sample attribution in FCM algorithm is set to 2. (4) SOM Clustering of self-organizing feature maps. In this study, the Euclidean distance is utilised to quantify sample similarity when the hierarchical clustering technique is applied. In the SOM algorithm, this paper uses a one-dimensional array of SOM network to become a one-dimensional SOM (1-D SOM). There are many parameters that can be selected in the SOM algorithm, and different parameters have a great impact on the results. The training parameters in this paper are shown in **Table 1**:

Table 1. SOM training parameters.

SOM training parameters	Parameter value
Initial value of learning rate in sorting stage	60
Steps in sorting stage	100
Initial value of learning rate in adjustment stage	0.02
Other	Default

In order to analyze the data characteristics of iris and wine, which are used for experiments, mainly mean value,

variance, peak value and frequency, etc. This paper uses its statistical analysis function, and finally outputs the desired results with clear and intuitive tables. The Waikato intelligent analysis environment is called Weka (Waikato environment for knowledge analysis). The Xmeans algorithm is one of the many data mining methods integrated by the open and free platform Weka[17].

3.2. Datasets

UCI is a database used exclusively for testing machine learning and data mining methods. Since the database's data sets are clearly categorised, external standards can be utilised to objectively and intuitively assess how well clustering algorithms perform. Two global data sets, Iris and Wine, show two distinct testing vantage points. In order to demonstrate the efficacy of the novel algorithm that is presented in this research, this paper examines and evaluates these two data sets. Iris includes 150 samples in three categories, namely, 1-setosa, 2-versicolor, 3-virginia. Each category includes 50 samples, namely, 1–50 is the first category, 51–100 is the second category, and 101–150 is the third category. The data contains four attributes, namely, calyx length, width, and petal length and width. The data are of numerical type. This article does not do any data processing for this, because after the data is normalized, it may destroy the relationship between the attributes of the data itself, using the most original data. Win includes 178 samples, which are divided into 3 categories. There are 13 attributes describing the samples, including 59 samples from 1–59 for the first category, 71 samples from 60–130 for the second category, and 48 samples from 131–178 for the third category. Like Iris dataset, it is numerical data, excluding name attribute and class attribute. This article also uses the most original Wine data set without any processing to test[18]. Because, different clustering algorithms have very different clustering effects for data sets with different characteristics. Therefore, it is necessary to analyze the mean, variance and other characteristics of the dataset. This paper uses SPSS data analysis software to analyze the mean, variance, peak and other parameters of Iris and Wine data sets, as shown in **Tables 2** and **3**. It is found that the difference between the average value and variance of the Iris data set is not very large, indicating that the Iris data set is intensive data, that is, the dispersion is small. The difference between the attribute values of the Wine data set, whether the average value or variance is large, is hundreds of times the smallest, and the variance is tens of thousands of times. For example, the variance of attribute 13 is 99167, while the variance of attributes 3, 8, 9, 11 is as low as 0, This indicates that the Wine dataset is decentralized, that is, highly dispersed.

Table 2. Iris data analysis.

Statistic					
		One	**Two**	**Three**	**Four**
N	Valid	150	150	150	150
	Defect	0	0	0	0
Mean value		5.8433	3.0567	3.7508	1.1993
Standard deviation		0.82807	0.4358	1.7653	0.7624
Variance		0.6860	0.1900	3.116	0.581

Table 3. Wine data aggregation analysis.

Statistic							
		One	**Two**	**Three**	**Four**	**Five**	**Six**
N	Valid	178	178	178	178	178	178
	Defect	0	0	0	0	0	0
Mean value		13	2	2	19	100	2
Standard deviation		1	1	0	3	14	1

Variance	1	1	0	17	204	0

3.3. Examination of test findings

According to the analysis and test results of iris and wine datasets, the total F-measure value of MKmeans algorithm on both datasets has been improved. As shown in **Figure 3**, the MKmeans algorithm is stable on both iris datasets with small dispersion and Wine datasets with large dispersion. The total Fmeasure value is above 93%, which is 8–20 percentage points higher than the original Kmeans algorithm. From the comparison of operation efficiency, as shown in **Figure 4**, the improved algorithm MKmeans consumes an average of 0.01–0.06 s more time than the Kmeans algorithm, but the total F-measure value increases by 8–20 percentage points, so it is worth the extra time.

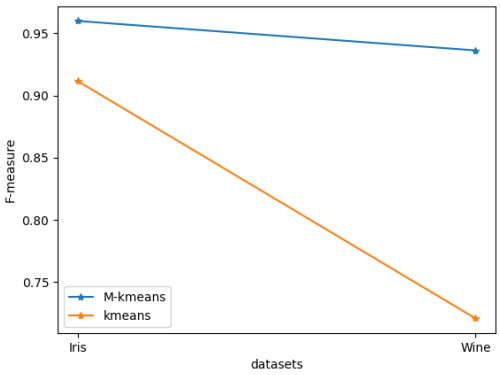

Figure 3. Comparison of total F-measure values.

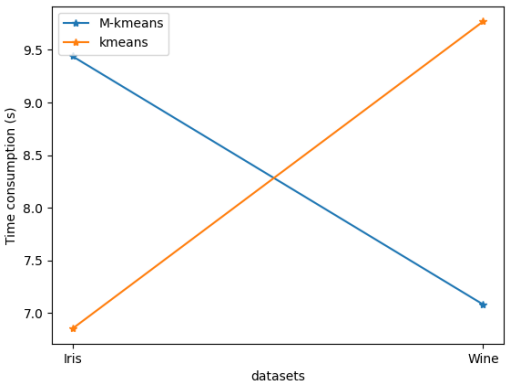

Figure 4. Time consumption comparison.

3.4. Application of MKmeans algorithm in teaching model

First, use Xmeans to get the optimal *k* value of 4. After that, the final dataset was clustered using MKmeans, Kmeans and two classical improved algorithms of Kmeans, K-mediums and Xmeans, respectively, Finally, as seen in **Figures 5–6**, the contour coefficient was used to analyse the clustering effect. Finally, the recognition ability grouping was analyzed using the clustering results.

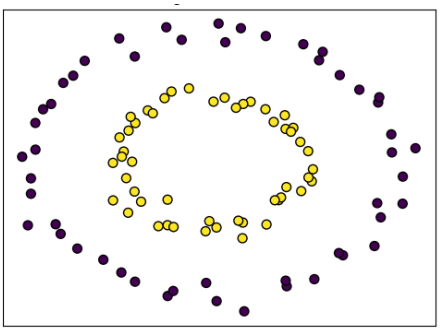

Figure 5. Profile coefficient of MKmeans.

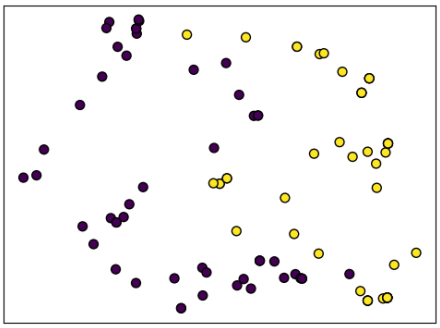

Figure 6. Profile coefficient of Kmeans.

To sum up the contour coefficient graphs of MKmeans, Kmeans, Xmeans, K-medoid algorithms, it can be found that the contour coefficients of all categories obtained by the new MKmeans algorithm proposed in this paper are close to 1, and there is no category that tends to –1. Other algorithms more or less have the contour coefficient values of partial data of some categories tend to –1, indicating that some data are not classified into the optimal category. As a result, the MKmeans method presented in this study has a better clustering effect than the Kmeans algorithm and its two classically modified algorithms. **Table 4** displays the final clustering centre determined by the MKmeans method. SPSS is utilised to analyse the clustering outcomes.

Table 4. Cluster Center of MKmeans in fmal.

	Memorizing	**Understand**	**Application**	**Comprehensive**	**Analysis**	**Evaluate**
Class 1	0.504941	0.504941	0.504941	0.437103	0.421745	0.305882
Class 2	0.70075	0.70075	0.68657	0.6525	0.661923	0.66
Class 3	0.25465	0.255655	0.219964	0.151078	0.15066	0.0597
Class 4	0.895926	0.895926	0.891228	0.902546	0.930484	0.912346

As shown in **Figures 7–10**, the analysis of four cluster centers obtained by MKmeans algorithm can explain the learning characteristics of learners. It can be seen from **Figures 5 and 6** that the cognitive abilities of category 1 are below 0.5, that of category 2 in **Figure 7** are between 0.65–0.70, that of category 3 in **Figure 8** are below 0.25, and that of category 4 in **Figure 9** are between 0.89–0.93. The aforementioned four categories allow learners' characteristics to be easily separated into four learning groups of varying degrees, which can greatly increase learning efficiency based on many aspects of targeted training and improvement.

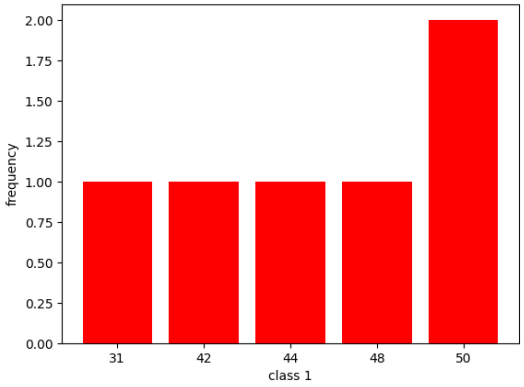

Figure7. Cluster center of class 1.

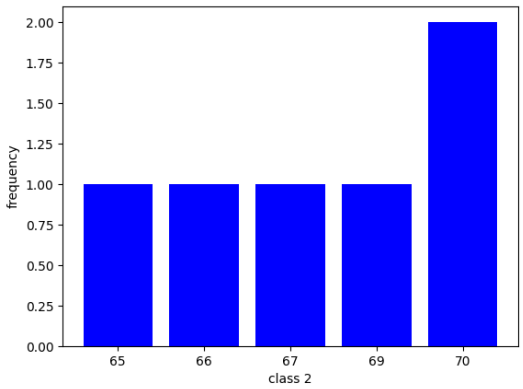

Figure 8. Cluster center of class 2.

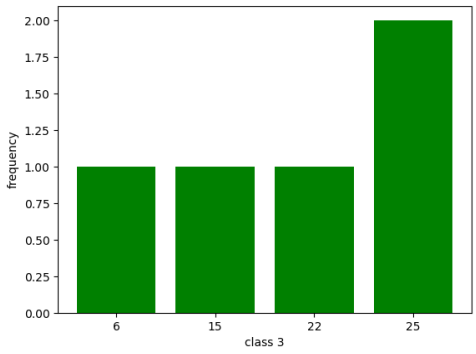

Figure 9. Cluster center of class 3.

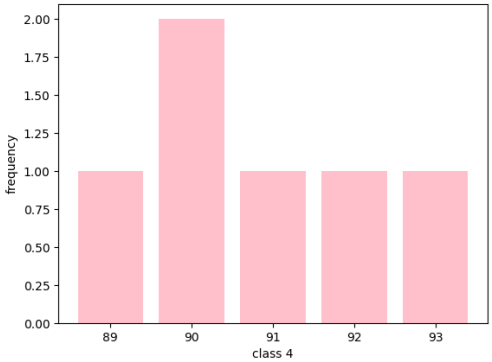

Figure 10. Cluster center of class 4.

4. Conclusion

This essay is based on Chinese instructors from abroad who teach online. This article builds an appropriate student model for online teaching of international Chinese teachers, and then examines the personalisation of online instruction by using the clustering algorithm theory in data mining. A traditional clustering technique built on the partition concept is the Kmeans algorithm. In order to create a student model appropriate for online instruction for foreign Chinese teachers, this paper applies the clustering algorithm to the personalised realisation of online instruction. Additionally, it makes improvements to the Kmeans algorithm to increase its efficiency, while also confirming the algorithm's superiority. The improved algorithm MKmeans is verified with the international classic datasets Iris and Wine. The final clustering result quality uses the F-measure value as the evaluation standard. It is found that MKmeans performs better than the algorithms of Kmeans, FCM, and SOM, indicating the value of

MKmeans. Finally, MKmeans and other algorithms are applied to the student feature data set corresponding to the student model, and learner groups with different levels of cognitive level and cognitive ability groups are obtained, as well as knowledge mastery level groups. It fully proves the value of cognitive model and knowledge model in student model.

Acknowledgments

The authors would like to show sincere thanks to those techniques who have contributed to this research.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Tang C, Zhang J. An intelligent deep learning-enabled recommendation algorithm for teaching music students. Soft Computing. 2022; 26(20): 10591-10598. doi: 10.1007/s00500-021-06709-x

[2] He P, Zheng C, Li T. Development and Validation of an Instrument for Measuring Chinese Chemistry Teachers' Perceived Self-Efficacy Towards Chemistry Core Competencies. International Journal of Science and Mathematics Education. 2021; 20(7): 1337-1359. doi: 10.1007/s10763-021-10216-8

[3] Zhang C, Li M, Wu D. Federated Multidomain Learning with Graph Ensemble Autoencoder GMM for Emotion Recognition. IEEE Transactions on Intelligent Transportation Systems. 2023; 24(7): 7631-7641. doi: 10.1109/tits.2022.3203800

[4] Sun, Y. (2023, December). Application research of online education clustering algorithm based on artificial intelligence. In *Proceedings of the 2023 International Conference on Information Education and Artificial Intelligence* (pp. 882-887).

[5] Ting, C. K., Ibrahim, N., Huspi, S. H., & Kadir, W. M. N. W. (2024). Multidimensional Context Clustering to Analyse Student Engagement in Online Learning Environment. *International Journal of Innovative Computing*, *14*(2), 89-96.

[6] Li, G., & Jamil, H. B. (2024). Teacher professional learning community and interdisciplinary collaborative teaching path under the informationization basic education model. *Yugoslav Journal of Operations Research*, (00), 29-29.

[7] Shi, H., Zhou, Y., Dennen, V. P., & Hur, J. (2024). From unsuccessful to successful learning: profiling behavior patterns and student clusters in Massive Open Online Courses. *Education and Information Technologies*, *29*(5), 5509-5540.

[8] Zhou, H. (2024, January). ASP Cluster analysis of humanistic quality evaluation data for private college students using NET technology and K-Means algorithm. In *2024 International Conference on Informatics Education and Computer Technology Applications (IECA)* (pp. 15-20). IEEE.

[9] He, Q. (2024). A personalised recommendation method of online and offline mixed teaching resources based on user preference behaviour. *International Journal of Reasoning-based Intelligent Systems*, *16*(5), 345-351.

[10] Ting, Y. (2024, June). Research on English Vocabulary Classification Method Based on Clustering Algorithm. In *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)* (pp. 1620-1624). IEEE.

[11] Jiang, D., He, Z., Chen, Y., Xu, L., & Lin, J. (2024). A robust and automatic method for the recognition of speech category in online learning discourse. *Journal of Ambient Intelligence and Humanized Computing*, 1-12.

[12] Huang, B., & Wang, C. (2023). RETRACTED ARTICLE: Research on Data Analysis of Efficient Innovation and Entrepreneurship Practice Teaching Based on LightGBM Classification Algorithm. *International Journal of Computational Intelligence Systems*, *16*(1), 145.

[13] Qi, X., Sun, G., & Yue, L. (2023). Applying Self-Optimised Feedback to a Learning Management System for Facilitating Personalised Learning Activities on Massive Open Online Courses. *Sustainability*, *15*(16), 12562.

[14] Ma, W., Yuan, Y., & Feng, J. (2023). Study on Predicting University Student Performance Based on Course Correlation. *Journal of Education and Educational Research*, *5*(3), 123-135.

[15] Ni, Q., Mi, Y., Wu, Y., He, L., Xu, Y., & Zhang, B. (2023). Design and Implementation of the Reliable Learning Style Recognition Mechanism Based on Fusion Labels and Ensemble Classification. *IEEE Transactions on Learning Technologies*, *17*, 241-257.

[16] ul Haque, M. M., & Kotaiah, B. (2023, November). A Cluster Based Recommendation System for MOOCs. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (pp. 406-411). IEEE.

[17] Kangaslampi, R., Rämö, J., Nokelainen, P., Hirvonen, J., Viro, E., Ali-Löytty, S., ... & Kaarakka, T. (2024). Changes in students' approaches to learning on engineering mathematics courses with two different instructional models. *International Journal of Education in Mathematics, Science and Technology*, *12*(3), 750-772.

[18] Dang, R. (2023, October). Construction of Online Teaching Effect Evaluation System Based on Complex Learning Behavior Data Analysis. In *Proceedings of the 2023 International Conference on Communication Network and Machine Learning* (pp. 289-293).