Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Cluster Analysis and Pattern Recognition of Students' Sports Performance Data during Physical Education Instruction

Yong Wang[1], Xu Wang[1], Zongshuai Hao[1,✉]

[1] *Department of Physical Education, Cangzhou Normal University, Cangzhou, Hebei, 061001, China*

ABSTRACT

In this paper, a K-Means clustering algorithm based on improved differential evolution (AGDE-KM) is proposed to design the adaptive operation operator, design the multi-variation strategy and introduce the weight coefficients in the variation stage to regulate the searching ability of the algorithm and accelerate its convergence speed. The Gaussian perturbation crossover operation based on the best individual of the current population is introduced, and the optimal solution output from the improved differential evolution algorithm is used as the clustering center to realize the cluster analysis of students' sports performance data. Afterwards, the hierarchical recognition algorithm and support vector machine are used to recognize students' sports patterns, and the wavelet transform algorithm is used to extract and select the students' sports feature quantities, so as to improve the accuracy of students' sports pattern recognition in sports teaching. In the process of physical education teaching, AGDE - KM algorithm is more pertinent to the clustering effect of students' sports performance, and its explanatory degrees of Calinski-harabasz metrics, profile coefficients, and Dunn metrics are 860.0276, 0.3928, and 0.0486, which are 19.0382, 0.0435, and 0.0099. In addition, the AGDE-KM algorithm achieves 95.7625%, 99.75%, and 99.85% of the mean value of step recognition accuracy for different testers in the 50m, 800m, and 1000m events, respectively, which is a good recognition effect.

Keywords: AGDE-KM, Adaptive manipulation operator, Multivariate strategy, Hierarchical recognition algorithm, Sports

1. Introduction

Physical education is an indispensable part of school education, and the evaluation of students' physical education performance is a crucial part of physical education teaching. Correct and effective

✉ Corresponding author.
E-mail address: czsfxytyx@163.com (Z. Hao).
Received 25 February 2024; Revised 05 April 2024; Accepted 15 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-498

evaluation of students' physical education performance can help to stimulate students' interest in learning and improve their physical education level [1-3].

When evaluating students' sports performance, diversified evaluation methods can be adopted to fully understand the actual situation of students. In sports competitions or examinations, students' athletic level can be evaluated by recording their results and observing their athletic performance [4-7]. Meanwhile, students' sports performance can also be evaluated comprehensively through face-to-face communication with them to understand their sports interests, participation and attitudes [8-9]. In addition, objective and fair evaluation criteria should be developed when evaluating students' sports performance. The evaluation criteria should be clear and clearly inform students of what they are being evaluated on. The evaluation criteria should also be fair and consistent, treating all students equally [10-13]. In addition, when evaluating students' physical education performance, encouragement and motivation should be emphasized to stimulate students' interest and motivation. By praising students' progress and excellent performance, they are motivated to persevere and improve themselves in physical education, not only focusing on students' performance, but also focusing on cultivating students' interest in physical education and health consciousness, so that every student can enjoy the fun of physical education [14-17]. Through the combination of multiple evaluation methods can effectively evaluate students' sports performance and promote students' overall development and progress [18].

In this study, adaptive operation operator is designed to improve the global search ability in the early stage and the convergence speed in the late stage of differential evolution algorithm. After that, a multi-variation strategy is designed and weight coefficients are introduced to play the advantages of different variation strategies in different evolutionary stages of the algorithm, to balance the global and local search ability of the algorithm and to accelerate the convergence speed of the algorithm. On this basis, a Gaussian perturbation crossover operation is proposed based on the best individual of the current population, which provides students with better clustering direction while maintaining the diversity of the population in the “dimension” to avoid the algorithm falling into the local optimum. The optimal solution output when the algorithm stops execution is used as the initial clustering center instead of the randomly selected clustering center of traditional K-Means. After that, the hierarchical recognition algorithm and support vector machine are used to recognize the students' movement patterns. Finally, the relevant feature quantities in the time and frequency domains of the motion are extracted by wavelet transform technique to provide a basis for the selection of motion features.

2. Clustering and pattern recognition methods for student athletic performance

2.1. K-Means clustering algorithm based on improved differential evolution

2.1.1. Improved differential evolutionary algorithm

1) Adaptive operation operator. The mutation and crossover operations in the differential evolution algorithm [19] are the center of the whole algorithm, and the mutation factor F and the crossover factor C_R have an extremely important influence on the algorithm. In the traditional differential evolution algorithm, the mutation factor F and crossover factor C_R are fixed values, which leads to the limitation of the algorithm's optimization ability and convergence performance, and is not conducive to the improvement of the algorithm's optimization performance. In the early stage of algorithm evolution, the diversity of population individuals should be maintained to improve the global optimization ability of the algorithm, while in the middle and late stages of the algorithm, the local search ability of the algorithm should be strengthened to improve the convergence speed of the algorithm.

In order to adapt F and C_R to the algorithm's optimization needs at different stages of evolution, this paper uses the adaptive operation operator to improve the algorithm, and adaptively balance the algorithm's global search ability and local search ability as the algorithm evolves. The adaptive strategy of the operation operator is as follows:

$$F = F_{\max} - (F_{\max} - F_{\min})(G / G_{\max})^2 \tag{1}$$

$$C_R = C_{R\max} - (C_{R\max} - C_{R\min})(G / G_{\max})^2 \tag{2}$$

where: $F_{\min}$ represents the lower limit of F, $F_{\min} = 0.3$; $F_{\max}$ represents the upper limit of F, $F_{\max} = 0.9$; $C_{R\min}$ represents the lower limit of C_R, $C_{Rmim} = 0.3$; $C_{R\max}$ represents the upper limit of C_R, $C_{R\max} = 0.9$; G represents the number of iterations of the current algorithm; $G_{\max}$ represents the maximum number of iterations specified by the algorithm.

Implementing adaptive adjustment of F and C_R according to the above equation ensures that the variation factor F and crossover factor C_R decrease linearly as the number of algorithm evolution increases.

2) Multivariate strategy with introduction of weighting factors. In order to improve the optimization seeking ability of the differential evolution algorithm, the diversity of the population should be maintained at the initial stage of the algorithm's evolution in order to improve the global optimization seeking ability of the algorithm. According to the characteristics of different mutation strategies, the DE/rand/l mutation strategy and the DE/current-to-best/2 mutation strategy were combined, and an improved differential evolution algorithm based on the multi-mutation strategy [20] (AGDE-KM) was proposed, and the weight coefficient w was introduced to adaptively adjust the proportion of different mutation strategies with the evolution times of the algorithm, so as to give full play to the advantages of different mutation strategies in different evolution stages of the algorithm. Here's what:

$$W = W_{\min} + (W_{\max} - W_{\min})(G / G_{\max}) \tag{3}$$

$$\begin{aligned} W = W_{\min} + (W_{\max} - W_{\min})(G / G_{\max}) V_{i,G+1} = (1-W)[X_{r_i,G} + F(X_{r_i,G} - X_{r_i,G})] \\ +W[X_{i,G} + F(X_{\text{bad},G} - X_{i,G}) + F(X_{r_sG} - X_{r_sG}) + F(X_{r_sG} - X_{r_sG})] \end{aligned} \tag{4}$$

where: W is the weight factor, $W_{\text{mim}} = 0$, $W_{\max} = 1$; G is the current number of iterations; $G_{\max}$ is the maximum number of iterations set by the algorithm; $V_{i,a+1}$ represents the mutated individuals in the $G+1$ th generation; $X_{\text{bent}G}$ represents the individual with the best fitness value in the current population; $r_1, r_2, r_3, r_4, r_5, r_6, r_7$ represents the 7 random numbers not equal to i and unequal to each other taken from [1, NP].

3) Gaussian perturbation crossover operation: Differential evolutionary algorithms develop and explore new solution spaces through variation and crossover operations [21], which are performed on individual “dimensions”, although the variation and crossover factors can be adaptively adjusted.

The new value on each dimension maintains the diversity of the population in the “dimension” as follows:

(1) For the clustering problem, for one of its dimensions n, take the random number rand(0, 1), and if $rand(0,1) \le C_R$, then $U_{i,c}^n = V_{i,c}^n$, and then end the crossover operation for that dimension; conversely, go to step 2).

(2) Again take the random number rand(0, 1), if rand(0, 1) < 0. 7, then $U_{i,G}^n = X_{lostG}^n$ if rand(0, 1) > 0. 7, then generate a random value in that dimension by Gaussian perturbation, the generation formula is:

$$U_{i,G}^{n} = X_{bost,G}^{n} \cdot \left[1 + C \cdot N\left(0,1\right) \right] \tag{5}$$

where: C is the control parameter for the Gaussian perturbation, $C = 0.1$; and the random value $N(0,1)$ indicates a normal distribution that satisfies a mean of 0 and a standard deviation of 1.

2.1.2. K-Means clustering algorithm based on improved differential evolution:

1) Fitness function. The improved differential evolutionary algorithm is integrated with the K-Means tight class algorithm, and the fitness function is used to calculate the fitness function value of each individual, and the iterative optimization search is carried out using the improved differential evolutionary algorithm. The formula is as follows:

$$\text{fitness} = \text{SSE} = \sum_{i,k} (x_{i,k} - c_k)^2 \tag{6}$$

where: x_{ik} denotes a data point; c_k denotes the center of the cluster to which the data point belongs. The smaller SSE is, the better the clustering effect is; and vice versa, the worse the clustering effect is.

2) Algorithm flow

(1) Set the population size NP as 10 times of the solution dimension D, the variation factor F and crossover factor C_R for adaptive taking, $F_{\min} = 0.3$, $F_{\max} = 0.9$, $C_{R\min} = 0.3$, $C_{R\max} = 0.9$, the current evolution number is G, the maximum evolution number of the algorithm is $G_{\max} = 1200$, and the Gaussian perturbation coefficient is $C = 0.1$.

(2) Initialize the population of size NP.

(3) For individual $X_{i,G}$ in the population, obtain mutant individual $V_{i,G}$ according to the improved mutation operation.

(4) For parent individual $X_{i,G}$ and variant individual $V_{i,G}$, obtain intermediate experimental individual $U_{i,G}$ according to the improved crossover operation.

(5) Calculate the fitness values of the parent individual $X_{i,G}$ and the intermediate experimental individual $U_{i,G}$, respectively, and compare the fitness values, and the individual with better fitness will enter into the next generation of the population to continue to participate in the evolution. Steps (3)~(5) are performed cyclically until the maximum number of evolutions of the algorithm is reached.

(6) The optimal individuals output from the algorithm are clustered as the initial clustering center of K-Means, and the cycle iterates until the stopping requirement of clustering or the maximum number of iterations is satisfied, the algorithm ends, and the clustering results are output.

The flow of K-Means clustering algorithm based on improved differential evolution is shown in Figure 1.

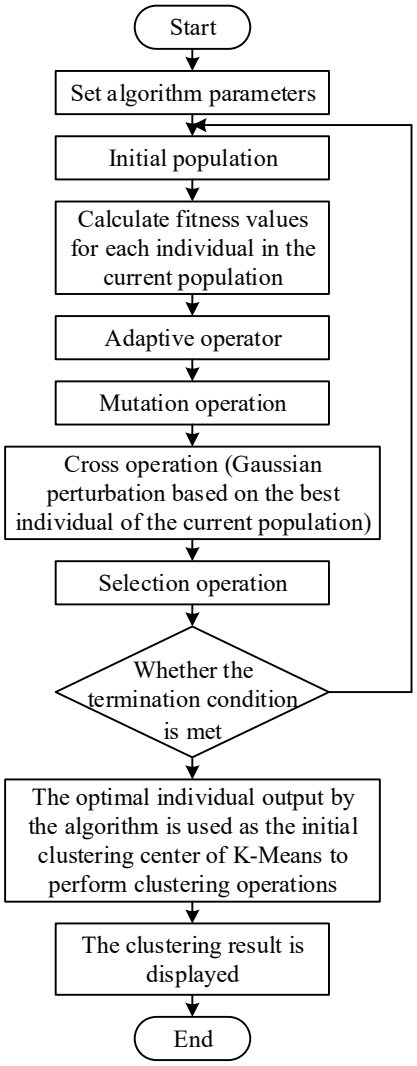


Fig. 1. Algorithm process based on improved difference evolution

2.2. Data sample and model design

2.2.1. Data samples. The data used in this study came from the 3-year physical fitness test data of all college students from 2021 to 2023 in the health database of a university, with a total of 60,000 records in the raw data (30,000 for men and 30,000 for women), and 6 attributes for each record. Its overall total score distribution was obtained by using SPSS Modeler software to preprocess the data of each field in the records. Figures 2 and 3 show the total score distribution and total score level distribution, respectively. In the past three years, the total scores of students' physical fitness test were mainly distributed between 60-90 points, and the total score grades of “Pass” and “Good” had the highest frequency, accounting for 51.27% and 36.71% of the total number of visits, respectively.

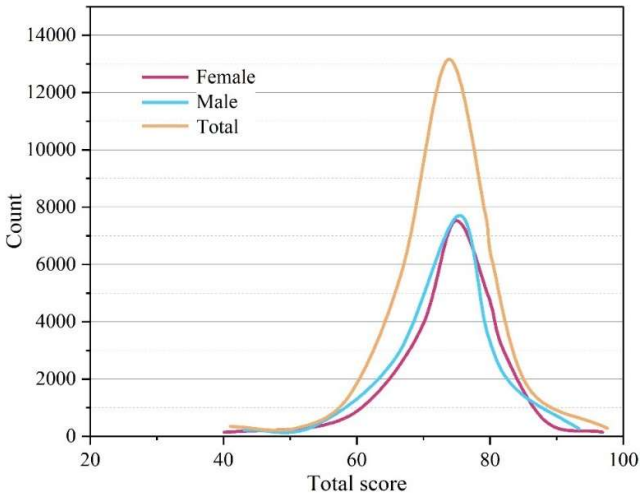


Fig. 2. General score profile

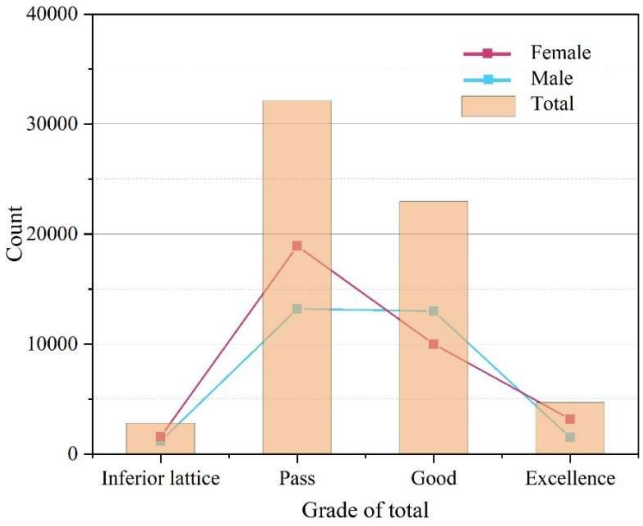


Fig. 3. The distribution of the total score

2.2.2. Data pre-processing. Generally speaking, the original data contains too much noise data and incomplete data, which need to go through the process of data cleaning, data screening, data processing and so on. Data cleansing is the process of re-examining and calibrating data with the aim of removing duplicate information, correcting existing errors, and providing data consistency. Data filtering is the process of removing unwanted fields according to the actual need. In this study, what is needed is the relationship between the total score of the physical fitness test and the degree of influence between the various test items, so the redundant fields such as year, grade number, class number, class name, ethnicity code, and date of birth are filtered, leaving only the useful fields. Data processing is based on the requirements of a specific data mining model, the data is processed into a form that meets the requirements of the model.

2.2.3. Modeling. SPSS Modeler software was used to design the clustering model for boys' and girls' records separately according to the requirements of K-means algorithm. For the boys' group, 7 fields of lung capacity score, 50m score, 1000m score, standing long jump score, sitting forward bends score, pull-ups score, and total score score were selected as input fields; for the girls' group, 7 fields of lung capacity score, 50m score, 800m score, standing long jump score, sitting forward bends score, one-minute sit-ups score, and total score were selected as input fields. After clustering the results are divided into 4 categories, the larger the average value of the ratings in these 4 categories of data

indicates that the students in this category in this index the better the athletic performance, and use this value to compare the same items in the 4 categories of data and analyze the running model to get the clustering results.

2.3. Student Movement Pattern Recognition

In this paper, a hierarchical recognition algorithm [22] and support vector machines are used for the recognition of multiple patterns, and the exercise patterns to be recognized are classified into two categories: one is common exercise behaviors; the other is behaviors aimed at developing healthy exercise habits.

In this paper, the support vector machine based on hierarchical recognition algorithm is used for the recognition of M pattern, then at most $M/2$ classifiers need to be constructed, which improves the complexity of support vector machine learning. The construction of the classifiers is shown below:

$$Minimize \frac{1}{2} \| \omega^2 \| + C \sum_{i=1}^{k+k} \xi_i , s \neq t, s, t \in \{1, 2, \cdots, N\} \tag{7}$$

$$subject\ to\ y_i(\omega_{st} \cdot x_i + b_{st}) \geq 1 - \xi_i \tag{8}$$

where: s and t are the category numbers of the two pattern samples. By solving this optimization problem, the following decision function model is obtained:

$$F^{s-t}(x) = \begin{cases} s & f_{s-t}(x) > 0 \\ t & \text{other} \end{cases} \tag{9}$$

where: $f_{s-t}(x)$ is the decision value of the support vector machine classification. In the identification of two types of patterns, the classifier identifies the sample category to which the feature parameter x_i belongs, and if the sample belongs to the pattern of class s, then the votes of class s patterns will be added 1, otherwise, the votes of class t patterns will be added 1, and finally, the "voting mechanism" is used, and the one that has the most votes is the final classification pattern.

The block diagram of the algorithm in this paper is shown in Figure 4. Firstly, the collected sensor data are filtered; then the time-domain features of the acceleration sensor and the angular velocity sensor for motion pattern recognition are calculated, and the extracted features are compared and analyzed; the final extracted time-domain features include the variance, quartile distance and peak value of the acceleration sensor, and the mean, variance and skewness of the angular velocity sensor.

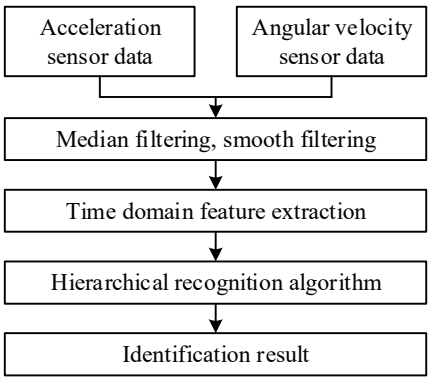


Fig. 4. Block diagram of algorithm

2.4. Data pre-processing methods

1) Acceleration data windowing. The acceleration signal is presented in the form of data stream in the time domain, which cannot be directly characterized by feature engineering. The conventional method is to set up a sliding window for processing, and this method has been widely used in human motion state recognition. The experiment comprehensively analyzes the characteristics of five kinds of motion modes and counts the time needed to complete each kind of action, and chooses 4s as the default length of a segment of the action, for a sliding window, the width of 800 sampling points, and the overlap of every two neighboring windows is 50%. Take the walking X-direction acceleration data as an example, 4s is chosen as the width of the window to regularize the length of the acceleration signals of different actions.

2) Filtering. Compared to machine motion, human motion is slower and contains mainly low-frequency signals. In practice, acceleration signals are inevitably affected by hardware circuits, transmission noise, perturbation frequency noise, and jitter, and the output is often interspersed with many irrelevant signals. When the acceleration is zero, the output is not zero, but a non-zero vector output a_{real}, measured value $a_{measured}$ is shown below:

$$a_{measured} = a_{real} + a_{error} \tag{10}$$

where a_{real} is the real acceleration value and a_{error} is the measurement error of the sensor. In order to make the measured acceleration signal closer to the real value, the experiment uses an adaptive smoothing filtering method.

There are two common smoothing filtering methods: the sliding average filtering method and the first-order inertia low-pass filtering method. Although the sliding average method is simple, the experiment proves that the practicality is poor, in the static situation, the first-order inertia low-pass filtering effect is obvious, but in the motion state, the phase lag, so the original method is not suitable for the dynamic signal processing, in order to adapt to the real-time nature of the signal, the increase of the net value is used to update the filtering parameters. Define the new output value as the current sampling value and the last output value are weighted to get the current output value. The algorithm is as follows:

$$Y(n) = m \cdot X(n) + (1 - m) \cdot Y(n-1) \tag{11}$$

where $Y(n)$ is the current filter output value; m is the filter coefficient, which has a great influence on the effect of filtering; $X(n)$ is the current sampling value; $Y(n-1)$ is the last filter output value. Next, the conditional threshold Δa for determining the motion state is introduced, and the filter coefficients m are modified according to the following equation, i.e.:

$$\Delta = Y(n) - Y(n-1) > \Delta_a \tag{12}$$

$$m = \left(1 - \frac{\Delta_a}{\Delta}\right) \cdot k_0 \tag{13}$$

where Δ is the difference between the output value of this filter and the output value of the last filter; Δ_a is the motion state threshold; k_0 is the default parameter.

2.5. Wavelet transform based data feature extraction and selection methods

2.5.1. Time Domain Feature Extraction Methods. After segmenting and smoothing the original acceleration signal, the relevant feature quantities are next extracted from the time and frequency domains.

The 3-direction acceleration mean value $\overline{a}_x, \overline{a}_y, \overline{a}_z$, which reflects the average state of the signal in each of the X, Y, and Z directions, can reflect the degree of intensity of the human body in the 3 directions during motion. The calculation formula is defined as follows:

$$\tilde{a}_x = \frac{\sum_{k=1} a_{x,k}}{n} \tag{14}$$

$$\tilde{a}_y = \frac{\sum_{k=1} a_{y,k}}{n} \tag{15}$$

$$\overline{a}_z = \frac{\sum_{k=1}^{n} a_{z,k}}{n} \tag{16}$$

where $a_{x,k}, a_{y,k}, a_{z,k}$ is the X, Y, and Z axis acceleration signal obtained at the k nd sampling point in a window; is the number of sampling points per window.

The synthetic acceleration mean value $\overline{a}$ is able to reflect the integrated intensity of a certain motion mode. The formula is as follows:

$$a_k = \sqrt{a_{x,k}^2 + a_{y,k}^2 + a_{z,k}^2} \tag{17}$$

$$\overline{a} = \frac{\sum_{k=1}^{n} a_k}{n} \tag{18}$$

where a_k the mode of the triaxial synthetic acceleration at the k nd sampling point in the window. The synthetic acceleration mean $\overline{a}$ can be obtained by bringing a_k into Eq. (18).

The standard deviation of the acceleration in the 3 directions, in the X direction for example, and the acceleration variance σ_x^2 are calculated as follows:

$$\sigma_x^2 = \frac{\sum_{k=1}^{n} (a_{x,k} - \overline{a}_x)^2}{n} \tag{19}$$

The same reasoning leads to σ_y^2, σ_z^2:

$$\sigma_y^2 = \frac{\sum_{k=1}^{n} (a_{y,k} - \overline{a}_y)^2}{n} \tag{20}$$

$$\sigma_z^2 = \frac{\sum_{k=1}^{n} (a_{z,k} - \overline{a}_z)^2}{n} \tag{21}$$

The synthesized acceleration variance σ^2 is given by:

$$\sigma^2 = \frac{\sum_{k=1}^{n} (a_k - \overline{a})^2}{n} \tag{22}$$

Acceleration peaks and valleys indicate the difference between the maximum and minimum values of acceleration in a window, expressed as a_{pv}:

$$a_{pv} = Max(a) - Min(a) \tag{23}$$

Using the above equation the peaks and valleys of the combined acceleration as well as the peaks and valleys of the acceleration in the 3 directions can be calculated.

The two-by-two correlation coefficients of the 3 directional accelerations are:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} \tag{24}$$

$$\rho(X,Z) = \frac{cov(X,Z)}{\sigma_x \sigma_z} \tag{25}$$

$$\rho(Y,Z) = \frac{cov(Y,Z)}{\sigma_y \sigma_z} \tag{26}$$

2.5.2. Wavelet transform based data distribution feature extraction. The components of wavelet are a set of wavelet basis functions [23], which can characterize the signal in the time and frequency domains in a local range, which is the advantage of wavelet transform over Fourier transform. There are 15 commonly used wavelet functions such as Haar, dbN wavelet and so on. After the signal is analyzed by wavelet, the energy characteristics in each level can be obtained. The binary wavelet decomposition of signal $f(t)$ can be expressed as:

$$f(t) = A^t + \sum D^t \tag{27}$$

where, A is the approximation signal, which is the low-frequency portion, and D is the detail signal, which is the high-frequency portion.

The total energy of the signal is:

$$E = EA_i + \sum ED_j \tag{28}$$

The energy of the approximated signal at layer j and the detailed signal at each layer are selected as features to construct the feature vectors:

$$F = [EA_j, ED_1, ED_2, \cdots, ED_j] \tag{29}$$

This experiment directly calls the PyWavelets signal processing library in Python, sets j to 2, performs feature extraction on the acceleration signal, and outputs the feature vector:

$$F = [EA_2, ED_1, ED_2] \tag{30}$$

2.5.3. Feature selection. Feature selection is an important "data preprocessing" process, which can remove the features that are irrelevant to the current learning task, in order to reduce the learning burden and learning difficulty. In this experiment, we adopt the feature selection algorithm based on information gain (IG), which is also a common and efficient feature selection algorithm.

Calculating information gain begins with defining information entropy as a metric generally used as a reflection of the purity of a sample set. Assuming that $p_k (k = 1, 2, \cdots, |y|)$ is the proportion of samples of type k in the current sample set D, the information entropy of D is defined as:

$$Ent(D) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \tag{31}$$

where $|y|$ represents the number of labeling categories and the smaller the value of information entropy $Ent(D)$, the higher the purity of set D. Information gain is defined below.

Since the features extracted from the signal are all continuous values and the number of fetchable values of continuous value attributes is not finite, the information gain cannot be calculated directly

based on the continuous values directly, the experiment uses continuous value discretization technique. Given the sample set D and continuous attribute v, assume v that n different values $\left\{v^1, v^2, \cdots, v^n\right\}$ appear on D, based on the division point t can be divided D into subsets D_t^- and D_t^+ for neighboring attribute values v^i and v^i+1, t taking any value in the interval $[v^i, v^{i+1})$ produces the same division results. Define the set of candidate division points containing $n-1$ elements. Then:

$$T_a = \left\{ \frac{v^i + v^{i+1}}{2} | 1 \le i \le n-1 \right\} \tag{32}$$

The median point of interval $[v^i, v^{i+1})$ is taken as a candidate division point, and the optimal division point is selected for the division of the sample set, i.e:

$$Gain(D, v) = Gain(D, v, t) \tag{33}$$

$$Gain(D, v, t) = Ent(D) - \sum_{\lambda \in \{-,+\}} \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) \tag{34}$$

where, $Gain(D, v, t)$ is the information gain of the sample set D based on the bisection of the division point t, such that we bring in each feature v to get the information gain value of all the features and thus calculate the importance value of each feature.

3. Analysis of clustering and identification results of students' sports performance

3.1. Analysis of clustering results of student athletic performance data

The improved K-Means algorithm is now compared with the original K-Mean method for quality and effectiveness. Generally, the clustering evaluation methods are categorized into two types according to whether there is a benchmark available or not: the extrinsic method is used to evaluate if there is an available benchmark, and the intrinsic method is used to evaluate if there is no benchmark available.

In this paper, there is no clustering benchmark, so the intrinsic method is used for evaluation, and the following parametric indicators are used for evaluation:

1) Calinski-Harabasz criterion formula is as follows:

$$\frac{TotalSSB}{TotalSSE} / \frac{n-K}{K-1} \tag{35}$$

Where, $(n-K)/(k-1)$ is the complexity. The Calinski-Harabasz formula, the larger the ratio, the better. The Calinski-Harabasz formula is simple to compute and fast to run, which will be used in this paper as a reference for determining a reasonable K-value.

2) The contour coefficient mathematical formula is as follows:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{36}$$

where $a(i)$ denotes the average value of the distance between object i and the other objects of the class to which i belongs. The smaller the value, the more compact the inverse cluster is. $b(i)$ denotes the minimum distance from object i to the class to which it does not belong to i. The larger the value, the more separated from other clusters. The contour coefficient takes values between 0 and 1, the larger the value the better the clustering effect.

3) Dunn's index, the mathematical formula is as follows:

$$DVI = \frac{\min_{0<m\neq n<K}\left\{\min_{\forall x,\in\Omega_+}\left\{x_i - x_j\right\}\right\}}{\max_{0<m\leq K}\max_{\forall x,\cdots\in\Omega_+}\left\{x_i - x_j\right\}} \tag{37}$$

Dunn's index first calculates the shortest distance between any two classes, and then divides the result by the maximum distance between objects in any class. A larger Dunn's index means better clustering.

3.1.1. Clustering results of performance data. The k-means algorithm clustered the data into four classes, and the clustering results of the k-means algorithm for students' exercise behavior are shown in Table 1.

Table 1. The k-means algorithm is a cluster of students' sports behavior

Cluster class	Sample size (number)	50 m (s)	Fixed jump (number)	Preflexion (cm)	800 m or 1000 m (s)	Sit-ups or Lead up (number)
1	22438	75.78	77.71	77.77	35.86/51.73	41.27/9.42
2	20574	69.96	56.02	15.29	38.24/71.49	38.08/7.36
3	10925	80.2	77.11	24.64	67.62/68.54	52.58/10.54
4	5938	85.24	78.75	76.09	65.47/59.82	49.24/11.39

The improved k-means algorithm divides the data into 4 classes, and the clustering results of the improved k-means algorithm for students' exercise behavior are shown in Table 2.

Table 2. Improved k-means algorithm is a clustering of students' sports behavior

Cluster class	Sample size (number)	50 m (s)	Fixed jump (number)	Preflexion (cm)	800 m or 1000 m (s)	Sit-ups or Lead up (number)
1	32242	79.13	79.04	77.06	35.82/52.42	43.77/10.11
2	20570	71.31	56.75	12.66	39.89/70.51	36.62/8.40
3	4482	79.82	77.37	25.31	70.11/69.43	51.74/9.89
4	2593	87.46	80.26	74.35	65.47/60.08	47.38/11.23

The results of the two clustering algorithms are compared parametrically.The comparison of the clustering results of the K-Means algorithm and the improved K-Means algorithm is shown in Table 3. It can be seen that the improved K-Means algorithm compared to the K-Means algorithm, its Calinski-harabasz index, profile coefficient and Dunn's index are 860.0276, 0.3928 and 0.0486 respectively, which are 19.0382, 0.0435 and 0.0099 more than the K-Means algorithm.It can be seen that the improved K-Means algorithm is more accurate in clustering the data of students' sports performance during physical education.

Table 3. The clustering of two algorithms is compared

Algorithm	Calinski-harabasz index	Contour coefficient	Dunne index
K-Means	840.9894	0.3493	0.0387
Improved k-means	860.0276	0.3928	0.0486

3.1.2. Characteristic presentation of clustering results. According to the improved K-Means algorithm to draw the 4 categories of change line graph is shown in Figure 5. By observing the graph, the major and minor weaknesses of the 4-class grouping can be seen. The results of the clustering structure characterization of the improved K-Means algorithm are shown in Table 4. The results of the study show that the improved K-Means algorithm clustering is more superior in terms of clustering

effect and clustering interpretation compared to the K-Means algorithm. In participating in the cluster analysis of each attribute, the differences between clusters are obvious and each grouping is easy to explain, and the overall clustering effect is good.

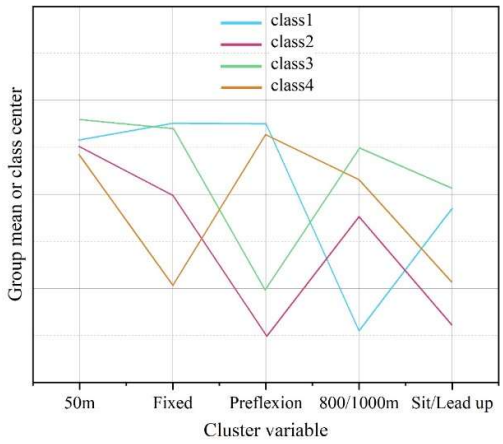


Fig. 5. Improved k-means clustering variable mean

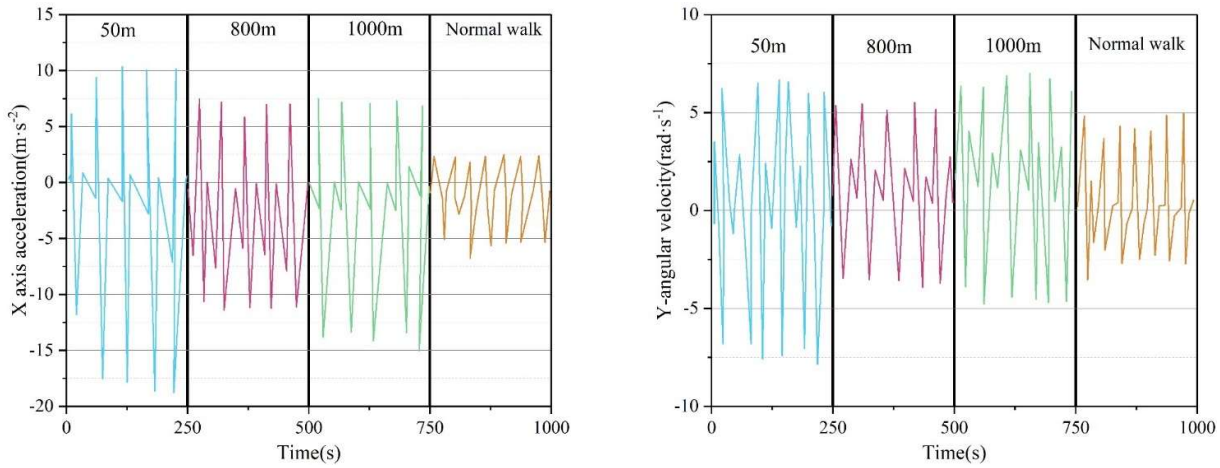
Table 4. Improved k-means algorithm clustering structure characteristics result

Cluster result	Major weakness	Minor weakness
1	800m/1000m	No
2	Preflexion	Sit-ups or Lead up
3	Preflexion	No
4	Fixed jump	Sit-ups or Lead up

3.2. Results and analysis of student athletic performance identification

The experimental equipment used was the MTi-G-710 MEMS combined navigation system with a gyroscope zero deviation of 10°/h and an accelerometer zero deviation of 40 μg . A pair of thick-soled shoes was selected, a groove was dug at the right heel, and the MTi-G-710 was embedded into the groove. Four people were selected for the test, and all of the testers were male. One of the teachers was 40 years old; all three students were 22 years old. The experiment was conducted in the laboratory building and student dormitory of a school. During the experiment, the testers performed 1000m, going up the stairs, and 800m3 motions randomly according to their own habits. Since the MEMS inertial device was only installed inside the heel of the right shoe, the right foot was only counted as one step of motion from lifting to landing.

A set of data from the teacher's walking is used as the sample data, in which the sliding window N and N_a take the value of 10, and the threshold TH_σ takes the value of 2.5×10^2 . After the sliding average, the output of the recognition model is shown in Fig. 6, and (a) and (b) are the output curves of the x-axis accelerometers and the y-axis gyroscopes, respectively. Based on this sample data to extract the eigenvalues, the judgment program is written to judge the data of other groups and test the algorithm. There were 16 groups of test data, of which 2 groups were tested by the teacher and 3 groups were tested by each of the 2 students. The trajectory of each group of data is not the same, and the number of steps of 50m, 800m and 1000m is not the same, and the number of steps of each movement is recorded in the experiment, so that it is easy to compare with the number of steps recognized by the algorithm.



(a)X-accelerometer (b)Y-axis gyroscope

Fig. 6. Identify the results of the model output

The recognition rate is defined as follows:

$$\text{Recognition rate} = \frac{\text{Number of correctly identified steps}}{\text{Actual number of steps}} \times 100\% \tag{38}$$

Table 5 summarizes the statistical results of the motor test data for the four testers. Positive numbers in parentheses in the table indicate the number of extra recognized steps, e.g., the teacher's average number of recognized steps is 97(-6) steps, which means that 97 steps are correctly recognized and 6 steps are less recognized, while + indicates the number of extra recognized steps. As can be seen from the table, the recognition rates for the 800m and 1000m exercise modes are very high, with all sets of data having a recognition rate of over 99%. Specifically, the recognition rate of 50m motion is greater than 95% for all human data except for the teacher (94.17%).The recognition rates of the testers in 800m and 1000m motion are between 99.44%-99.88% and 99.70%-100%, respectively. It can be seen that the use of accelerometers and gyroscopes wave peaks and troughs output sequential logical order can recognize the 50m, 800m, 1000m and other movement patterns, the recognition rate of which is more than 94.17%.

The average value of the recognition rate of different testers' motion data is taken as the total recognition rate, then the recognition rates of 50m, 800m and 1000m are 95.7625%, 99.75% and 99.85% respectively. In the experiment, because the inertial sensor is installed at the heel of the right foot, when the left foot is far away from the next step and the right foot takes the next step, it should be the next step state according to the motion state, but at this time the algorithm is easy to judge it as the previous step state. In the transition process between the previous step and the next step, if the left foot is far away from the ground of the next step, and the next right foot leaves the ground to reach the next step, according to the state of motion should be the next step state, but at this time, the logic of the peaks and valleys of the output of the accelerometers and gyroscopes is similar to that of the previous step, and the algorithm is easy to judge it as the previous step state. For these two cases, inertial sensors can be installed on the left foot as well, which in turn improves the recognition efficiency.

Table 5. Statistics of four test personnel exercise test data

Test content		Tester			
		Teacher	Student A	Student B	Student C
50m	Average number of steps	97(-6)	103(+3)	103(+4)	102(+4)

	Mean of actual steps	103	100	99	98
	Average recognition rate(%)	94.17	97.00	95.96	95.92
800m	Average number of steps	1605(-2)	1604(-9)	1598(-3)	1597(+2)
	Mean of actual steps	1603	1595	1595	1599
	Average recognition rate(%)	99.88	99.44	99.81	99.87
1000m	Average number of steps	1999	1996(6)	1999(-1)	1998(-5)
	Mean of actual steps	1999	2002	1998	2003
	Average recognition rate(%)	100.00	99.70	99.95	99.75

4. Conclusion and outlook

4.1. Conclusion

The article proposes a K-Means clustering algorithm based on Improved Differential Evolution (AGDE-KM) to cluster and analyze students' sports performance data, and then fuses the feature quantities of acceleration and angular velocity sensor data to identify students' multiple movement patterns.

The AGDE-KM algorithm has more accurate clustering effect on students' sports performance data in the process of physical education, and its Calinski-harabasz metrics, profile coefficients, and Dunn metrics are increased by 19.0382, 0.0435, and 0.0099, respectively, compared with the K-Means algorithm.The AGDE -KM algorithm has obvious differences between clusters, and the individual subgroups are easy to be interpreted, and its clustering The average total recognition rate of AGDE-KM algorithm reaches more than 95% for different testers' motion data in 50m, 800m and 1000m events, and the recognition effect is good.

4.2. Discussion

The method proposed in this paper has a high recognition rate but is limited in the motion patterns it can recognize. For other motion modes such as lateral movement, backward movement, jumping, etc., although the waveforms output from gyroscopes and accelerometers are different in each direction, the recognition rate is not high and may be difficult to recognize if the logical order of wave peaks and troughs outputs is used again for judgment. Therefore, in the later research, for multiple motion modes, it is recommended to adopt a combination of the waveform output from the inertial device and the variance of the output data, wave peaks and valleys, and other features for identification.

References

[1] Kewei, S., Díaz, V. G., & Kadry, S. N. (2022). Evaluating the Efficiency of Student Sports Training Based on Supervised Learning. International Journal of Technology and Human Interaction (IJTHI), 18(2), 1-17.

[2] Pestano, R. D., & Ibarra, F. P. (2021). Assessment on the implementation of special program in sports and student-athletes performance in sports competition. International Journal of Human Movement and Sports Sciences, 9(4), 791-796.

[3] Wang, G. (2022). Design of College Students' Sports Assessment System Based on Data Mining. Scientific Programming, 2022(1), 1337048.

[4] Luo, M., Yu, Y., Xu, W., Zhang, Q., & Xia, J. (2020, November). Analysis and prediction on sports performance of primary school students. In 2020 Chinese Automation Congress (CAC) (pp. 1041-1045). IEEE.

[5] Liu, J., & Wang, Y. (2022). Study on Academic Evaluation Practice of Theoretical Courses of Sports Basketball Education under the Concept of Quantitative Evaluation. Mathematical Problems in Engineering, 2022(1), 9151943.

[6] Tang, Y. (2024). Quality Evaluation of College Students' Sports Work Based on Intellectual or Intuitive Fuzzy Information in Language. Journal of Electrical Systems, 20(7s), 1584-1594.

[7] Cruz, R. B. (2022). Teaching competence and students' performance in sports class in physical education. International Journal of Advanced Multidisciplinary Studies, 2(6), 48-61.

[8] Soheili Pishkenari, S., Hematinejhad, M., & Gholizade, M. H. (2021). Analysis of Factors Affecting Performance Management in Sports Organizations (Case Study: Student Sports Federation and Physical Education and Sports Activities Office). Research on Educational Sport, 9(24), 241-272.

[9] Zaman, S., Mian, A. K., & Butt, F. (2018). Attitude of young students towards sports and physical activities. Global Management Journal for Academic & Corporate Studies, 8(1), 33-42.

[10] Huang, X., Yan, Z., Ma, Y., & Liu, H. (2023). The influence of different levels of physical activity and sports performance on the accuracy of dynamic lower limbs balance assessment among Chinese physical education college students. Frontiers in Physiology, 14, 1184340.

[11] Fradejas Medrano, E., & Espada Mateos, M. (2018). How do psychological characteristics influence the sports performance of men and women? A study in school sports. Journal of Human Sport & Exercise, 13(4), 858-872.

[12] Vitale, J. A., & Weydahl, A. (2017). Chronotype, physical activity, and sport performance: a systematic review. Sports medicine, 47, 1859-1868.

[13] Demetriou, Y., Bachner, J., Reimers, A. K., & Göhner, W. (2018). Effects of a sports-oriented primary school on students' physical literacy and cognitive performance. Journal of Functional Morphology and Kinesiology, 3(3), 37.

[14] Xuanchen, Z. (2022). How Does School Physical Education Stimulate Students' Interest In Sports. J Adv Sport Phys Edu, 5(4), 68-72.

[15] Ganciu, O. M. (2020). Ways to Stimulate Students' Participation in Physical Education and Sports Activities. Bulletin of the Transilvania University of Braşov. Series IX: Sciences of Human Kinetics, 127-132.

[16] Long, M. (2019). Key Factors of Students' Participating in Ecology Sports Motivation. Ekoloji Dergisi, (107).

[17] Samoling, A., Sutoro, E. S., & Mandosir, Y. M. (2022). Survey of sports science students' interest in petanque sports in 2021. International Journal of Physical Education, Sports and Health, 9(6), 34-38.

[18] Yüce, A., Balcı, V., & Katırcı, H. (2019). Evaluation of sports education at higher education level in the scope of total quality management. Spor Bilimleri Araştırmaları Dergisi, 4(2), 167-180.

[19] Seyed Jalaleddin Mousavirad,Gerald Schaefer,Khosro Rezaee,Diego Oliva,Davood Zabihzadeh,Ripon K. Chakrabortty... & Mehdi Pedram. (2024). A novel metaheuristic population algorithm for optimising the connection weights of neural networks. Evolving Systems(1),18-18.

[20] M. A. Attia,M. Arafa,E. A. Sallam & M. M. Fahmy. (2019). An Enhanced Differential Evolution Algorithm with Multi-mutation Strategies and Self-adapting Control Parameters. International Journal of Intelligent Systems and Applications(IJISA)(4),26-38.

[21] Qingzhen Shang,Jinfu Yang & Jiaqi Ma. (2024). Multi-branch evolutionary generative adversarial networks based on covariance crossover operators. Knowledge-Based Systems112527-112527.

[22] Yu Chuan,Zheng Shijie & Zhao Xie. (2024). A novel version of hierarchical genetic algorithm and its application for hyperparameters optimization in CNN models for structural delamination identification. Journal of the Brazilian Society of Mechanical Sciences and Engineering(8),

[23] Abdallah Azzouz,Billel Bengherbia,Patrice Wira,Nail Alaoui,Abdelkerim Souahlia,Mohamed Maazouz & Hamza Hentabeli. (2024). An efficient ECG signals denoising technique based on the combination of particle swarm optimisation and wavelet transform. Heliyon(5),e26171-.