International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6001-6011

https://doi.org/10.70517/ijhsa463516

Research on clustering model construction of youth sports training data in the Internet of Things environment

Xuemin Han[1], Peng Guo[2,*], Ziqi Deng[1], Xu Han[3] and Hong Wang[1]
[1] Hainan University, Danzhou, Hainan, 573717, China
[2] Cangzhou Transport University, Huanghua, Hebei, 061199, China
[3] School of International Education, Henan University, Zhengzhou, Henan, 450000, China
Corresponding authors: (e-mail: hainanxuemin@126.com).

Abstract Youth sports training is crucial for physical development, and Internet of Things (IoT) technology can realize scientific training, but it faces challenges such as high-dimensional data, noise interference, and unreasonable training intensity, so it is necessary to explore the data fusion model with higher adaptability to improve the training effect. This study proposes an adaptive bat algorithm optimized fuzzy clustering algorithm model for the characteristics of youth sports training data in the IoT environment. The model effectively avoids the problem that the traditional fuzzy C-mean clustering algorithm is prone to fall into local optimization by improving the velocity update formula in the bat algorithm and introducing the inertia weight coefficient adjustment mechanism based on the distribution entropy and average bit distance. Through the validation on the Iris and Wine datasets of UCI database, the results show that the clustering correct rate of this algorithm on the Iris dataset reaches 96.24%, which is 6.87% and 3.57% higher than that of the traditional FCM algorithm and GAKFCM algorithm, respectively, and that the correct rate on the Wine dataset reaches 94.76%, which is also better than that of the other two algorithms. Applying this algorithm to the analysis of 35,659 adolescents' exercise behavior data, the algorithm successfully classified them into five class clusters and identified that 34.47% of the adolescents had regular exercise habits, 19.92% belonged to the category of frequent exercise but very seldom attendance, and only 10.02% hardly exercised. In the comparison of evaluation metrics, the proposed algorithm reaches 75.46%, which is significantly higher than 63.36% for K-Means and 67.23% for K-Means++. The study shows that the fuzzy clustering model optimized by the proposed adaptive bat algorithm can effectively deal with the complex data of youth sports training in the IoT environment, providing a reliable tool for data mining and personalized training program development.

Index Terms Internet of Things environment, youth sports training, fuzzy C-mean clustering, adaptive bat algorithm, data clustering, sports behavior analysis

I. Introduction

Youth sports training is a very important physical exercise activity that can promote the physical development of children and adolescents, help to improve physical fitness and agility, enhance physical endurance, release stress, relieve anxiety and tension, improve the immune system, reduce the risk of obesity and metabolic diseases, and realize the all-around healthy development of body and mind [1]-[5]. It should be noted that adolescents are in a special developmental stage, their bone growth and other age differences, the training program needs to be dynamically adjusted, and excessive training is likely to lead to irreversible injuries, which is not conducive to the long-term development of adolescents [6]-[8]. The status of traditional experience-based training methods is gradually declining. And the commonality of science and technology-assisted sports training is the new era background of the current development of youth sports, has been the youth sports training as one of the high concerns in the field of competitive sports, the reality of its special homogeneity early and the sensitive period of the physical development of the contradictions need to be solved [9]-[11]. Among them, the Internet of Things (IoT) technology, as a hot direction in the research branch of sports training, has been emphasized by more and more sports practitioners in terms of its empowering sports training [12], [13]. In the IoT environment, youth sports training data, in addition to the dynamic nature, also presents high-dimensional data characteristics, serious noise interference, data bias due to unreasonable time allocation of training intensity, and differences in different training data capture devices, which bring challenges to the application of training data [14]-[16]. Demonstrating the need to explore the fusion model of youth sports training data with high adaptability.

Adolescent sports training is an important part of physical exercise activities and is significant in promoting the physical development of children and adolescents. Appropriate sports training can improve the physical fitness and

agility of adolescents, enhance physical endurance, release stress, relieve anxiety and tension, improve the function of the immune system, reduce the risk of obesity and metabolic diseases, and realize the all-round healthy development of body and mind. However, adolescents are in a special stage of growth and development, and there are age differences in their bone growth, so training programs need to be constantly adjusted, and over-training can easily lead to irreversible injuries, which will adversely affect the long-term development of adolescents. At present, the status of traditional experience-based training methods is declining, and technology-assisted sports training has become a new era of youth sports development. In the field of competitive sports, youth sports training has been one of the focuses of high attention, and the real problems of early homogenization of its specialties and the contradiction between the sensitive period of physical development need to be solved. Internet of Things (IoT) technology, as a hot direction of sports training research, has been emphasized by more and more sports practitioners in empowering sports training. In the IoT environment, youth sports training data are not only dynamic, but also present high-dimensional data characteristics, serious noise interference, unreasonable time allocation of training intensity leading to data bias, and differences in different training data capture devices, which bring challenges to the application of training data. Therefore, it becomes necessary to explore the fusion model of youth sports training data with high adaptability.

The traditional clustering algorithm divides the samples in too absolute a way, which is a hard division and cannot reflect the actual situation. This problem is solved by the proposal of fuzzy clustering methods, among which the fuzzy C-mean (FCM) clustering algorithm is widely used. However, the FCM algorithm has the risk of falling into a local optimum during the iteration process, which may cause the algorithm to converge in that local region and fail to obtain the global optimum if the randomly specified clustering center is near the local minimum. To overcome this problem, this study proposes an adaptive bat algorithm optimized fuzzy clustering model. Firstly, on the basis of the classical bat algorithm, the speed update formula is improved by considering the joint searching influence of the algorithm's optimal solution and the local optimal solution on the possible global optimal solution; secondly, inertia weight coefficients adjusted based on the distribution entropy and the average bit distance are introduced to regulate the algorithm's optimal searching ability; finally, the improved adaptive bat algorithm is applied to the fuzzy clustering, and a new optimization model is formed. The model searches the local optimal solution through the adaptive bat algorithm and applies it to the fuzzy mean algorithm, which effectively overcomes the problems that the traditional method is greatly affected by the initial center point, easy to fall into the local optimal solution and unable to obtain the optimal clustering in the face of high-dimensional data. In this study, experimental validation was carried out on the standard dataset and actual youth sports behavior data respectively, and the effectiveness and superiority of the proposed model was verified through comparative analysis with other algorithms, which provides a new method and idea for mining and analyzing youth sports training data in the environment of Internet of Things.

II. Clustering model design for youth sports training data

II. A.Fuzzy C-mean clustering algorithm

The traditional clustering algorithm is too absolute in the way of division of samples, is a hard division, can not reflect the actual situation in life. The proposal of fuzzy clustering method makes this problem solved, so that the cluster analysis can be better combined with the actual situation to solve more complex practical problems, a very typical fuzzy clustering method [17].

FCM clustering algorithm can be regarded as an optimization problem to minimize the objective function, which is widely used in various fields because of its easy operation and few parameters, and its basic idea is to convert the clustering problem into a mathematical problem, and then continuously loop iterations over and over again to update the affiliation matrix and the clustering centers iteratively, so as to allow the objective function to reach the minimum value [18]. The basic structure is as follows: let the sample dataset be $X=\{x_1,x_2,\cdots,x_n\}$, where the element $x_i(i=1,2,\cdots,n)$ has k attributes, $x_i=(x_{i1},x_{i2},\cdots,x_{ik})$, and let x_i be called the eigenvector, and c denote the number of clustering The number of categories to be delineated.

The objective function of the FCM algorithm is shown in Equation (1):

$$J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m(d_{ij})^2 \tag{1}$$

The constraints are:

$$0\le u_{ij}\le 1,1\le j\le n,1\le i\le c \tag{2}$$

$$\sum_{i=1}^{c}u_{ij}=1,1\le j\le n \tag{3}$$

$$\sum_{j=1}^{n}(u_{ij}) > 0, 1 \le i \le c \tag{4}$$

where $U=[u_{ij}]$ denotes the affiliation matrix of order $c \times n$, indicating the affiliation of each sample with the center of each class of clustering in the same principal set. u_{ij} is the degree of affiliation of the j th sample within the i th class about this class. $V=[v_i]$ represents the clustering center matrix of order $c \times k$. v_i represents the center of the $i(i=1,2,\cdots,c)$ th class. $m \in [1,+\infty)$ is the fuzzy weighting index, and the general value of m is 2. $d_{ij}=\left\|x_j - v_i\right\|$ represents the distance of the sample x_j from the center of the $i(i=1,2,\cdots,c)$ th class v_i.

The FCM clustering algorithm is the process of minimizing the objective function solution when clustering is continuously performed, and during the solution process, the affiliation values between the samples and each clustering center are independent of each other, and the solution of the objective function for minimizing the FCM is shown in Equation (5):

$$\min\left\{J_m(U,V)\right\} = \sum_{j=1}^{n}\min\left\{\sum_{i=1}^{c}(u_{ij})^m d_{ij}^2\right\} \tag{5}$$

Updating the fuzzy center affiliation u_{ij} as well as the clustering center v_i is iterated using the Lagrange multiplier method as shown in Eqs. (6) and (7):

$$u_{ij} = \left(\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{\frac{2}{m-1}}\right)^{-1} \tag{6}$$

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m} \tag{7}$$

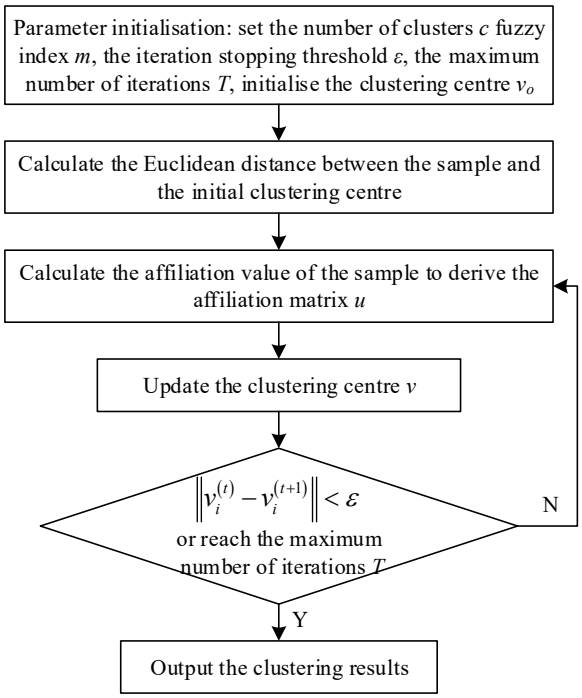

Figure 1: FCM algorithm process

Iteratively update u_{ij}, v_i until the objective function reaches convergence.

The steps of FCM algorithm are:

Step 1: Initialize the algorithm parameters. Initialize the number of categories in the clusters c, the number of samples n, the fuzzy weighting index m, the maximum iteration threshold ε, the maximum number of iterations T, and the number of iterations t, and randomly select the cluster center v_0 when $t=0$.

Step 2: Calculate the Euclidean distance $d_{ij}^{(t)}$ between the sample and the original clustering center.

Step 3: Based on the Euclidean distance $d_{ij}^{(t)}$, update the affiliation value $u_{ij}^{(t)}$ according to Eq.

Step 4: Update the clustering center $v_i^{(t+1)}$ based on the affiliation matrix $u_{ij}^{(t)}$ according to Eq.

Step 5: Calculate the distance between two similar iterations of clustering centers, if $\left\| v_i^{(t)} - v_i^{(t+1)} \right\| < \varepsilon$ or exceeds the maximum number of iterations T, the iteration stops. Otherwise, $t = t+1$ and go to Step3 until the condition is met. The flow of the FCM algorithm is shown in Figure 1.

The stopping condition of the FCM clustering algorithm is: the distance between the cluster centers of two adjacent iterations is smaller than a set threshold, which represents that the cluster centers have not changed significantly, and this process is regarded as a convergence process, or when the number of iterations reaches the pre-set maximum number of iterations. In the iterative process, the solution goal is to obtain the minimum value of the objective function, but the objective function There is a risk that the objective function falls into the local optimum in the iterative process, if in the iteration, the randomly specified clustering center is in the vicinity of the local minimum, it may lead to the convergence of the algorithm in the local region, which will have an impact on the entire clustering process, and can not obtain the global optimal effect, resulting in poor clustering results. To address this problem, the performance of the traditional FCM clustering algorithm is improved by the improved bat algorithm, so that it is more targeted when selecting the initial clustering center, and more excellent clustering results are obtained.

II. B.Bat Algorithm

II. B. 1) Classical Bat Algorithm

The Bat Algorithm (BA) is an optimization algorithm based on the localization behavior of bats. The frequency, velocity and wavelength of the current solution of the algorithm are updated iteratively until the optimal solution is found [19]. The method of updating the frequency, velocity and wavelength of the bat population is shown in Eqs. (8) to (10).

$$v_i^{t+1} = v_i^t + (x_i^t - x_*) f_i \tag{8}$$

$$f_i = f_{\min} + (f_{\max} + f_{\min})\beta \tag{9}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{10}$$

where the frequency of the i th bat is denoted as f_i, the global best position is x_*, and v_i^t and x_i^t represent the velocity and position of the i th bat, respectively.

If a bat chooses one of the current solutions as the best solution, each bat generates a new position according to Equation (11).

$$x_{new} = x_{old} + \varepsilon A_i^{(t)} \tag{11}$$

where $\varepsilon \in [-1,1]$, loudness A_i^t and firing rate r_i are updated by Eq:

$$\begin{cases} A_i^{t+1} = \alpha A_i^t \\ r_i^{t+1} = r / [1 - \exp(\gamma t)] \end{cases} \tag{12}$$

where $0 < \alpha < 1$ and $\gamma > 0$.

II. B. 2) Adaptive bat algorithm

Considering the joint search effect of the algorithmic optimal solution x_* and the local optimal solution x_r on the global optimal solution x_f that may exist in the neighborhood of the local optimal solution x_r, the speeds in Eq. are updated as in Eq. (13):

$$v_i^{t+1} = w(t)v_i^t + \lambda_1(x_i^t + x_*)f_i + \lambda_2(x_i^t + x_*)f_i \tag{13}$$

Inertia weighting factors based on distributional entropy and mean locus adjustment were incorporated into the algorithm evolution to regulate the algorithm's ability to find optimality. Mean site distance is often used to express the level of distributional dispersion among individuals in a population, however, the use of mean site distance alone to describe the diversity of a population is limiting. For this reason, the concept of distributional entropy is introduced into the algorithm, which represents the distribution of individuals in the population across search intervals. With the combined effect of average bit distance and distribution entropy, the algorithm can achieve the purpose of avoiding falling into local minima and obtaining the global optimum.

The current bat population solution space R_t^n is divided into Q equal regions, each with area S. The number of bats contained in each region is $N_1, N_2, \cdots, N_n$, and the probability that a bat occurs in the k th region is $q_k = \frac{N_k}{S}$, $k = 1, 2, \cdots, Q$, and Eqn. (14) defines the entropy of the distribution of the population.

$$E(t) = -\sum_{k=1}^{Q} q_k \ln q_k \tag{14}$$

If the bats are assumed to be uniformly distributed and the total number of bats is N, then the probability of each bat appearing in the k th region is equal and $q = \frac{N}{SQ}$, then Equation (15) defines the uniformly distributed entropy of the population.

$$\bar{E}(t) = -\sum_{k=1}^{Q} q \ln q \tag{15}$$

Using $\bar{E}(t) = \frac{E(t)_{\max}}{2}$ as the distributional entropy threshold for individuals of the population according to the definitions of population distributional entropy and uniform distributional entropy in Eqs. The inertia weights are updated by applying Eq. (16) to the bat algorithm computed in any iteration.

$$\begin{cases} E(t) > \bar{E}(t), w = -(w_{start} - w_{end})\lambda^2 + w_{start} \\ E(t) < \bar{E}(t), w = -(w_{start} - w_{end})\lambda^2 + (w_{end} - w_{start})(2\lambda)^2 + w_{start} \end{cases} \tag{16}$$

where $\lambda = \frac{t}{t_{\max}}$.

Let L be the maximum length of the solution space, S be the size of the population size, n be the number of solution spaces, the value of the d th dimensional coordinate of the i th bat is denoted as P_{id}, and the mean value of the d th dimensional coordinates of all the bats is denoted as $\bar{P}_d$, and the mean locus distance is as shown in Eq. (17).

$$\begin{cases} D_{mean}(t) = \frac{1}{SL}\sum_{i=1}^{S}\sqrt{\sum_{d=1}^{n}(P_{id} - \bar{P}_d)^2} \\ D_{\max}(t) = \max\left|P_{id}\right| \end{cases} \tag{17}$$

Here $J = \frac{D_{\max}(t) - D_{mean}(t)}{D_{\max}(t)}$ is introduced to denote the aggregation of the current bat state, and the inertia weight coefficients adjusted according to Eq. (18).

$$\begin{cases} w(t) = ww_{start}, J \geq \alpha \\ w(t) = \frac{1}{w_{start}}, J \leq \beta \end{cases} \tag{18}$$

where $w_{start}=0.95$, $w_{end}=0.05$, $\lambda_1+\lambda_2=1$ and $\lambda_1>0$, $\lambda_2\geq 0$, t is the current number of iterations, $t_{max}=1000$, and ζ is a real number between $[0,1]$.

The basic flow of the adaptive bat algorithm is shown in Figure 2.

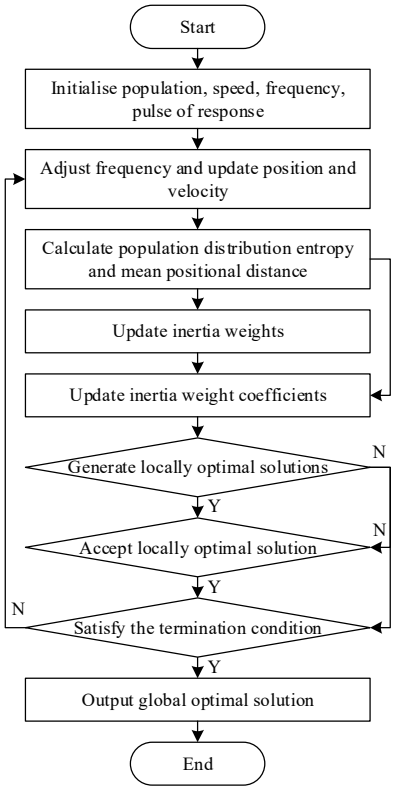

Figure 2: Adaptive bat algorithm basic process

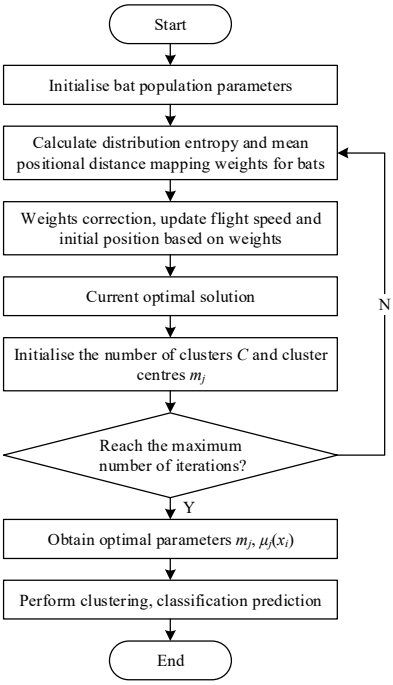

Figure 3: The fuzzy clustering algorithm process of adaptive bat algorithm optimization

II. C.Fuzzy clustering algorithm model optimized by adaptive bat algorithm

In order to overcome the problem that the fuzzy clustering algorithm optimized by evolutionary algorithm is greatly affected by the initial centroid, easy to fall into the local optimal solution and unable to get the optimal clustering in the face of high-dimensional data, the proposed adaptive bat algorithm is used to search for the local optimal solution and apply it to the fuzzy mean algorithm [20]. The flow of fuzzy clustering algorithm optimized by adaptive bat algorithm is shown in Fig. 3.

III. Experimental results and analysis

III. A. Algorithm Performance Analysis

In order to verify the performance of this paper's algorithm, the FCM algorithm, GAKFCM algorithm and this paper's algorithm are compared respectively, and 30 experiments are conducted on each of the Iris and Wine datasets in the UCI database. The clustering center comparison results of the three algorithms after running on the Iris dataset are shown in Table 1. The location of the true clustering center of the Iris dataset is: $p_1 = (5.02, 3.43, 1.56, 0.26), p_2 = (5.96, 2.79, 4.36, 1.38), p_3 = (6.36, 2.98, 5.53, 2.06)$. From the table, it can be seen that on the Iris dataset, the clustering center obtained by this paper's algorithm through the comparison experiments has the smallest error sum of squares of 0.016 with the real clustering center, which is closest to the real center.

Table 1: Comparison of clustering centers

Algorithm	FCM			GAKFCM			Ours		
Cluster center	5.01	5.92	6.73	5.11	6.57	5.89	5.06	6.01	6.59
	3.41	2.74	3.02	3.38	3	2.77	3.4	2.79	3.02
	1.45	4.45	5.61	1.51	5.38	4.28	1.44	4.28	5.56
	0.24	1.42	2	0.27	1.97	1.38	0.24	1.41	1.99
Error sum of squares	0.077			0.023			0.016		

In order to visualize the clustering effect, the real distribution results of the first two-dimensional sample points of the Iris dataset are selected, and the real distribution of the first two-dimensional sample points of Iris is shown in Fig. 4. The clustering results of the FCM algorithm, the GAKFCM algorithm and the algorithm of this paper are shown in Figs. 5~7, respectively. It can be seen that compared with FCM algorithm and GAKFCM algorithm, the clustering results of this paper's algorithm are better and closer to the actual center.

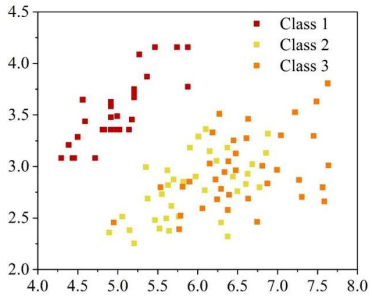

Figure 4: The real distribution of the front two dimensional sample points of iris

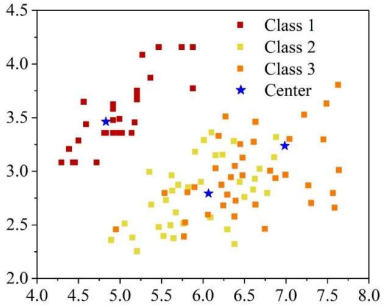

Figure 5: FCM algorithm clustering results

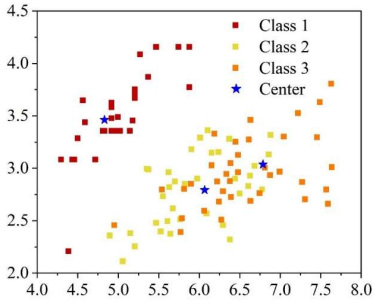

Figure 6: GAKFCM algorithm clustering results

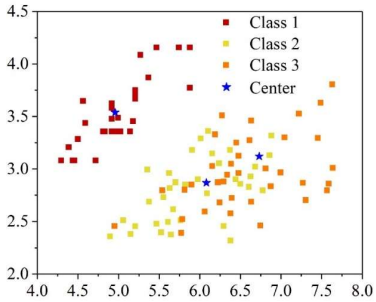

Figure 7: Clustering results of this algorithm

In order to further verify the clustering quality of this paper's algorithm, the three algorithms are compared experimentally in terms of correct rate and running time, and the comparison results of the three algorithms in terms of clustering correct rate and running time are shown in Table 2. As can be seen from the table, on the Iris dataset, the correct rate of this paper's algorithm is 96.24%, while the correct rate of the traditional FCM algorithm is 89.37%, which is lower than that of this paper's algorithm. The correct rate of the GAKFCM algorithm is 92.67%, which is higher than that of the traditional FCM algorithm but lower than that of this paper's algorithm. On the Wine dataset, the highest correct rate of this paper's algorithm is 94.76%, while the correct rates of traditional FCM algorithm and GAKFCM algorithm are 68.53% and 81.82% respectively, which are lower than this paper's algorithm correct rate. Secondly, comparing the running time of the algorithms, the running time consumption on Iris and Wine datasets is slightly higher than that of the FCM algorithm due to the inclusion of the optimization seeking process of the improved bat algorithm in this paper's algorithm, but the running time consumption of this paper's algorithm is less compared with the GAKFCM algorithm. In summary, this paper's algorithm achieves high clustering accuracy under reasonable time consumption conditions.

Table 2: Comparison of clustering accuracy and running time of the three algorithms

	FCM		GAKFCM		Ours	
	Accuracy (%)	Running time (s)	Accuracy (%)	Running time (s)	Accuracy (%)	Running time (s)
Iris	89.37	12	92.67	30	96.24	22
Wine	68.53	25	81.82	36	94.76	29

III. B. Cluster Analysis of Youth Sports Behavior

In this section, a clustering algorithm is used to cluster 35,659 adolescents in a school, and the results of the clustering of the adolescent sports behavior indicator system are shown in Table 3. The results of the evaluation indicators of this paper's algorithm on adolescent sports behavior data are shown in Figure 8. As can be seen from the figure, when the density threshold is 1618, the evaluation indicators are optimal and the clustering effect is the best. Therefore, the youth sports behavior data is divided into five classes, and calculate the number of youth in each class cluster, the clustering of the youth classification cluster heart for comparative analysis, compare the average value of the indicators in each class cluster with the average value of each indicator of the overall youth, use “1” to indicate that the indicators in the class cluster are greater than or equal to the overall youth indicators The use of “1” means that the indicators in the cluster are greater than or equal to the average value of the indicators

of the overall adolescents, and the use of “0” means that the indicators in the cluster are smaller than the average value of the indicators of the overall adolescents.

As can be seen from the table, the algorithm of this paper divides the sports data of 35659 adolescents into five classes, of which the largest proportion of adolescents is Class Cluster 5, with 12,292 adolescents, accounting for 34.47% of the total number of athletic adolescents, and the heart of this class of clusters is greater than the average value of the overall adolescents' sports behavior indexes for the three indexes, and a comprehensive observation of the indexes of the heart of the clusters can be found that the adolescents in the class of clusters are belong to the group of adolescents who exercise regularly and take attendance frequently and have better exercise habits.

The least number of adolescents is class cluster 2 with only 3573 adolescents, accounting for 10.02% of the total number of athletic adolescents, and the observation of the cluster heart's indicators of attendance, running kilometers, and overall performance reveals that all three of these indicators are the lowest among all the class clusters. Therefore, it can be inferred that very few adolescents in this cluster are in the habit of exercising.

There are 7103 adolescents in Class Cluster 1, accounting for 19.92% of the total number of athletic adolescents, and the average number of attendance of cluster hearts in this cluster is only 0.57, but their running kilometers and overall performance indicators are the highest among all class clusters. Therefore, although the youth in this cluster rarely take attendance, they belong to the group of youth who exercise frequently, and youth managers should urge the youth in this cluster to participate in attendance regularly.

There are 6044 adolescents in Cluster 3, accounting for 16.95% of the total number of athletic adolescents. Observation of the indicators in Cluster 3 reveals that the heart of this cluster is in the middle of all the clusters in terms of the number of kilometers run and the overall performance indicators, but the number of times of attendance in this cluster is on the low side. Therefore, the youths in this cluster belong to the group of moderately active youths, and the youth managers also need to urge the youths in this cluster to participate in the attendance regularly.

There were 6647 youths in cluster 4, which accounted for 18.64% of the total number of athletic youths, and the values of the kilometers run and overall performance indicators of this cluster were low, but when observing the attendance indicator, the value of the attendance indicator of this cluster was the highest of all the clusters in this cluster. Therefore, it can be inferred that the youth in this cluster belongs to the group of youths who exercise occasionally, and there may be a situation of participating in attendance but not in exercise, and youth managers should pay attention to the exercise of youths in this cluster.

In addition, it can be found through the table that most of the adolescents out of 35659 adolescents have good exercise habits. About 34.47% or so of the adolescents participated in sports on a regular basis and had more regular sports habits. As a result, there are a majority of adolescents who are in the moderate or higher level of exercise and have a higher average exercise performance. However, there were a small number of adolescents who did not exercise regularly.

Table 3: Adolescent movement behavior index system clustering results

Cluster number	Student ratio	Times	Mileage (KM)	Comprehensive results (points)	Comparison result	Cluster label
1	19.92%	0.57	116.52	75.81	011	Frequent motion
2	10.02%	1.42	11.24	8.68	000	Minimal motion
3	16.95%	1.28	70.29	48.03	011	Medium motion
4	18.64%	12.13	35.82	32.68	100	Occasional motion
5	34.47%	11.93	72.16	58.13	111	Regular motion
Average		5.43	61.21	44.67		

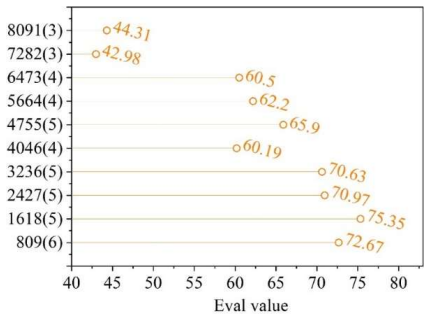

Figure 8: The results of this algorithm are evaluated in the data of adolescent motion behavior

The results of the comparison of the evaluation metrics of this paper's algorithm with the K-Means algorithm and the K-Means++ algorithm are shown in Table 4. The evaluation index value of this paper's algorithm is the lowest when the density threshold is 9726 and the number of class clusters is 3. At this time, the inter-cluster dissimilarity is smaller and the intra-cluster cohesion is poorer. The evaluation index eval of K-Means and K-Means++ clustering algorithms under the same conditions is lower than that of this paper's improved K-Means clustering algorithm. In summary, this paper's algorithm is more advantageous than K-Means and K-Means++ algorithms for clustering youth sports behavior data under the same conditions, and the method improves the applicability of K-Means and K-Means++ algorithms for clustering youth sports behavior data. In this paper, the algorithm is tested on different behavioral data of adolescents, and it is found that the operation results of the algorithm basically match the data distribution of the target dataset. Therefore, in this paper, the algorithm is applied to adolescent behavioral data mining so as to improve the efficiency of data mining to a certain extent.

Table 4: Evaluation index comparison results

Algorithm name	Eval index (%)
K-Means	63.36
K-Means++	67.23
Ours	75.46

IV. Conclusion

The clustering of youth sports training data in the Internet of Things environment is a study of great significance. The fuzzy clustering model optimized by the adaptive bat algorithm shows excellent performance through experimental validation. In the standard UCI dataset test, the sum of squares of the errors between the clustering centers and the actual centers of this algorithm on the Iris dataset is only 0.016, which is significantly lower than the 0.077 of the FCM algorithm and the 0.023 of the GAKFCM algorithm; and the clustering correctness rate reaches 96.24%, which exceeds the FCM algorithm by 6.87 percentage points. Although the running time of this algorithm is 22 seconds, which is 10 seconds more than the FCM algorithm, it is 8 seconds less than the GAKFCM algorithm, which achieves a higher accuracy rate with reasonable time consumption.

Applying this model to the analysis of 35,659 adolescents' exercise behavior data, the adolescents were successfully classified into five categories: frequent exerciser (19.92%), very infrequent exerciser (10.02%), moderate exerciser (16.95%), occasional exerciser (18.64%), and regular exerciser (34.47%). Among them, the regular exercise type adolescents were above average in the three indicators of the number of attendance, kilometers run, and overall performance, and performed the best; while the rarely exercise type adolescents were the lowest in all three indicators, which need to be focused on.

A comparison of the evaluation metrics reveals that the present algorithm achieves 75.46% of the metric values, which is significantly higher than 63.36% for the K-Means algorithm and 67.23% for the K-Means++ algorithm. This proves that the fuzzy clustering model optimized by the adaptive bat algorithm has higher applicability and accuracy in adolescent sports behavior data mining. Based on the clustering results, a differentiated exercise training program can be developed for adolescents, thus improving the training effect and promoting the healthy development of adolescents.

Funding

Funded by the Natural Science Foundation of Hainan Province (Project No.: 320MS013).

References

[1] Marin, A., Stefanica, V., & Rosculet, I. (2023). Enhancing Physical Fitness and Promoting Healthy Lifestyles in Junior Tennis Players: Evaluating the Influence of “Plyospecific” Training on Youth Agility. Sustainability, 15(13), 9925.

[2] Bouamra, M., Zouhal, H., Ratel, S., Makhlouf, I., Bezrati, I., Chtara, M., ... & Chaouachi, A. (2022). Concurrent training promotes greater gains on body composition and components of physical fitness than single-mode training (endurance or resistance) in youth with obesity. Frontiers in physiology, 13, 869063.

[3] Zou, C., Hu, X., & Huang, L. (2022). INFLUENCE MECHANISM OF LEISURE SPORTS TRAINING ON RELIEVING TEENAGERS'NEGATIVE EMOTIONAL PRESSURE. Psychiatria Danubina, 34(suppl 2), 310-310.

[4] Qomariah, N., Biworo, A., Ahdiya, W., Ridhoni, M. H., & Amaliea, N. N. R. (2023). CRP as a mediator for the innate immunity system and blood glucose levels in football trained adolescents. Revista Latinoamericana de Hipertension, 18(9), 417-422.

[5] Branco, B. H. M., Mariano, I. R., De Oliveira, L. P., Bertolini, S. M. M. G., De Oliveira, F. M., Araújo, C. G. A., & Adamo, K. (2021). Sports and functional training improve a subset of obesity-related health parameters in adolescents: A randomized controlled trial. Frontiers in Psychology, 11, 589554.

[6] Cheng, X. (2022). Effects of sport on skeletal development in adolescents. Revista Brasileira de Medicina do Esporte, 28(6), 679-681.

[7] Brenner, J. S., Watson, A., Brooks, M. A., Carl, R. L., Briskin, S. M., Canty, G., ... & Emanuel, A. (2024). Overuse injuries, overtraining, and burnout in young athletes. Pediatrics, 153(2).
[8] Heilmann, F., Memmert, D., Weinberg, H., & Lautenbach, F. (2023). The relationship between executive functions and sports experience, relative age effect, as well as physical maturity in youth soccer players of different ages. International journal of sport and exercise psychology, 21(2), 271-289.
[9] Malina, R. M., Cumming, S. P., Rogol, A. D., Coelho-e-Silva, M. J., Figueiredo, A. J., Konarski, J. M., & Kozieł, S. M. (2019). Bio-banding in youth sports: background, concept, and application. Sports Medicine, 49(11), 1671-1685.
[10] Yao, J., & Li, Y. (2022). Youth sports special skills' training and evaluation system based on machine learning. Mobile Information Systems, 2022(1), 6082280.
[11] Van Hooren, B., & Croix, M. D. S. (2020). Sensitive periods to train general motor abilities in children and adolescents: do they exist? A critical appraisal. Strength & Conditioning Journal, 42(6), 7-14.
[12] Lu, S., Zhang, X., Wang, J., Wang, Y., Fan, M., & Zhou, Y. (2021). An IoT‐Based Motion Tracking System for Next‐Generation Foot‐Related Sports Training and Talent Selection. Journal of Healthcare Engineering, 2021(1), 9958256.
[13] Di Palma, D., Cusano, P., Russo, C., & Ascione, A. (2019). New technologies in sport through the internet of things systems. Research Journal of Humanities and Cultural Studies, 5(1), 16-22.
[14] Wang, Z., & Gao, Z. (2021). Analysis of real‐time heartbeat monitoring using wearable device Internet of Things system in sports environment. Computational Intelligence, 37(3), 1080-1097.
[15] Yao, W., & Zhihai, Z. (2022). Design of sports training data monitoring system based on wireless internet of things. Mobile information systems, 2022(1), 4162088.
[16] Wu, Q., Tang, P., & Yang, M. (2020). Data processing platform design and algorithm research of wearable sports physiological parameters detection based on medical internet of things. Measurement, 165, 108172.
[17] Song Liu,Di Liu & Meilong Le. (2025). Multi-UAV Delivery Path Optimization Based on Fuzzy C-Means Clustering Algorithm Based on Annealing Genetic Algorithm and Improved Hopfield Neural Network. World Electric Vehicle Journal,16(3),157-157.
[18] Peng Li,Tianqi Chen,Yan Liu,Meifeng Cai,Liang Sun,Peitao Wang... & Xuepeng Zhang. (2025). Automatic Identification of Rock Discontinuity Sets by a Fuzzy C-Means Clustering Method Based on Artificial Bee Colony Algorithm. Applied Sciences,15(3),1497-1497.
[19] Mohamed AmineLaamari & NadjetKamel. (2025). A New Multi‐Objective Binary Bat Algorithm for Feature Selection in Intrusion Detection Systems. Concurrency and Computation: Practice and Experience,37(4-5),e70000-e70000.
[20] Lijun Liu,Chang Yin,Yonghui Su,Yinghai Lin & Ying Lei. (2025). Optimization of Covariance Matrices of Kalman Filter with Unknown Input Using Modified Directional Bat Algorithm.Buildings,15(2),196-196.