Chen, Z.; Li, S. (2022) Applications of continual construction of sports training system based on data mining technology. Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte vol. 22 (85) pp. 335-353
DOI: https://doi.org/10.15366/rimcafd2022.85.021

ORIGINAL

Applications of continual construction of sports training system based on data mining technology

Zhengtian Chen[1], Shuting Li[1*]
[1]School of Physical Education and Health Science, GuangXi MinZu University, Nanning, 530006, China
Email: chenzhengtian111@163.com

UNESCO Code / UNESCO Code:
Council of Europe classification / Council of Europe classification:

Recibido 26 de febrero de 2021 **Received** February 26, 2021
Aceptado 26 de febrero de 2022 **Accepted** February 26, 2022

Abstract

Objective: In competitive sports, one of the most effective ways to improve athletes' competitive level is intensive training. The current training pro-cess is mainly based on the coach’s existing experience and the athlete’s personal physical condition to formulate a training plan, but the formulation of this training plan obviously lacks scientific basis. This paper combines data mining technology and image recognition tech-nology to construct a sports training system. In order to better deal with the over-dispersed and multi-peak count data, this paper proposes a mixed integer generalized autoregressive conditional heteroscedasticity model. The model uses virtual reality technology to construct train-ing scenes to improve athletes' training immersion, and can use digital technology to identify sports training processes, correct actions, and improve sports training effects. Finally, this article combines with experimental research to verify the training effect of the system con-structed in this article.

Keywords. Data mining, sports training, system construction, intelligent system

1. INTRODUCTION

With the rapid development of modern technology, the development of various disciplines has gradually merged together. In order to effectively improve the athlete's performance, the support system combined with computer technology can make the athlete's training more systematic, scientific and standardized (Abanazir, 2019). Data mining technology is a key development field in my country at this stage, and it has a wide range of applications in many industries (Barbosa, 2018). Moreover, the analysis of the sports training model based on data mining technology can provide a corresponding decision-making system and better improve the traditional sports training model.

Physical fitness training has become an important research hotspot in implementing the nationwide fitness program and promoting the development of competitive sports. With the advancement of science and technology and the integration of multiple disciplines, the scientific level of physical fitness training has rapidly improved. In addition, sports academies, as an important way and an important base for cultivating all kinds of sports professionals, provide talent support and scientific and technological support for the research and practice of physical fitness training in China. Specialized physical fitness is an important research field of physical fitness training. With the advent of the era of big data, the relationship between data information and training decisions in the field of physical training is getting closer. Moreover, it is gradually changing from the development mode of "intuition and experience-driven decision-making" and "data-centric decision-making" to "data-driven decision-making" (Catro-Sánchez, Zurita-Ortega, & Chacón-Cuberos, 2019). In addition, data collection, screening, analysis, and management for the evaluation, monitoring, program implementation and modification of special physical training will provide effective support for training decision-making, and it can accelerate the scientific level of physical fitness training and sports training practice in my country, and it has a wide range of application value in the field of national fitness. At the same time, data mining and the establishment of a normal cloud model provide data support and application scenarios for the combination of cloud computing, AI and sports.

Physical fitness in competitive sports training is generally divided into basic fitness and specific fitness due to different training tasks and goals. Basic physical stamina is the foundation and support of special physical stamina. It is the athletic ability displayed when completing non-special sports. It is also the foundation and guarantee of the special physical stamina level. Special physical fitness is the ability required for a specific sport to effectively complete the special skills and tactics and improve the athletic ability. The improvement of specific physical fitness not only helps to withstand more load in training and competition, but also can make up for the lack of technology. With the scientific and refined development of competitive sports training, appropriate physical training based on the laws and characteristics of different sports has become the current trend of physical training in competitive sports. The systematic training system, concepts and methods with physical training as the entry point, as well as core strength training, functional training, digital physical training and other means, instruments are used in the practice of high-level athletes’ physical training, and scientific and training through evaluation methods The systematization of the means enables the effective implementation of targeted and personalized training programs. At the same time, the orientation of physical training content, the quantification of training load, and the individualization of the implementation of the training process, which are closely integrated with special sports, have promoted the improvement of the theoretical system, training concepts and methods of special physical training. In particular, the practice trend of physical training that drives decision-making with data presentation has accelerated the scientificization of physical training not only in the establishment of a physical training data management platform, but also in the realization of athletic ability status evaluation, real-time monitoring of training quality, key data collection, and training information analysis. Degree.

According to the current training process, in the process of determining training content, physical training must be based on strength, speed, endurance, etc., but there is a lack of tactical analysis in some targeted training. How to collect the actual combat data of the athletes and formulate their adjustment countermeasures afterwards is particularly important. In the context of database, data mining, and computer technology advancement, materials on the personal growth of athletes are collected. After sorting and analyzing, a variety of data are obtained. Through the processing and application of these data, it provides a reference basis for subsequent analysis of athletes.

2. Related work

Scientific sports training requires health education and certain knowledge of sports, nutrition, hygiene, physiology, and psychology, and it needs to carry out physical training activities that can enhance physical fitness based on its own health status and exercise prescription guidance (Cristiani, Bressan, Longarela Pérez, Galatti, & Reverdito, 2017). The literature (Du Plessis & Berteanu, 2020) pointed out that health should be the indicator to measure the improvement of people's living standards, and analyzed the viewpoint that medical treatment promotes health, and believed that scientific sports training is the effective way to promote health. In the definition of scientific sports training in the literature (Emery & Pasanen, 2019), the emphasis on medical knowledge is more prominent. It points out that scientific sports training requires the mastery of hygiene knowledge, physiological anatomy knowledge, health care knowledge, etc., as well as medical examination to determine its own health.

The definition of scientific sports training proposed in the literature (Ferguson, Carlson, & Rogers, 2019) reflects the dependence on medicine, which reflects the characteristics of society, and reflects the general view that health is promoted by medical means. The concept of scientific sports training in literature (Gerke, Babiak, Dickson, & Desbordes, 2018) highlights the purpose of promoting health, which is fundamentally different from blind sports training and unscientific sports training. Similar to this definition is Zhang Bing's definition of scientific sports training. These definitions all emphasize that scientific sports training requires certain basic knowledge, including nutritional knowledge, physiological knowledge, and health care knowledge. It is also necessary to understand one's own physical condition under medical examination, and carry out sports training activities according to one's own physical condition combined with exercise prescriptions, and all emphasize the role of scientific sports training in promoting physical health (Giulianotti & Numerato, 2018).

Literature (Gurinovich & Petrova, 2019) believes that scientific sports training is to improve the individual's resistance and adaptability, thereby overcoming the danger of insufficient exercise brought about by modern life, and being able to invest in the process of production and reproduction with a healthy body and abundant energy. The definition of scientific sports training in the literature (Hadlow, 2018) highlights the basic situation of people’s lack of exercise (Stylianou, Hogan, & Enright, 2019). The current reality of sedentary and insufficient exercise has led to the emergence of many unhealthy states. Based on this reality, the concept of scientific sports training is carried out. Defined. In addition to emphasizing the purpose of scientific sports training to promote physical health, this definition also points out that scientific sports training can promote individual energetic and highlights the characteristics of scientific sports training to promote health in many aspects. Literature (Hadlow, 2018; Richmond et al., 2020) gives a relatively brief definition of scientific sports training, and believes that scientific sports training is sports training that arranges sports reasonably and avoids sports injuries. Compared with the definition of other people's scientific sports training, the definition in the literature (Happ, Schnitzer, & Peters, 2021) highlights the skills of scientific sports training, emphasizes the skills of rational arrangement of sports in scientific sports training, and points out that the sports carried out by scientific sports training must be reasonable; Emphasizes the need to avoid sports injuries. Because unscientific sports training is very easy to cause sports injuries, the occurrence of sports injuries runs counter to the purpose of sports training. The definition of scientific sports training in the literature (Happ et al., 2021) emphasizes the awareness, knowledge, understanding and habits of scientific sports training. In terms of @, pointed out that scientific sports training includes the awareness of scientific sports training, mastering the theoretical knowledge of scientific sports training and sports practice, fully understanding various specific factors affecting physical health, and achieving the goal of improving health while cultivating lifelong sports habits. . The definition in the literature (Ilies et al., 2018) emphasizes the habit of lifelong sports, which is quite different from the previous definition of scientific sports training. It is believed that scientific sports training not only has the purpose of promoting health, but also promotes the development of lifelong sports habits. Lifelong sports generally emphasizes two content. One is to emphasize that sports should run through the entire life cycle, and that sports should be an important life content that is indispensable for individuals from the beginning to the end of life; the other is to emphasize the practicality of sports, which is different for individuals. The opportunities, content, and the results achieved for participating in sports activities in different areas of life and age vary. Individual sports activities should be planned under the guidance of the goal of lifelong sports integration. The literature (Kimasi, Shojaei, & Boroumand, 2019) pointed out that people must abandon the previous view that sports are only played for a short period of time in their lives (Riley & Callahan, 2019).

Literature (Kondrukh, 2017) believes that scientific sports training is based on the basic theories and methods of sports science and the basic laws of scientific sports training, using scientific sports training methods and scientific sports training methods to achieve the purpose of improving health and strengthening physical fitness. He literature (Kondrukh, 2017) explored the relationship between scientific sports training and human health, scientific sports training and exercise prescriptions, discussed the scientific sports training methods of individuals in different body parts, different periods and different groups of people, and emphasized more on science. Sports training practice methods and methods, etc (Shulyatyev & Bulavina, 2019).

2.1 Data mining algorithm for sports training data

This paper studies the application of data mining algorithms in sports training data. The data mining process is shown in Figure 1 (Mountjoy et al., 2018).

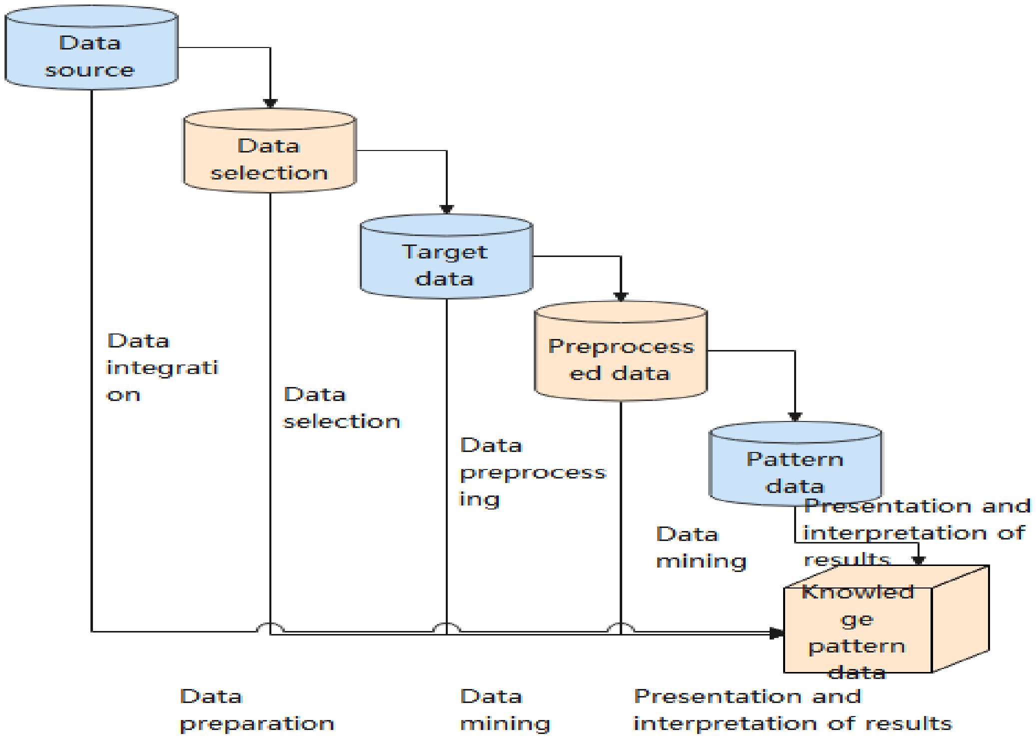


Figure 1. Process diagram of data mining

The conditional marginal distribution is Poisson distribution. The model form is as follows:

$$\begin{cases} X_t | F_{t-1}: P(\lambda_t), \\ \lambda_t = \alpha_o + \sum_{i=1}^{p} \alpha_i X_{t-1} + \sum_{i=1}^{q} \beta_j \lambda_{t-i}, \end{cases} \quad (1)$$

Among them, $\{X_t, t \in N_0\}$ is the counting process $N_0 \in \{0\} \cup N$, $P(\cdot)$ is the Poisson distribution, F_{t-1} is the σ-domain generated by $X_{t-1}, X_{t-2}, \ldots$, and $\alpha_0 > 0, \alpha_i \geq 0, \beta_j \geq 0, i = 1,2.\ldots, p, j = 1,2,\ldots, q$. This kind of model can well describe the conditional heteroscedasticity occasions of counting process with Poisson deviation. Research has found that this model can also handle the overdispersed counting process.

The so-called overdispersion means that the variance of the sequence is greater than its mean. In fact, in addition to non-negative integers, counting time series are often over-dispersed. There may be many reasons for over-dispersion, but one of the main reasons frequently mentioned in a large amount of literature is the positive correlation of the process. As far as the distribution itself is concerned, the Poisson distribution is always suitable for unconditional over-dispersion, but not for conditional over-dispersion, because the mean variance of the Poisson distribution is always equal. As we all know, the negative binomial distribution is a natural extension of the Poisson distribution, and it is easy to handle and suitable for over-dispersion, so the negative binomial INGARCH model was established. The model is defined as (Mountjoy et al., 2018):

$$\begin{cases} X_t|F_{t-1}: NB(r, p_t), \\ \frac{1-p_t}{p_t} = \lambda_t = \alpha_o + \sum_{i=1}^{p} \alpha_i X_{t-1} + \sum_{i=1}^{q} \beta_j \lambda_{t-i}, \end{cases} \quad (2)$$

Among them, $NB(\cdot,\cdot)$ represents negative binomial distribution, and other symbols are the same as model (1). Studies have shown that over-dispersed count data with very many end values can be well described by such an integer model.

In the real world, there are many time series with multi-peak marginal distribution or conditional distribution, and the mixed time series model is suitable for handling multi-peak situations. For the count sequence, there is also a multi-peak conditional distribution or marginal distribution, and the mixed marginal conditional distribution is Poisson. The model has the following form:

$$\begin{cases} X_t = \sum_{k=1}^{K} \Pi(\eta_t = k) Y_{kt} \\ Y_{kt}|F_{t-1}: P(\lambda_{kt}) \\ \lambda_{kt} = \beta_{ko} + \sum_{i=1}^{p} \beta_{ki} X_{t-i} \end{cases} \quad (3)$$

Among them, $\{X_t, t \in N_0\}$ is still the counting process, $\Pi(\cdot)$ is the indicative function, pk is the autoregressive order of the k-th mixing process, and $\eta_t \sim P(\eta_t = k) = \alpha_k, \alpha_k > 0, k = 1, \dots, K (i.i.d.)$,and there is$\alpha_1 + \cdots + \alpha_K = 1$. The definitions of F_{t-1} and $P(\cdot)$ are similar to those of model (1). In addition, for any t and j> 0, X_{t-j} and η_t are independent, and Y_{kt} and η_t are independent of F_{t-1} conditions. For model (3), the intercept term and coefficient in the third formula satisfy: $\beta_{ko} > 0, \beta_{ki} > 0, i = 1, \dots, p_k, k = 1, \dots, K$.

In order to better deal with the over-dispersed and multi-peak count data, a mixed integer generalized autoregressive conditional heteroscedasticity (MINGARCH) model is proposed. Among them, the conditional distribution of each mixing process is negative binomial. With the advantages of negative binomial distribution and mixed distribution, such mixed integer GARCH model can not only deal with multi-peak and uneven components, but also deal with excessive dispersion. Next, this article first discusses the stationarity of the model, and then conducts parameter estimation, simulation and empirical research (Pogrebnoy & Komlev, 2018).

$\{X_t\}$ is the counting process, which is said to obey a mixed negative binomial integer generalized autoregressive conditional heteroscedasticity (MNBINGARCH) model. Among them, the mixing process number and order are respectively K and $p_1,..,p_K;q_1, ...,q_K$, if it has the following form:

$$\begin{cases} X_t = \sum_{k=1}^{K} \Pi(\eta_t = k) Y_{kt} \\ Y_{kt}|F_{t-1}: NB(\lambda_{kt}) \\ \frac{1-p_{kt}}{p_{kt}} = \lambda_{kt} = \beta_{ko} + \sum_{i=1}^{p_k} \beta_{ki} X_{t-i} + \sum_{i=1}^{q_k} \alpha_{kj} \lambda_{k,t-i}, \end{cases} \quad (4)$$

Among them, K is the number of mixing processes, p_k and q_k are the autoregressive and sliding orders of the k-th mixing process respectively, $\eta_t \sim P(\eta_t = k) = \alpha_k, \alpha_k > 0, k = 1, \dots, K(i.i.d.)$and there is $\alpha_1 + \cdots + \alpha_K = 1$, $F_{t-1} = \sigma(X_s, s \le t-1), NB(r_k, p_{kt})$ represents the negative binomial distribution, and its probability distribution is:

$$p(Y_{kt} = y|F_{t-1}) = \binom{y + r_k - 1}{r_k - 1} p_{kt}^{r_k} (1 - p_{kt})^y, y = 0,1.. \quad (5)$$

In addition, for any t and j> 0, X_{t-j} and η_t are independent, and Yx and Ft-: conditions are independent. For the constant term and coefficient in the third formula of model (1), satisfy: $\beta_{ko} > 0, \beta_{ki} \ge 0, \alpha_{ki} \ge o, i = 1, \dots, p_k, k = 1, \dots, K$. The abbreviated process {Xt} obeys the model, and {Xt} is also called MNBINGARCH (K; $p_1,..,p_K;q_1,...,q_K$) process.

It can be seen from the model definition that if r_k= 1, k = l,...,K, model (4) becomes a mixed integer GARCH model based on geometric distribution. When K=1, model (4) degenerates to model (2). If the conditional distribution N B(r_k, P_{kt}) of model (4) is replaced by P(λ_{kt}), and q = 0,k = 1,...,K, model (3) is obtained. Therefore, in a sense, model (4) is a generalization of models (2) and (3). Therefore, it combines the advantages of the two models, and can better handle multi-peak, non-stationary mixing process and over-dispersed counting process (Reinhart & Wichmann, 2020).

First, we discuss the over-dispersion of Y_{kt} in each mixing process in the model. According to the conditional probability distribution of Y_{kt}, its conditional expectation and variance are calculated,

$$E(Y_{kt}|F_{t-1}) = \frac{r_k(1 - p_{kt})}{p_{kt}} = r_k \lambda_{kt}$$

$$Var(Y_{kt}|F_{t-1}) = \frac{r_k(1 - p_{kt})}{p_{kt}} = r_k \lambda_{kt}(1 + \lambda_{kt}) > r_k \lambda_{kt}$$

Further, its variance is calculated,

$$\begin{aligned} Var(Y_{kt}) &= E\big(Var(Y_{kt}|F_{t-1})\big) + Var\big(E(Y_{kt}|F_{t-1})\big) \\ &= E\big(r_k\lambda_{kt}(1 + \lambda_{kt})\big) + Var(r_k\lambda_{kt}) \\ &= r_k E(\lambda_{kt}) + r_k\big(E(\lambda_{kt})\big)^2 + (r_k + r_k{}^2) Var(\lambda_{kt}) \\ &> r_k E(\lambda_{kt}) = E(Y_{kt}) \end{aligned}$$

The above two equations show that each mixing process {Y_{kt}} in the model (4) is a conditional and unconditional over-dispersion process. Next, the expected and variance changes of the sequence {Xt} after mixing are discussed.

$$E(Y_{kt}|F_{t-1}) = \sum_{k=1}^{K} \alpha_k E(Y_{kt}|F_{t-1}) = \sum_{k=1}^{K} \alpha_k r_k \lambda_{kt}$$
$$= \sum_{k=1}^{K} \alpha_k r_k \beta_{k0} + \sum_{k=1}^{K} \alpha_k r_k \sum_{i=1}^{p_k} \beta_{ki} X_{t-i} + \sum_{k=1}^{K} \alpha_k r_k \sum_{j=1}^{q_k} \alpha_{ki} \lambda_{k,t-j}$$
$$= \textstyle\sum_{k=1}^{K} \alpha_k r_k \beta_{k0} + \sum_{i=1}^{p} (\sum_{k=1}^{K} \alpha_k r_k \beta_{ki}) + \sum_{j=1}^{q} (\sum_{k=1}^{K} \alpha_k r_k \alpha_{ki})\, \lambda_{k,t-j} \quad (6)$$

$$Var(X_{kt}|F_{t-1}) = E\left(\sum_{k=1}^{K} 1(\eta_t = k) Y_{kt}^2 | F_{t-1}\right) - \left(E(X_t|F_{t-1})\right)^2$$
$$= \sum_{k=1}^{K} \alpha_k E(Y_{kt}^2|F_{t-1}) - \left(E(X_t|F_{t-1})\right)^2$$
$$= \sum_{k=1}^{K} \alpha_k [r_k \lambda_{kt}(1+\lambda_{kt}) + (r_k \lambda_{kt})^2] - \left(E(X_t|F_{t-1})\right)^2$$
$$= \sum_{k=1}^{K} \alpha_k r_k \lambda_{kt} + \sum_{k=1}^{K} \alpha_k r_k \lambda_{kt}^2 + \sum_{k=1}^{K} \alpha_k r_k^2 \lambda_{kt}^2 - \left(\sum_{k=1}^{K} \alpha_k r_k \lambda_{kt}\right)^2$$
$$= E(X_t|F_{t-1}) + \textstyle\sum_{k=1}^{K} \alpha_k r_k \lambda_{kt}^2 + \sum_{k=1}^{K} \alpha_k \left(r_k \lambda_{kt} - E(X_t|F_{t-1})\right)^2 \quad (7)$$

It is easy to see that $Var(X_{kt}|F_{t-1}) > EX_t|F_{t-1}$, that is, the process {Xt} is also conditionally over-dispersed. Therefore, similar to the general GARCH model, one of the important features of the mixed negative binomial integer GARCH model is that it can better handle the changing conditional variance.

The over-dispersion and multi-peak conditions of the process {Xt} are discussed below. Since the conditional expectation and variance of each process of mixing in the model depend on the past value of the sequence, to a large extent the shape of the conditional distribution of the sequence after mixing also changes with time. Even

tually, the conditional distribution of the sequence is bimodal or multimodal. Therefore, for {Xt}, its conditional expectations may not be the best prediction of future values. The third term in formula (7) is non-negative, and the condition for the equality sign is $r_1\lambda_{1t} = r_2\lambda_{2t} =\ldots= r_k\lambda_{kt}$. At this time, the conditional variance is likely to be the smallest. On the contrary, when the difference between $r_1\lambda_{1t},\ldots,r_k\lambda_{kt}$ is large, the conditional variance may be large, and the conditional distribution should also be bimodal or multimodal. Therefore, the MNBINGARCH model is suitable for describing excessively bimodal or multimodal counting sequences. Further, we derive the expectation and variance of the process,

$$E(X_t) = \textstyle\sum_{k=1}^{K} \alpha_k r_k E(\lambda_{kt}) \quad (8)$$

$$Var(X_t) = E\left(Var(X_t|F_{t-1})\right) + Var\left(E(X_t|F_{t-1})\right)$$

$$= E(X_t) + \sum_{k=1}^{K} \alpha_k (r_k + r_k^2) E(\lambda_{kt}^2) - E\left(\left(E(X_t|F_{t-1})\right)^2\right) + Var\left(E(X_t|F_{t-1})\right)$$

$$= E(X_t) + \sum_{k=1}^{K} \alpha_k r_k E(\lambda_{kt}^2) + \sum_{k=1}^{K} \alpha_k r_k^2 E(\lambda_{kt}^2) - \left(E(X_t)\right)^2$$

$$= E(X_t) + \sum_{k=1}^{K} \alpha_k (r_k + r_k^2) E(\lambda_{kt}^2) - \left(\sum_{k=1}^{K} \alpha_k r_k E(\lambda_{kt})\right)^2$$

$$= E(X_t) + \sum_{k=1}^{K} \alpha_k \left(r_k E(\lambda_{kt}^2) + r_k^2 Var(\lambda_{kt})\right) + \sum_{k=1}^{K} \alpha_k \left(k E(\lambda_{kt}) - E(X_t)\right)^2 \quad (9)$$

The above formula shows that $Var(X_t) > E(X_t)$, that is, the process {Xt} is still unconditionally over-dispersed, so the MNBINGARCH model is suitable for describing the over-dispersed counting sequence. It can be seen from the definition of model (4) that the MNINGARCHI model can actually be approximated as a mixture of a finite number of NBINARCH models.

In addition, according to the definition of model (4), the conditional probability distribution of process {Xt} with respect to Ft-1 can be obtained as:

$$p(X_t = x|F_{t-1}) = \sum_{k=1}^{K} \alpha_k P(Y_{kt} = x|F_{t-1})$$

$$= \sum_{k=1}^{K} \alpha_k \binom{x + r_k - 1}{r_k - 1} p_{kt}^{r_k} (1 - p_{kt})^x, x = 0,1\ldots \quad (10)$$

In this section, we discuss the problem of maximum likelihood estimation of MNBINARCH model parameters. We first make the following notation: $\alpha = (\alpha_1, \ldots, \alpha_K)^T, \beta = \left(\beta_{k0}, \beta_{k1}, \ldots, \beta_{K_{pK}}\right)^T, \theta = (\partial^T, \beta_1^T, \ldots, \beta_K^T)^T$First, the conditional likelihood function at time t is

$$L_t = \sum_{k=1}^{K} \alpha_k \binom{X_t + r_k - 1}{r_k - 1} p_{kt}^{r_k} (1 - p_{kt})^{X_t}$$

$$= \sum_{k=1}^{K} \alpha_k \binom{X_t + r_k - 1}{r_k - 1} \left(\frac{1}{\lambda_{kt} + 1}\right)^{r_k} \left(\frac{\lambda_{kt}}{\lambda_{kt} + 1}\right)^{X_t}$$

It is almost impossible to maximize such a likelihood function. Next, we introduce unobservable random variables and use the EM algorithm to achieve maximum likelihood estimation.

In fact, the commonly used method for maximum likelihood estimation of the parameters of the mixed distribution is the EM algorithm. The unobservable random variable Z=(Z,...,Zn) is introduced, where Z is a K-dimensional random vector, and the k-th component Zxa is defined as follows

$$Z_{kt} = \begin{cases} 1 \text{If Xt comes from the kth distribution, } 1 \le k \le K, \\ 0 \text{Other wise.} \end{cases}$$

The probability distribution of $Z_t = (Z_{1t}, \ldots, Z_{Kt} \quad)^T$ is $P(Z_t = (1,0,\ldots,0,0)^T) =$

$$\alpha_1,\ldots,P(Z_t=(0,0,\ldots,0,1)^T)=\alpha_K.$$

In fact, Z_t is independent of each other, and X and Z are also independent of each other. According to the MNBIARCH(K;p,..,p) model definition, the conditional distribution of the complete data (X_t,Z_t) is:

$$\prod_{k=1}^{K}\left(\alpha_k\binom{X_t+r-1}{r-1}\frac{1}{(1+\lambda_{kt})^T}\left(\frac{\lambda_{kt}}{1+\lambda_{kt}}\right)^{X_t}\right)^{Z_{kt}}$$

Therefore, the conditional log-likelihood function at time t is:

$$l_t=\sum_{k=1}^{K}Z_{kt}\,log\,\alpha_t+\sum_{k=1}^{K}Z_{kt}\,rlog\,p_{kt}+X_t\sum_{k=1}^{K}Z_{kt}\,log(1-p_{kt})+log\binom{X_t+r-1}{r-1}$$
$$=\sum_{k=1}^{K}Z_{kt}\,log\,\alpha_t-\sum_{k=1}^{K}Z_{kt}(r+X_t)\,log(1+\lambda_{kt})+X_t\sum_{k=1}^{K}Z_{kt}\,log\,\lambda_{kt}+log\binom{X_t+r-1}{r-1}$$

Therefore, the conditional log-likelihood function is:

$$l=\sum_{t=p+1}^{n}l_t=\sum_{t=p+1}^{n}\left\{\sum_{k=1}^{K}Z_{kt}(r+X_t)\,log(1+\lambda_{kt})+X_t\right.$$
$$\sum\nolimits_{k=1}^{K}Z_{kt}\,log\,\lambda_{kt}+log\binom{X_t+r-1}{r-1}\quad(11)$$

To find the first derivative of the log likelihood function with respect to 0, we get:

$$\frac{\partial l}{\partial\alpha_k}=\sum\nolimits_{t=p+1}^{n}\left(\frac{Z_{kt}}{\alpha_k}-\frac{Z_{kt}}{\alpha_K}\right),k+1,\ldots,K-1\quad(12)$$

$$\frac{\partial l}{\partial\beta_{ki}}=\sum\nolimits_{t=p+1}^{n}Z_{kt}\left(\frac{X_t}{\lambda_{kt}}-\frac{r+X_t}{1+\lambda_{kt}}\right)u(X_t,i),i=0,1,\ldots,p,k=1,\ldots,K\quad(13)$$

Among them,

$$u(X_t,i)=\begin{cases}1i=1,\\X_{t-i}i>0.\end{cases}$$

Since the process {Z_t} is not observable, the data we observe is incomplete data (missing data), so it cannot be directly used to estimate the parameter θ. The commonly used processing method in the parameter maximum likelihood estimation similar to the mixed distribution is used to maximize the conditional likelihood function by using the EM algorithm. The E-step and M-step included in the EM algorithm are as follows:

E-step: If it is assumed that the parameter θ is known, the missing data Z is estimated using their conditional expectations about the parameter θ and the observable data X. According to the definition of Z_t, the conditional expectation of the k-th component of Z_t is exactly equal to the conditional probability of X with respect to 6 and the k-th distribution of X from the mixed distribution. If $\tau_{kt} = E(Z_{kt}|X,\theta)$, the E-step equation is:

$$\tau_{kt} = \frac{\alpha_k \binom{X_t + r - 1}{r-1} p_{kt}^{\tau}(1-p_{kt})^{X_t}}{\sum_{k=1}^{K} \alpha_k \binom{X_t + r - 1}{r-1} p_{kt}^{\tau}(1-p_{kt})^{X_t}} = \frac{\alpha_k \lambda_{kt}^{\tau}(1+\lambda_{kt})^{-(r+X_t)}}{\sum_{k=1}^{K} \alpha_k \lambda_{kt}^{X_t}(1+\lambda_{kt})^{-(r+X_t)}}$$

Among them, k= 1,.. .,K, and t = p+1,... ,n.

M-step: The missing data Z is replaced by their conditional expectations about the parameters é and the observed data $X_1,\dots,X_n$, that is, it can be assumed that the missing data Z is known. The parameter e can be estimated by maximizing the log-likelihood function to obtain l. If the first derivative (12)-(13) of the log likelihood function l is zero, then the M-step equation is:

$$\widehat{\alpha_k} = \frac{1}{n-p}\sum_{t=p+1}^{n} \tau_{kt}, k = 1,\dots,K, \quad (14)$$

$$\sum_{t=p+1}^{n} \frac{\tau_{kl} X_t}{\hat{\lambda}_{kt}} u(X_t,i) = \sum_{t=p+1}^{n} \frac{\tau_{kt}(r+X_t)}{1+\hat{\lambda}_{kt}} u(X_t,i)$$
$$i = 0,1,\dots,p, k = 1,\dots,K. \quad (15)$$

Further, by introducing $\hat{\lambda}_{kt}$ into equation (15), we get:

$$\sum_{t=p+1}^{n} \frac{\tau_{kl} X_t}{\sum_{j=0}^{p} \hat{\beta}_{kj} u(X_t,j)} u(X_t,i) = \sum_{t=p+1}^{n} \frac{\tau_{kt}(r+X_t)}{1+\sum_{j=0}^{p} \widehat{\beta_{kj}} u(X_t,j)} u(X_t,i)$$
$$i = 0,1,\dots,p, k = 1,\dots,K. \quad (16)$$

Obviously, the non-linear equation (16) group has no display solution, so the numerical solution can only be obtained by numerical methods. The following uses the standard Newton-Raphson algorithm to achieve. For k, k* = 1,...,K and k≠k*, $\frac{\partial^2 l}{\partial \beta_k \partial \beta_{k^*}} = 0$can be obtained. Therefore, we can consider to estimate the parameter β_k separately in each mixing process.所以For each given k, use the Newton-Raphson algorithm to find the numerical solution of (16). We first take the derivative of (13), and get:

$$-\frac{\partial^2 l}{\partial \beta_{ki} \partial \beta_{ki}} = \sum_{t=p+1}^{n} Z_{kt}\left[\frac{X_t}{\lambda_{kt}^2} - \frac{r+X_t}{(1+\lambda_{kt})^2}\right] u(X_t,i) u(X_t,j)$$
$$i = 0,1,\dots,p, k = 1,\dots,K. \quad (17)$$

If $\beta_k^{(0)}$ is the initial value of β_k, then β_k iterates as follows:

$$\beta_k^{(i+1)} = \beta_k^{(i)} - \left\{\frac{\partial^2 l}{\partial \beta_{ki} \partial \beta_{ki}} |_{\beta_k^{(i)}}\right\}^{-1} \frac{\partial l}{\partial \beta_k} |_{\beta_k^{(i)}}$$

Among them, $\beta_k^{(0)}$ is the i-th iteration value. In addition, Z_{kt} is replaced by τ_{kt} calculated by E-step in the EM algorithm. E-step and M-step are iterated alternately in two steps until the iteration converges, and the estimation of the parameter θ is obtained. If $\theta_j^{(i)}$ is the jth component of $\theta^{(i)}$, the iterative convergence criterion of the EM algorithm is:

$$max\left\{\left|\frac{\theta_j^{(i+1)}-\theta_j^{(i)}}{\theta_j^{(i)}}\right|, j, i \geq 1\right\} \leq 10^{-5}$$

The mean absolute deviation er-ror (MADE) is adopted, that is

$$\frac{1}{m}\sum_{j=1}^{m}\left|\hat{\theta}_j-\theta\right|$$

is the evaluation standard of parameter performance, where m is the number of repeated simulations, and $\hat{\theta}_j$ is the estimated value of parameter θ in the jth simulation.The data flow diagram is shown in Figure 2.

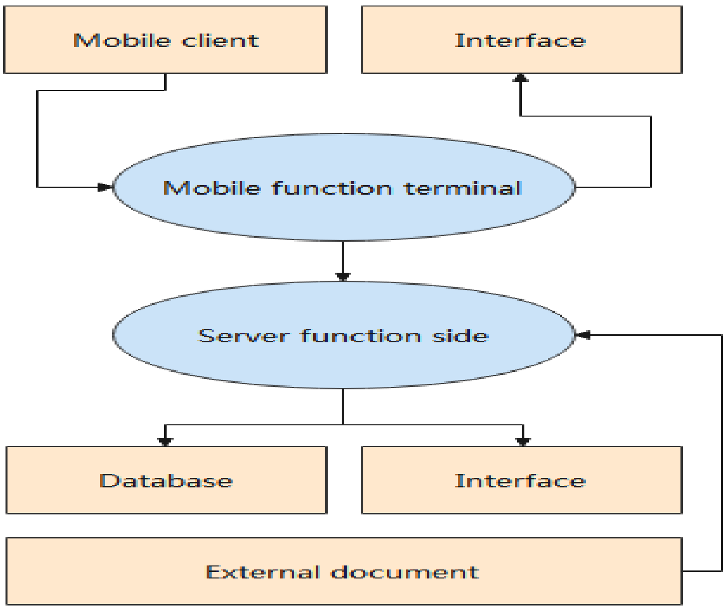


Figure 2. Data flow diagram

2.2 Design of sports training system based on data mining technology

The network topology diagram of the sports training system design is shown in Figure 3. It can be seen that in the wireless network environment, the coaching staff can enter the training status and competition status of their athletes based on the mobile terminal. Subsequently, network technology is used to realize the collection and sorting of the entire data in the process of using data resources. In this way, the collected data will also be uploaded to the server, and the server will store a large amount of information such as trainers, athletes, etc. In order to ensure the integrity of subsequent information collation, a related firewall must be set up on the front end of the server (Richmond et al., 2020).

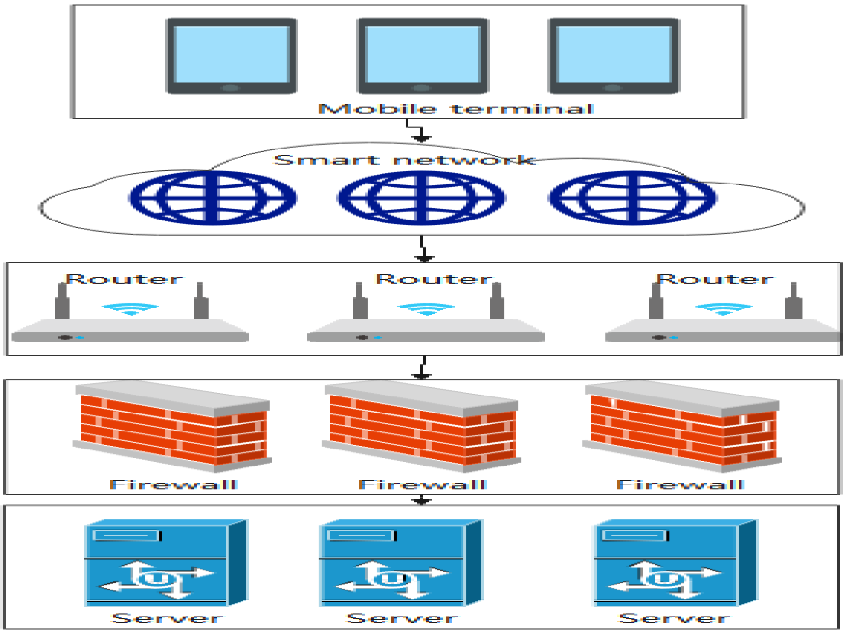


Figure 3. Network topology diagram

According to the design status of the sports training system, the basic system framework of this design is sorted out, and the system framework is also shown in Figure 4.

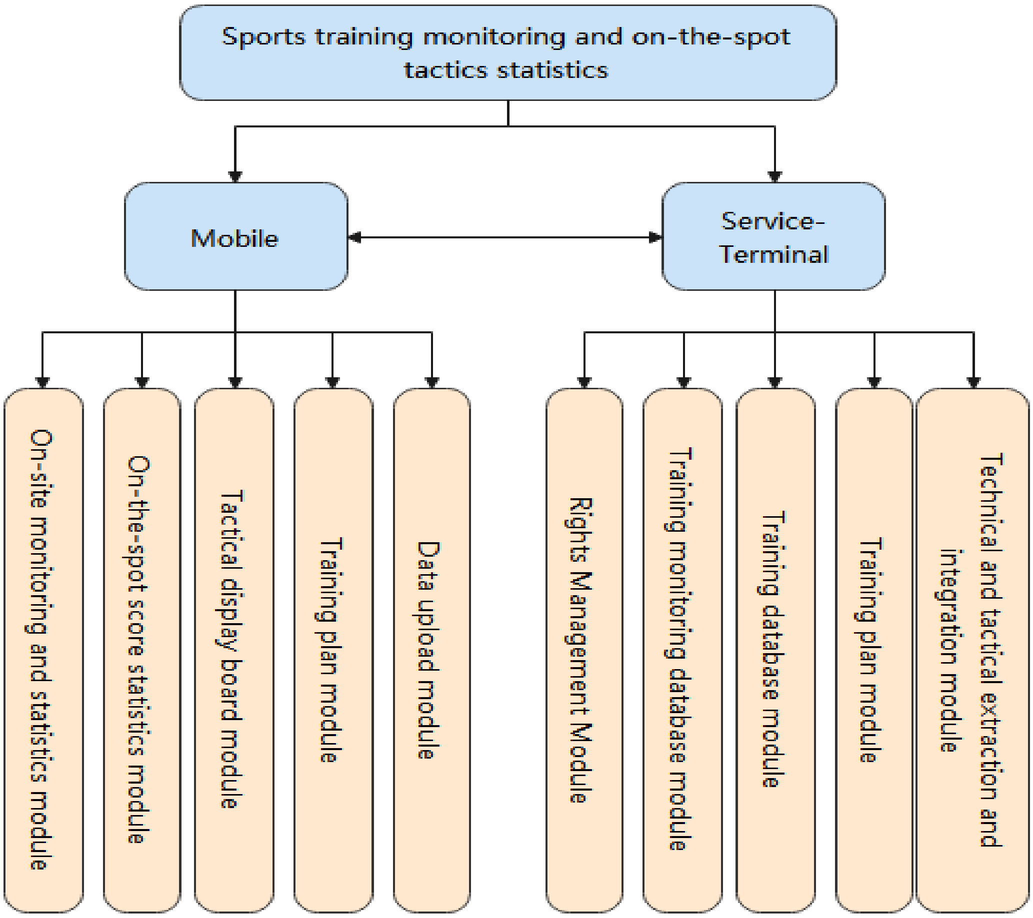


Figure 4. System frame diagram

It can be seen in Figure 4 that the main design modules of the system include two parts, the mobile terminal and the server, and include five main application modules.

The application module analyzes its indicators for the athletes' training status, game scores, tactical execution status, and service score status. In the process of data sorting, it is helpful for athletes and coaches to adjust it, improve its follow-up effect, and also improve the overall competition performance.

The detailed flow chart is shown in Figure 5, and the server is in the process of establishing a comprehensive information database. For the daily training data, training data, and game data received by the mobile terminal, the coach adjusts the data to make data storage. In the process of efficient information management, it solves its in-depth mining problems and provides reliable data analysis results.

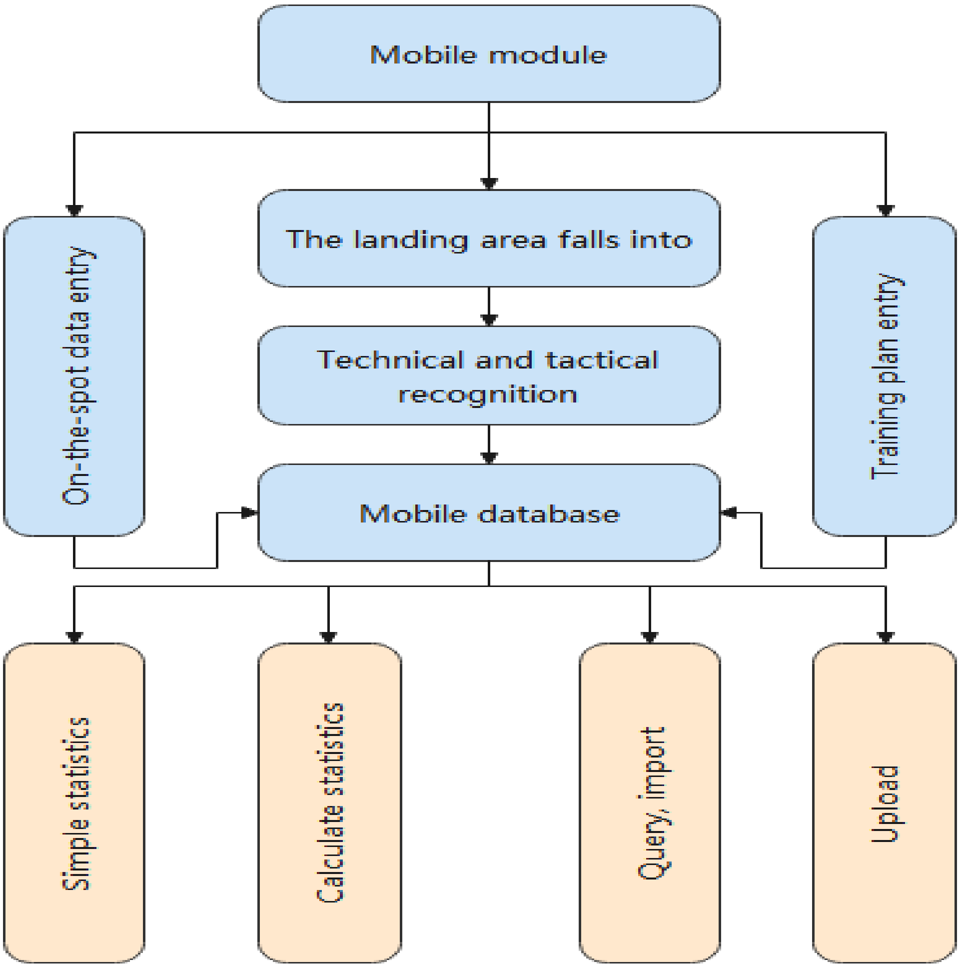


Figure 5. Mobile terminal flow chart

The physiological index sensor connected to the individual terminal is used to monitor the physiological index in real time and transmit it to the terminal, and the terminal completes the real-time exercise intensity evaluation function locally. The overall structure of the system is shown in Figure 6.

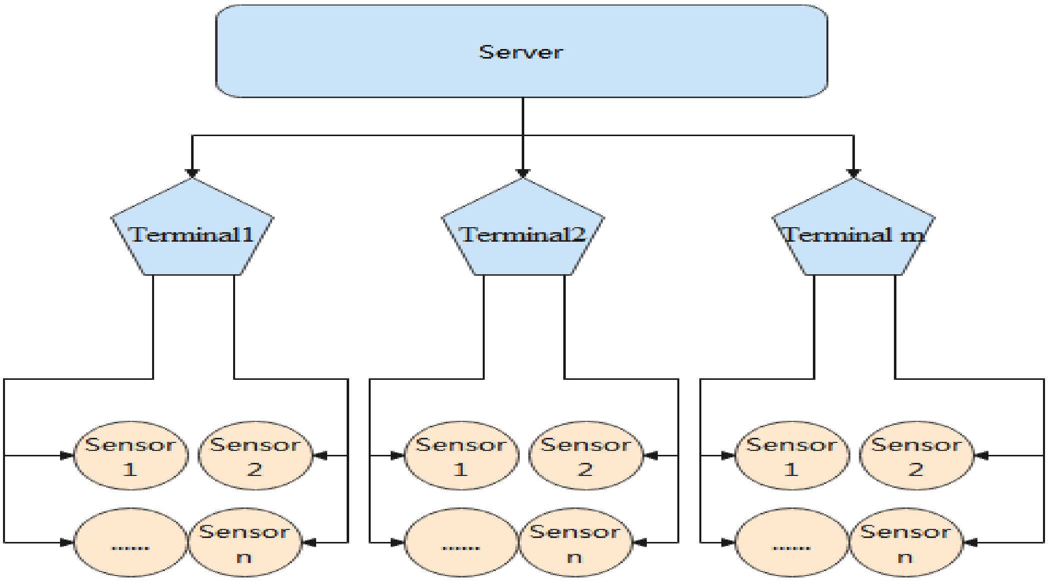


Figure 6. Basic structure of the system

This article builds an environment of virtual scenes, and takes indoor bicycle training as an example to construct a bicycle riding site, which is an important part of realizing system functions. The elements of the environment setting in the virtual scene include terrain, buildings, trees, grass, characters, lights, etc. Whether the riding environment is realistic or not will directly affect the exerciser's sense of immersion. The main content of building a scene in 3D is drawing terrain, modeling, texture import, and lighting rendering. The flow of the specific virtual 3D scene is shown in the figure 7.

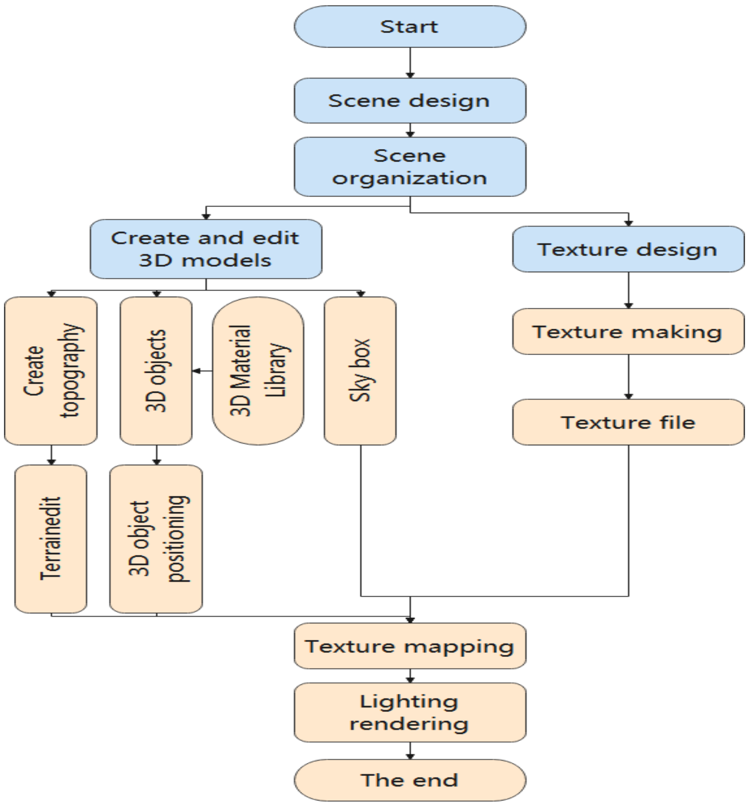


Figure 7. Flow chart of building a virtual scene

The software design of the lower computer of this system mainly includes the main controller program, the data transmission program, the speed acquisition program, the heart rate acquisition program, and the turning angle acquisition program. The overall process of the lower computer software system is shown in Figure 8.

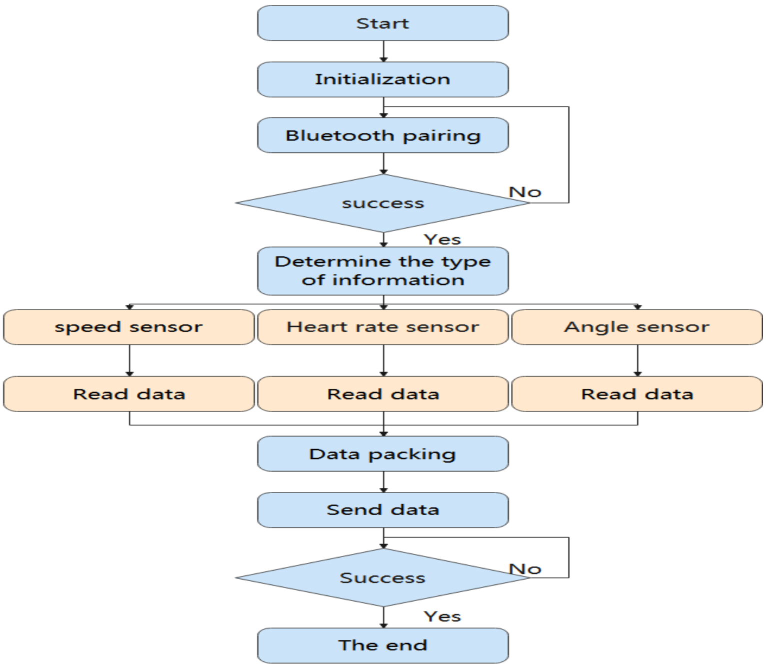
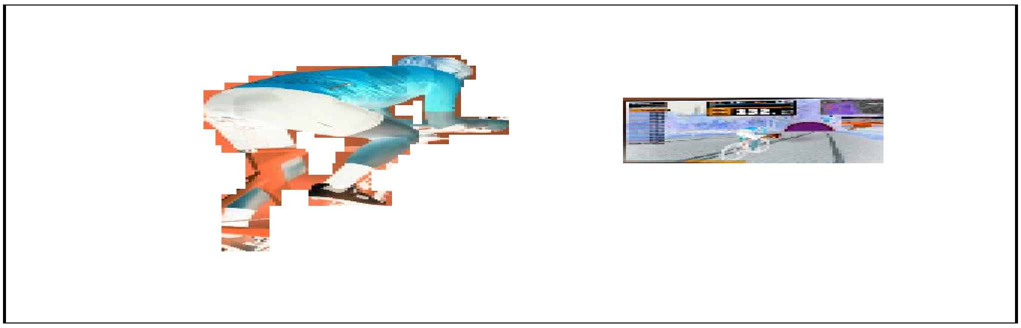


Figure 8. Flow chart of software system design of lower computer

On the basis of the above system construction, design experiments to verify system performance, taking bicycle exercise as an example, use image recognition combined with big data algorithms to process sports training data.The resulting image and feature extraction results are shown in Figure 9.

(a) Test scene

(b) Feature recognition image



(c) Sports training image recognition after background removal

Figure 9. Image and feature extraction results

It can be seen from Figure 9 that the continual construction of the sports training system in this paper has a good feature recognition effect. On this basis, a simulation study of the sports training effect is carried out. The results are shown in Table 1.

Table 1. Training effect of sports training system based on data mining technology

NO	Sports training effect	NO	Sports training effect	NO	Sports training effect
1	96.2	21	79.7	41	95.7
2	93.9	22	86.3	42	89.6
3	91.5	23	85.1	43	87.1
4	81.0	24	88.5	44	95.4
5	86.0	25	94.1	45	83.7
6	79.5	26	92.5	46	91.7
7	87.9	27	97.9	47	88.3
8	84.8	28	92.1	48	86.2
9	89.8	29	94.4	49	94.1
10	96.3	30	97.7	50	81.8
11	84.2	31	82.0	51	94.7
12	95.3	32	97.3	52	88.2
13	82.2	33	90.4	53	94.7
14	83.1	34	87.2	54	83.6
15	86.6	35	82.7	55	88.2
16	92.9	36	90.9	56	88.9
17	97.8	37	83.4	57	84.7
18	80.4	38	86.6	58	88.8
19	87.8	39	84.0	59	87.0
20	82.9	40	93.6	60	86.2

From the above research, it can be seen that the sports training system based on data mining constructed in this paper has a good training effect.

3. Conclusion

In recent years, scientific researchers in sports training have been emphasizing the value of data during training practice. The implementation of personalized and precise training programs is inseparable from the collection and analysis of sports evaluation data of athletes in each period and the file management.

The information-based training platform based on various data analysis during the training process can monitor the special sports training process in real time, and can timely test and feed back the quality, effect and individual status of the training, and then make targeted adjustments. With the acceleration of the integration of science and technology, advanced instruments, equipment and training platforms have become more and more powerful for data analysis, and can instantly extract information reflecting the true state of the object, so that the practice process of sports training can be continuously implemented and revised.

Moreover, coaches and scientific researchers have not only improved the effects of special sports training, but also have a clear understanding of the development level of the athlete's individual athletic ability. Finally, this article combines data mining technology to construct a sports training system. The research results show that the sports training system constructed in this paper has good practical effects.

4. Acknowledgement

The 2022 special project of Guangxi Education Science "14th Five Year Plan" titled "Research on Strategies for Improving the Quality of Physical Education Courses in Private Colleges and Universities in Guangxi during the" 14th Five Year Plan "(Project No.: 2022ZJY3254)

References

Abanazir, C. (2019). E-sport and the EU: the view from the English Bridge Union. *The International Sports Law Journal, 18*(3-4), 102-113.

Barbosa, T. (2018). Smart sport equipment: reshaping the sports landscape. *Motricidade, 14*(2-3), 1-2.

Catro-Sánchez, M., Zurita-Ortega, F., & Chacón-Cuberos, R. (2019). Motivation towards sport based on sociodemographic variables in university students from Granada. *Journal of Sport and Health Research, 11*(1).

Cristiani, J., Bressan, J. C. M., Longarela Pérez, B., Galatti, L. R., & Reverdito, R. S. (2017). Clubs socio-deportivos en un municipio brasileño: espacio, equipos y contenidos.

Du Plessis, J. H., & Berteanu, M. (2020). The importance of prosthetic devices in sport activities for Romanian amputees who compete in Paralympic competitions. *Medicina Sportiva: Journal of Romanian Sports Medicine Society, 16*(1), 3197-3204.

Emery, C. A., & Pasanen, K. (2019). Current trends in sport injury prevention. *Best Practice & Research Clinical Rheumatology, 33*(1), 3-15.

Ferguson, L. J., Carlson, K. T., & Rogers, D. (2019). Moving towards reconciliation through sport: Sharing our process of exploring team saskatchewan experiences at the North American Indigenous Games. *Journal of Exercise, Movement, and Sport (SCAPPS refereed abstracts repository), 51*(1).

Gerke, A., Babiak, K., Dickson, G., & Desbordes, M. (2018). Developmental processes and motivations for linkages in cross-sectoral sport clusters. *Sport Management Review, 21*(2), 133-146.

Giulianotti, R., & Numerato, D. (2018). Global sport and consumer culture: An introduction. *Journal of consumer culture, 18*(2), 229-240.
Gurinovich, A., & Petrova, G. (2019). Key priorities of physical education and sport sector budgeting laws and regulations in the Russian Federation. *Theory and Practice of Physical Culture*(4), 34-34.
Hadlow, S. (2018). Modified perceptual training in sport: A new class.
Happ, E., Schnitzer, M., & Peters, M. (2021). Sport-specific factors affecting location decisions in business to business sport manufacturing companies: a qualitative study in the Alps. *International Journal of Sport Management and Marketing, 21*(1-2), 21-48.
Ilies, D., Buhas, R., Ilies, M., Ilies, A., Gaceu, O., Pop, A., . . . Baias, S. (2018). Sport activities and leisure in Nature 2000 protected area–Red Valley, Romania. *Journal of Environmental Protection and Ecology, 19*(1), 367-372.
Kimasi, K., Shojaei, V., & Boroumand, M. R. (2019). Investigation of safety conditions at gymnasia in different organizations. *Journal of Humanities Insights, 3*(02), 70-74.
Kondrukh, A. (2017). Practical shooting sport in Russian sport system: essential specifications and features. *Theory and Practice of Physical Culture*(5), 27-27.
Mountjoy, M., Costa, A., Budgett, R., Dvorak, J., Engebretsen, L., Miller, S., . . . Carr, J. (2018). Health promotion through sport: international sport federations' priorities, actions and opportunities. *British journal of sports medicine, 52*(1), 54-60.
Pogrebnoy, A., & Komlev, I. (2018). Sport institutions reporting to Ministry of Sport of Russian Federation: intellectual property, invention activity, patenting and legal consulting service analysis. *Theory and Practice of Physical Culture*(2), 2-2.
Reinhart, K., & Wichmann, B. (2020). The TuS Fortschritt Magdeburg-Neustadt (soccer section) in the GDR–an example of amateur socialist sport. *Soccer & Society, 21*(4), 408-420.
Richmond, S. A., Donaldson, A., Macpherson, A., Bridel, W., van den Berg, C., Finch, C. F., . . . Emery, C. A. (2020). Facilitators and barriers to the implementation of iSPRINT: a sport injury prevention program in junior high schools. *Clinical journal of sport medicine, 30*(3), 231-238.
Riley, A. H., & Callahan, C. (2019). Shoulder rehabilitation protocol and equipment fit recommendations for the wheelchair sport athlete with shoulder pain. *Sports medicine and arthroscopy review, 27*(2), 67-72.
Shulyatyev, V., & Bulavina, M. (2019). Sport interview: strategy, design and content. *Theory and Practice of Physical Culture*(9), 18-18.
Stylianou, M., Hogan, A., & Enright, E. (2019). Youth sport policy: The enactment and possibilities of 'soft policy'in schools. *Sport, Education and Society, 24*(2), 182-194.