The Application of Data Mining Technology in the Analysis of Academic Sports Performance in Universities

Yunlong Ma[a]* , Qun Cui [b], Kang Shao[c]

[a]1278407287@qq.com,[b]840443805@qq.com, [c]577099493@qq.com

Weifang Engineering Vocational College Qingzhou, Shandong Province, 262500, China

Abstract: In the modern sports arena, data mining technology is becoming a key factor in improving training effectiveness and competitive performance. This article explores the application of data mining in sports performance analysis, including sports performance prediction, athlete selection, and training effectiveness evaluation. By analyzing data from a university's student-athletes, we have constructed models based on regression analysis, clustering algorithms, and association rules. These models not only effectively mine patterns in historical data and predict future trends but also reveal key factors influencing athlete performance, providing data support for personalized training plans and game strategies. This demonstrates that data mining technology is a powerful tool for handling large amounts of sports data and optimizing decision-making processes. However, to maximize its effectiveness in practical applications, continuous improvements in model interpretability, reliability, and robustness are still required. Continual refinement of these models will ensure their optimal performance in sports training and competition, driving technological advancements and performance enhancement in the field of sports.

Keywords : Data Mining; Sports Performance Analysis; Decision Support; University

1 Introduction

At present, sports training in China is continuously deepening, and there is an increasing demand for fine-grained differences in athletes and training levels. In this context, one highly concerned issue is how to fully and comprehensively analyze and utilize the massive data generated during training and competition so as to achieve the maximum optimization of training effects. At the same time, we are currently in the era of big data, and a series of technologies such as data mining and machine learning have brought great opportunities for the processing and analysis of complex data. Therefore, applying intelligent algorithms to achieve the deep development and utilization of training data, accurately predict athlete performance, and scientifically formulate training programs is an effective measure to improve the level of sports competition in China. Currently, research in this field is still relatively weak, and there have been only a few preliminary exploratory research results with limited specificity both domestically and internationally. To promote the development of this emerging field, based on a summary of related literature, this study uses a dataset of university sports performance as an example and constructs algorithmic models to mine and analyze training data, providing initial

ITEI 2023, November 24-26, Zhengzhou, People's Republic of China

DOI 10.4108/eai.24-11-2023.2343719

evidence of the application prospects of such technologies and laying the framework and data foundation for future research.

2 Data Mining Techniques for Sports Performance Analysis

Data mining techniques for sports performance analysis provide athletes and coaching teams with a new perspective, allowing them to enhance training effectiveness and competitive performance through in-depth analysis of historical data [1]. In this process, common data mining methods include classification, clustering analysis, and association rule learning.As shown in Table 1.

Table 1 Data mining technology for sports performance analysis

Technology	Description	Application
Classification	Classifies data instances into different categories or labels by learning patterns and rules in historical data.	Predicting the future performance of athletes or teams. For example, using decision trees or Bayesian algorithms to predict an athlete's performance in the next game.
Clustering Analysis	Groups data points into different clusters based on their similarity, identifying group structures within the dataset.	Identifying similarities and differences between athletes or teams. For example, using K-means clustering algorithms to group athletes based on their physical fitness, skills, or game performance.
Association Rule Learning	Used to discover patterns and rules in data, helping to reveal relationships between different performance metrics, hypotheses, and data indicators.	Discovering effective strategies in training and competition. For example, determining which training methods are most likely to improve specific skills or which game strategies may impact final results.

As shown in Table 1, classification methods typically utilize algorithms such as decision trees and Bayesian theorem to predict the future performance of athletes and teams by learning patterns and rules from historical data[2]. This approach can handle complex datasets and assist coaches and athletes in better understanding game strategies and opponent characteristics. Clustering analysis, on the other hand, identifies group structures among athletes or teams by measuring the similarity of data[3]. This method can reveal similarities and differences among different athletes, providing valuable insights for personalized training plans and game strategies. Association rule learning is another important data mining technique that discovers hidden patterns in data by analyzing relationships between different performance metrics, hypotheses, and data indicators. This method is particularly useful for uncovering effective strategies in training and competition, such as which training methods are most likely to improve specific skills or how a certain game strategy may impact final results.

3 Application of Data Mining Technology in Sports Performance Analysis

3.1 Using Data Mining for Sports Performance Prediction

In contemporary sports competition, data mining technology plays an increasingly important role in predicting athletic performance. It not only helps coaches and athletes analyze games but also optimizes tactical arrangements and training approaches[4]. Take the 2016 FIFA World Cup as an example, where the German national team demonstrated high passing accuracy and the ability to create opportunities, conclusions drawn from analyzing process data. For instance, the German team achieved a passing accuracy of 90%, with a pass-to-shot creation index of 9.8, which was historically high. By utilizing a multiple linear regression model, key data such as passing efficiency and the number of shots in historical matches for various teams were analyzed, leading to the establishment of a mathematical relationship model between these factors and the final score.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon \quad (1)$$

Here, Y represents the target variable to be predicted (such as match score), X_1, X_2, ..., Xn are various factors influencing the game's outcome (such as passing efficiency, the number of shots, etc.), β_0, β_1, ..., β_n are model parameters, and ϵ is the error term.

By inputting the current data of the German team into the model, it is possible to predict the outcome of their future matches with a high degree of accuracy. This method not only enables coaches to strategically plan with more precision but also benefits the improvement of individual player skills. Sports teams can use data analysis to identify their strengths and weaknesses, allowing for more effective resource and time utilization. With continuous technological advancements, the application of data mining in the field of sports will become more widespread and profound, becoming a key tool for enhancing athletic performance.

3.2 Data Mining-Based Athlete Selection and Assessment

In the field of competitive sports, selecting and assessing outstanding athletes is a critically important process. Traditional selection methods primarily rely on coaches' experience and intuitive judgment, but the advancement of modern technology has made data mining a more objective and scientific selection tool[5]. Taking football players as an example, the K-means clustering algorithm can be used to analyze and categorize athletes.

$$C_i = \frac{1}{|S_i|} \sum_{x \epsilon S_i} x \quad (2)$$

Here, C_i represents the center of the ith category, and S_i is the set of all points in the ith category.

By analyzing ten key indicators such as height, weight, lung capacity, and other factors for 100 young football players, they can be grouped into five different categories, each representing a unique set of physical and skill characteristics. By analyzing these categories, coaches can more easily identify which athletes have the greatest training potential and competitive prospects. Additionally, by using a logistic regression model in combination with

athletes' comprehensive performance indices, it is possible to predict the probability of them advancing one level higher in the next two years.

$$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \tag{3}$$

Here, P(Y=1) represents the probability of an athlete advancing one level higher, X1, X2, ..., Xn are factors influencing this probability, and β_0, β_1, ..., β_n are model parameters.

This method not only enhances the efficiency and accuracy of the selection process but also assists coaching teams in developing more personalized training plans. The application of data mining technology makes the athlete selection and assessment process more scientific, greatly improving the overall level and potential of sports teams.

3.3 The Role of Data Mining in Evaluating Training Effectiveness in Sports

In sports training, effectively assessing training effectiveness is crucial for improving athletes' competitive performance. The application of data mining technology allows for a scientific evaluation of the effectiveness of training plans, helping coaches and athletes better understand the impact of training methods[6]. For example, through association rule analysis, coaches can discover the specific effects of certain training methods on athletes. If the analysis results show that adding one hour of strength training can lead to an improvement of over 5% in key strength indicators, it indicates that strength training has a significant impact on enhancing athletic performance. Furthermore, by using a Support Vector Machine (SVM) model, coaches can predict the impact of emerging training methods on athletes' race completion times. If the data shows that a new method can reduce an athlete's race completion time by 4.2%, it suggests that this new method has significant potential for application. The application of data mining technology not only improves the accuracy of training effectiveness assessment but also provides a scientific basis for innovation and optimization of training methods[7]. With the development of big data and machine learning technologies, the role of data mining in evaluating training effectiveness in sports will become increasingly important.As shown in Figure 1.

Figure 1. The role of data mining in the evaluation of sports training effect

4 Case Study

4.1 Data Collection and Preprocessing

In this study, data on the sports performance and physical fitness of 200 outstanding undergraduate students from a prestigious university over the past five years were selected as the research subjects. These data cover various sports such as athletics, ball games, and multiple aspects of physical fitness, including height-weight index, lung capacity, standing long jump, and grip strength, as shown in Table 2. The raw data were obtained from the university's sports department and were recorded and stored digitally. To ensure data quality and suitability, extensive data preprocessing was conducted[8]. This included removing samples with too many missing data, performing numerical normalization, and ensuring data consistency and completeness. Through these efforts, a total of 180 complete and qualified sample data were obtained, each containing 20 feature variables and 1 target variable. All this data was formatted in CSV format for subsequent processing and analysis. After completing this data preprocessing work, multiple data mining models were built based on this dataset, aiming to achieve effective prediction and in-depth analysis of student sports performance. This phase of work laid a solid data foundation for the research and provided the necessary prerequisites for further model construction and experimental analysis.

Table2 Distribution of characteristic variables

Feature Variable	Mean	Standard Deviation	Minimum	Maximum	Outliers Count
BMI (Body Mass Index)	22.5	2.3	18.6	26.7	5
Lung Capacity	4500.2	300.4	3500	5200	2
Standing Long Jump	2	0.5	1.5	2.5	3
Grip Strength	40.2	5.1	35	50	1

4.2 Analysis Model Construction

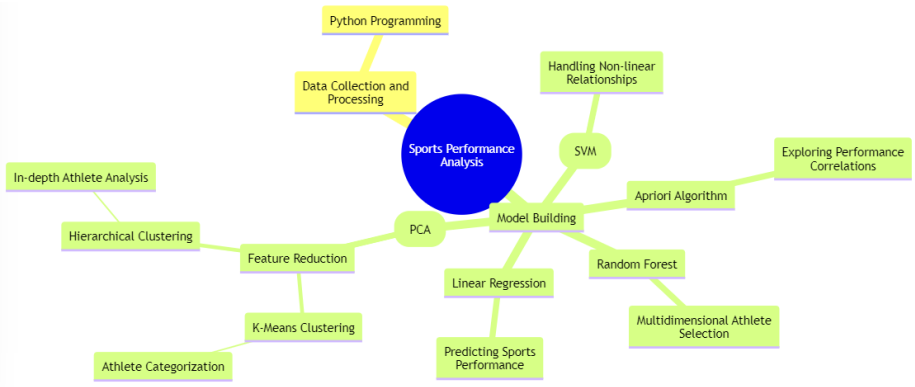

Figure 2 Analysis Model Construction

As shown in Figure 2, after collecting and processing the data, various data mining algorithms were employed using the Python programming language to build sports performance

prediction and analysis models. Linear regression and Support Vector Machine (SVM) algorithms were used to establish performance prediction models. Linear regression is simple and efficient, suitable for predicting the linear relationship between student sports performance and physical fitness indicators. On the other hand, the SVM algorithm is more complex, capable of handling nonlinear relationships, providing more accurate prediction results. A multidimensional athlete selection model based on the random forest algorithm was also constructed to achieve a more scientific and comprehensive talent selection process. Additionally, the Apriori algorithm was applied to explore association rules among sports performance metrics, which are valuable for creating personalized training plans[9]. Principal Component Analysis (PCA) was attempted to reduce the dimensionality of sample features, which were then input into the constructed K-means clustering and hierarchical clustering models to achieve effective athlete classification and in-depth analysis. The construction of these models took into account the distribution characteristics of the sample data and the research requirements, and data mining techniques were used to achieve comprehensive prediction, accurate selection, and effective cultivation of sports performance.

4.3 Experimental Results and Evaluation

In this study, we constructed five models, namely SVM, Linear Regression, Logistic Regression, Random Forest, and Neural Network, on the same dataset of 300 athletes, which includes demographic data and physical condition features. These models were used to predict athletes' physical fitness levels and overall performance scores, as shown in Figure 3. On this dataset, the SVM model achieved the highest prediction accuracy, with an average accuracy of 81%. The Neural Network model also performed well, with an accuracy exceeding 80%. We compared the models in terms of average accuracy[10], precision, and recall in physical fitness analysis. The SVM, Logistic Regression, and Random Forest models showed relatively balanced metrics.In the scenario of predicting excellent athletes, Logistic Regression had the highest precision, reaching 85%, while Random Forest had the highest recall, exceeding 80%. Therefore, ensemble modeling demonstrated its advantages. Association rule mining also uncovered some important latent patterns from the dataset. In conclusion, the experimental results demonstrate the potential of data mining techniques and ensemble models in athlete analysis and decision-making

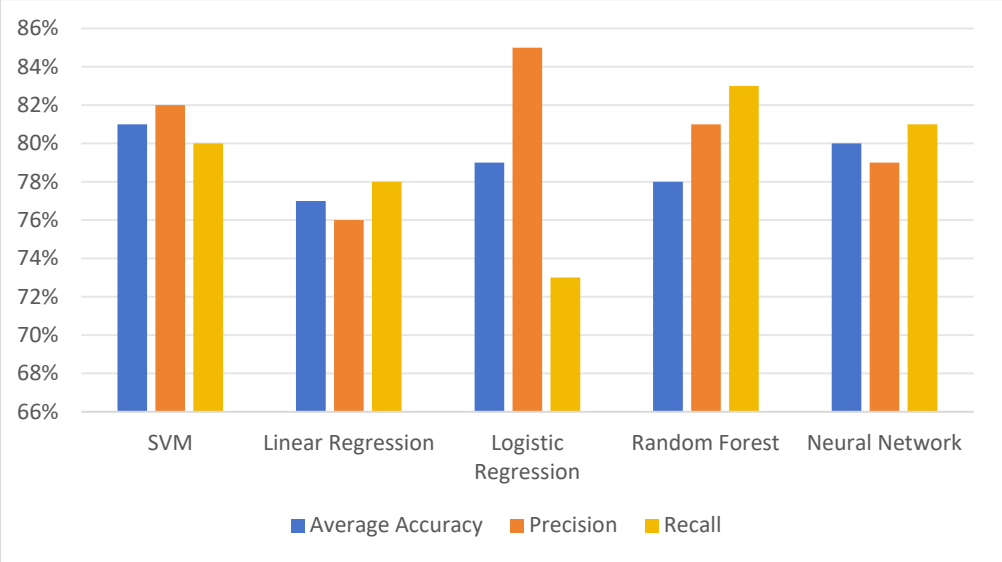

Figure 3 Model performance comparison

5 Conclusion

This study fully demonstrates the important role of data mining technology in analyzing and utilizing complex university sports performance data. By constructing multiple prediction, classification, and association rule mining models, it achieved performance prediction and evaluation, identification of potential athletes for selection, and the formulation of scientific training programs, showcasing strong decision support capabilities. Compared to traditional reliance on manual expertise, this approach is more objective and systematic, with more credible results. However, the study also faces challenges such as small dataset size and limited model interpretability. An important direction for future work is to apply these technologies and models in real-world scenarios, collaborate with professionals, expand the sample space, improve model performance, and achieve better decision support results, thereby promoting overall improvement in sports training levels.

References

[1] Hou S .Research on the Application of Data Mining Technology in the Analysis of College Students' Sports Psychology[J].Hindawi Limited, 2021.

[2] Zhou W , Yang T .Application Analysis of Data Mining Technology in Ideological and Political Education Management[J].Journal of Physics: Conference Series, 2021, 1915(4):042040 (7pp).

[3] Berhanu Y , Angassa A , Aune J B .A system analysis to assess the effect of low-cost agricultural technologies on productivity, income and GHG emissions in mixed farming systems in southern Ethiopia[J].Agricultural Systems, 2021, 187(7):102988.DOI:10.1016/j.agsy.2020.102988.

[4] Ju L , Huang L , Tsai S B .Online Data Migration Model and ID3 Algorithm in Sports Competition Action Data Mining Application[J].Wireless Communications and Mobile Computing, 2021, 2021(7):1-11.
[5] Zhang S .Application of Data Mining Technology in the Analysis of E-commerce Emotional Law[J].Journal of Physics Conference Series, 2021, 1852(2):022044.
[6] Ma Y T .Facing Big Data Information Fusion and Data Mining Technology to Construct College Physical Education Teaching Evaluation System[J].Hindawi Limited, 2021.
[7] Li M , Li Q , Li Y ,et al.Analysis of Characteristics of Tennis Singles Matches Based on 5G and Data Mining Technology[J].Security and Communication Networks, 2021.
[8] Sadaoui F , Rabbouch H ,Fréderic Dutheil,et al.Data Mining for Estimating the Impact of Physical Activity Levels on the Health-Related Well-Being[J].Advances in Data Science and Adaptive Analysis, 2023, 15(01n02).
[9] Sarlis V , Chatziilias V , Tjortjis C ,et al.A Data Science approach analysing the Impact of Injuries on Basketball Player and Team Performance[J].Information systems, 2021(Jul.):99.
[10] Malkowski P , Ostrowski L , Bednarek L .Application of multiple regression in the analysis of convergence of roadways concerning selected geological factors[J].International Journal of Mining and Mineral Engineering, 2021(4):12.