Application of Classification Algorithm Based on Naive Bayes in Data Analysis of Fitness Test

Yang Sun¹*, YueHong Ma¹,²

¹HEILONGJIANG UNIVERSITY OF TECHNOLOGY Heilongjiang Jixi 158100
²Pai Chai University

*Corresponding author e-mail: jxdxtxy@163.com

**Abstract.** An important aspect of physique data analysis is to provide a basis for developing national sports and improving physical education. This article aims to explore the application of Naive Bayesian classification algorithm in physical test data analysis, in order to conduct differentiated teaching and training from person to person, so as to effectively improve the physical health of students. This article uses the real physical test data of our school in 2018 and 2019 to form a complete data set after coding to construct a student physical classifier. In this way, the correct judgment of the student's physical condition can be realized with a certain probability, and the problems of the student's physical health can be detected and warned in time. Using the slicing method to cut out 20% as the test data, the classifier is tested. From the test results, it can be seen that the comprehensive correct rate of the physique classifier based on the naive Bayes algorithm reaches 81.02%. The research in this article provides help to better promote the physical fitness testing of colleges and universities, and provides a new way of thinking and practical methods for improving the quality of physical education teaching.

**Keywords:** Naive Bayesian Classification Algorithm, Physique Analysis, School Physical Education, Physical Health

1. Introduction

In recent years, domestic and external physical fitness data analysis mainly focuses on testing methods and testing instruments for physical fitness data acquisition, but mathematical analysis is rarely used to analyze and compare data processing methods and results represented by data¹². Study of physical
fitness test data of students present research field of coverage, the research content close to the actual characteristics of the descriptive research, physical health intervention experiment, some groups but less empirical research to the individual. Sports science has accumulated a large number of test results, the traditional research method is lack of deep analysis of the data and knowledge hidden behind the data and related useful information which are not well mining\cite{1,2}. At present, statistics has shown unprecedented achievements in sports science research, but it also exposes its own limitations, making relevant data analysis work unsatisfactory\cite{3}.

At present, all kinds of schools at all levels in China carry out physical health test for all students after the opening season of each year, and report the collected data to "Student Physical Health Network", realizing the data-based management of students' physical health test scores\cite{4}. Although continuous monitoring is becoming more and more scientific and standardized in the process of continuous implementation, the significance of physical fitness monitoring lies not only in the testing process, but also in the results of testing, so as to guide the practice of physical health improvement with the application of data and better guide the improvement of students' physical health level\cite{5}. Due to some intuitive judgment of the human body physique condition index of the test are relatively complicated, and the test methods have higher requirements, but in actual regular physical judgment we often want to according to some simple test data to determine the subjects' physical quality \cite{6,7,8}. Therefore, how to judge the physique condition according to the simple measurement index is an important subject in the sports data analysis.

This paper mainly adopts the methods of literature investigation and mathematical statistics. The real body test data of students in 2018 and 2019 were used to test the classifier. Through the analysis of the current situation of physical health testing and put forward the rationalization proposals for the development of school physical education in the future, so as to timely find and early warning of students' physical health problems and other aspects to provide help.

2. Naive Bayes Algorithm and Fitness Classifier

2.1. Naive Bayes Algorithm

Naive Bayes algorithm is an important algorithm in the field of machine learning and data mining. It is mainly used in the classification field of machine learning, and has a wide range of applications in text analysis, public opinion analysis, medical diagnosis, user preference analysis and other fields. Its main advantages are highly stable, simple, efficient and strong theoretical foundation. What type of physique a student belongs to is like a classification problem.

Given the training data set $(X, Y)$, where each sample $X$ includes n-dimensional features, and
the class label set contains $k$ categories, if a new sample $x$ is given, the total probability formula is obtained:

$$P(y_k \mid x) = \frac{P(x \mid y_k)P(y_k)}{\sum_k P(x \mid y_k)P(y_k)}$$

(1)

Therefore, the conditional probability $P(x \mid y_k)$ can be transformed into:

$$P(x \mid y_k) = P(x_1, x_2, \cdots, x_n \mid y_k) = \prod_{i=1}^{n} P(x_i \mid y_k)$$

(2)

So the naive Bayes classifier is expressed as:

$$f(x) = \arg \max P(y_k)\prod_{i=1}^{n} P(x_i \mid y_k)$$

(3)

2.2. Constitution Classifier

In view of the physical problems of students, according to the National Student Physical Health Standards, 12 physical characteristics are determined: gender, age, grade, hometown, height, body mass, body mass index (BMI), vital capacity, speed quality, explosive power quality, flexibility, Sexual quality, endurance quality, strength quality. The classification group is defined as: excellent, good, pass, and fail. In order to facilitate the calculation, it is converted to a value of 1/2/3/4, and the conversion of height and weight is combined into one BMI data element. During the training process, the interval probability is calculated. Use Python coding to obtain data preprocessing, use the slicing method, and divide it into 80% training data and 20% test data. The first classifier is a forward application of the Naive Bayes algorithm for predicting physical fitness classification. The second classifier in turn uses the physique classification result as a feature item to predict the attribute value interval.

Characteristics of classifier 1: gender, height, body mass, vital capacity. Gender ($x_1$) has two values for men and women, and grade ($x_2$) has two values (a freshman, a second, a third and a fourth). Height and weight are converted into BMI ($x_3$) 4 value ranges ($\leq 17.1, 17.2 \sim 23.9, 24.0 \sim 27.9, \geq 28.0$), and vital capacity ($x_4$) is divided into 20 value ranges.

The physique level is excellent, good, passing, and failing, expressed as ($y_1, y_2, y_3, y_4$). Classifier 2 features gender, grade, height, body mass, vital capacity, speed, explosive power, and physical fitness
level. Category: Endurance level.

3. Research Methods

This paper mainly adopts the methods of literature investigation and mathematical statistics. Through collecting and reading about the naive bayesian classification algorithm and physical test and analysis of the literature, books, newspapers, with the relevant policy documents issued by the state, and through all kinds of text information network, sorts through comprehensive analysis, summarizes the research. The data in this paper are from the actual body measurement data of the students in our school in 2018 and 2019. According to the national criteria of student physique healthy standard, the score of each student, total score and evaluation grade. A complete data set with physical classification results is formed and the classifier is tested. Through the analysis of the current situation of physical health testing and put forward the rationalization proposals for the development of school physical education in the future. In order to timely find and early warning of students' physical health problems and other aspects of the help, the effective development of physical education in our school will play an important guiding significance.

4. Student Physique Analysis and Classifier Detection

4.1. Physical Distribution Map of Our school Students

Through the archives of the school, the real body profile data of our school in 2018 and 2019 are obtained, and the scores and ratings of each student are calculated by writing a computer program, as shown in Table 1 and Figure 1.

| Rating   | Percentage |
|----------|------------|
| Failed   | 8.36%      |
| Pass     | 79.31%     |
| Good     | 12.03%     |
| Excellent| 0.30%      |

Table 1. Physical distribution of students in our school
Figure 1. Physical distribution map of our school students

It can be seen that the physical condition of the students in our school is not optimistic. Although most of the students have reached the passing line, and the passing grade accounts for 79.31%, there are still 8.36% of the students who failed the physical examination. To achieve a good physique accounted for 12.03%, to achieve excellent physical fitness is very small, accounting for only 0.30%.

4.2. Classifier Detection

Using slice method to cut out the other 20% of the student's real body side data as test data, four evaluation indicators of Precision, Recall, Fb-score and Accuracy are used. Fb-score is the harmonic average of accuracy and recall, and the test results are shown in Table 2.

| Physique classification | Sample | Accuracy/% | Recall rate/% | Harmonic average/% |
|-------------------------|--------|------------|---------------|--------------------|
| 1 (Excellent)           | 5      | 0          | 0             | 0                  |
| 2 (good)                | 298    | 0          | 0             | 0                  |
| 3 (pass)                | 1548   | 0.82       | 1             | 0.91               |
The test results show that the comprehensive correct rate of the classifier is 81.02%. The trained classifier can be used directly by inputting the physical data of the students, and the corresponding classification results can be obtained. These results can be used to predict the physical condition of the students. Based on the basis of classification 1, classification and prediction can be made for physical fitness for all students, according to the classification result, students will be divided into different groups. For students who may fail the physical fitness test, differentiated teaching and training can be carried out from person to person to improve their physical fitness. Based on the basis of classification 2, as part of the students want to achieve higher levels of physical fitness, it is necessary to strengthen the endurance training program. At the same time, on the basis of the prediction and judgment evaluation of the physique classification model based on the naive Bayes algorithm, the relevant school departments or physical education teachers can provide training and guidance for students' physical health problems. Continue to follow up to collect data for the next year, measure the physical health of students, analyze the physical changes in the longitudinal direction of time, and the effectiveness of exercise intervention.

5. Conclusion

In summary, there are many factors that affect the physical health of students. The analysis of physical data based on the Naive Bayesian classification algorithm allows teachers to formulate personalized student physical health improvement plans based on the result analysis, thereby providing nutrition guidance, provide assistance in physical exercise guidance, timely detection and early warning of students' physical health problems. Differentiated teaching and training from person to person enables the organic integration of the various components of our school sports, so that the physical health of students can be effectively improved. Make full use of data as much as possible to provide help to better promote the physical fitness testing of colleges and universities, and to provide new ways of thinking and practical methods to improve the quality of physical education teaching. The ultimate goal and significance is to tap the value of data.

Acknowledgments

2020 Basic Scientific Research Business Expense Project of Provincial Undergraduate Universities in Heilongjiang Province.

References
[1] Liu P, Zhao H H, Teng J Y, et al. Parallel naive Bayes algorithm for large-scale Chinese text classification based on spark[J]. Journal of Central South University, 2019, 26(1):1-12.

[2] Ning B, Junwei W, Feng H. Spam Message Classification Based on the Nave Bayes Classification Algorithm[J]. IAENG Internaitonal journal of computer science, 2019, 46(1):46-53.

[3] Ximeng X, Rennong Y, Ying F. Situation assessment for air combat based on novel semi-supervised naive Bayes[J]. Journal of Systems Engineering and Electronics, 2018, 29(4):768-779.

[4] Ferdina V, Kristanda M B, Hansun S. Automated complaints classification using modified nazief-adriani stemming algorithm and naive bayes classifier[J]. Journal of Theoretical and Applied Information Technology, 2019, 97(5):1604-1614.

[5] Al-Batah M S, Mrayyen S, Alqaqebah M. Arabic Sentiment Classification using MLP Network Hybrid with Naive Bayes Algorithm[J]. Journal of Computer Science, 2019, 14(8):1104-1114.

[6] H. A, P. O, Ugwu C. An Improved Traffic Crime Predictive System using Multinomial Naive Bayes Text Classification Algorithm[J]. International Journal of Computer Applications, 2019, 182(39):24-31.

[7] Shang F, Li Y, Deng X, et al. Android malware detection method based on naive Bayes and permission correlation algorithm[J]. Cluster Computing, 2017, 21(8):1-12.

[8] Xue Z, Wei J, Guo W. A Real-time Naive Bayes Classifier Accelerator on FPGA[J]. IEEE Access, 2020, 8(99):40755-40766.

[9] Manna S. PERFORMANCE ANALYSIS OF CLASSIFICATION ALGORITHM ON DIABETES HEALTHCARE DATASET[J]. International Journal of Research Granthaalayah, 2017, 5(8):260-266.

[10] Jafarizadeh V, Keshavarzi A, Derikvand T. Efficient cluster head selection using Naive Bayes classifier for wireless sensor networks[J]. Wireless Networks, 2017, 23(3):779-785.