College Students’ Physical Fitness Test Data Analysis, Visualization and Prediction Using Data Mining Techniques

Yanqin Liu*
Department of Sports, Xi’an Fanyi University, Xi’an, China
Email: liuyanqin86@163.com

Abstract. Physical Fitness test analysis is crucial in the field of kinematics and health science, the success of which can potentially help efficiently explore students’ physical issue and promote better physical education as well. Considering that the physical fitness of many college students is declining, it has become urgent to conduct a scientific, rational, and exploratory analysis of physical fitness testing data. In this paper, we visualized and explored several data mining algorithms to learn the relationship among different test items and uncover some potential patterns. Experiments are conducted on a real dataset from department of sport, Xi’an Fanyi University. The experimental results show that Catboost outperforms existing approaches in terms of prediction accuracy and F measure. The promising results will effectively evaluate students’ fitness condition, provide insights on correlation of different test items and assist educators for decision making.

1. Introduction

Physical fitness data analysis is a very important problem in kinematics and health science, the success of which can help both educators and students to better understand their physical fitness and enhance education as well. Although universities take physical fitness tests every year to promote students actively participate in sports activities, the physical fitness of many college students is still not optimistic. According to a research [1], the physique of college students is declining, and there still exists some problems in physical fitness test methods. Now, with the development of big data analytics and massive data storage technology, more and more data is obtained and stored for researchers, effectively visualize and analyze these data can provide new insights for better exploring physical fitness patterns and correlations.

College students physical fitness test is held every year to evaluate students’ physical health, which includes body mass index (BMI), lung’s capacity, sprint (50 meters) and Long-distance run (1000 meters for male and 800 meters for female), standing long jump, sit and reach, pull-up (male)/sit-up (female).

Data Mining approaches are based on the theory of various machine learning algorithms [2]. As a fundamental technique of big data analytics, data mining is useful and effective in many fields such as crime classification [3], crime trends forecasting [4], traffic accident prediction [5], traffic association rules analysis [6], air quality analysis [7] and so on. Data mining is significant in all walks of life, it can help us better explore the unknown fields, but also let us have a deeper understanding of what is known. By utilizing data mining techniques, we can easily identify potential patterns for different genders or majors and how they are related each other. The implementation of data mining and statistical techniques on physical fitness test data will enable the extraction and understanding of
associated patterns and correlations, ultimately enhancing education and students’ physique health at the same time.

In this paper, we employ visualization methods on the data and state-of-the-art classification algorithms are compared for the mining and prediction of students’ physical testing data from Xi’an Fanyi University. We first preprocess the data, which includes data cleaning and attributes normalization, test items are visualized to explore statistical results and correlations between different items. Several of classification methods are employed for physical ability prediction.

2. Related Work
For physical fitness data analysis, researchers have spent some effort by using data mining and visualization techniques. Tong et al. [8] developed a visualization system to analyze physical activity data from sensors. The system can effectively emphasize the browsing of sports activity data and give a clear interpretation of the data by using visualization. Ma et al. [9] utilized a tree-based classification model named ID3 to extract physical fitness patterns hidden in the data. Wang et al. [10] combined big data analytics and data mining to establish a model that can effectively evaluate physical education teaching quality in college. Zhu et al. [11] applied association rule mining algorithms for students majored in physical education for their training, they proposed a method called AD-Apriori. Zhou et al. [12] utilized ensemble learning method on physical health data, by combining different classification algorithms they improved the results. Sarode et al. [13] discovered key factors that affect student athletes’ performance through data mining techniques. Wang et al. [14] conducted a research on finding the association between athletic ability and physical health by using data mining approaches. Based on gradient boosting machine (GBM) method, Guo et al. [15] proposed a model to evaluate people’s physical fitness condition. Besides, they also improved the model by using feature engineering and Bayes hyper-parameter optimization operation. Fang et al. [16] proposed a BP network based model to promote the quality of school physical education, they successfully improved the management ability of school sports department.

In summary, data mining is powerful and strong in handling multi-attributes datasets. By training optimal parameters and using the best algorithms, it can predict the unknown and interpret the trends. So in this paper, we utilized classification methods to predict students’ performance in certain sports items given the results of other test items.

3. Methods
We explored 3 state-of-the-art tree-based classification models for physical ability prediction, i.e. Random Forest, CATBoost and LightGBM.

Random Forest [17] is an ensemble learning method. Unlike traditional ensemble learning, it integrates different decision trees to improve the results. For each decision tree, it utilizes different parts of the data, thus enhance the accuracy.

CatBoost [18] is a very successful technique for solving classification problems. It is a vital classification algorithm in machine learning and has achieved great success in many competitions.

Lightgbm [19] is a highly efficient gradient boosting decision tree. It utilized two techniques: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). As a result, it speeds up the training process of conventional GBDT by up to over 20 times while achieving almost the same accuracy.

4. Experimental Results
4.1. Data and Preprocessing
The data we studied in this article comes from the physical fitness test data of fresh students in 2019 by the Department of Sports, Xi’an Fanyi University. The data contains 6873 test records with 26 attributes.
Before implementing any algorithms on the dataset, we conducted a series of preprocessing tasks in order to get better visualization and classification results. These steps include:

1. We eliminated some attributes like name, student ID, class ID which are privacy information;
2. We eliminated records which are incomplete;
3. We merged 800 meters run and 1000meters run into one attribute named LongRun;
4. We merged sit-ups and pull-ups into one attribute named Others;
5. LongRun is classified as “Fail”, “Pass”, “Good”, “Excellent” according to the grades of long run;

After pre-processing, there remains 6124 items, a sample of the dataset is shown in table 1.

| Gender | Male     | Male     | Female   |
|--------|----------|----------|----------|
| Level  | Pass     | Qualified| Good     |
| Score  | 62.9     | 58.3     | 80.1     |
| Height | 179      | 178      | 164      |
| Weight | 80       | 68       | 62       |
| LungCapacity | 3992 | 3536     | 2815     |
| Sprint | 4.05     | 4.23     | 3.16     |
| LongRun| 7.59     | 8.28     | 7.86     |
| Longjump| 228    | 220      | 170      |
| Other  | 8        | 6        | 32       |
| Lefteyesight | 4.7   | 5        | 4.3      |
| Righteyesight | 4.6   | 5        | 4.2      |
| SitandReach | 17.9  | 23.7     | 20       |
| College | College of English| College of English | College of Commerce |
| Major  | English  | Germany  | Marketing |
| Run Level | Good   | Pass     | Good     |

4.2. Visualization and Analysis

Figure 1 shows the distribution of students taking the test by college and major. From the chart we know that College of English language is the college has the most students taking physical test, while preschool education is the major which has the largest number. Figure 2 demonstrates the level distribution of test results, the graph indicates that 80% of the students pass the test (scores over 60), 9.94% of the student are qualified (scores between 50 and 60), 5.5% of the students are good (scores above 80), 4.08% of the students failed the test, while only 0.44% of the student got excellent (scores above 90) for the test. The results suggest that university should encourage student to take more activity outdoors to strengthen their body.

The score summary by college reveals some interesting facts as shown in figure 3. We can conclude that the overall score of college of public health is the best among other colleges. And to our surprise, college of engineering and technology is the worst while this college has more boys than girls.

In figure 4, we calculated the correlations between different test items, we can know from the results that scores have negative correlation with sprint and long run, which means if one wants to get better test marks, he must improve his sprint and long run scores. Height has positive correlation with long jump. Weight has positive correlation with lung capacity.
4.3. Prediction
For prediction task, we divided long run into four categories i.e., “Failed”, “Pass”, “Good”, and “Excellent” according to long run scores. Then with other test items' results we predict their long run results using classification algorithms. We explored three state-of-the-art models, Random Forest, LightGBM, and CatBoost. Finally we compare the results in terms of accuracy, recall and F1-score. The results are shown on table 2. From the table, we can conclude that CatBoost outperforms the other two models.

Figure 1. Summary of student by college and major using Sunburst plot.

Figure 2. Statistical of levels using pie chart.

Figure 3. The distribution of score by college using box plot.
Figure 4. Correlations of each items using heatmap plot.

Table 2. Comparison of different models by accuracy, recall and F1-score.

| Models       | Accuracy | Recall | F1-score |
|--------------|----------|--------|----------|
| RandomForest | 0.732    | 0.74   | 0.736    |
| LightGBM     | 0.816    | 0.79   | 0.803    |
| CatBoost     | 0.867    | 0.82   | 0.843    |

5. Conclusion
In this paper, we analyzed college students’ physical fitness test data to study the patterns and accomplished two tasks: (1) visualizing the data to get statistical summary of the data and correlations among different test items; (2) Predicting students’ test performance using data mining techniques. We visualized the data to obtain basic information among different colleges and majors and get correlation among different test items. We explored 3 state-of-the-art classification algorithms to predict physical fitness test results. We found CatBoost is the best model in our experiments. In the future, we will establish a data analytics platform to process and analyze the data. Besides, more machine learning and deep learning models will be explored to get better results.

Acknowledgments
This work has been supported by the 13th Five-Year Plan of Education Science in Shaanxi Province (Program No. SGH18H457) and Key Project in Xi’an Fanyi University (Program No. J18A07).

References
[1] Hu Q, Chen L, et al. 2018 Promotion of college students’ physical health: A case study on physical education intervention 2018 9th Int. Conf. on Information Technology in Medicine and Education (ITME) pp 540-544.
[2] Han J, Pei J and Kamber M 2011 *Introduction, Data Mining: Concepts and Techniques* (Elsevier) chapter 1 pp 5-7.

[3] Feng M, Zheng J, et al. 2018 Big data analytics and mining for crime data analysis visualization and prediction *Int. Conf. on Brain Inspired Cognitive Systems* pp 605-614.

[4] Feng M, Zheng J, et al. 2019 Big data analytics and mining for effective visualization and trends forecasting of crime data *IEEE Access* 7 106111-106123.

[5] Feng M, Zheng J, et al. 2020 Towards big data analytics and mining for UK traffic accident analysis, visualization & prediction *Int. Conf. on Machine Learning and Computing* pp 225-229.

[6] Feng M, Zheng J, et al. 2019 Association rule mining for road traffic accident analysis: A Case Study from UK *Int. Conf. on Brain Inspired Cognitive Systems* pp 520-529.

[7] Yadav M, Jain S and Seeja K R 2019 Prediction of air quality using time series data mining *Int. Conf. on Innovative Computing and Communications* pp 13-20.

[8] Tong C, Zhang J, et al. 2019 An interactive visualization tool for sensor-based physical activity data analysis *Proc. of the Australasian Computer Science Week MultiConference* pp 1-4.

[9] Ma G, Zhang L and Li S 2020 ID3-based classification of college students’ physical fitness data *Recent Trends in Intelligent Computing, Communication and Devices* 245-252.

[10] Wang C 2020 Analysis method of college student physical education quality based on big data analysis *The Int. Conf. on Cyber Security Intelligence and Analytics* pp 576-581.

[11] Zhu S 2019 Research on data mining of education technical ability training for physical education students based on Apriori algorithm *Cluster Computing* 22 (6) 14811-14818.

[12] Zhou Z, Zheng J and Meng F 2019 Physical health test analysis based on adaptive proportional fusion ensemble classifier for college students 2019 15th Int. Conf. on Computational Intelligence and Security (CIS) pp 282-286.

[13] Sarode R, Muley A, et al. 2019 Discovery of variables affecting performance of athlete students using data mining *Computational Intelligence in Data Mining* 711 449-458.

[14] Wang J, Xie Z, et al. 2020 Relationship between health status and physical fitness of college students from South China: An empirical study by data mining approach *IEEE Access* 8 67466-67473.

[15] Guo J, Yang L, et al. 2019 An XGBoost-based physical fitness evaluation model using advanced feature selection and Bayesian hyper-parameter optimization for wearable running monitoring *Computer Networks* 151 166-180.

[16] Fang W 2019 Analysis of BP neural network model with human-autonomous devices for school physical education management *Journal of Ambient Intelligence and Humanized Computing* 1-9.

[17] Svetnik V, Liaw A, et al. 2003 Random forest: A classification and regression tool for compound classification and QSAR modeling *Journal of Chemical Information and Computer Sciences* 43 (6) 1947-1958.

[18] Prokhorenkova L, Gusev G, et al. 2018 CatBoost: unbiased boosting with categorical features *Advances in Neural Information Processing Systems* pp 6638-6648.

[19] Ke G, Meng Q, et al. 2017 Lightgbm: A highly efficient gradient boosting decision tree *Advances in Neural Information Processing Systems* 3146-3154.