The influence of team feature vectors on NBA championship team prediction

Xinyi Li*

TianJin University, Tianjin, China

* Corresponding Author Email: xinyili@tju.edu.cn

Abstract. Predicate National Basketball Association championship has always been the focus of people's attention. This paper makes an in-depth analysis of the prediction of the 2021-2022 National Basketball Association championship team. Select 22 possible influencing factors as feature vectors, such as 3-pointers, rebounds, and steals. SVM, KNN, and three different grouping methods were used to predict performance, including accuracy, recall, and F1-score. The result shows that the grouping method of a solid team and a weak team has a remarkable effect. Then principal component analysis was used to visualize the grouping methods of the four categories in two and three dimensions. The results showed that the champion team did not have significant characteristics. However, it was found that the teams were divided into two clusters, so the hypothesis was proposed: whether the league ball style changed with the year. Further verification found that the two clusters were not significantly related to the year. The conclusion is that under this method, the teams entering the playoffs can be selected with high accuracy, that is, the strong and weak teams in the experiment, but the characteristics of the champion team are not evident and difficult to predict.

Keywords: Championship Predication, Machine Learning, SVM, KNN.

1. Introduction

Predicting winners has always been a hot topic in competitive sports such as the Champions League, World Cup, and National Basketball Association (NBA). However, accurate prediction is a challenging task. Take the famous "March Madness" event, where you can win $1 billion from stock god Warren Buffett by guessing the results of all the national college basketball teams. However, since 2014, no one has been able to guess and walk away with $1 billion accurately. Therefore, accurate prediction is not a simple matter. Nowadays, with the development of science and technology and the changes of the times, the traditional way of cognition is gradually replacing. Data analysis are useful in every industry. Due to the practicability of machine learning, its research and application are highly valued in the world. Machine learning can learn and optimize prediction models step by step based on new data during application. Applying machine learning to the prediction of sports champions, mining key information which is hidden in the data from historical data of major events. This can satisfy fans' curiosity, help them predict champions and provide decision-making support for corporate leaders. The role of artificial intelligence is immeasurable.

Wang, K. C., & Zemel, R. focus on the classification of offensive tactics. While good individual players are crucial to a team's success, the strategy a team can execute and understand the opponent's approach can also significantly impact the outcome of a match [1]. Cheng, G., Zhang, Z. et al. applied the maximum entropy principle to construct an NBA maximum entropy (NBAME) model suitable for discrete NBA game statistics [2]. Barnaghi, P., Ghaffari, P., & Breslin, J. G. extracted positive or negative emotions from Tweets to find a correlation with the 2014 World Cup [3]. Thabtah, F., Zhang, L., & Abdelhamid, N. have compared performance and models of different feature sets associated with basketball games to see key features that contribute to better performance, such as the accuracy and efficiency of prediction models [4]. Patel R, Passi K. Sentiment filtered and analyzed the data through natural language processing technology and calculated the emotional polarity according to the inspirational words detected in users' tweets [5]. Basit A, Alvi M B, Jaskani F H, et al. of Bahawalpur compared popular machine learning techniques to predict T20 Cricket World Cup winners. The random forest proved to be the best machine learning algorithm using custom precision

metrics in developed models. Custom accuracy up to 80.86% [6]. Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H., & Zeileis, A. combined two different ranking methods with several other predictors of football game scores in a joint random forest approach. The proposed combination method was applied to the data of the previous two Women's World Cups. Based on the estimated results, the 2019 Women's Football World Cup was repeatedly simulated to get the probability of all teams winning [7]. Fan, M., Billings,et al. have focused on 'social identity theory, the emotional attitudes of live fans, including 'BIRGing' and 'CORFing [8].' Wu, W. considered the impact of injuries as an uncertain factor [9] Bai, Y., & Zhang, X. used k-mean and DPC clustering algorithm to analyze and predict the match results of the last World Cup and established a football field analysis and prediction model. [10]. Georgievski, B., & Vrtagic, S. changed in performance through data simulation to see how a difference in performance is affected by actual points [11]. Stephanos, D. K., Husari, G., Bennett, B. T., & Stephanos, E. according to the location tracking SportVU introducing technology provides the number of field data collection, discussed the athletes all sorts of motion detection and analysis method, and finally put forward a kind of recognition and analysis of the depth of the learning methods, to produce a searchable data set, can assist and automation professionals in the NBA and now most of the work done [12].

Different from the above work, this paper analyzes whether there are significant differences between the strong and weak teams in the NBA by analyzing the correlation between each feature vector of the team and the champion team. A total of 20 years of team data from 2000 to 2021 were collected from basketball-reference, including rebound, assist, 3-point, and other feature factors for analysis. This paper used two classification methods, including Support Vector Machine (SVM) and K-Nearest Neighbor (KNN); The experiment was divided into three groups: four categories (Championship team, playoff teams except for the championship team, the bottom eight teams and the rest teams; champion and others; Strong and weak teams (Playoff teams and non-playoff teams) The experimental results showed that the champion & others had no good distinction, while the other two had significant effects. In addition, Principle Component Analysis (PCA) was used for 2-dimensionnal and 3- dimensionnal visualization of the experimental data. The results showed that there was no outstanding factor among all the features. Moreover, after 3d visualization, it was found that these features were not distinguishable in years.

The rest of the content be divided into three parts. The second section describes the methods that will be used, including support vector machines, KNN, and PCA. In section 3, the performance comparison and visual analysis of the three improved classification methods. In section 4, the work is summarized.

2. Methods

KNN. KNN algorithm is a non-parametric classification algorithm (no training parameters are required), which can be seen from its name (k nearest neighbors means to find K neighbors). It is affiliated with supervised learning, and its core idea is: "If you keep company with people, you will be red." Defining KNN method is simple, easy to understand, easy to implement, without estimating parameters. According to the principle of minority obeying majority, the test sample points are classified into the category with the highest proportion among K points. Because the KNN method of sample is nearest neighbor used in several points, rather than relying on discriminant method to determine the class field, so the KNN method is compared with other methods are more suitable for class field or overlap of sample set. Since this method must calculate all the samples to compare the nearest K points, and thus classify them according to the distance, the large amount of calculation is the most significant and inevitable disadvantage of this method. Another disadvantage is that when the samples are unbalanced, such as the sample size of one class is much larger than that of other classes, the samples of other classes can be ignored, which may also exist. Several methods can be used to measure distance in space, such as Euclidean distance calculation. In fact, in the usual KNN

algorithm, Euclidean distance is always preferred. Take a two-dimensional plane as an example. The Euclidean distance of two points in two-dimensional space is calculated as follows:

$$\rho = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

When extended to multi-dimensional space, the formula becomes:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

SVM shows many unique advantages in solving small samples, nonlinear and high-dimensional pattern recognition. SVM use the analogy of the equation $Ax + By + Cz + D = 0$ for a 3-dimensional plane, and $Ax + By + D = 0$ for a 2-dimensional line. Then, by analogy, SVM can obtain a higher-dimensional space called the hyperplane:

$$\omega^T + b = 0 \tag{3}$$

The goal of SVM is to find a hyperplane, which can effectively solve the problem of dichotomy. Therefore, the closest point of the sample point of each classification to the hyperplane should be found first to maximize the distance between the point and the hyperplane. Established in the hyperplane $\omega^T + b = 0$ cases, $|\omega X + b|$ can represent point x distance to the hyperplane, and by observing the $\omega X + b$ sign symbols are consistent with the class mark y can determine classification is correct, so, SVM can use $y(\omega^T + b)$ plus or minus of the correctness of the classification to determine or said.

PCA is a multivariate statistical analysis method that selects several important variables through the linear transformation of multiple variables. In practical projects, to comprehensively analyze problems, many variables (or factors) related to this are often put forward because each variable reflects some information about the project to varying degrees. However, too many variables can increase the complexity of the subject. In fact, in many cases, there will be a specific correlation between variables, which can be explained as the subject information reflected by the two variables having an inevitable overlap. Principal component analysis is the removal of redundant (closely related) variables and the creation of as few new variables as possible so that these new variables are irrelevant. These new variables as possible reflects the information of the participants, to keep the original information. This is also a mathematical way of reducing dimension.

Selection is the most classic way with F1 (first linear combination, that is the first comprehensive index) of the variance. Var (F1), the greater the F1 to contain more information.

3. Experiment results and analysis

Data sources and preprocessing methods. Integrate valid data collected from the Internet from 2000 to 2020.There are four categories: 1 represents the champion teams, 2 represents the 15 teams that made the playoffs but not the champion teams, 3 represents the weak teams, namely the bottom eight teams, and 4 represents the remaining teams. Twenty-two attributes are selected in the table, which is: MP for minutes played, FG for field goals, FGA for field goal attempts, FG% for field goal percentage, 3P for 3-points field goals, 3PA for 3-point field goal attempts, 3P% for 3-point field goal percentage, 2P for 2-points field goals, 2PA for 2-point field goal attempts, 2P% 2-point field goal percentage, eFG% for effective field goal percentage, FT for free throws, FTA for free throw attempts, FT% for free throw percentage, ORB for offensive rebounds, DRB for defensive rebounds, TRB for total rebounds, AST for assists, STL for steals, BLK for blocks, TOV for tumovers, PF for personal fouls, PTS for points, including 626 pieces of instances. The experiment use three evaluation criteria: precision, recall and F1-score.

Correlation analysis of attributes. The similarity matrix between different data attributes can be obtained based on cosine similarity, as shown in Figure 1. The darker the color, the stronger the correlation, and vice versa. Therefore, the diagonal element has the darkest color because the corresponding two elements are the same. In addition, FG and FGA, 3P and 3PA are also highly

correlated, which is very reasonable. For example, FG and FGA mean-field goal and field goal attempts. When the value of field goal attempts increases, the typical field goal situation will increase accordingly.

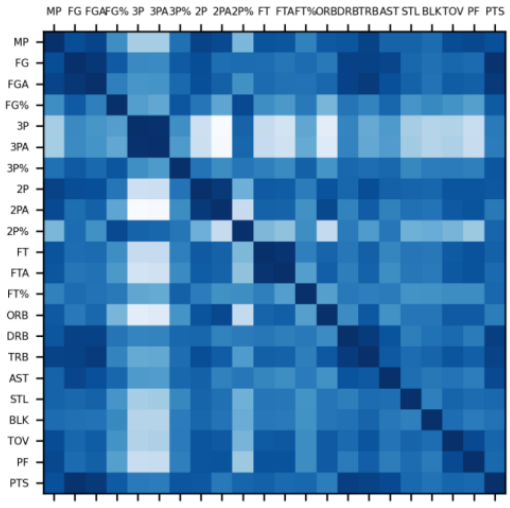

Figure 1. Correlation between attributes

Performance comparison of different classification methods. The specific content of the experiment can be divided into three parts, which are classified in three ways, respectively. (Inside the brackets is the classification code). The first is the Champion team (0) or other teams except for the champion team (1); The second is the Strong team (team entering the playoffs (0)) or weak team (team not entering the playoffs (1)). The last one is the most original four categories: the champion team (0), playoff teams except the championship teams (1), the lowest eight teams (2), and other teams (3).

Table 1. Performance of Champion or Other teams

classes	SVM			KNN		
	precision	recall	F1-score	precision	recall	F1-score
0	0	0	0	0	0	0
1	0.96	1	0.98	0.96	0.99	0.98
Macro avg	0.48	0.50	0.49	0.48	0.50	0.49

As shown in Table 1, whether SVM or KNN, the accuracy of predicting the champion team is 0. There are two possibilities for this result. The imbalance of data causes one. After all, there is only one champion team every year, while there are nearly 30 other teams. In the second case, the champion team has no significant characteristics, which can help the model distinguish and make it stand out from many teams. To verify the reason, we carried out the second classification: strong team or weak team prediction.

Table 2. Performance of Solid Teams (0) and Weak Teams (1)

classes	SVM			KNN		
	precision	recall	F1-score	precision	recall	F1-score
0	0. 71	0.84	0.77	0.74	0.77	0.75
1	0.80	0.65	0.71	0.75	0.73	0.74
Macro avg	0.76	0.74	0.74	0.75	0.75	0.45

As shown in Table 2, Under this classification method, there are 16 solid teams and 14 weak teams, which solves the problem of data imbalance under the former classification method. It can be seen

that the effect is remarkable. The precision of both methods is more significant than 0.70, and the performance of KNN is better than SVM, indicating that the data used is a high-quality evaluation index to distinguish strong teams from weak teams. Next, four classifications are further carried out.

Table 3. Performance of Four categories

classes	SVM			KNN		
	precision	recall	F1-score	precision	recall	F1-score
0	0. 00	0.00	0.00	0.00	0.00	0.00
1	0.59	0.92	0.72	0.63	0.88	0.73
2	0.68	0.45	0.54	0.69	0.66	0.68
3	0.33	0.12	0.18	0.17	0.04	0.07
Macro avg	0.40	0.37	0.36	0.37	0.40	0.37

As shown in Table 3, the distinction between precision strength is apparent. However, we still cannot rule out the impact of data imbalance on performance. The precision of champion teams is still as low as 0, while the precision of 1 and 2 is higher. However, there are few absolute strong team samples for the particular field of competitive sports, so data imbalance is a common situation. At the same time, due to the significant differences in the team-building concepts and styles of solid teams in different periods, it is difficult to solve the problem of data imbalance in this field by general data enhancement methods. In this regard, we carried out a further visual analysis.

Visual analysis of different classification methods. The first step is to standardize features to have an average of 0 and a variance of 1. This process is essential because principal component analysis works by maximizing the variance explained by principal component analysis. Some features will naturally have higher variances because they are not standardized; for example, a distance measured in centimeters will have higher variances than the same distance measured in kilometers. PCA will be "attracted" to those features with high variance in the absence of scaling features. According to different categories, use the Boolean index to obtain different categories of data, different categories of data with different colors to mark, to carry out visual operations.

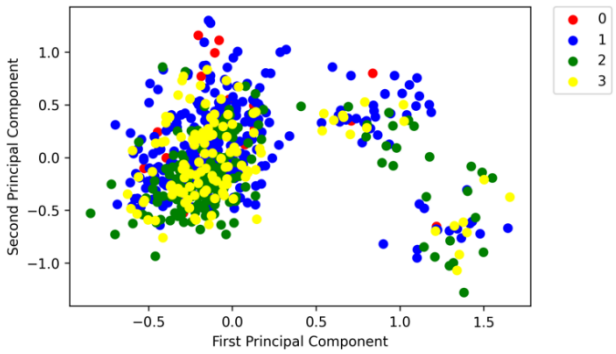

Figure 2. Four categories of two-dimensional visual

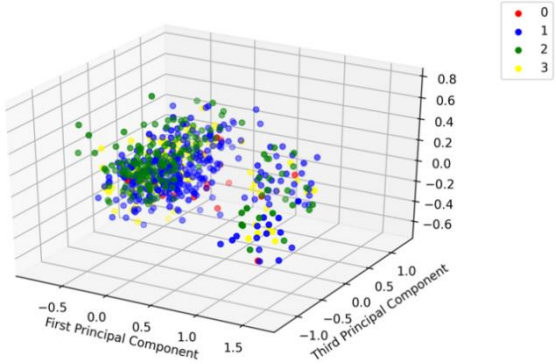

Figure 3. Four categories of three-dimensional visual

As shown in Figure 2 and Figure 3, the four-color categories are not easy to distinguish. There are no apparent features and specific points in the two-dimensional and three-dimensional views. However, through careful observation, we can find that all the data show two clusters in the three-dimensional view of solid and weak teams. We preliminarily believe that the two clusters are related to the year (the overall style change of the team in the league, but after further confirmation, it is found that the formation of the two clusters has nothing to do with the year. This phenomenon confirms our analysis in Table 3 that the style difference of the team is not the decisive factor in the strength of the team).

4. Conclusion

There are many possible attributes of a team in the national basketball association. The attributes of a championship team are of particular concern. However, with the change of league style of play, the difference in team style is not the decisive factor in team strength. Therefore, it is not advisable to use SVM, KNN, and PCA to analyze the characteristic factors of the champion team to predict the champion team.After all, there is only one champion team, compared with 29, and the imbalance makes predicting the winner extremely difficult. And in the actual situation, there are many uncertain factors, such as the injury of players, which are uncontrollable, but the impact is very large. Predicting championship teams is a great curiosity for viewers and a passion for basketball fans. It abstractly displays an unknown result with the constructed model and code and connects the developed technology with reality.

References

[1] Wang, K. C., & Zemel, R. (2016, March). Classifying NBA offensive plays using neural networks. In *Proceedings of MIT Sloan Sports Analytics Conference* (Vol. 4).

[2] Cheng, G., Zhang, Z., Kyebambe, M. N., & Kimbugwe, N. (2016). Predicting the outcome of NBA playoffs based on the maximum entropy principle. *Entropy*, *18*(12), 450.

[3] Barnaghi, P., Ghaffari, P., & Breslin, J. G. (2015, August). Text analysis and sentiment polarity on FIFA world cup 2014 tweets. In *Conference ACM SIGKDD* (Vol. 15, No. 2015, pp. 10-13).

[4] Thabtah, F., Zhang, L., & Abdelhamid, N. (2019). NBA game result prediction using feature analysis and machine learning. *Annals of Data Science*, *6*(1), 103-116.

[5] Patel R, Passi K. Sentiment analysis on Twitter data of world cup soccer tournament using machine learning[J]. IoT, 2020, 1(2): 218-239.

[6] Basit A, Alvi M B, Jaskani F H, et al. ICC T20 Cricket World Cup 2020 Winner Prediction Using Machine Learning Techniques[C]//2020 IEEE 23rd International Multitopic Conference (INMIC). IEEE, 2020: 1-6.

[7] Groll, A., Ley, C., Schauberger, G., Van Eetvelde, H., & Zeileis, A. (2019). Hybrid Machine Learning Forecasts for the FIFA Women's World Cup 2019. *arXiv preprint arXiv:1906.01131*.

[8] Fan, M., Billings, A., Zhu, X., & Yu, P. (2020). Twitter-based BIRGing: Big Data analysis of English national team fans during the 2018 FIFA World Cup. *Communication & Sport*, *8*(3), 317-345.

[9] Wu, W. (2020). Injury Analysis Based on Machine Learning in NBA Data. *Journal of Data Analysis and Information Processing*, *8*(4), 295-308.

[10] Bai, Y., & Zhang, X. (2021). Prediction Model of Football World Cup Championship Based on Machine Learning and Mobile Algorithm. *Mobile Information Systems*, *2021*.

[11] Georgievski, B., & Vrtagic, S. (2021). Machine learning and the NBA Game. *Journal of Physical Education and Sport*, *21*(6), 3339-3343.

[12] Stephanos, D. K., Husari, G., Bennett, B. T., & Stephanos, E. (2021, April). Machine learning predictive analytics for player movement prediction in NBA: applications, opportunities, and challenges. In *Proceedings of the 2021 ACM Southeast Conference* (pp. 2-8).