Research on CBA League Championship Based on LSTM Neural Network

Chenkai Ma[a], Guozhang Zhao, Wenxian Feng

Jilin University, Changchun, China

[a]15110778625@163.com

Keywords: Monte Carlo tree, computer simulation, system clustering, LSTM neural network, principal component analysis

Abstract: With the continuous development of the CBA league, the demand for team level analysis and evaluation is also increasing. This paper aims at the level of each team's history, combined with the actual situation of the domestic CBA league, gives the mathematical model and programming realization, making the calculation of the probability of winning, predicting the team's ranking, analyzing the team's level and providing reasonable suggestions for the team. may. This paper first uses the LSTM neural network to predict the score of each team's next game, and then uses the principal component analysis method to comprehensively consider the prediction score, the standard deviation and the mean of each team's historical data. Sort the 14 teams by these three indicators. Finally, a comprehensive evaluation of 14 team levels was obtained.

1. Introduction

2019 is undoubtedly the year of radical reform for Chinese basketball. The CBA has changed a lot in the new season, whether it is the number of foreign aid replacements, the length of the league, or the increase in the number of playoffs. These changes have prompted the team to use more young players and rotate personnel. The competition between the various teams is also more intense, and the team seeks to evaluate and guide itself in a scientific way. So people feel that the CBA is getting better and better, and it is predicted that the results of the CBA game will once be on the hot search.

2. Related work

2.1 Restatement of the Problem

There are 14 teams in the basketball game, the game is divided into two stages of the regular season and the playoffs. Each game must have a winner and loser, and each team number is fixed. Under the condition that each team's 100-point comprehensive score data is known, the probability of winning each team is estimated, and the top four teams are predicted. The level of 14 teams was qualitatively analyzed under the condition of a comprehensive score of nearly 100 historical records of 14 teams.

2.2 Problem analysis

The title requires estimating the probability of each team winning the championship based on historical score data for each team and predicting the top four teams. According to the historical score data of each team, the winning percentage of the match between any two teams can be calculated separately. Using the winning rate, based on the actual rules of the regular season and the playoffs, a mathematical model can be established to obtain the probability of winning each team. At the same time, the four teams with the highest probability of entering the top four are calculated as the top four of the forecast. However, because the calculation is too large, the computer can simulate the actual game to obtain the approximate probability of winning, and predict the top four. The name of the team. In order to test the correctness of the prediction of the top four models, the R language can be used, and the clustering method in unsupervised learning is used to program the system clustering method to calculate the results and verify.

DOI: 10.25236/mmmce.2019.154

2.3 Assumptions and Justifications

In order to simplify the problem and make it easy to solve, consider the actual situation and the accuracy required to solve the problem.

In this case, the following reasonable assumptions are made:

1) Assume that the historical scores of all teams given are given under normal competition (eg no fakes).

2) Suppose the article uses the actual rules of the domestic CBA league as a template, combined with the established conditions in the title, to formulate the following competition rules. The game is divided into two stages, the regular season and the playoffs. Each game must have a winning and losing, and each team is fixed in number. In the regular season, all participating teams will play a double-loop competition, and the total number of regular seasons will be determined according to the ratio of the number of wins to the number of negative games. The winners will be ranked first. The top eight players in the regular season are required to enter the playoffs, and the playoffs are cross-takeout. Among them, the quarter-finals and the semi-finals all adopt the five-game three-win system, and the championship finals adopts seven-game four-win system.

3. The Model

3.1 Part one

● Mathematical model for calculating the probability of team i winning

(1) When calculating the team i and team j matches, the victory probability Pij of team j

Let the team i compete with the team i at all the level of play at a fixed level of the fixed level, and think that the team with a high overall score can win. This process is repeated until the overall score of all the presence levels of team i is traversed. Pj is obtained by substituting the obtained data into the calculation.

$$P_{ij} = \mathrm{s} \div \left(\mathrm{hnum}_i \times \mathrm{hnum}_j\right)$$

Among them, the s here indicates the number of times the team i wins in the $\left(\mathrm{hnum}_i \times \mathrm{hnum}_j\right)$ team i and team j competitions.

(2) Calculate the probability that team i wins in the regular season

- When s=26:

Equivalent to the team i all win, that is, l=0. According to the knowledge of random mathematics, the probability is:

$$\prod_{k=1}^{i-1} P_{ik}^2 \times \prod_{h=i+1}^{(cnum-1)\times 2} P_{ih}^2$$

- When s=25:

Equivalent to team i has one and only one failure, that is, l=1.

Suppose team i loses when playing against team g (g is any integer from 1 to 14, and $\mathrm{g} \neq \mathrm{i}$).

When the values of g are different, replacing the corresponding $\mathrm{P}_{ik}(\mathrm{k} = \mathrm{g})$ in the equation with $\left(1 - P_{i\mathrm{g}}\right)$, respectively, will result in 13 equations of the following form.

$$\prod_{k=1}^{\mathrm{g}-1} P_{ik}^2 \times \left(1 - P_{i\mathrm{g}}\right) \times P_{i\mathrm{g}} \times \prod_{q=\mathrm{g}+1}^{i-1} P_{iq}^2 \times \prod_{h=i+1}^{(cnum-1)\times 2} P_{ih}^2$$

$$\prod_{k=1}^{i-1} P_{ik}^2 \times \prod_{h=i+1}^{\mathrm{g}-1} P_{ih}^2 \times \left(1 - P_{i\mathrm{g}}\right) \times P_{i\mathrm{g}} \times \prod_{q=\mathrm{g}+1}^{(cnum-1)\times 2} P_{iq}^2$$

The probability is obtained by adding all the 13 expressions obtained.

- When s=24, 23...0:

As each s number decreases, the number of team i failures increases. For each occurrence of a decrease in s, a replacement operation similar to the previous step is performed for the probability formula obtained in the previous step. Adding the results of the replacement adds the probability that s will get a specific value. The iterative calculation is continued until s = 0, and the probability that team i wins the s field in the regular season (s is an integer from 0 to 26) is obtained.

(3) The calculation team's ranking in the regular season is the top eight, thus obtaining the probability of each ranking of the qualifications for participating in the playoffs.

A method for calculating the probability P that the team i obtains the alphath in the regular season

According to the different values of the number of wins obtained when the team β obtains the alphath in the regular season, the probability of each β is calculated separately. Add these probabilities to get the solution.

Repeat the previous step to get the team's ranking in the regular season as the top eight, the probability of each ranking.

(4) Calculating the winning probability of the team

Calculate the probability of team i qualifying for the second round after the playoffs according to the rules of the playoffs

Analyze the actual game situation, using knowledge such as the full probability formula, you can get the following formula to calculate the probability:

$$\sum_{j=1}^{i-1}\sum\nolimits_{k=1}^{8}\left\{P_k \times P'_{9-k} \times \left[P_{ij}^3 + C_3^1 \times P_{ij}^3 \times \left(1-P_{ij}\right) + C_4^2 \times P_{ij}^3 \times \left(1-P_{ij}\right)^2\right]\right\}$$
$$+\sum_{r=i+1}^{cnum}\sum\nolimits_{h=1}^{8}\{P_h \times P'_{9-h} \times [P_{ir}^3 + C_3^1 \times P_{ir}^3 \times (1-P_{ir}) + C_4^2 \times P_{ir}^3 \times (1-P_{ir})^2]\}$$

According to the rules of the playoffs, calculate the probability that the team i will advance to the third round after the playoffs. Analogous to the calculation method of the previous step, the following calculation formula can be used to calculate the probability:

$$\sum_{j=1}^{i-1}\sum_{k=1}^{4}\left\{P''_k \times P'''_{5-k} \times \left[P_{ij}^3 + C_3^1 \times P_{ij}^3 \times \left(1-P_{ij}\right) + C_4^2 \times P_{ij}^3 \times \left(1-P_{ij}\right)^2\right]\right\} +$$
$$\sum_{r=i+1}^{cnum}\sum_{h=1}^{4}\{P''_h \times P'''_{5-h} \times [P_{ir}^3 + C_3^1 \times P_{ir}^3 \times (1-P_{ir}) + C_4^2 \times P_{ir}^3 \times (1-P_{ir})^2]\}$$

Calculate the probability of team i winning according to the rules of the playoffs, Analogy to the calculation method in the first two steps, you can get the following formula to calculate the probability:

$$\sum_{j=1}^{i-1}\sum\nolimits_{k=1}^{2}\left\{P''_k \times P'''_{3-k}\right.$$
$$\left.\times\left[P_{ij}^4 + C_4^3 \times P_{ij}^4 \times \left(1-P_{ij}\right) + C_5^3 \times P_{ij}^4 \times \left(1-P_{ij}\right)^2 + C_6^3 \times P_{ij}^4 \times \left(1-P_{ij}\right)^3\right]\right\}$$
$$+\sum_{r=i+1}^{cnum}\sum\nolimits_{h=1}^{2}\{P''_h \times P'''_{3-h}$$
$$\times [P_{ir}^4 + C_4^3 \times P_{ir}^4 \times (1-P_{ir}) + C_5^3 \times P_{ir}^4 \times (1-P_{ir})^2 + C_6^3 \times P_{ir}^4 \times (1-P_{ir})^3]\}$$

- Predicting the mathematical model of the top four teams

Looking at step 1) in step (4) above, we know that the probability of the team making the second round after the playoffs is the probability that the team will become the top four. Compare the size of this probability value and sort it. Take the top four, which is the predicted top four teams.

Key Points and Difficulties in Realizing the Algorithm of Calculating the Probability of Winning

(1) Difficulties in the algorithm

Analysis of the calculation process, we can see that the difficulty of the algorithm lies in the probability that the calculation team will win the game in the regular season. This is not only because the amount of calculation involved in this step is much larger than the other steps, but more importantly the complexity of the actual game process itself.

The algorithm implementation of the other steps can be easily implemented by programming according to the language description of the mathematical model.

(2) The focus of the algorithm

The focus of this algorithm is also on the probability that the computing team will win in the regular season. There are three reasons for this:

1) This calculation process is located at the relative lower level of the entire calculation, which makes the frequency of use of the data obtained at this step higher than other calculation processes. According to the actual situation analysis, the data generated in this step does have a higher usage rate.

2) This calculation process essentially simulates the outcome of all teams in the regular season. This physical process occupies most of the entire physical process, making him more important than other processes.

3) Although it is not necessary to retain the intermediate value when calculating the probability, in order to enhance the practicability, functional diversity, maintainability and efficiency of the algorithm, the algorithm also supports the retention of intermediate values. Due to the large proportion of the physical process, the number of intermediate values generated during the calculation is also large.

Use the tree as a data structure for calculating and preserving the probability that the team will win in the regular season

1) Monte Carlo tree [1]

If you use the traditional calculation method to calculate the probability of winning, you will face the dilemma of calculation. On the one hand, in order to get the most accurate probability possible, it is necessary to simulate the scene of the real game to calculate the probability; on the other hand, because the amount of calculation involved in the simulation is too large, it is necessary to avoid the scene when simulating the real game. A closer look at this dilemma reveals that the most critical issue when using traditional computational methods is the theoretical need to simulate actual physical processes and the practical infeasibility.

One of the greatest advantages of the Monte Carlo tree search method is that the optimal solution can be obtained because the actual physical process can be realistically simulated. It is possible to simulate the actual physical process that is the greatest characteristic of the Monte Carlo tree. This article is inspired by the use of a tree similar to the Monte Carlo tree as a data structure used to simulate the actual game.

2) Huffman Tree

According to the pre-established coding rules, the Huffman tree can be used for lossless compression of data. Among them, one of the biggest theoretical cores of the Huffman tree is to obtain Huffman coding by using the links of the nodes in the Huffman tree to the two child nodes.

In order to reduce the computational complexity of each time finding the probability that the team will lose to a specific numbered team in the regular season, and in order to improve the rationality of the storage and the gracefulness of the code, this paper uses Huffman in the established tree. The idea of the tree, with '0' and '1', respectively represents the failure and victory of a team. Combine all '0's with '1' to make a string that represents the winnings of a team. By storing the string in the node of the tree, it is possible to store the winning and losing situation of any team to any stage of the regular season. This facilitates future access and extends the functionality of the algorithm.

- Algorithm description of forest construction

(1) Construction of the forest

This article will build a tree for each team. Since the team is cnum, the algorithm will get the num tree. Subsequent analysis will be built into a collection of num trees, which is the forest that you expect to acquire. Thus, as long as each tree is obtained, the forest for subsequent analysis is naturally obtained.

(2) Tree construction

1) Symmetry between different trees

According to the actual situation, due to the symmetry between different teams, this paper only needs to give a tree construction method for a certain team.

2) Nodes in the tree

In each node in the tree, there are the following variables:

a) A string variable: team. This variable is used to save the number of the team represented by this node. For the root node, this variable is used to save the number of the team's rootteam. The team refers to the team that built the tree for that team. For other nodes, this variable is used to save the team otherteam that matches the team saved by the root node of the tree.

b) A double-precision floating-point variable: currentPij. This variable is used to save Pij with ootteam as the team and otherteam as the team j to facilitate the calculation when the node is generated. The analysis and calculation process shows that this variable has a special position in the calculation of the model and is very important. Therefore, this paper gives all possible values, that is, the values of all Pij, as shown in the following table:

c) A pointer to the left son node: leftchildren. This document stipulates that the process from the parent node to the left son node represents the process of the team rootteam saved by the root node and the team otherteam saved by the left son node. The result of the game was winning by rootteam.

Table 1. The winning percentage of the match between any two teams (the winning percentage of the vertical axis versus the horizontal axis)

i \ j	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A		0.281	0.255	0.335	0.415	0.330	0.383	0.321	0.369	0.331	0.321	0.194	0.393	0.175
B	0.719		0.444	0.537	0.698	0.561	0.620	0.534	0.653	0.581	0.557	0.398	0.648	0.375
C	0.745	0.556		0.583	0.725	0.604	0.653	0.581	0.688	0.623	0.606	0.464	0.680	0.445
D	0.665	0.463	0.417		0.634	0.517	0.570	0.494	0.596	0.530	0.513	0.373	0.594	0.350
E	0.585	0.302	0.275	0.366		0.383	0.445	0.347	0.444	0.379	0.358	0.199	0.448	0.161
F	0.670	0.439	0.396	0.483	0.617		0.555	0.473	0.571	0.507	0.489	0.344	0.570	0.319
G	0.617	0.380	0.347	0.430	0.555	0.445		0.420	0.505	0.448	0.432	0.286	0.519	0.257
H	0.679	0.466	0.419	0.506	0.653	0.527	0.580		0.611	0.537	0.524	0.370	0.611	0.341
I	0.631	0.347	0.312	0.404	0.557	0.429	0.495	0.389		0.433	0.405	0.243	0.500	0.205
J	0.669	0.419	0.377	0.469	0.621	0.493	0.552	0.463	0.567		0.478	0.316	0.570	0.282
K	0.679	0.443	0.394	0.487	0.642	0.511	0.568	0.476	0.595	0.522		0.344	0.589	0.314
L	0.807	0.602	0.536	0.627	0.801	0.656	0.714	0.630	0.757	0.684	0.656		0.744	0.482
M	0.607	0.352	0.320	0.406	0.552	0.430	0.481	0.389	0.500	0.430	0.411	0.256		0.214
N	0.825	0.625	0.555	0.650	0.839	0.681	0.743	0.659	0.795	0.718	0.686	0.518	0.786	

d) A pointer to the right son node: righchildren. This document stipulates that the process from the parent node to the right son node represents the process of the team's rootteam saved by the root node and the team otherteam saved by the right son node. The result of the match was won by otherteam.

e) A double-precision floating-point variable: currentProbability. This variable is used to save the probability that the ootteam will be determined by several opponents and the result of the match is also determined. The initial value of this variable is 1.0. Each time you go from the parent node to the left son node, the value of currentProbability*Pij is calculated (where i is the number of the rootteam, j is the number of the otherteam), and is reassigned to currentProbability; each time from the parent node to the right son At the time of the node, the value of currentProbability*(1-Pij) is calculated (where i is the number of the rootteam and j is the number of the otherteam) and is reassigned to currentProbability.

f) A string variable: process. This variable is used to save the results of the game between rootteam and otherteam. If the rootteam wins, add 1 to the end of the string. If the rootteam is defeated, add 0 to the end of the string.

3.2 Result

The same results were obtained using the three methods described above, respectively. The probability of winning the championship is shown in figure 1:

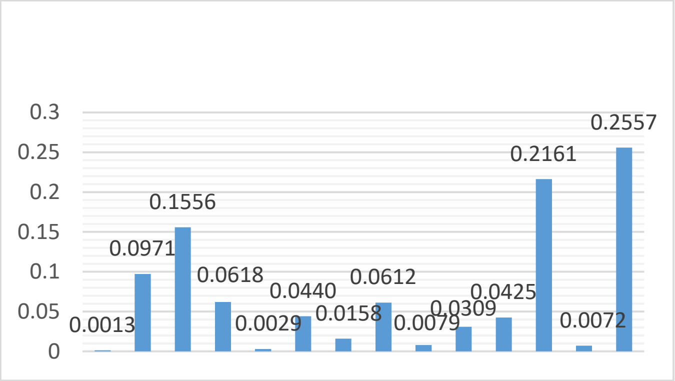

Figure 1

The team that is most likely to be the top four is: N team, L team, C team and B team.

3.3 Compare and test results by cluster analysis

In this paper, the clustering method in unsupervised learning [2] is used to realize the method of squared deviation in system clustering using R language programming [3][4], which is applied to this problem and is applied to the first four predictions. The test of the name problem, the results obtained are shown below:

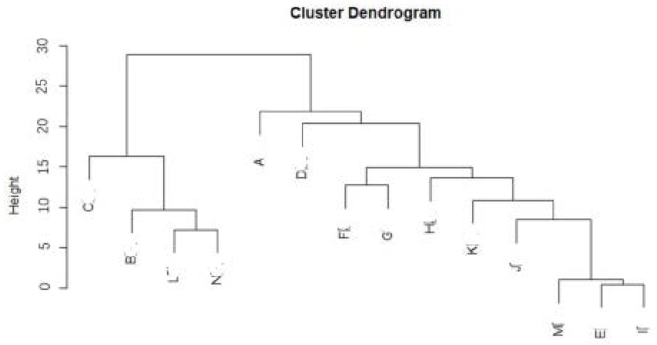

Figure 2 classifies teams by the method of squared deviation in system clustering

As can be seen from the figure, the results are divided into two categories. The C, B, L, and N on the left side of the figure are the top four teams, and the top four balls predicted by theoretical calculations and computer simulations above. The team matched and verified the correctness of the top four models predicted above.

References

[1] D. E. KNUTH The TEXbook the American Mathematical Society and Addison- Wesley Publishing Company , 1984-1986.

[2] Lamport, Leslie, LATEX: “ A Document Preparation System “, Addison-Wesley Pub- lishing Company, 1986.

[3] http://www.latexstudio.net/

[4] http://www.chinatex.org/

[5] Zheng Xiaoping, Liu Mengting. Forecasting model for pedestrian distribution under emergency evacuation [J]. Reliability Engineering and System Safety (S0951-8320), 2010, 95(11): 1186-1192.