Pisco Med Publishing

Original Research Article

Data Mining of NBA

Jiamin Liu

Noncommissioned Officer Academy of PAP,Hangzhou Zhejiang 310000

Abstract: With the development of scientific and technology, data mining is more and more popular. Data mining is the process of finding hidden information in the material such as trend, pattern and relationship. It is gradually applied in various kinds of fields. More and more sports event such as NBA use this technology to predict team's performance. Data mining has been applied in NBA for many years. And it has brought many benefits to NBA. This project will introduce data mining's application in NBA. The project is divided into five parts: introduction, literature search and selection, recommendation of suitable algorithms for the data mining tasks, sample analysis of data input, data preprocessing, data output, application.
Keywords: NBA; Classification; Data mining

Introduction:

When data mining is applied in NBA in the first time, many professionals doubt it. But data mining is very popular in NBA nowadays. It only takes less than ten years from suffering doubt to getting wide application. It has been confirmed that data mining played an important role in 2010. According to the wall street journal's survey, more than half of NBA teams have at least one data analyst. These teams all have good grades. The most representative team is Oklahoma Thunders. Its general manager is a professional of data analysis. It behaves greatly in every season.

Data mining's functions mainly behave in two parts. On the one hand, teams can know which attribute is useful to wining rate through data mining. Then teams can make use of this attribute to increase wining rate in next season. On the other hand, coaches can matchup players better through data mining. According to data mining's results, coaches can get players' advantage and disadvantage. Thereby, coaches can make use of every player's advantage to arrange strategy and avoid players' disadvantage.

Literature search and selection:

We searched the data from the following website. We can search all data about different teams, different seasons and players' information and so on. Then, we selected about 30 teams' data . These data contains the following attributes: age, 2-point fields goal, offensive rebounds, defensive rebounds and so on. We input these data to excel. Then we produce data set. There are 810 instances in total. At last, we enter these data into WEKA.

1. Recommendation of suitable algorithms for the data mining tasks

In order to predict the winning percentage of a team for a new season based on the player's performance in the last season, we can use Classification to finish this job.

Specifically, we will use the three common Classification algorithm to build the model. There are Naïve Bayes, J48, and SMO. These algorithms will be tested in WEKA software, and choose the best algorithm based on the training and testing result.

2. Sample analysis of data input, data preprocessing, data output

2.1 Data Input[1]

Attributes	Attributes details
Age	Average age of players
3P	3-Point Field Goals
3PA	3-Point Field Goals Attempts
2P	2-Point Field Goals
2PA	2-Point Field Goals Attempts
FT	Free Throws
FTA	Free Throws Attempts
ORB	Offensive Rebounds
DRB	Defensive Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
PF	Personal Fouls
Class	Classification based on the winning percentage

2.2 Preprocessing:

Choose data from the past thirteen seasons, and do preprocessing as followed:

Eliminate the data in season 1998-1999, 2011-12 due to the incomplete season at that year.

Eliminate the attributes that cannot contribute to the winning percentage, such as Team name, Season, League, Total minutes.

Eliminate the attributes that are not independent and are calculated by other attributes, such as Total Rebounds, which is the sum of Offensive Rebounds and Defensive Rebounds; 3-Point Field Goals Percentage, which is being calculated by 3-Point Field Goals and 3-Point Field Goals Attempts; Points, which is the sum of 3-Point Field Goals and 2-Point Field Goals.

For the Class,

The winning percentage above 60% converted to Class A

The winning percentage between 40% and 60% converted to Class B

doi: 10.18282/l-e.v10i1.2199

The winning percentage less than 40% converted to Class C

After preprocessing the data, it can get 810 instances with 14 attributes and 3 class. Then Input those data into WEKA software.

2.3 Classification algorithm

Run the WEKA by using the Classification algorithms, set the test options to the Percentage split, the value is 66%. The results of the three common classifiers are shown as below:

2.3.1 Naïve Bayes

false

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         173               62.9091 %
Incorrectly Classified Instances       102               37.0909 %
Kappa statistic                          0.4122
Mean absolute error                      0.3051
Root mean squared error                  0.3956
Relative absolute error                 70.3715 %
Root relative squared error             85.3926 %
Total Number of Instances              275

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.651    0.384      0.6       0.651     0.625      0.698    B
                 0.587    0.1        0.688     0.587     0.633      0.876    A
                 0.634    0.127      0.634     0.634     0.634      0.86     C
Weighted Avg.    0.629    0.24       0.633     0.629     0.629      0.788

=== Confusion Matrix ===

  a  b  c   <-- classified as
 84 20 25 |  a = B
 30 44  1 |  b = A
 26  0 45 |  c = C
```

2.3.2 J48

false

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         152               55.2727 %
Incorrectly Classified Instances       123               44.7273 %
Kappa statistic                          0.3164
Mean absolute error                      0.3185
Root mean squared error                  0.4978
Relative absolute error                 73.4529 %
Root relative squared error            107.4646 %
Total Number of Instances              275

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.504    0.295      0.602     0.504     0.549      0.605    B
                 0.587    0.225      0.494     0.587     0.537      0.701    A
                 0.606    0.172      0.551     0.606     0.577      0.742    C
Weighted Avg.    0.553    0.244      0.559     0.553     0.553      0.667

=== Confusion Matrix ===

  a  b  c   <-- classified as
 65 37 27 |  a = B
 23 44  8 |  b = A
 20  8 43 |  c = C
```

2.3.3 SMO

false

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         185               67.2727 %
Incorrectly Classified Instances        90               32.7273 %
Kappa statistic                          0.4818
Mean absolute error                      0.3046
Root mean squared error                  0.3956
Relative absolute error                 70.2676 %
Root relative squared error             85.399  %
Total Number of Instances              275

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.69     0.342      0.64      0.69      0.664      0.674    B
                 0.64     0.105      0.696     0.64      0.667      0.847    A
                 0.676    0.093      0.716     0.676     0.696      0.86     C
Weighted Avg.    0.673    0.213      0.675     0.673     0.673      0.769

=== Confusion Matrix ===

  a  b  c   <-- classified as
 89 21 19 |  a = B
 27 48  0 |  b = A
 23  0 48 |  c = C
```

2.3.4 Algorithms comparison

Algorithm	Accuracy
Naive Bayes	62.9091%
J48	55.2727%
SMO	67.2727%

Based on the testing results, the SMO classifier is the best algorithm of Classification with the highest accuracy.

2.3.5 Data output:

Based on a team's new player average data, it can predict the team's performance by classy the team with A, B, or C.

3. New applications

In our case, the winning rate analyzed by data mining technology is used for two purposes. First of all, it can predict the winning rate of next season. In the previous analysis, we conclude that accuracy rate of the classification is up to 70%. Based on analyzing the winning rate of this season, the coach may know the weaknesses of the team and adjust the strategy, arrange the players for the next season. Apart from this, they need to use these data to understand the characteristics of the players and help them avoid the weaknesses. In addition, in order to maintain long-term competitiveness, the coach can develop tactics to promote the development of team according to the composition of the team.

Secondly, it has great impact on player transition. The player's ability and the team cooperation translate into the data on the court and the coach can manage the team and bring players based on these data. In addition, based on the data of a player, it can predict the performance of a new combination and promote the transition of players. For example, Phil Jackson applied a "triangle" strategy in Bulls and Lakers' systems and achieved success. Kings fully used "Princeton system" and the team had outstanding performance in the court. Suns created 7 seconds fast-break by using of Nash and Amare Stoudemire's coordinate. In NBA, the teams need to transfer the players in each seasons, the players' adjustment may has huge impact on the result. Sometimes a famous player cannot work well in a team, but if he changes another team the performance will be good due to the coordination of players.

4. Conclusion

In conclusion, we select the previous data of NBA and rank the team based on their performance. We use the data to predict the winning rate of the team in the next season and promote the transition of players by using data mining technology classification. The results can be applied in business fields, such as exploring the player's business value and developing games.

Reference:

[1]"[NBA China Official Website | League Player Data Rankings]",https://m.china.nba.com/statistics/