"Moneyball" in NBA to predict the performance of the players

Kirill Novik
kirill.novik@colorado.edu

Krishna Chaitanya Sripada
krishna.sripada@colorado.edu

Yu-Ching Kuo
yuching.kuo@colorado.edu

Task Overview

The first goal of this project is to assess the goodness of a team based on whether it reaches the playoff or not. By finding the attributes that represent good teams, we can achieve our primary goal which is evaluating the performance of individual players. To accomplish this task, we need to identify parameters that affect the goodness. The datasets we are working with have 26 parameters to choose from, which after reduction becomes only 9. A decision tree approach is applied to find out the most important attributes form the 9 attributes we choose. When the most important attributes are determined, we merge them into one single factor after applying the min-max normalization to these attributes and use it for describing the goodness of a team.

Keywords

Data Mining, Information Gain, Decision Tree, Min-Max Normalization, Accuracy

Decision Tree Approach

From the 26 attributes, we reduce the attributes to 9 attributes which are FG%, ORB, DRB, AST, STL, BLK, TOV, PF and PTS/G and rule out the attributes that are either redundant or relevant. Here the decision is made based on if a team can make to playoff or not. Information gain for all the 9 attributes we choose are calculated to determine the importance of each attribute. We split the past 30 seasons of NBA league into 20 seasons as a training dataset and 10 seasons as the testing database which has been shown in Figure 1. Within the training dataset, we calculate the information gain for each attribute from the decision tree we construct in Figure 2 for each season and take the average over the past 20 seasons. The results are shown in Figure 3. As can be seen from Figure 3, the top five most critical attributes that can determine the goodness of a team is FG%, DRB, TOV, PTS/G, and AST respectively due to their high information gain.

$$Attribute_n = (Attribute - Min)/(Max - Min) \quad (1)$$

Merge the attributes into a factor (Attribute n+1): With the goal of using one single factor to describe a team, we need to merge several most critical attributes. Due to the variety of the scale of different attributes, a min-max normalization approach which is shown in Equation 1 where the subscribed N denotes normalized attribute, is applied to unify the scale of various attributes (scale are in $0 \sim 1$ after the normalization) and make the merging applicable. An example of normalization is shown in Figure 4 where we normalize the top five most important attributes in the season 2004-2005.

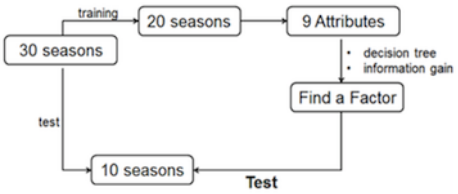


Figure 1: Overview of the analytical model

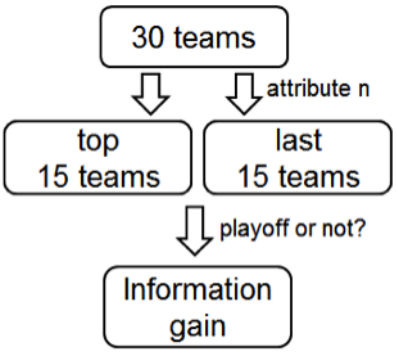
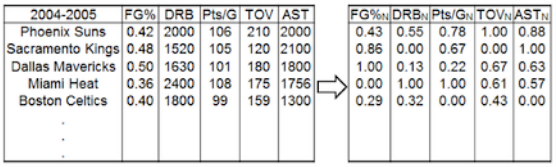


Figure 2: Decision Tree Layer-1 for our dataset

Information gain	Attributes								
season	FG%	DRB	TOV	PTS/G	AST	ORB	BLK	PF	STL
2013-2014	0.218	0.013	0.013	0.052	0.119	0.119	0.052	0.119	0.013
2012-2013	0.218	0.119	0.013	0.119	0.052	0.000	0.052	0.013	0.119
2011-2012	0.119	0.353	0.013	0.013	0.052	0.013	0.052	0.000	0.052
2010-2011	0.119	0.119	0.119	0.013	0.013	0.052	0.013	0.052	0.052
2009-2010	0.119	0.119	0.119	0.052	0.013	0.013	0.119	0.000	0.013
2008-2009	0.218	0.013	0.119	0.013	0.013	0.000	0.013	0.013	0.013
2007-2008	0.218	0.052	0.218	0.119	0.052	0.013	0.052	0.119	0.013
2006-2007	0.052	0.119	0.052	0.119	0.119	0.000	0.052	0.013	0.052
2005-2006	0.000	0.218	0.052	0.052	0.052	0.000	0.013	0.119	0.000
2004-2005	0.013	0.052	0.052	0.013	0.013	0.013	0.000	0.000	0.000
2003-2004	0.073	0.073	0.007	0.001	0.073	0.042	0.001	0.042	0.073
2002-2003	0.073	0.023	0.042	0.155	0.073	0.001	0.007	0.007	0.155
2001-2002	0.155	0.073	0.202	0.274	0.023	0.001	0.007	0.042	0.155
2000-2001	0.073	0.073	0.202	0.073	0.073	0.007	0.073	0.001	0.023
1999-2000	0.155	0.023	0.202	0.023	0.274	0.042	0.023	0.001	0.001
1998-1999	0.155	0.073	0.007	0.073	0.023	0.042	0.073	0.042	0.007
1997-1998	0.155	0.274	0.106	0.155	0.073	0.042	0.007	0.023	0.001
1996-1997	0.155	0.073	0.106	0.073	0.073	0.106	0.001	0.007	0.073
1995-1996	0.023	0.073	0.106	0.023	0.023	0.007	0.007	0.042	0.001
1994-1995	0.089	0.089	0.124	0.189	0.089	0.008	0.008	0.008	0.089
Average	0.120	0.101	0.094	0.080	0.065	0.026	0.031	0.033	0.045

Figure 3: The results of information gain in the past 20 seasons



2004-2005	FG%	DRB	Pts/G	TOV	AST
Phoenix Suns	0.42	2000	106	210	2000
Sacramento Kings	0.48	1520	105	120	2100
Dallas Mavericks	0.50	1630	101	180	1800
Miami Heat	0.36	2400	108	175	1756
Boston Celtics	0.40	1800	99	159	1300

$FG\%_N$	DRB_N	Pts/G_N	TOV_N	AST_N
0.43	0.55	0.78	1.00	0.88
0.86	0.00	0.67	0.00	1.00
1.00	0.13	0.22	0.67	0.63
0.00	1.00	1.00	0.61	0.57
0.29	0.32	0.00	0.43	0.00

Figure 4: An example of min-max normalization

After the normalization, we used Equation 2 to merge the important attributes to generate a mixed factor to describe the goodness of a team which we will name it $Attribute_{n+1}$. When the new attribute ($Attribute_{n+1}$) is found, it will be used to evaluate the NBA teams in each season. For each season, the 30 NBA teams will be split into the top 15 which have the highest value of $Attribute_{n+1}$ and the last 15. In this work, we aim at finding an $Attribute_{n+1}$ that can ensure 80% of the top 15 teams make to the playoff over the past 30 years.

$$Attribute_{n+1} = aFG\%_n + bDRB_n + cTOV + dPTS/G_n + ... \quad (2)$$

To find the $Attribute_{n+1}$, the unknown parameters in Equation 2 are required to be found. Since the scale of the normalized attributes are in 0 $\sim$1, we postulate that the unknown parameters are ranging from 0 to 1. For each parameter, a trial and error approach with an interval of 0.01 from 0 to 1 is used for detecting the optimized parameter pair. Here the top two attributes are used for constructing the new factor due to its better accuracy compare to other models which can be found in Figure 5. Within the 10^4 possible combinations of the unknown parameters, the optimized pair is found to be (1.0, 0.6) with the training accuracy to be 79.4% for 20 seasons (1995-2014) and the prediction accuracy to be 82.0% for the rest of the 10 seasons.

Number of parameters	a	b	c	d	Training (20 seasons)	Testing (10 seasons)
1 parameter (a)	1.0	-	-	-	0.747	0.80
2 parameters (a,b)	1.0	0.6	-	-	0.794	0.82
3 parameters (a,b,c)	1.0	0.6	0.02	-	0.790	0.82
4 parameters (a,b,c,d)	1.0	0.6	0.02	0.02	0.783	0.82

Figure 5: The optimized parameters and accuracy for various models

Remaining Work

- We will further optimize the $Attribute_{n+1}$ with 5 unknown parameters to check if we can find a better model for finding the new attribute. However, the result from the two-parameter model we have is satisfactory and acceptable for the prediction.

- Since several important attributes including FG%, DRB, TOV, PTS/G, and AST and an optimized factor is found for finding a good team, we will use these attributes to evaluate an individual player. With similar approach and by the evaluation of these attributes and the player's annual income, we may be able to find a new index to describe the goodness of an individual player. We believe that by the new index we are going to find, we will be able to form a cheap but competitive team that is guaranteed to be in the playoff.