Using Bayesian Network for Creating NBA's MVP Award Prediction Model

Dong Joon ”DJ” Kim
Computer Science Department
University of California, Los Angeles
djkim@cs.ucla.edu

Abstract – **In this paper, I will discuss applying Bayesian network to create prediction models that predict the winner of NBA's MVP award for each season, using NBA statistics. After parsing NBA statistics from 1997-98 season through 2016-17 season, four different Bayesian network prediction models were trained. The accuracy of the Bayesian network prediction models was compared and analyzed using k-fold cross validation with prediction models created with other Machine Learning algorithms, such as Linear Regression and Support Vector Machine (SVM).**

I. INTRODUCTION

NBA gives out the Most Valuable Player award to the best player of each season. However, many NBA fans constantly argue as they all have different perspective on who the best player is. For example, in recent years, many NBA fans have argued between Kobe Bryant and LeBron James, or Stephen Curry and LeBron James, as the best player in the league. In addition, this season, the race for NBA's MVP award is considered as the tightest NBA MVP race ever, with Russell Westbrook, James Harden, LeBron James, and Kawhi Leonard, all having historic statistical season, on both ends of the floor.

Bayesian network is a widely-used probabilistic graphical model used for representing the dependency (and independency) between different random variables. In this project, I constructed four different NBA MVP award prediction models, which were learned from NBA regular season statistics. Using the prediction models created with Bayesian network, we can analyze how the MVP is chosen for each season, based on the accuracy of the prediction model: If the accuracy is high, the MVP is chosen solely based on each player's statistics, otherwise, there are factors other than statistics that affect chance of becoming the MVP.

The rest of this report is organized as follows: Section 2 provides a brief overview of NBA MVP, NBA regular season statistics, and Bayesian networks; Section 3 explains how NBA statistics were parsed and different Bayesian networks were created for different prediction models; Section 4 provides results and analysis; Section 5 concludes the paper with possible future work ideas.

II. BACKGROUND

In this section, I briefly introduce how NBA's MVP is decided, different types of NBA statistics that were considered, and Bayesian networks.

A. NBA Most Valuable Player Award

NBA Most Valuable Player Award is an annual award given to the best performing player of the regular season. Before 1980-81 season, the MVP was voted by NBA players. Since 1980-81 season, the MVP was voted by sportswriters and broadcasters. Starting in 2016-17 season, sportswriters and broadcasters who are affiliated with an NBA team are not eligible to vote for the MVP. Instead, 100 voters will be chosen by NBA from independent media members who are not affiliated with teams [1].

However, what does “Most Valuable” even mean? Voting for the MVP is much more complicated than simply voting for the scoring leader of the season, or picking the best statistical player of the team. To win the MVP award, not only a player's regular season statistics have to excel compared to other players, his team must have good record as well. For example, since 1985-86 season, every single MVPs played for a playoff team [2]. In addition, bias of the voters make it even harder to predict the MVP. In previous seasons, the voters who worked for an NBA team were biased towards to voting for a player who played for the same team.

B. NBA Regular Season Statistics

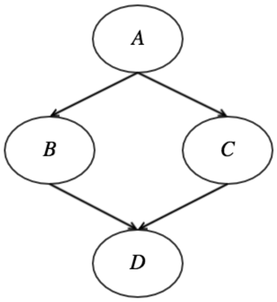
NBA started collecting player statistics since 1946-47 season. There are two different types of statistics: base stats and advanced stats. Base stats include points, rebounds, assists, steals, blocks, shooting percentage, etc. Advanced stats include categories that are computed by normalizing base stats based on pace (how fast a team plays) and efficiency (calculating statistics per-possession instead of per-shot attempt). Advanced stats were first introduced in 2013 and the data is only available from 1996-97 season.

Due to various rule changes and introduction of new statistical categories, some base stats, including 3-point field goals, which were officially introduced in 1979-80 season, and steals, which were officially introduced in 1973-74 season, are only available in modern NBA era.

C. Bayesian Networks

This section provides background knowledge on Bayesian networks from Chapter 6.7, 3.7, and 17.2 of [3].

In a probabilistic graphical model, a node represents a random variable and an edge between two nodes represent the relationship between two nodes. Bayesian network is a probabilistic graphical model represented with a directed acyclic graph and a set of conditional probability table. In a Bayesian network, parents of a node *N* denote the direct causes of *N* and descendants of a node *N* denote the effects of *N*. As shown in *Figure 1,* variable *A* is a direct cause of *B* and *C*, and *D* is an effect of *B* and *C*.



A	Pr(*A)*
true	0.35

A	*B*	Pr(*B*|*A*)
true	true	0.8
false	true	0.7

A	*C*	Pr(*C*|*A*)
true	true	0.3
false	true	0.2

B	*C*	*D*	Pr(*D*|*B,C)*
true	true	true	0.85
true	false	true	0.7
false	true	true	0.9
false	false	true	1

Figure 1. Example Bayesian Network

1. Inference

Given any Bayesian network, one can compute marginal probability over any variables, even if CPT does not specify it. For example, one might want to compute the probability of *A* = true, when we know through observation that *B* = false, *C* = true, and *D* = false. The marginal probability could be computed by an inference technique called variable elimination. Variable elimination consists of two parts: multiplication and summation. Multiplying factors combines two tables for union and yields a joint probability distribution, and summing out a variable from a factor removes the variable from a Bayesian network while maintaining its ability to answer queries of interest. Calculating Pr(*A*=true | *B*=false*, C*=true*, D*=true) is shown as follows:

$$\Pr(A = \mathrm{T} \mid B = \mathrm{F}, C = \mathrm{T}, D = \mathrm{T})$$

$$= \frac{\Pr(A = \mathrm{T}, B = \mathrm{F}, C = \mathrm{T}, D = \mathrm{T})}{\Pr(B = \mathrm{F}, C = \mathrm{T}, D = \mathrm{T})}$$

$$= \frac{\begin{array}{c}\Pr(A = \mathrm{T}) \times \Pr(B = \mathrm{F} \mid A = \mathrm{T}) \\ \times \Pr(C = \mathrm{T} \mid A = \mathrm{T}) \times \Pr(D = \mathrm{T} \mid B = \mathrm{F}, C = \mathrm{T})\end{array}}{\sum_a \begin{array}{c}\Pr(A = a) \times \Pr(B = \mathrm{F} \mid A = a) \\ \times \Pr(C = \mathrm{T} \mid A = a) \times \Pr(D = \mathrm{T} \mid B = \mathrm{F}, C = \mathrm{T})\end{array}}$$

Equation 1. Computing Pr(*A*=true | *B*=false*, C*=true*, D*=true) using variable elimination

Using variable elimination for calculating marginal probability is limited to Bayesian network only consisting discrete variables. However, certain application areas, including working with NBA regular season statistics, require the use of continuous variables. Use of continuous variable could be avoided by emulating hard evidence using soft evidence on a discrete variable.

Using the "nothing else considered" method, suppose we obtain soft evidence on an event α and suppose we express the strength of our obtained soft evidence as *k*. Then, using Jeffrey's Rule, we can compute the updated probability of an event α as follows:

$$\Pr{}'(\alpha) = \frac{k \Pr(\alpha)}{k \Pr(\alpha) + \Pr(\neg\alpha)}$$

Equation 2. Probability of an event α given the Bayes factor of the evidence

Using the Gaussian density *f(y)*, an evidence *y* observed from a continuous variable *Y* can be viewed as soft evidence on a propositional variable X and the Bayes factor k could be computed as follows:

$$f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/2\sigma^2}$$

$$k = \frac{f(y \mid X = true)}{f(y \mid X = false)}$$

Equation 3. Emulating hard evidence $Y = y$ by using soft evidence on a propositional variable X

2. Learning

In the context of Bayesian network, learning refers to constructing the structure of the graph and conditional probability table. If the training data is complete, we can scan the dataset and simply count the number of occasions of variable instantiations in the dataset. The following method is called the Maximum likelihood method.

$$\Pr_{\mathfrak{D}}(a, b, c) = \frac{\mathfrak{D}\#(a, b, c)}{n}$$

$$\Pr_{\mathfrak{D}}(a \mid b) = \frac{\mathfrak{D}\#(a, b)}{\mathfrak{D}\#(b)}$$

Equation 4. Estimating CPT using Maximum Likelihood Method

III. Creating Prediction Model Using Bayesian Network

In this section, I describe how I designed the program that created my prediction model using Python. The program consists of three modules: NBA statistics parser module, maximum likelihood module for constructing CPT, and inference module.

I considered four different Bayesian networks for creating my prediction models. Two of the models used continuous variables for representing statistics and the other two models used discretized ranks of each statistical categories of players. In addition, two of the models take account of each player's position because some statistical categories were more important than others for each position. For example, point guards, who focus on ball handling, passing, and perimeter defense, tend to have higher assist and steal numbers compared to other positions. In contrast, centers, who focus on rebounding, low-post offense and defense, tend to have higher field goal percentage, rebounds, and blocks. The graph structure could have been learned through structure learning algorithms. However, I chose a Naïve-Bayes structure due to its simplicity and effectiveness.

Statistics that I used to create different prediction models are as follows: Win, win percentage, minutes, field goal percentage, 3-point field goal percentage, free throw percentage, rebounds, assists, turnovers, steals, blocks, points, plus-minus, double-doubles, triple-doubles, and player impact estimate. Four different Bayesian networks are shown in *Figure 2*.

To train and test the prediction model, I needed to get NBA statistics. First, I tried to use a Python library for parsing NBA statistics at stats.nba.com called nba_py. However, nba_py was extremely inefficient due to how it is implemented. Every time a single player's statistics is queried, nba_py library fetches statistics of all NBA players who played in that season, and then fetches each player's statistics for that season. In addition, using nba_py library, statistical ranks of players had to be manually computed. Therefore, I decided to implement my own parser module.

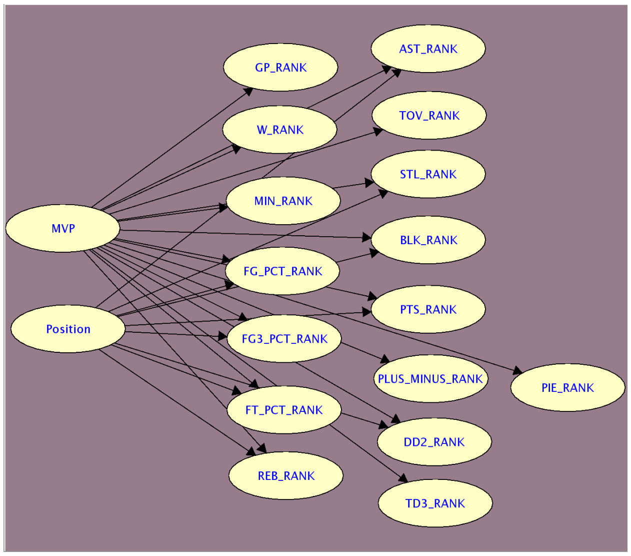
First, I parsed top 50 scoring leaders' player IDs, names, and teams that they played for in a given season from nba.com. I wanted to restrict the training data to players who have legitimate chance of winning the MVP award, so my training data consisted of top 50 scoring leaders for every season. In addition, the game of basketball has evolved throughout seasons so I restricted my dataset to modern NBA era, starting from 1996-97 season. Then, I got both base stats and advanced stats of everyone that played in a certain season from stats.nba.com and stored in a dictionary where keys correspond to seasons and values correspond to an array of players' statistics for each season. Finally, I looped through each season statistics and filtered out the players who were not top 50 scoring leaders for the season to create my dataset.

Surprisingly, statistics from stats.nba.com did not include player's position. Therefore, I created a crawler which crawls through Wikipedia pages of NBA players and manually parsed their positions. In addition, there are five different positions in basketball: point guard, shooting guard, small forward, power forward, and center. However, only one center and two shooting guards have won the MVP award since 1996, while six point guards won the MVP award since 1996. Therefore, I decided to use three positions instead of five for the Bayesian network model that will support players' position: point guard, swingman (shooting guards and small forwards), and bigman (power forwards and centers).

Year	MVP	Original Position	New Position
1996-97	Karl Malone	PF	Bigman
1997-98	Michael Jordan	SG	Swingman
1998-99	Karl Malone	PF	Bigman
1999-00	Shaquille O'Neal	C	Bigman
2000-01	Allen Iverson	PG	PG
2001-02	Tim Duncan	PF	Bigman
2002-03	Tim Duncan	PF	Bigman
2003-04	Kevin Garnett	PF	Bigman
2004-05	Steve Nash	PG	PG
2005-06	Steve Nash	PG	PG
2006-07	Dirk Nowitzki	PF	Bigman
2007-08	Kobe Bryant	SG	Swingman
2008-09	LeBron James	SF	Swingman
2009-10	LeBron James	SF	Swingman
2010-11	Derrick Rose	PG	PG
2011-12	LeBron James	SF	Swingman
2012-13	LeBron James	SF	Swingman
2013-14	Kevin Durant	SF	Swingman
2014-15	Stephen Curry	PG	PG
2015-16	Stephen Curry	PG	PG

6 point guards, 2 shooting guards, 5 small forwards, 6 power forwards, and 1 center have won the MVP award using traditional positions of basketball.
6 point guards, 7 swingmans, and 7 bigmans have won the MVP award using the reclassification of positions.

Table 1. List of Most Valuable Players from 1996-97 season and their positions.



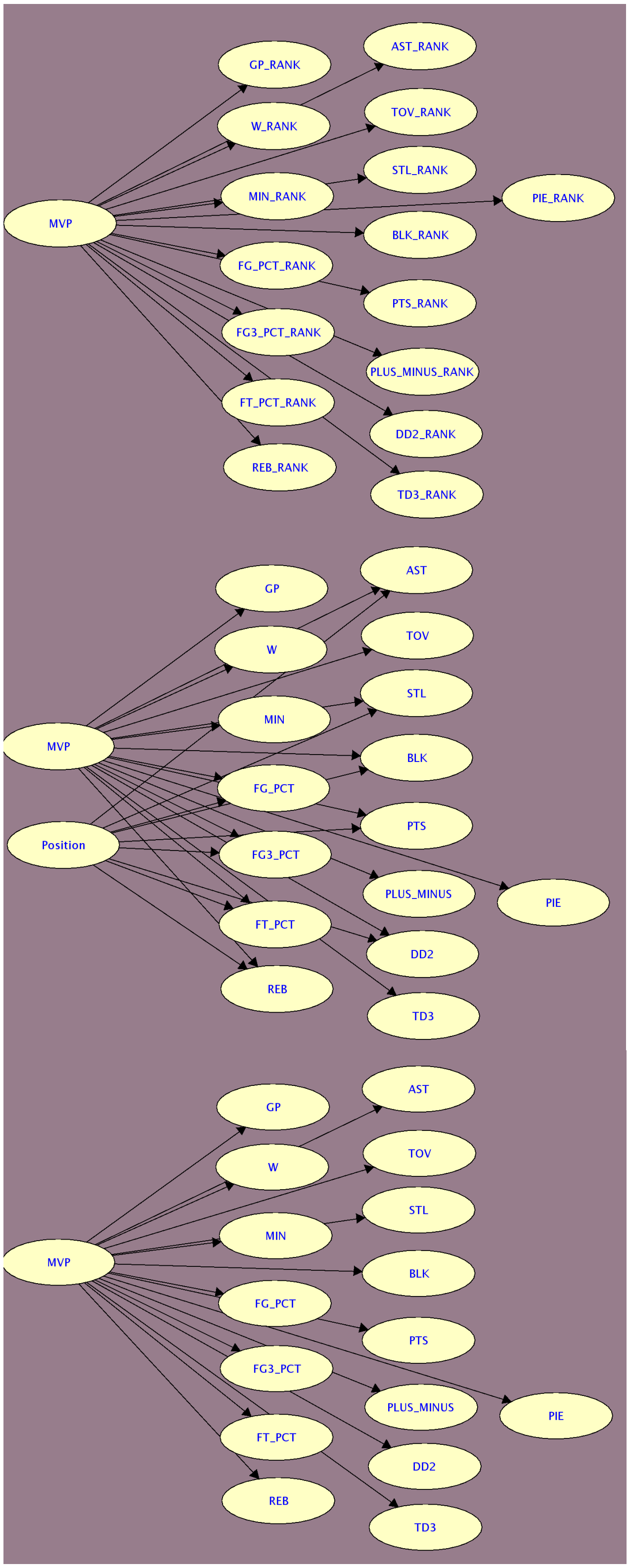


Figure 2. Different Bayesian Networks Considered for Prediction Model. Bottom two networks contain continuous variables.

Then, for Bayesian network with discrete variables, The CPT was learned through maximum likelihood approach described in section *II.C.2*. Each statistical rank category was scanned and discretized to four instantiations based on their rank as shown in *Table 2.* The cutoff for each instantiation was selected by how well players are distributed for each statistical category. Then, the discretized dataset was scanned and all possible instantiations were counted to create the CPT. Advanced Stats could not be used for Bayesian network with discrete variables because advanced stats could not distinguish superstars that constantly perform well with players who do not play as much but performs well whenever they get a chance to play. For example, Javale McGee, who played 9.6 minutes per game this season, has higher offensive efficiency as Stephen Curry, who played 33.4 minutes per game and is regarded as the best NBA player at the moment. Finally, there were some parameters that had zero probability. However, zero probability is extremely rare in sports. For example, Russell Westbrook is top 10 in rebounding this season, but no other point guards in the history of NBA were top 10 in rebounding. Therefore, all parameters with zero probability were substituted with a very small $\varepsilon = 0.001$ and normalized.

Statistical Categories	Top Rank	Above Average	Average	Below Average
GP_RANK	Top 50	51~100	101~150	151~
W_RANK	Top 25	26~50	51~75	76~
MIN_RANK	Top 25	26~50	51~75	76~
FG_PCT_RANK	Top 50	51~100	101~150	151~
FG3_PCT_RANK	Top 25	26~50	51~75	76~
FT_PCT_RANK	Top 50	51~100	101~150	151~
REB_RANK	Top 25	26~50	51~75	76~
AST_RANK	Top 15	16~30	31~45	46~
TOV_RANK	Top 25	26~50	51~75	76~
STL_RANK	Top 25	26~50	51~75	76~
BLK_RANK	Top 25	26~50	51~75	76~
PTS_RANK	Top 15	16~30	31~45	46~
+/-_RANK	Top 15	16~30	31~45	46~
DD2_RANK	Top 25	26~50	51~75	76~
TD3_RANK	Top 5	6~10	11~15	16~
PIE_RANK	Top 5	6~10	11~15	16~

Table 2. Discretization of Statistical Rank Category

For the Bayesian networks with continuous variables, the data had to be normalized for 1998-99 season and 2011-12 season for W, GP, DD2, and TD3 because the seasons were shortened due to NBA lockout.

For Bayesian networks with continuous variables, instead of computing the CPT, I used emulating hard evidence by using soft evidence method described in section *II.C.1*. In order to compute the Bayes Factor, the entire dataset was scanned in order to obtain each statistical category's mean and standard deviation. Both mean and standard deviation of each statistical category were computed using mean and std functions in numpy library.

Then, the following query was asked to Bayesian network for all top 50 scorers of each season.

$$\Pr(MVP = true \mid statistics\ of\ each\ player)$$

Equation 5. Query to Determine the MVP for each Season

For each season, a player who had the highest probability is predicted as the winner of the MVP award for Bayesian network with discrete variables. For Bayesian network with discrete variables, variable elimination method described in section *II.C.1* was used to compute the marginal explained above.

For Bayesian network with continuous variables, the Bayes factor of each player's chance of winning the MVP award given their statistics was calculated. The Bayes factor was calculated as shown in the following equation using the fact that each statistical category is independent with each other because I used Naïve-Bayes structure.

$$k = \prod_{y \in Stats} \frac{f(y \mid MVP = true)}{f(y \mid MVP = false)}$$

Equation 6. Calculating Bayes Factor for each Player

After all Bayes factors have been calculated, the player with largest Bayes factor is predicted as the winner of the MVP award for Bayesian network with continuous variables.

IV. Results and Analysis

To determine which Bayesian network yields the most accurate prediction, 7-fold cross validation method was used, where 13 seasons were picked as training data and the other 7 seasons were used as test data. Model 1 uses Bayesian network with continuous variables and have an additional root node for position. Model 2 uses Bayesian network with continuous variables and does not have an additional root node for position. Model 3 uses Bayesian network with discrete variables and have an additional root node for position. Model 4 uses Bayesian network with discrete variables and does not have an additional root node for position. Model 5 uses linear regression and model 6 uses support vector regression.

The accuracy of each model is calculated based on two different methods: first method is the percentage where the predicted MVP winner is indeed the MVP for a certain season, and the second method is the percentage where the MVP for a certain season is one of the top 3 predicted MVP winners. The accuracy of different models is illustrated in the following table:

Model	**% of Correctly Predicted MVPs**	**% of MVP in Top 3**
1	25%	51.79%
2	26.79%	58.93%
3	21.43%	46.42%
4	23.21%	42.84%
5	25%	82.14%
6	48.21%	62.5%

Table 3. Accuracy of different models based on 7-fold cross validation

As shown on the table above, prediction models based on Bayesian network with continuous variables performed better compared to Bayesian network with discrete variables. In addition, Bayesian network models seem to underperform compared to model based on linear regression and SVM.

However, *7*-cross validation method was biased against models with an additional root variable for player's position. *Table 1* shows that 6 bigmans, 1 swingman, and 1 point guard have won the MVP award from 1996-97~2003-04 season, and 1 bigman, 6 swingmans, and 5 point guards have won the MVP award from 2004-05~2015-16 season. The game of basketball has evolved from being extremely physical and focusing more on post scoring in the late 90's and early 2000's to being guard-focused and focusing more on three-point scoring since late 2000's. When the dataset was divided into training data and test data, the models that take position into factor did not perform well compared to other models because the training dataset simply did not contain any bigman or point guard in the training dataset whereas the test dataset included seasons where bigmans or point guards won multiple MVP award.

In addition, SVM model does not perform as well as even though the accuracy seems higher than other models, since it only picks four players to win the MVP award: LeBron James, Kevin Durant, Karl Malone, and Tim Duncan, and these four players have won 9 MVP awards combined in last 20 years.

For this season's MVP, model 1 and 2 picked Russell Westbrook as MVP, and model 3 and 4 picked LeBron James as MVP. Model 1 and 2 is believed to be more accurate compared to model 3 and 4 so this season's MVP award will be Russell Westbrook.

V. Conclusion

In this project, I designed and compared different prediction models created using Bayesian network, using maximum likelihood method for parameter learning and variable elimination and emulating continuous hard evidence by soft evidence techniques for inference. Although some Bayesian network model did not perform as well as simple machine learning algorithm such as linear regression model, the accuracy of prediction models using Bayesian Network with continuous variables were comparable. As there were only 20 MVPs in the dataset, it is highly likely that the prediction models were suffering from the "small data problem". In the future, as more data is accumulated, the prediction models will perform better.

In the future, I would like to create a "five-thirty-eight-like" graphical visualization and publish the daily day-to-day update of top 5 players who are predicted as MVP and publish on my own website. In addition, I would like to compare the result with a new Bayesian network model where different structure learning algorithms were applied.

VI. References

1. Nathan, Alec. "How the NBA MVP Voting Process Works, Announcement Date." Bleacher Report. Bleacher Report, 14 Apr. 2017. Web. 13 May 2017.
2. Schuhmann, John. "The New NBA.com/stats: Advanced Stats All Start With Pace And Efficiency « NBA.com | Hang Time Blog." NBAcom Hang Time Blog. N.p., 15 Feb. 2013. Web. 14 May 2017.
3. Darwiche, Adnan. Modeling and reasoning with Bayesian networks. New York, NY: Cambridge U Press, 2014. Print.

VII. Appendix

Source Code:
https://github.com/djkim02/NBA_MVP_Predictor