A Bayesian Method to Estimate the Strength of NCAA Division I Men's Basketball Teams

August 4, 2018

1 Introduction

The National Collegiate Athletic Association Division I (NCAA D-I) mens basketball tournament is one of the major sports events in the United States that draws attention to not only sports fans and betters, but also statisticians and operations researchers to use the regular season results and other data to predict the outcome of the tournament. A NCAA D-I basketball regular season usually start in mid-November and is played through early March, which consists of regular inter-conference and intra-conference games, special inter-conference tournament games and single-elimination conference tournament games. and the latter two are usually played in neutral stadiums. There are 353 colleges playing in 32 conferences in 2018-2019 season. After the conclusion of the regular season, conference champions and best teams (judged by the tournament selection committee), 68 in total, are selected to participate in the tournament, which is nicknamed *March Madness*. The teams play single elimination games until the final championship is decided and all games are played in neutral stadiums as well.

Since there are billions of dollars wagered on the tournament's outcome [4], statistician work on different types of prediction models using regular season data of various fidelities. In terms of ranking systems, Ken Pomeroy uses the Pythagorean Expectation, while Jeff Sagarin, Ken Massey or Raymond Cheung also post ranking results on their websites without entirely disclosing the methodology [3, 5]. Kvam and Sokol [2] first present a logistic regression/Markov chain (LRMC) model that uses only regular season game scores to rank the teams. Based on this ranking, the tournament results are predicted by simply choosing the higher seed to proceed in every round. Brown and Sokol [1] improve this method with two Bayesian update methods in the place of logistic regression. The major critique about LRMC models is that they only use game scores as a measure for team strength. Yuan et al. [6] provide a mixture of models ranging from logistic regression to neural network based on the features of actual game statistics and strength and ranking information compiled by experts. Since they use aggregated data for the regular season, the progress of regular season of teams are ignored. Zimmerman et al. [7] use machine learning techniques, including decision trees, neural networks, naive Bayes and ensemble learners such as random forest. This paper points out that there is no consistency of opponents' quality since teams can pick their own opponents in college basketball.

We address these issues by using more thorough game statistics and trying to incorporate the

trend of team performance in our learning process. The outcome of our method reflects multiple dimension of each team's strength and based on those sophisticated properties, we further train a prediction model for the NCAA tournament. In Section 2 we describe the data collected. We create multiple measures to evaluate the strength of each team, and we propose a Bayesian method to learn them. Next in Section 3, we use these measures as features and train models to predict the outcome of the NCAA tournament in 2012-2017.

2 Learning Teams' Strength Measures

- What data do we have (years/category)? The source of data. How do we abstract the data?
- List the measures.
- Update method.

3 Predict the NCAA Tournament Outcome

4 Conclusions

References

[1] M. Brown and J. S. Sokol. An improved LRMC method for NCAA basketball prediction. *Journal of Quantitative Analysis in Sports*, 6(3).

[2] P. Kvam and J. S. Sokol. A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NrL)*, 53(8):788–803, 2006.

[3] K. Massey. College basketball ranking composite, 1999.

[4] E. Matuszewski. March Madness gambling brings out warnings from NCAA to players, 2011.

[5] J. Sagarin. Jeff sagarin's college basketball ratings, 2018.

[6] L. Yuan, A. Liu, A. Yeh, A. Kaufman, A. Reece, P. Bull, A. Franks, S. Wang, D. Illushin, and L. Bornn. A mixture-of-modelers approach to forecasting ncaa tournament outcomes. *Journal of Quantitative Analysis in Sports*, 11(1):13–27, 2015.

[7] A. Zimmermann, S. Moorthy, and Z. Shi. Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned. *arXiv preprint arXiv:1310.3607*, 2013.