Big Data Analytics in NCAA Football

Nsikan Udoyen
School of Informatics and Computing, Indiana University
P.O. Box 1212
Dublin, Indiana 43017-6221
nudoyen@iu.edu

ABSTRACT

This paper provides an overview of applications of big data in NCAA football by surveying current research and development work that supports the increased application of big data analytics to various aspects of NCAA football. The focus of current research is support for player performance management, injury prevention, and the use of predictive analysis to predict outcomes of games. However, the nature of interactions between players in football limit the efficacy of big data techniques in other areas such as strategy.

KEYWORDS

i523,hid342, big data, analytics, NCAA football

1 INTRODUCTION

National Collegiate Athletics Association (NCAA) football is one of the most widely watched sports in the United States. The size of the fan base and the profits that can be derived from televised games incentivize universities and other interested parties to invest in the application of big data analytics and data science methods in general to improve on-field outcomes by enabling better management of player well-being and performance. The purpose of this paper is to provide an overview of the use of data science in National Collegiate Athletics Association (NCAA) football. Recent research on the use of data science to improve various aspects of NCAA football will be surveyed, while current trends and their implications will be discussed.

2 BIG DATA ANALYTICS IN NCAA FOOTBALL

2.1 Predictive Analytics

NCAA football analysts invest a significant amount of time trying to forecast performance of various teams throughout the season. Their analysis fuels sports talk shows and other mass media programs that target dedicated fan bases, giving them a deeper understanding of the game and allowing them to learn more about their teams. Data used to support NCAA football analysts' predictions is drawn from a mix of sources such as coaches' polls, and detailed and routinely updated data on players' performance. Some of this data is combined to create composite indexes, such as ESPN's Football Power Index (FPI)[1], which are used to rank teams based on thousands of simulations of their game outcomes, and updated weekly, based on available data. Composite indexes such as the FPI support broader discussion of matchups every week, and encourage analysts to ask broader questions in previewing games, but typically are not used in any systematic way to predict outcomes.

Several researchers have applied data mining methods towards the prediction of NCAA football scores[4],[2]. Various research efforts have focused on the scope of relevant data, and how to model such data. In their paper comparing NCAA football game outcome prediction methods, Delen et al. used data on NCAA teams from 244 bowl games between 2002 and 2009 to generate and compare several predictive models[2]. They compared the performance of the models by using them to predict 2010-11 bowl game scores and found that classification-based models were better than regression-based classification methods at predicting game outcomes.

2.2 Performance Management & Player Safety

Several data mining methods have been developed to monitor athletes' performance and enable coaches to make data-driven decisions to improve results and avoid injuries. Platforms such as Microsoft's Sports Performance Platform [3] enable the collection and aggregation of biometric and other data that can be used to monitor performance. The use of wearable technology devices such as Fitbit to monitor NCAA football players has been proposed. Most efforts to apply data analytics to performance management in NCAA football focus on the evaluation and management of individual players, rather than the use of data mining to drive strategic decisions for teams during games.

In support of performance management, groups such as the NCAA Sports Science Institute gather data on injuries to college athletes and have used findings from their studies based on that data to advise the NCAA on issues such as the optimal frequency of football practices[7]. By analyzing data from the Big 12 conference, scientists at the NCAA Sports Science Institute were able to determine that the majority of injuries (and 58% of concussions) occurred during preseason practice. Their suggested guidelines, which were endorsed by 16 medical organizations, called for a reduction in the frequency of preseason practice sessions and less full-contact practice sessions.

In their paper, Ofoghi et al. describe how performance analysis requirements influence data gathering in their presentation of a general framework that applies data mining methods to sports [5]. The authors attempt to describe in their framework the most important features needed to categorize sports to enable data mining. Through their framework, Ofoghi et al. discuss the types of data that can be collected, depending on the nature of the sport being studied, and list important considerations.

Schumaker et al., list several standard data-driven metrics used to assess football teams and individual players[6]. The listed metrics include:

- *Defense-Adjusted Value Over Average (DVOA)*, which measures the success of a particular play against a defense and compares it to the average.

- *Defense-Adjusted Points Above Replacement (DPAR)*, which evaluates individual players by assessing their contribution (in points) compared to a replacement player.
- *Adjusted Line Yards (ALY)*, which assigns credit to an offensive line based on how far the ball is carried

While abundant data exists to compute the listed metrics and compare teams using them, their subjective nature makes them unreliable. DVOA, for instance, accounts for variables such as time remaining in the game, field position, and the quality of the opponent. There is no guidance on how such variables are computed or the weights assigned to each one. The ALY measures the contribution of the offensive line and the running back by rewarding the running back's individual effort for successful carries and punishing the offensive line for failed attempts. The ALY is adjusted based on league averages, which do not account for issues such as weather or bad officiating, which may have impacted a team's performance.

When used together, these metrics give a detailed view of a team's past performances. There is however, no evidence of successful use of such detailed assessments of a team's past performances to support strategic decisions during a game. The metrics are more suitable for highlighting areas of concern than predicting how well one team will fare against another before they play.

3 DISCUSSION

Research on predictive models that predict outcomes of NCAA football games illustrates the difficulty involved in capturing the nuances and complexity of the sport in a model. It also illustrates problems with the use of historical data for predictive purposes in NCAA football. For example, the data mined for the study by Delen et al., which was used to predict 2010-11 bowl games, included data points from as early as 2002, when none of the players in the 2010-11 bowl games were even eligible to play college football. It is difficult to determine how much data is sufficient to produce accurate predictions, and current data alone may not be sufficient, since some NCAA football teams may play as few as eleven games in a season.

Several features of the metrics used to describe and rate NCAA football players and teams make it difficult to use them for predictive purposes, despite the abundance of data to be collected. These include

- *The subjective nature of the metrics*
  To account for the context-specific nature of the data being gathered to describe individual and team performances, some metrics are weighted to reflect factors such as the quality of the opponent. Such subjective factors are usually not evenly considered by different evaluators, and may change as the season progresses.
- *Focus on outcome-based metrics, such as ALY*
  By relying on metrics that report only the outcomes of individual plays, data that reflects the tactics used and other technical aspects of the game are overlooked. Such metrics also ignore an opponents ability to learn and improve after a football game.
- *Inability to aggregate metrics*
  No single metric effectively describes a football team's performance well enough to enable comparison to other teams. When different metrics are combined to describe a football team's performance, the manner in which they are combined is subjective. When the metrics are combined to create a composite index used to compare teams and predict outcomes, they do not provide a complete picture of potential interactions and mismatches between teams that could influence the outcome of the game between them. A prime example of this is the Bowl College Series (BCS) formula used to select the teams that would play for the NCAA Football National Championship from 1998 to 2013.
- *Lack of context*
  When metrics are used to rate individual players, they often do not account for teammates' inputs. An example is yards-after-catch (YAC), often used by scouts to rate wide receivers. YAC reports the amount of additional yards a player gains after catching a pass from the quarterback, and should measure individual effort of the player that catches the ball. However, additional yards gained by a player after catching the ball may be due to defensive errors or assistance from teammates who block players on the opposing team. Likewise, other metrics used to rate receivers such as yard-per-catch or total yards are computed without considering the quality of the quarterback's decision-making or the defensive schemes employed by the opponent.

The use of data mining to manage player performance raises concerns over privacy and the ownership and potential misuse of the data collected[8]. The scope and amount of data collected about players has increased with the proliferation of the use of data mining methods to study player performance. In some cases, the harvesting of data collected by wearable technology devices by sportswear companies is permitted under the terms of the agreements between universities and the sportswear companies that sponsor their football teams. While companies such as Nike have stated that they have not yet begun harvesting players' biometric data, at least some of the data they could collect would not be covered by United States federal HIPA (Health Information Portability and Accountability Act) laws[9].

4 CONCLUSION

The use of data mining and analytics in NCAA football is increasing, as it has in other sports. However, due to the complexity of the game, practical uses of data analytics currently available and under exploration are in individual and team performance management and prevention of injuries. Research on data analytics, and current applications of technology to NCAA football have focused on techniques to extract meaningful information from gathered data, rather than the explanation and use of such information for predictive purposes.

The inability to account for context in data makes the use of data science to predict outcomes and influence strategy in NCAA football games difficult. The use of data primarily to compile metrics that describe past outcomes and average individual and team performance levels does not enable an understanding of their true

capabilities. There is thus a need to continue to rely on qualitative assessments by experts when making predictions or scouting individual players, and use data analytics as a supporting tool to provide relevant information to guide the discussion.

REFERENCES

[1] 2017. ESPN Football Power Index - 2017. ESPN Online. (Oct. 2017). http://www.espn.com/college-football/statistics/teamratings

[2] Dursun Delen, Douglas Cogdell, and Nihat Kasap. 2012. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting* 28 (2012), 543–552. https://doi.org/10.1016/j.ijforecast.2011.05.002

[3] Jeff Hansen. 2017. Sports Performance Platform puts data into play fi?! and action fi?! for athletes and teams. Official Microsoft Blog. (June 2017). https://blogs.microsoft.com/blog/2017/06/27/sports-performance-platform-puts-data-play-action-athletes-teams/

[4] Carson K. Leung and Kyle W. Joseph. 2014. Sports data mining: predicting results for the college football games. *Procedia Computer Science* 35, special issue of KES 2014 (2014), 710–719.

[5] Bahadorreza Ofoghi and John Zeleznikow. 2013. Data Mining in Elite Sports: A Review and a Framework. *Measurement in Physical Education and Exercise Science* (July 2013), 171–186. http://dx.doi.org/10.1080/1091367X.2013.805137

[6] Robert P. Shumaker, Osama K. Solieman, and Hsinchun Chen. 2010. *Sports Data Mining*. Springer.

[7] Jon Solomon. 2017. NCAA recommends ending two-a-day football practices and reducing tackling. CBS Sports Online. (Jan. 2017). https://www.cbssports.com/college-football/news/ncaa-recommends-ending-two-a-day-football-practices-and-reducing-tackling/

[8] Tom Taylor. 2017. Football's Next Frontier: The Battle Over Big Data. (June 2017). https://www.si.com/2017/06/27/nfl-football-next-frontier-battle-big-data-whoop-nflpa

[9] Mark Tracy. 2016. With Wearable Tech Deals, New Player Data Is Up for Grabs. The New York Times. (Sept. 2016). https://nyti.ms/2creZ4t

5 BIBTEX ISSUES

Warning–no key, author in espn2017

Warning–no author, editor, organization, or key in espn2017

Warning–to sort, need author or key in espn2017

Warning–no key, author in espn2017

Warning–no key, author in espn2017

Warning–no key, author in espn2017

Warning–no author, editor, organization, or key in espn2017

Warning–empty author in espn2017

Warning–no number and no volume in Ofoghi2013

Warning–empty address in Shumaker2010

(There were 10 warnings)

6 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

6.1 Structural Issues

Acknowledgement section missing