Machine Learning and Data Mining for Sports Analytics

Jan Van Haaren, Albrecht Zimmermann, Joris Renkens,
Guy Van den Broeck, Tim Op De Beéck, Wannes Meert, and Jesse Davis

DTAI, Department of Computer Science, KU Leuven
{firstname.lastname}@cs.kuleuven.be

Abstract

Sports analytics had its public breakthrough as early as the 1970s when baseball enthusiasts started developing a range of statistical tools for analyzing players, teams, and strategies. Due to a combination of early successes, increased computational power and advanced, automated data collection methods, sports analytics has been a steadily growing area in the last decade. The discipline is no longer restricted to designing new statistics and building simple statistical models. Moreover, sports analytics has found its way in other professional sports as well. Many professional sports clubs have started hiring performance analysts whose main task is to analyze the large quantities of data that are being collected nowadays, including play-by-play data, video tracking data, and sensor readings.

In hindsight, the success of sports analytics in baseball is not surprising. Baseball matches can easily be split into distinct events, which can be analyzed using simple statistical tools. However, more sophisticated techniques are needed to analyze more fluid sports such as basketball and soccer. In these sports, it is much harder to distinguish between events which are also not necessarily sequential and have more complex interrelations. In order to address this additional complexity and to deal with the large quantities of data that are available, we are applying machine learning and data mining techniques to this type of data.

In terms of soccer, we are looking at predicting match outcomes using raw statistics and match ratings [1], predicting and preventing injuries using test results and workloads, automatically generating match reports from play-by-play data, and identifying players' playing styles in play-by-play data. In terms of basketball, we are looking at predicting match outcomes using raw statistics [2], identifying playing styles of both teams and players, and determining optimal player rotations using play-by-play data as any number of substitutions can be made during a basketball match.

References

[1] Van Haaren, J., Van den Broeck, G. (2011). Relational Learning for Football-Related Predictions. Latest Advances in Inductive Logic Programming. Twenty-First International Conference on Inductive Logic Programming (ILP-2011). Windsor Great Park, United Kingdom, 31 July - 3 August 2011.

[2] Zimmermann, A., Moorthy, S., Shi, Z. (2013). Predicting NCAAB Match Outcomes Using ML Techniques – Some Results and Lessons Learned. Machine Learning and Data Mining for Sports Analytics Workshop (MLSA-13). Prague, Czech Republic, 27 September 2013.

[3] Zimmermann, A., Davis, J., and Van Haaren, J. (2013). Machine Learning and Data Mining for Sports Analytics Workshop (MLSA-13) at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-2013). Prague, Czech Republic, 27 September 2013, http://dtai.cs.kuleuven.be/events/MLSA13/.