Table of Contents

List of tables

List of figures

1. aims and objectives.
2. Background and Context

There are two main approaches for player evaluation. The first is to use various statistics to sum up ‘‘the total contributions of a player to his/her team’’ into a number. The second approach is to assign values to the actions performed during a match. In the second approach, traditional methods (e.g., [1] J. Ensum, R. Pollard, and S. Taylor, ‘‘Applications of logistic regression to shots at goal at association football,’’ in Proc. 5th World Congr . Sci.Football, 2005, p. 214.) The associate editor coordinating the review of this manuscript and approving it for publication was Guangcun Shan. demonstrate significant limitations as they can only evaluate the actions that directly lead to a score (e.g., shooting), but are unable to evaluate those actions that indirectly lead to score. Recently, Markov models have been used to address this issue, which have the advantage of unified evaluation criteria (actions are evaluated in the same scale by anticipating expected outcomes). These approaches are based on the analysis of event stream data (including optical data) that describe the actions performed in a game. However, in racket sports, the task of action value evaluation of players is almost unexplored because, in such a sport, technical whole-body movements have to be evaluated in addition to the tactics.

* 1. Basic concepts of basketball
     1. Court
     2. Team structure
     3. Basic rules
  2. Related work

The related work can be divided into ranking and prediction in basketball. In ranking the performance of basketball players is evaluated by using various statistics whereas in prediction the outcome of the basketball game is predicted using machine learning classifiers.

* + 1. Ranking basketball players

NBA game statistics like points, blocks, rebounds, field goals etc2 are widely used for rating the basketball players. John Hollinger (per) introduced a formula that uses player box score statistics to measure the efficiency of the player. To know how much a player is efficient, the idea of on and off the court was proposed by [15]( P. Fearnhead, B.M. Taylor, On estimating the ability of nba players, J. Quant.Anal. Sports 7 (3) (2011).). They observed whether the team performance increases or decreases when a specific player is on the court or off the court. Both offensive, defensive and combination of both were used to measure the strength of NBA players. Using data from the 2008–2009 season, LeBron James was considered the best player. The impact of an NBA team player is evaluated by [16](S.K. Deshpande, S.T. Jensen, Estimating an NBA player’s impact on histeam’s chances of winning, J. Quant. Anal. Sports 12 (2) (2016) 51–72.), they used a Bayesian linear regression model for finding an individual player impact on the team winning. The said research ranks the players with respect to their team and across the leagues. Slack based measure method is used by [17] (F. Asghar, M. Asif, M.A. Nadeem, M.A. Nawaz, M. Idrees, A novel approach to ranking national basketball association players, J. Glob. Econ. Manage.Bus. Res. (2018) 176–183) to rank players in NBA. They compared their ranking with player impact measure approach. They conclude that even the players who are ranked top by slack based measure approach are ranked at bottom by player impact measure approach. The reason for getting different ranking on same data is because both methods work in different manner. Authors in [18] ( J. Koster, B. Aven, The effects of individual status and group performance on network ties among teammates in the national basketball association,PLoS One 13 (4) (2018) e0196013.) shows how network ties among players are affected through individual status and group performance. Relationship between game statistics and match outcome is explored by [19]( S. Zhang, A. Lorenzo, C. Zhou, Y. Cui, B. Gonçalves, M. Angel Gómez, Performance profiles and opposition interaction during game-play in elite basketball: evidences from national basketball association, Int. J. Perform. Anal. Sport 19 (1) (2019) 28–48.). They also consider how player’s technical and physical performance is affected by interaction of opposition.(补充一下那个可视化的东西，20%，然后PER不可靠这玩意！)

2.2.2 Match outcome prediction

The fuzzy rule-based system (FRBS) is proposed by [20]（K. Trawinski, A fuzzy classification system for prediction of the results of the basketball games, in: International Conference on Fuzzy Systems, IEEE, 2010, pp. 1–7.） for the prediction of the basketball match outcome. Feature selection was applied for the selection of best features and various fuzzy models were used for the prediction of match outcome. A model for college basketball was proposed by [21]（F.J. Ruiz, F. Perez-Cruz, A generative model for predicting outcomes in college basketball, J. Quant. Anal. Sports 11 (1) (2015) 39–52.） that combined a simple soccer model and Poisson factorization. The simple soccer model identifies each team by its attack and defense coefficients whereas the Poisson factorization considers the elements of the matrix that are independent of the Poisson random variables. For match outcome prediction in basketball an integrated model called HSVMDT(Hybrid Support Vector Machine and Decision Tree) is proposed by [22]（P.-F. Pai, L.-H. Chang Liao, K.-P. Lin, Analyzing basketball games by a support vector machines with decision tree model, Neural Comput. Appl. 28 (12) (2017) 4159–4167.）. Feature selection was used to select the best features (7 features were selected out of 17). HSVMDT was tested on both selected features and on all 17 features HSVMDT achieved 82.25% with feature selection and without feature selection, the accuracy was 67%. The decision tree generates many rules that can cause confusion for decision makers. Rules pruning was used to limit the number of rules. For measuring the quality of decision rules, the sum of testing accuracy and coverage index was used. The results showed that decision rules have better quality after pruning. The rules generated by said model aim to help coaches to identify which factors are affecting match outcome. Analysis based on classification and regression tree was performed by [23] (M. A Gómez, S. J Ibáñez, I. Parejo, P. Furley, The use of classification and regression tree when classifying winning and losing basketball teams, Kinesiology: Int. J. Fundam. Appl. Kinesiol. 49 (1) (2017) 47–56.) to find best predictor in order to classify teams as winning or losing teams. Their analysis showed that in fast paced games the importance of defensive rebounds is 100%, importance of free throws is 94.7%, assists 86.1% and importance of fouls is 55.9%. On the other hand the importance of variables in slow paced games are: free throws is 100%, defensive rebounds 82.3%, fouls 68.4%, assists 66.9%, 2-points 62.2% and importance of 3-point field goals is 62.1%. Data driven and data envelopment analysis-based techniques were used by [24](Y. Li, L. Wang, F. Li, A data-driven prediction approach for sports team performance and its application to national basketball association, Omega(2019) 102123.) for predicting the performance of sports team. They used multi-variate logistic regression to find relationship between winning probability and match outcome. Their study suggests that team coaches and managers should focus on communication and co-operation of team. Various machine learning models are used by [25]( F. Thabtah, L. Zhang, N. Abdelhamid, NBA game result prediction using feature analysis and machine learning, Ann. Data Sci. 6 (1) (2019) 103–116.) for the prediction of match outcome in basketball. They examined the strength of various features for match outcome prediction. The defensive rebound was observed to be the most suitable feature for match outcome prediction. Discrete-time and finite-state Markov chain has been used by [26](J. Shi, K. Song, A discrete-time and finite-state Markov chain based in-play prediction model for NBA basketball matches, Comm. Statist. Simulation Comput. (2019) 1–9.) to predict the outcome of the match when the game is in progress. The aim of the said model is to model the difference between the home team and the visiting team score at some time point. The predictions for the ongoing match can be made on the current score of the team instead of past data.

2.2.3 Applications of machine learning techniques

Machine learning models have wide range of applications. Here we give an overview of some of the application of machine learning techniques.

SVM [27] (C. Cortes, V. Vapnik, Support-Vector Networks Machine Learning, vol. 20, Kluwer Academic Publisher, Boston, MA, 1995.)and Naive Bayes [28] （I. Rish, et al., An empirical study of the naive Bayes classifier, in: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, vol. 3, no.22, 2001, pp. 41–46.）techniques have been used by [29]（S.V. Wawre, S.N. Deshmukh, Sentiment classification using machine learning techniques, Int. J. Sci. Res. (IJSR) 5 (4) (2016) 819–821.） for classification of movie reviews. Text classification based on document embedding is used by [30]（R.A. Sinoara, J. Camacho-Collados, R.G. Rossi, R. Navigli, S.O. Rezende, Knowledge-enhanced document embeddings for text classification, Knowl.-Based Syst. 163 (2019) 955–971.）. One of the application of machine learning in the domain of legal documents is presented by [31]（I. Chalkidis, M. Fergadiotis, P. Malakasiotis, I. Androutsopoulos, Large-scale multi-label text classification on eu legislation, 2019, arXiv preprintarXiv:1906.02192.）, where the authors applied various models for multi-label text classification on legislation documents. Words in pair neural networks is presented by [32]（W. Yujia, L. Jing, S. Chengfang, J. CHANG, et al., Words in pairs neural networks for text classification, Chin. J. Electron. 29 (3) (2020) 491–500.） for text classification that overcome the limitation of text classification based on single word with multiple meanings. Novel machine learning model SS3 proposed by [33]（S.G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Syst. Appl. 133 (2019) 182–197.） for text classification that have the ability of early risk detection on social media. Siamese capsule networks that are based on local and global features for text classification has been used by [34]（Y. Wu, J. Li, J. Wu, J. Chang, Siamese capsule networks with global and local features for text classification, Neurocomputing (2020).） Machine learning has also been actively used for classification of spam messages. A review of soft techniques for classification of sms spam is presented by [35]（O. Abayomi-Alli, S. Misra, A. Abayomi-Alli, M. Odusami, A review of soft techniques for SMS spam classification: Methods, approaches and applications, Eng. Appl. Artif. Intell. 86 (2019) 197–212.）. Discrete Hidden Markov Model is used by [36]（T. Xia, X. Chen, A discrete hidden Markov model for SMS spam detection,Appl. Sci. 10 (14) (2020) 5011.） for spam detection that has the capability to exploit the order of words and can handle the problem of low term frequency. Rule based algorithm with the ability of constant time complexity has been used for detection of spam [37]（T. Xia, A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems, IEEE Access 8 (2020)82653–82661.）classical machine learning technique are not much efficient in situations where the decisions are time-dependent, for such situations, [38]（Y. Chen, Y. Zhou, Machine learning based decision making for time varying systems: Parameter estimation and performance optimization, Knowl.-Based Syst. 190 (2020) 105479.） presented a machine learning model that have the ability to work in time varying systems. Machine learning techniques based on evolutionary frame-work has been used in medical domain on clinical data [39].（J.A. Castellanos-Garzón, E. Costa, J.M. Corchado, et al., An evolutionary framework for machine learning applied to medical data, Knowl.-Based Syst. 185 (2019) 104982.） For prediction of breast cancer, Support Vector Machines and Artificial Neural Networks has been applied by [40]（E.A. Bayrak, P. Kırcı, T. Ensari, Comparison of machine learning methods for breast cancer diagnosis, in: 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science, EBBT, IEEE,2019, pp. 1–3） on Wisconsin Breast Cancer dataset.