Literature Review

Project: Who Will be the Next Football Superstar?

--Using FIFA game's data to predict young player's potential

Siyuan Peng

The evaluation of match result and athlete performance is a widely discussed topic in sport science. Due to the tremendous amount of money on sports lottery, lots of attention and research resource have been put into the prediction of the competition result in prior years. However, with the emphasis of the quantitative research in the work of sport managers and the sharp rise of players' market value, the number of researches about the player's performance has also increased.

With the evolution of technologies and statistical tools, sports scientists are intensively using data mining and machine learning skills. At first, the technique is rather simple and straight-forward. Purucker(1996) [1]'s paper is one of the earliest research which used ML to predict sports result. He tried to predict the score of the National Football League (NFL) using an Artificial Neural Network (ANN) model. Instead of using the supervised neural network which is the mainstream method of now, he used unsupervised neural networks based on clustering to distinguish between good and weak teams. He used the best four teams and the worst four teams as the training set and conducted score prediction based on this ANN. His method gives me some hint that I might make an arbitrary prediction on my player data to judge whether a player would be a good or bad player in the future. However, this classification is slightly rough, and the evaluation of the player is not that straight forward as the team. Thus, this method could only be used as a pre-test to the data.

Seven years later, Kahn (2003) [2] extended Purucker's work with the same data source. He switched the model to back-propagation multi-layer perceptron (BP MLP) network,

which is a supervised ANN. This switch makes sense that too many modes influence the result of a football game and exposure to a large number of possibilities under supervised learning conditions would yield a network that had the better predictive capability. After carefully selection, Kahn used Total yardage differential, rushing yardage differential, Time of possession differential, Turnover differential and Home or away as model parameters and found the best model structure with a learning coefficient (α) of 0.01, momentum (μ) of 0, and a network structure of 10-3-2. As for the prediction accuracy, while Purucker's model is around 60%, Kahn's is about 75% which is even higher than the ESPN experts' prediction accuracy (70%). Kahn's model's success further proved the advantages of supervised ANN in sport science domain to dealing with the highly non-linear relationship between game result and sports statistical data.

McCabe (2008) [3] further improved the model, and he attempted to predict results in four different sports: NFL (Rugby League), AFL (Australian Rules football), Super Rugby (Rugby Union), and English Premier League Football (EPL). Besides back-propagation, he also used the conjugate gradient as the learning algorithm. The inputs of his model are shared with all these sports, such as points-for, points-against, home performance, away performance and so on. With his three-layer MLP with nineteen input units, ten hidden units and a single output unit, the average performance of his model prediction was around 65%. A highlight in his model is one parameter called Player Availability. That parameter shows the availability of the critical player in the team, and it just gives me some hint that there might be some indispensable parameter that the full status (like a whole team's performance) could be influenced hugely with or without it. In the case of the player potential, that parameter might be the height of the goalkeeper or the speed of the forward.

As for prediction in single player's performance, in Iyer's (2009) [4] paper, he used three different ANN to suggest the coach on whether a player should be invited to the national team or not. The model classified players as "Performer," ‘‘Moderate," and

‘‘Failure” and the standard for this classification is different within different positions. He used ANN to evaluate the performance of each player in each year, and those who are above the threshold (a specific number of “performer” or “moderate”) are qualified for the national team. As for data processing, the author used two different approaches. He cleaned the data and kept all the years in the first database. As for the second one, he segregated the data into different years and further modified the threshold to make sure it is fair for those new face (the data ends at 06/07 season, and some young players join the league at 05/06 or even 06/07). Thus, his model could also recommend promising rookies. Finally, he used MLP, linear neural network and Radial basis function (RBF) neural networks to predict the performance of the player and his model’s average accuracy is about 80% (in which MLP performs the best). The most inspiring part of his paper is not the method like RBF he introduced, but different ways of data processing and segregation. In my model, I will also try to segregate my data in different ways (like years, position, age group and so on) and find the best way with the highest accuracy.

Maszczyk’s (2014) [5] further compared the neural networks and non-linear regression to predict the distance of Javelin throws. The data he used involved a group of 116 javelin throwers, aged 18±0.5 years. Using correlation matrix and regression analysis, they found four significant predictors of javelin throw and used them to build the non-linear regression and MLP neural network. The results of the investigation into the group of 18-year-olds javelin throwers show that the created neural models offer a much higher quality of prediction than the nonlinear regression model (absolute network error 16.77m versus absolute regression error 29.45 m), which again shows the superiority of ANN towards other methods on sport analysis.

In Rory’s (2019) [5] paper, authors give a machine learning framework for sports result prediction. This paper intensively focused on the classification of the result of the sports match (win, loss or tie) using Artificial Neural Network (ANN). Author decomposes the framework into six steps: Domain Understanding, Data Understanding, Data

Preparation & Feature Extraction, Modelling, Model Evaluation and Deploy Model. I will follow this ‘SRP-CRISP-DM’ framework in my project.

As for other method used in the sports analysis, Miljković’s (2010) [7] paper used the Naïve Bayes model to predict the results of NBA matches. The classifiers are host win or visiting win. Besides, they used multivariate linear regression to examine the relationship between the point difference and all other attributes on the spreadsheet. Finally, using K fold cross-validation as the evaluation system, their model predicts the result with a 67% accuracy, which is not quite ideal. Agarwal’s (2017) [8] paper shows the necessity of using Map-Reduce Jobs when facing big data. They used Hive Framework over Hadoop to deal with a large cricket sports dataset. Zdravevski’s (2009) [9] and Trawinski’s (2010) [10] paper used WEKA (Waikato Environment for Knowledge Analysis, a non-commercial and open-source data mining system) to predict the match result. With Fuzzy system, decision trees, random forest, and other methods, their prediction accuracy is around 70%.

Stanojevic’s (2016) [8] paper is the one that most similar to my project. In his paper, he experimented with several ensembles supervised learning methods including random forests, gradient boosting trees regression (GBT) as well as generalized linear models for assessing the player’s market value using players’ performance data. He then compared the transfermarkt.com (most widely used market value estimates website) market value estimate (TMVE) and his performance-driven market value estimate (PDMVE). The median difference between the TMVE and PDMVE is around 34%, while the mean difference is about 60%. He then used these two market value indicators as the performance predictor and tried to find which one of the two metrics is a better predictor of the game outcome. His result showed that PDMVE (the market value calculate by his model) has higher Pearson correlation, lower median error, and lower RMSE. I think the whole structure of his paper is inspiring. Using the parameter made by the expert as the supervised parameter (for him is the market value in transfermarket.com, for me is the potential value made by EA) to generate the model,

then, try to compare the original parameter and the model calculated parameter.

When I was searching for papers related to my project, I didn’t find any document mainly focused on the prediction of the player’s potential or using game data. Their research primarily focused on the result of the game or the qualification and market value for that player. My project will be the first one to discuss such a question. Different from real-world data, game data has more attributes which I could select. The model in those papers usually has 5 to 8 parameters while I have more than 20 characteristics which descript the physical, skill and psychological level of the player. Besides, I admit that the prediction of the player potential might not give you an instant profit just like the result of the match or the underestimated market value of the player. However, it could help the manager of the team to make the correct judgment which would benefit the future. A considerable gap between the potential and current performance could also be viewed as an underlying investment, and the rate of return could be huge as well.

Reference

[1] M.C. Purucker, Neural network quarterbacking, IEEE Potentials 15 (1996) 9–15.

[2] Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9-15.

[3] McCabe, A., & Trevathan, J. (2008, April). Artificial intelligence in sports prediction. In Fifth International Conference on Information Technology: New Generations (itng 2008) (pp. 1194-1197). IEEE.

[4] Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), 5510-5522.

[5] Maszczyk, A., Gołaś, A., Pietraszewski, P., Roczniok, R., Zając, A., & Stanula, A. (2014). Application of neural and regression models in sports results prediction. Procedia-Social and Behavioral Sciences, 117, 482-487.

[6] Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), 5510-5522.

[7] Miljković, D., Gajić, L., Kovačević, A., & Konjović, Z. (2010, September). The use of data mining for basketball matches outcomes prediction. In IEEE 8th International Symposium on Intelligent Systems and Informatics (pp. 309-312). IEEE.

[8] Agarwal, S., Yadav, L., & Mehta, S. (2017). Cricket Team Prediction with Hadoop: Statistical Modeling Approach. Procedia Computer Science, 122, 525-532.

[9] Zdravevski, E., & Kulakov, A. (2009, September). System for Prediction of the Winner in a Sports Game. In International Conference on ICT Innovations (pp. 55-63). Springer, Berlin, Heidelberg.

[10] Trawinski, K. (2010, July). A fuzzy classification system for prediction of the results of the basketball games. In International conference on fuzzy systems (pp. 1-7). IEEE.

[11] Stanojevic, R., & Gyarmati, L. (2016, December). Towards data-driven football player assessment. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW) (pp. 167-172). IEEE.