Automated Player Selection for a Sports Team using Competitive Neural Networks

Rabah Al-Shboul
Al Al-Bayt University,
Mafraq, Jordan

Tahir Syed
National University
of Computer and &
Emerging Sciences, Pakistan

Jamshed Memon
Barrett Hodgson University,
Pakistan

Furqan Khan
French Institute for Research
in Computer Science
and Automation (INRIA), France

***Abstract*—The use of data analytics to constitute a winning team for the least cost has become the standard *modus operandi* in club leagues, beginning from Sabermetrics for the game of basketball. Our motivation is to implement this phenomenon in other sports as well, and for the purpose of this work we present a model for football, for which to the best of our knowledge, previous work does not exist.**

The main objective is to pick the best possible squad from an available pool of players. This will help decide which team of 11 football players is best to play against a particular opponent, perform prediction of future matches and helps team management in preparing the team for the future. We argue in favour of a semi-supervised learning approach in order to quantify and predict player performance from team data with mutual influence among players, and report win accuracies of around 60%.

Keywords*—Team selection; match outcome prediction; neural networks

I. Introduction

Predicting game outcomes in sports is both challenging and interesting for their potential value for betting houses, team managements, sports fans, sports media, etc. More specifically, management of football teams are interested in selecting the best team to play to maximize their chances of winning a game, hence optimizing their return on investments. Betting houses on the other hands, would like to have this ability so that they could adjust the betting odds to maximize their profits. Sports media can optimize their contract values for teams and players based on their likelihood of winning. We would not be a miss if we consider sports prediction market to be multi-billion and still not well tapped in.

Sports including baseball and basketball have used analytics based on past records and statistics to analyze and produce results for future use. It is a realistic expectation nowadays that automated systems should be used to predict results. Sabermetrics [8] (the baseball analytic model) is one such practical implementation. Basketball coaches and managers have used analytics regularly to maximize their results.

Creating an optimal lineup of players that is capable of winning over another lineup is a major challenge [2]. The difficulty comes from different player positions requiring different skills. This means that taken collectively, players performing optimally on as individuals does not necessarily translate to an optimal combination, because of new team dynamics.

This work aims to aid team managers and selectors in identifying the best possible team to play that has the best odds at winning. We focus on team games and on player statistics and analytics computed from past data. We also perform match-by-match analytics of matches played between specific opponents. Specifically, our objective is two-fold:

1) Selecting the best possible team combination for a specific match given the knowledge of a particular opponent, and
2) Predicting the likelihood of victory in a match given the knowledge of the two lineups.

This paper is concerned with creating a tool that allows football team management to do analysis on their pool of players and thereby generate a ranking based on that analysis. This give rise to two main questions, the first of which being, "What is being analyzed and whether it could help construct a ranking?" and the second being, "How is this analysis being performed?"

We focus on football (soccer) for it is arguably the most popular and one of the most unpredictable games in the world where the occurrence of an upset is arguably more likely than any other sport. The unpredictability factor in this sport will make this project challenging for us to complete.

Van Haaren et al. [1] have discussed the limitation of data available for the purpose of analysis in football. Most techniques for predicting football match outcomes have been derived from methods regularly used by statisticians. Most of these include estimation techniques that used parameterized models. The values of these parameters are learned (or estimated, from a statistical standpoint) from scores of a history of football matches. The authors give two difficulties with this approach:

1) Match statistics for club football are usually not publicly available, in contrast to American sports like basketball and baseball, where they are available in detail online. That has led to a profusion of interest in those sports than in football.
2) It is not obvious how to derive meaningful measures and statistics from football matches [1].

Our interpretation of the second problem lies in the fact that these sports are fundamentally different. In baseball it is relatively simple because it is a series of individual match-ups between a hitter and a pitcher. In basketball, it is a more

challenging than in baseball because there is a lot of scoring and rolling substitutions. Still, this leaves us with a wealth of data of different groups of players on the pitch. The problem is substantially more complex in football due to:

- the dearth of scoring,
- the dynamic nature of the play, and
- data sparsity.

For example, no two teams play 4-4-2 the same way. Some play narrow, some wide - some use a diamond in the midfield, while some use a double-pivot. Things get even more complex with the variations of 4-5-1, which is the most used formation across the top 5 leagues in Europe in recent years. So if we account for all the variations, we will obtain with very sparse datasets.

In recent past, association football has undergone a metamorphosis in terms of professionalization and the incorporation of technical advances [3]. The previous generation of predictive models for football almost exclusively worked on the basis of the number of goals scored in a match. As an example, Maher's [4] model predicts outcomes of football matches given two lineups. The model proposed uses individual Poisson variables to calculate the score for both teams separately. Dixon and Coles [5] adopt Maher's model while proposing some changes. They show that there exists a strong dependency between individual scores in low-scoring football matches.

Current football-related prediction techniques are typically applications of statistical methods that fail to exploit the full range of information in the available data and are limited to learning from football match histories [6]. In spite of the vast popularity of football, research has is lacking in terms of more sophisticated models that take the numerous events that influence a match result (e.g., a red card) as well as time-dependent and positional information (e.g., dribbles and passes) into account.

The biggest challenge is that we do not know which data are trustworthy - "clean" in machine learning parlance, and which contain duplicates, invalid records and inaccurate data entries. According to Gartner, a shocking proportion of all business data are inaccurate [7].

II. Methodology

The work presented here is mainly in two major directions:

- Player rating.
- Team predictions.

The main goal is what we call *team selection*. Team selection includes mainly both tasks but player rating has to be changed from independent player rating to relational player rating that contribute in team winning and Team predictions are less important and instead we have to find a combination that maximizes the team's winning chances. This approach is converse of team prediction. A major challenge is to compete with the human mind. Coach will think of a player in a different way rather than statistics but we want to provide a statistical comparison to use the power of stats in determining the importance of a player. We take advantage of neural networks to learn relative ranking criteria from individual players' statistics.

In the following we first described our prediction work and then introduced our learning methodology for the given predictive model and then finally we selected an optimal squad for a given team.

A. Predictive Model

The predictive model is used to generate player contribution for an individual player relative to others in his/her team. However a naïve approach of using the winning ratio might be misleading for it might be too similar for multiple players. What is the legitimation of winning ratio arguably being the correct true label? To address this concern, we used a semi-supervised approach to find player importance. For this purpose we trained a neural network. Neural network analyzes the input features of all the players, separately for selected and opponent team and uses the final outcome to generate two individuals team scores. Both outcome are combined to get the final win/loss. The learning process assigns weights to the input links of each player. So in this manner we would be able to get an evaluation measure of each player with respect to his position. These weights are kept non-negative by saturating them at 0 so the impact of the player is in favor of his team rather than with the opponent.

B. Team Selection

In the team selection using the weights learned by the neural network team we generate player rating according to its playing position in the team. The quantified attribute of players are first multiplied by the weights generated by the first visible and first hidden layer of neural network. This will generate scores for individual player for his performance in a particular team.

$$P_{si} = \Sigma_{j=1}^{n}(D_j \times {}_1j)_i \tag{1}$$

$$\mathbf{P}_{si} = \mathbf{D}_i.\Theta_1 i \tag{2}$$

where, P_s = Player score, ${}_j$ = Attribute of one player, θ_1 = Weights from first hidden layer of the neural network.

The individual player scores are then multiplied by the weights generated by the first and the second hidden layer of the neural network.

$$T_s = \Sigma_{i=1}^{n}(P_{si} \times \theta_2 i) \tag{3}$$

$$\mathbf{T}_s = \mathbf{P}_i\Theta_2 \tag{4}$$

where, T_s=Team Score and θ_1= Weights from second hidden layer of the neural network.

After summing up the calculated scores we will get magnitude of how good a particular team is. This procedure is initially applied for the opponent team to get its best team score, secondly we apply the same method for subject team which will give us all the possible combination rated higher

than the opponent team. By this we achieve our target in two ways, firstly selecting a better subject team then its opponent and secondly giving multiple team combinations for subject team.

III. Experiments

The dataset comprises four categories of players with the features listed. Note that the three categories sharing the complete feature set are mentioned together.

- *Keepers*: The feature set of keepers include age, game starts, substitute ins, saves, goals conceded, fouls committed, fouls suffered, yellow cards, red cards and wins. The last one is binary, while the rest are real-valued.
- *Defenders*, *Midfielders*, *Strikers*: The feature set of keepers include age, goals scored, goals conceded, shots taken, assists, fouls committed, fouls suffered, yellow cards, red cards and wins.

The data for Keepers include 285 samples, Defenders have 957 samples, and Midfielders have 1026 samples and Forwards have 549 samples in all. This dataset comprises 10 years of data from the English Premier League (EPL).

A. Network Architecture for Player Rating

We use neural networks for semi-supervised learning using data of matches played during the past 10 years. Random weights are initialized to the neural networks and data inputs are the features of players that played the game. The input to neural network is a set of matches played. The target variable is the match outcome for the subject team. The architecture consists of 11 input nodes one for each player. The first hidden layer has the same number of neurons as input layer with one to one connection and it receives the transformed sigmoid values of the initial attributes. Lastly, the output layer is the combination of both of these. The accuracy is calculated thus: (true positives + true negatives) / total test data.

The results show 54% accuracy with the given set of data. With more data, this accuracy is likely to improve. Fig. 1 shows the architecture. Our process is justified due to the fact that we do not need to rate individual players on their own individual abilities rather we have to maximize the team winning probability. The train and test accuracies on 5-fold cross validation and their averages are shown below:

TABLE I: Accuracies for the player rating neural network.

ALPHA	1	0.1	0.3
a1	52.34	51.44	56.431
a2	54.32	54.22	55.034
a3	55.65	55.63	55.697
a4	53.23	52.43	49.877
a5	54.54	51.522	53.454
AVG	54.016	53.0484	54.0986

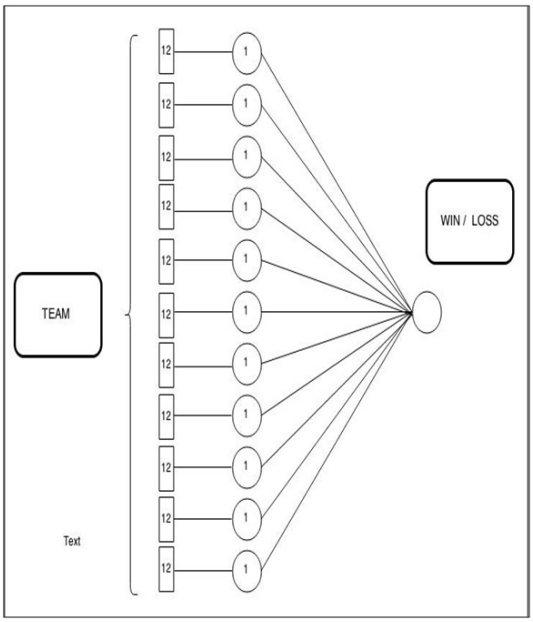

Fig. 1: The player selection neural network architecture.

B. Network Architecture for Team Predictions

We then revised our architecture by considering the players of the opponent team as well. New architecture consists of 22 input nodes, 11 for each team's players and an additional neuron for home/away value. First hidden layer also has the same number of neurons and it receives the transformed sigmoid values of the initial attributes. The second hidden layer then combines all players of subject team into one neuron and all players of opponent team into another neuron. Lastly, the output layer is the combination of both of these. Fig. 2 shows the neural network architecture. Fig. 2 illustrates the network. In general we can introduce arbitrary number of hidden layers before L1 and L2 of Fig. 2. Hence our model is extensible and can learn non-linear player features.

TABLE II: Accuracies for the Team Prediction Neural Network

ALPHA	1	0.1	0.3
a1	61.6774	61.2903	58.8710
a2	61.6774	55.6452	59.6774
a3	61.6774	54.8387	59.6774
a4	61.8710	55.6452	59.6774
a5	60.8065	59.6774	59.6774
AVG	60.741	57.4194	59.5161

IV. Results

In the first experiment, we used the neural network with 11 players in the input layer only considering the subject team. It gave a training accuracy of 54% with 5-fold cross validation.

In the second experiment, we use the neural network with 22 players, 11 for each side (subject and opponent teams) with an additional hidden layer. It gave training accuracy of 5-fold cross validation of 60%, improving by approximately 6%.

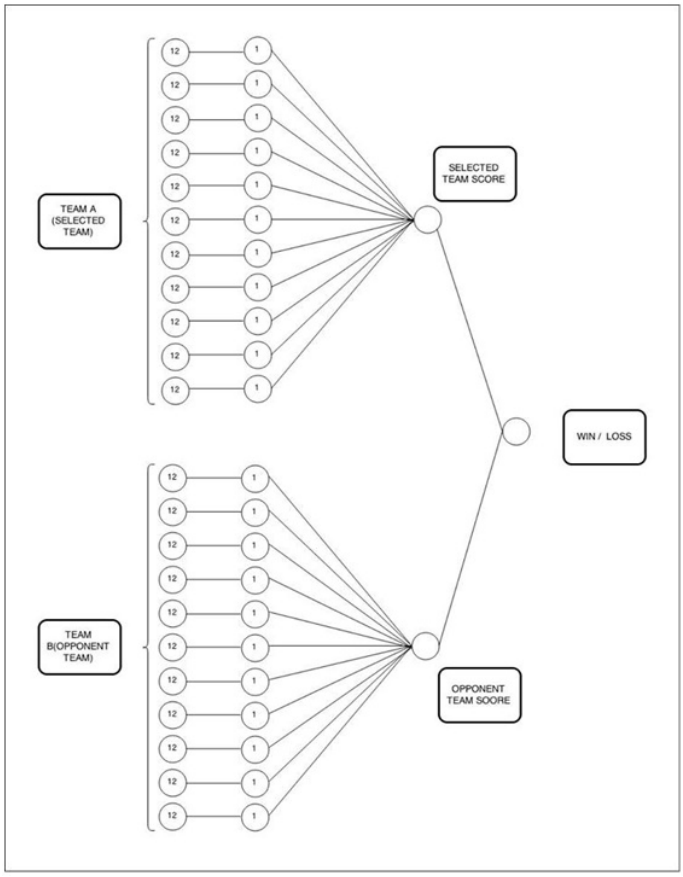

Fig. 2: The team selection neural network architecture.

V. Conclusions

We present a design of an ANN that is tailored to assist team managers in the selecting a team that will provide the best performance against a given opposition. The architecture is currently designed for a $(4, 4, 2)$ team combination which means 4 defenders, 4 mid-fielders and 2 forwards. Secondly the system will be able to help team managers select players to buy and drop from their own and other teams in order to form team combination that can provide best possible performance.

A future direction of research could be to make this problem generic to work with all possible team formations. Another could be working on a methodology for new players that do not have previous records, to help predict their future performance.

Individual player ratings can be useful in the proper context. If a team is trying to replace a player with someone who is very similar, ratings like these can be used to short-list transfer targets. Better still, if and when we get to a point where youth and academy players have the same level of detailed data as professionals, the ratings can be used to gauge the progress and map the career trajectory of academy players.

Acknowledgements

The authors appreciate support from M. Hani, M. Jhone, M. Raza.

References

[1] Jan Van Haaren, Albrecht Zimmermann, Joris Renkens, Guy Van den Broeck, Tim Op De Beck, Wannes Meert, and Jesse Davis, “Machine Learning and Data Mining for Sports Analytics,” Department of Computer Science, KU Leuven.

[2] Ohlmann, Michael J. Fry and Jeffrey W., “Introduction to the Special Issue on Analytics in Sports Part I: General Sports Applications,” Department of Operations, Business Analytics and Information Systems, University of Cincinnati.

[3] Jan Van Haaren and Guy Van den Broeck, “Relational Learning for Football-Related Predictions,” Katholieke University Leuven, no. 2010.

[4] Maher, “Modelling Association Football Scores,” Statistica Neerlandica 36(3), 1982.

[5] Dixon, M, Coles, “Modelling Association Football Scores and Inefficiencies in the...,” Journal of the Royal Statistical Society: Series C (Applied, 1997.

[6] Thomas H. Davenport and Jeanne G. Harris, “Competing on Analytics,” The New Science of Winning. Harvard Business School Press, March 2007.

[7] J. V. Haaren, “Analyzing Football Matches Using Relational Performance Data,” Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium.

[8] James Albert, Jay M. Bennett, “Curve Ball: Baseball, Statistics, and the Role of Chance in the Game”. Springer. pp. 170171, 2001