Modeling Team-Compatibility Factors Using a Semi-Markov Decision Process: A Data-Driven Framework for Performance Analysis in Soccer

Ali Jarvandi
Department of Engineering Management and Systems Engineering
George Washington University
Washington, DC USA
ajarvandi@gmail.com

Thomas Mazzuchi, Shahram Sarkani
Department of Engineering Management and Systems Engineering
George Washington University
Washington, DC USA
emseocp@gwu.edu

***Abstract*—Player selection is one of the great challenges of professional soccer clubs. Despite extensive use of performance data, a large number of player transfers at the highest level of club soccer have less than satisfactory outcome. This paper proposes a Semi-Markov Decision Process framework to model the dependencies between players in a team and its effects on individual and team performance. The study uses data from the English Premier League to artificially replace players in their prospective teams and approximate the outcome using a simulation. A comparison between the expected and actual performance determines the goodness of the model. In early experiments, the output of the model has correctly identified the trend changes in scoring and conceding goals and provided a high correlation with actual performance.**

Keywords-Markov Modeling; Decision Process; Sports; Simulation.

I. INTRODUCTION

Player transfers are a significant part of each soccer club's plan of action towards technical and financial success. This process has been traditionally guided by observation-based scouting. With the significant progress in data collection tools and methods in the recent years, professional soccer clubs have been collecting large amounts of data on their prospective players in order to make better transfer decisions. However, a preliminary study shows that about 40% of major signings by top European clubs between years 2010 and 2011 have been unsuccessful [1]. This translates to millions of dollars of loss for clubs. While detailed data is available on different aspects of players' performance, the challenge is to accurately utilize large amounts of data on various parameters towards making a single decision as to whether or not hire a player. This problem is particularly difficult due to the highly stochastic nature of the game as well as the interdependence between players, known as the compatibility factor [2]. While the challenge remains to develop an end-to-end model that addresses the dynamics of the game fairly accurately, there has been much progress in developing statistical models describing individual aspects of the game such as formation [3], possession [4], and goal scoring [5]. Though these models have offered great value to clubs in terms of making tactical decisions and designing training sessions, they have not been as useful in player selection. This is primarily because these models have not been applied in a context that addresses the two important issues of stochastic behavior and team compatibility.

This study is proposing a data-driven framework that incorporates existing sub-models in an end-to-end simulation in order to approximate player and team performance. Using a Semi-Markov Decision Process, this approach provides a stochastic foundation that reflects the nature of the game. In addition, the proposed approach offers the capability to artificially replace players in various teams and therefore include the effect of team compatibility in the estimated team performance. The historic performance data on players' technical and decision making attributes are used to construct the decision likelihoods and Transition Probability Matrices (TPM) needed for a Semi-Markov Decision Process. Finally, a strategy is developed to both test model accuracy and translate the output into specific predictions within the obtained accuracy. The ultimate goal of this study is to provide a decision support tool that assists decision makers in achieving a higher success rate in transfer decisions.

This paper provides an overview of the problem, state of the art, proposed approach, modeling strategies, and accuracy measures. Finally, the preliminary results have been listed and potential future directions have been discussed.

II. STATE OF THE ART

The studies of team selection in sports have been highly influenced by deterministic operations research. Boon and Sierksma modeled the team selection process using an integer programming approach in 2003 [6]. Also Dobson and Goddard computed the expected payoff associated with each playing strategy in 2010 [7]. In 2011, William A. Young developed a Knapsack model to approximate the compatibility between players in American Football using players' key individual attributes. While all of these studies provide valuable insight into the expected team performance, none of them has utilized players' decision making data or attempted to model the flow of the game.

This study considers the effects of individual decisions and interactions and offers the following advantages over the other methods:

 ISBN: 978-1-61208-242-4

A. *Using each player's decision making attributes in addition to technical attributes to model the flow of the game*

B. *Applying different sub-models to model various parts of the game with higher accuracy*

C. *Utilizing data in a stochastic framework, which is more reflective of the game*

D. *The ability to provide insight into the details of team and player performane as a result of modeling the entire flow of the system in approximately 1500 iterations per game*

III. METHODOLOGY

This study utilizes player performance data to model the interactions between players in a network. This leads to an approximation of the context generated for each player, which is then used to compute expected contribution to team performance in a simulation. The model is using a discrete time Semi-Markov Decision Process that provides players with a set of available decisions in each state at time t. Then the probability of transitioning to each state at time t+1 is computed by the probability of taking a decision multiplied by the corresponding value in the TPM. For example, a player in state one has four available decisions consisting of short pass, long pass, shoot, and dribble. The player's historical data provides the percentage of the time that each of these decisions is made and the success rate associated with each decision. This feeds the likelihood fields and the TPM needed for that player in state one. Fig. 1 shows the transition for a given decision at time t. In the figure, $P(D_1, S_k)$ represents the probability of transitioning to state k under decision 1.

The possible states for each player at a given instant in the game have been defined as following:

A. *Player has possession of the ball*

B. *A teammate has possession of the ball*

C. *The opposition has possession of the ball*

D. *Neither team has full possession of the ball*

Each of these states is associated with a set of available decisions that is mutually exclusive from the decisions available in other states. This results in 4 TPMs for each player, corresponding to each possible state. Table I shows a sample TPM from the model.

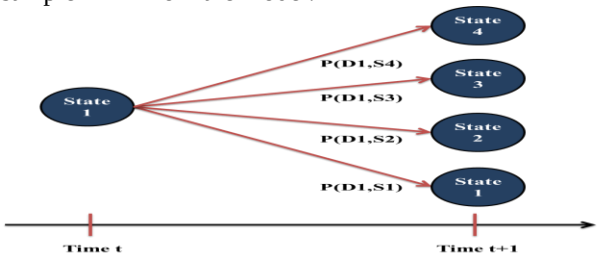

Figure 1. Transition from state 1 under decision 1.

TABLE I. TPM AND DECISION LIKELIHOOD IN STATE 1

Likelihood	Cesc Fabregas	State			
	Decision	1	2	3	4
0.8564	Short Pass	0	0.8241	0.1759	0
0.0765	long Pass	0	0.5954	0.4046	0
0.0414	Shoot	0	0	0.7042	0.2958
0.0257	Dribble	0.55	0	0.45	0

In addition to the transition probability matrices, which reflect players' individual trends, this study uses dependency matrices for various decisions to capture the effects of individual decisions on team trends. For instance, an 11x11 matrix represents the probability of each prayer playing a short pass to each teammate, given that a decision has been made to play a short pass.

Finally, two logical flows are developed to capture the processes for scoring and conceding goals. For goal scoring, two attributes of chance creation and conversion are captured for each player. These strictly individual attributes in the context of Transition and Dependency Matrices previously formed, will determine the team's expected scoring record, goal scorers, and more frequent plays. The process of conceding goals however, cannot be modeled in a similar way due to the lack of contribution measures for individual players. Therefore, it is significantly more relevant to model the process of conceding goals as a team process rather than sum of individual contributions. Previous studies on goal scoring in soccer have suggested different probability models based on two parameters: the position in which the scoring team gains possession of the ball, and the number of passes that are exchanged among the scoring team that lead to a goal. Using these attributes in a reverse way can generate a model for conceding goals. As a result, each possession for the opponent carries two probabilities that contribute to the chance of conceding a goal: Starting Probability, which is defined, based on the position where possession of the ball has been lost; and Carrying Probability, which is a function of the number of passes exchanged within the opponent. A model containing these two probabilities can lead to an accurate estimation of conceding goals in different conditions.

IV. MODEL OUTPUT

A regression analysis of the goal differential and number of points obtained in three of the top European leagues (Spain, England, and Italy) shows that the number of points obtained by each team can be predicted with a large degree of confidence (r^2>0.9 in each of the three leagues) based on goal differential by using historical data (Fig. 2).

 ISBN: 978-1-61208-242-4

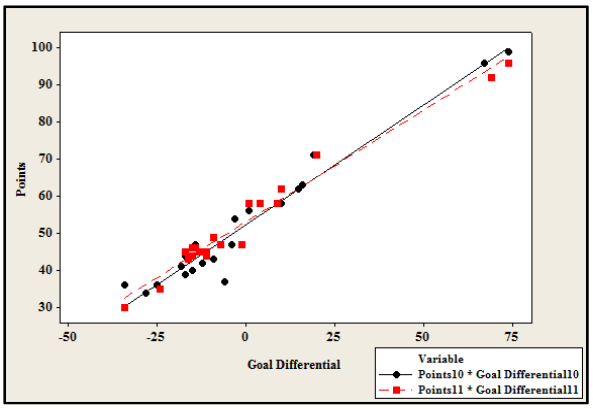

Figure 2. Points versus goal differential - Spain, 09/10 and 10/11 seasons.

The high correlation between points and goal differential in consecutive years enables the model to use expected number of goals scored and allowed as the ultimate team performance measure. Therefore, the significance of each player's technical and decision-making attributes is measured by its estimated contribution to scoring or conceding goals. The final model output will then be the expected number of goals scored and conceded in a large number of games. The difference between the expected changes in the number of scored and conceded goals and the team's previous record determines the estimated contribution of the new players to the team performance.

V. DATA AND RESULTS

The data used in this study includes game-by-game and season-by-season performance data for all players in the English Premier League between 2008/09 and 2011/12 seasons. The data also includes each player's overall playing minutes in the season, which is used to normalize player performance data with respect to other players. Also, while accounting for injuries is not in the initial scope of this research, it is important to note that in some cases raw data without normalization for injury periods can provide insight into the impacts of players' injury prospect on their overall performance. To analyze each transfer from team A to team B, three sets of data will be used: 1) Data from team A in year n, 2) Data from team B in year n, 3) Data from team B in year n+1. Using this criteria, 116 transfers have been identified that can be analyzed using the current dataset.

To approximate the expected team performance, the data for the player of interest from team A is replaced in the same playing position in team B and the player's impact will be measured by the difference in the expected number of goals scored and conceded compared to the output of the model for the original dataset 2. Finally, a comparison between the expected change in goal differential and the difference between the normalized goal differential from team B in year n+1 with the player of interest on and off the pitch determines the model accuracy. In the rare cases when a player plays the entire season, model accuracy is simply measured by a comparison between the expected and actual change in goal differential. For each transfer, model accuracy is measured using two parameters: Trend Identification and Correlation. Trend Identification is the measure for the success rate of predictions made by the model and refers to the overall effect of each transfer on team performance. The values for this attribute can be "Positive", "Negative", or "Insignificant". Assuming that the model correctly predicts this value, Correlation is computed. This measure determines the accuracy of the model in approximating the magnitude of the difference in the expected performance caused by the introduction of the new player. Currently, the model has been run for 5 transfers and has provided 100% success in trend identification and 81.3% average correlation with the actual performance. While these numbers can change with a larger dataset, the obtained accuracy is significantly higher than the one in the current transfer market.

VI. CONCLUSION AND FUTURE WORK

The proposed study provides a framework for utilizing player performance data for approximating the expected performance of a prospective player within a given context. As a decision support tool, this model will assist clubs in increasing their success rate in player transfers and reaching higher efficiency in budget allocation. Potential future work on this topic include a risk analysis of squad selection with this method with respect to potential injuries, the effect of substitute players on model accuracy, and the study of transfers from a league to another.

ACKNOWLEDGMENT

The authors would like to thank Dr. David F. Rico for his continuous support in proposing the original idea.

REFERENCES

[1] J. Ali, "A System Approach to Team Building in Soccer", Unpublished.

[2] W.A. Young, "A Team-Compatibility Decision Support System to Model the NFL Knapsack Problem: An Introduction to HEART," unpublished.

[3] N. Hirotsu and M. Wright, "Determining the best strategy for changing the configuration of a football team," Journal of the Operational Research Society, vol. 54, pp. 878-887, 2003.

[4] M. Shafizadeh, S. Gray, J. Sproule, and T. McMorris, "An exploratory analysis of losing possession in professional soccer," International Journal of Performance Analysis in Sport, vol. 12, pp. 14-23, 2012.

[5] A. Tenga and E. Sigmundstad, "Characteristics of goal-scoring possessions in open play: Comparing the top, in-between and bottom teams from professional soccer league," International Journal of Performance Analysis in Sport, vol. 11, pp. 545-552, 2011.

[6] B. Boon and G. Sierksma, "Team formation: Matching quality supply and quality demand," European Journal of Operational Research, vol. 148, pp. 277-292, 2003.

[7] S. Dobson and J. Goddard, "Optimizing strategic behavior in a dynamic setting in professional team sports," European Journal of Operational Research, vol. 205, pp.661-669, 2010.

[8] P. O'Donoghue, K. Papadimitriou, V. Gourgoulis, and K. Haralambis, "Statistical methods in performance analysis: an example from international soccer," International Journal of Performance Analysis in Sport, vol. 12, pp. 144-155, 2012.

 ISBN: 978-1-61208-242-4