Relational Learning for Football-Related Predictions

Jan Van Haaren and Guy Van den Broeck

jan.vanhaaren@student.kuleuven.be, guy.vandenbroeck@cs.kuleuven.be
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A, 3001 Heverlee, Belgium

Abstract. Association football has recently seen some radical changes, leading to higher financial stakes, further professionalization and technical advances. This gave rise to large amounts of data becoming available for analysis. Therefore, we propose football-related predictions as an interesting application for relational learning. We argue that football data is highly structured and most naturally represented in a relational way. Furthermore, we identify interesting learning tasks which require a relational approach, such as link prediction or structured output learning. Early experiments show that this relational approach is competitive with a propositionalized approach for the prediction of individual football matches' goal difference.

Keywords: Statistical Relational Learning

1 Introduction

Association football is becoming increasingly competitive. The financial stakes of football clubs are enormous as their budgets have increased enormously over the past two decades due to increasing revenues from gate sales, merchandising, broadcasting rights and price money [6]. Football clubs and football leagues have become more professional as a result. Football clubs have to adopt a well-thought selling and buying policy. Their managers have to exploit the potential of the players they have at their disposal the best they can. Modern technological tools to track football players are generating large amounts of data which are being used by experts to analyze matches and players' performance [9].

We will show that machine learning techniques can be applied to this data. Current approaches to football-related predictions do not use the rich data that is available nowadays. The techniques applied until now come from statistical modeling, not machine learning. We will argue that football-related data is relational and that therefore relational learning is particularly suited for football-related predictions. The rise of the Internet has made football betting increasingly popular. An interesting learning task therefore is the prediction of football match results. Despite the simplicity of both rules and objectives the results of football

matches are highly uncertain and difficult to predict. Football matches are typically low scoring, which makes it hard to identify events that have an immediate impact on the final score. We report on early experiments with kLog that show that our relational approach is competitive with a propositionalized approach for the prediction of individual football matches' goal difference.

2 Related Work

Football analytics has been given little attention in academic literature due to the limited availability of match statistics. A number of descriptive and predictive models for the outcomes of football matches have been proposed over the years though. The first generation of football-related models was mainly concerned with the distribution of the number of goals scored in a football match. Moroney [8] shows that the Poisson distribution is applicable to the results of football matches. He also shows that improvements are possible by using the negative binomial distribution. Maher [7] presents a model for outcomes of football matches between two specific teams. This model represents the score for both teams separately by means of independent Poisson variables. The model is able to take the respective qualities of both teams into account. Dixon and Coles [3] propose a number of adaptations to Maher's model. They show there exists a strong dependency between the individual scores in low-scoring football matches. Independent Poisson distributions are not able to take this dependency into account. Dixon and Coles therefore suggest to make direct modifications to the marginal Poisson distributions for low-scoring football matches. Baio and Blangiardo [2] suggest a Bayesian hierarchical model for the individual scores in football matches. They also show that there is no need for bivariate Poisson variables to capture the correlation between individual scores. The correlation is taken into account when assuming two conditionally independent Poisson variables for the number of goals scored as the observable variables influence each other at an higher level.

3 Current Challenges

Most of the available techniques for predicting football match results are applications and extensions of well-known statistical methods. These techniques learn models with a limited expressivity as they represent each team's strength by a limited set of model parameters. Typically these techniques solely learn from the final scores of previously played football matches. Two important reasons can be distinguished:

1. Until recently, *match statistics were usually not publicly available.* In contrast, for popular American sports (e.g., basketball and baseball) it is common that detailed match statistics are available both in print and online. Consequently, there has been an explosion in interest for analytics in these sports by academic researchers and fans alike.

2. It is not obvious how to derive meaningful measures and statistics from football matches. *Football has a very complex structure*, which cannot easily be captured by a fixed set of parameters. Football teams have a lot of freedom in the tactics they use and football players are almost free to perform the actions they want. Match results are typically very low as well. One touch of the ball can turn a draw into a home win in just a matter of seconds. Therefore it was not until about a decade ago that the registration of match statistics became possible. Modern technological tools are required in order to register match statistics for football matches on a large scale. Camera-based tracking systems are nowadays able to measure player and ball movements with a very high accuracy. Detailed match statistics for both football players and football teams can be easily derived from these measurements.

The discussion in the previous section shows however that the currently available approaches for modeling football matches are nowhere near capable of handling the huge amounts of data made available by tracking systems. Despite the immense popularity of football not much research is conducted on more sophisticated models. Two major challenges for such models can be distinguished:

Flexibility A model should be able to take the numerous aspects that influence the result of a football match into account. Match statistics do not only contain information on the final scores of football matches but also hold interesting details on the way these scores came about. Knowing for example that the referee pulled a red card for the home team might help explain a surprising win for the away team.

Rich data A model should be able to take time-dependent and positional information into account. An obvious example are player and team forms. Football teams and players are rarely able to perform at the same level for a long period of time. At a lower level, the passes and tackles performed during a football match are also time-dependent. When taking positional details into account as well interesting patterns regarding a team's playing style or complementarity between players can be revealed.

4 Relational Representation

Relational models possess interesting properties allowing them to handle the current challenges. An important asset of a relational model is the *flexibility* with which data can be represented. A relational representation of a football match should not adhere to a fixed form but can vary according to the events that happened during a football match. This way special events such as a red card or an own goal can be stressed. Relations also allow for the representation of data with a more complex structure such as a team's lineup or the transfer of a player between two teams. All of this is less obvious or even impossible in a propositional structure such as the attribute-value format.

The currently available approaches for modeling football matches have difficulties with taking *time-dependent and positional information* into account as

they hold their knowledge in a limited number of model parameters. These models implicitly assume that football teams and football players constantly perform at the same level. This is certainly not a valid assumption as football teams rarely perform at the same level for a long period of time. For example, the form of a team may help explain an apparently anomalous result. A relational model is however able to represent each player, team and match explicitly. Relations between matches can be defined for preserving the order in which these matches have been played. Hence performance gaps and form fluctuations can be represented in the data and captured by the model.

A relational model also allows for *learning from interpretations*, a key concept for many machine learning techniques. One approach could be to represent each individual football match as a single interpretation. Another approach could be to represent all football matches from one football season as one large interpretation. The latter approach offers the advantage that it also allows for relations between football matches.

5 Learning Tasks

Relational models provide the ability to describe plenty of interesting learning tasks. Both descriptive and predictive learning tasks can be performed. Descriptive tasks focus on assessing past performances (e.g., to identify who was the most efficient player in a football match) whereas predictive tasks are mainly concerned with analyzing past performances to predict future behavior (e.g., to predict how many goals a team will score in its next match). Traditional learning tasks include:

- **Regression:** e.g., to predict the number of yellow or red cards the referee will pull during a football match;
- **Classification:** e.g., to classify a football match according to its outcome: home win, away win or draw.

Besides these traditional learning tasks, relational models also allow for more complex learning tasks requiring rich data representations. Moreover, such models support structured output learning tasks which are very common in football. The output of these learning tasks can be represented in a straightforward fashion by means of entities and relations. Some interesting learning tasks made possible by relational models include:

- **Collective regression:** e.g., to jointly predict the statistics of players in a match;
- **Collective classification:** e.g., to identify which players will be selected in a team's starting lineup;
- **Link prediction:** e.g., to predict who passes the ball to whom during a match.

6 Learning Example

In this section we will illustrate the applicability of a relational model for football-related predictions using a simple learning task. This learning task comprises the prediction of the goal difference for individual football matches. The goal difference is determined by the home score minus the away score. The huge amount of relationships in football allows for a large number of different relational topologies. The relational model we discuss here focuses on the *explicit representation of individual football matches* and the relationships which exist amongst them. Besides its *goal difference* each football match is represented through a number of *performance measures*. These performance measures can be derived directly from match statistics. We consider thirteen different performance measures in five different domains per team: ball possession, discipline, defending, crossing and passing. Each of the performance measures can take five different values indicating the level of performance in that specific area.

We use kLog [4] for the implementation of the relational model. kLog is a language for *kernel-based relational learning*. kLog relies on several simple but powerful concepts including learning from interpretations, data modeling through entities and relationships, deductive databases and graph kernels. Unlike other models based on probabilistic logic, kLog derives features from a *ground entity-relationship diagram*. It is therefore not directly concerned with the representation of probability distributions. Numerical and symbolic data can be jointly used in the same model. Several statistical techniques are available to fit the model parameters.

Due to lack of comparable models we use propositionalizations of our relational model to assess its performance. We use the Weka Toolbox [5] to conduct experiments with *two different propositionalizations*. The first represents each football match using its 26 performance measures and its goal difference. The second also considers the performance measures and the goal difference for each team's two previous matches. We use the match statistics available at The Guardian's website [1] for the first nineteen match days of the 2010-2011 season of the Premier League for our experiments.

Table 1. Support vector regression (ϵ loss function) in kLog.

	linear	**polynomial**	**radial basis**	**sigmoid**
MAE	1.230	1.234	1.236	1.228
RMSE	1.725	1.728	1.729	1.728

Table 1 shows the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for four different kernel functions in kLog. These error values have been obtained through ten-fold cross validation. The table shows that the choice of the kernel function hardly has an impact on the performance of our kLog model.

Table 2. Support vector regression (ϵ loss function) in Weka.

	linear (1)	**linear (2)**	**sigmoid (1)**	**sigmoid (2)**
MAE	1.195	1.414	1.111	1.112
RMSE	1.532	1.769	1.484	1.485

Table 2 shows the MAE and RMSE for two different kernel functions and two different models in Weka. As for the kLog model these error values have been obtained by ten-fold cross validation with random folds. The table shows that the choice of the kernel function is important. The sigmoid function clearly delivers better results than the linear function for the second propositionalization.

Our preliminary experiments show that the relational approach we propose is competitive and yields interesting results. However, more experiments are needed before being able to draw any profound conclusions.

7 Conclusion

We have proposed football-related predictions as an interesting application for relational learning. We have argued that football data is highly structured and most naturally represented in a relational way. Interesting learning tasks requiring a relational approach include link prediction and structured output learning. Early experiments with a relational model yield promising results.

References

1. Guardian Interactive Chalkboards. `http://www.guardian.co.uk/football/chalkboards`
2. Baio, G., Blangiardo, M.: Bayesian Hierarchical Model for the Prediction of Football Results. Journal of Applied Statistics 32, 253–264 (2010)
3. Dixon, M., Coles, S.: Modelling Association Football Scores and Inefficiencies in the Football Betting Market. Journal of the Royal Statistical Society: Series C (Applied Statistics) 46(2), 265–280 (1997)
4. Frasconi, P., Costa, F., De Raedt, L., De Grave, K.: kLog - A Language for Logic-Based Relational Learning with Kernels. Tech. rep. (2011), `http://www.dsi.unifi.it/~paolo/ps/klog.pdf`
5. Holmes, G., Donkin, A., Witten, I.: Weka: A Machine Learning Workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems. pp. 357–361 (1994)
6. Lago, U., Simmons, R., Szymanski, S.: The Financial Crisis in European Football. Journal of Sports Economics 7(1), 3–12 (2006)
7. Maher, M.: Modelling Association Football Scores. Statistica Neerlandica 36(3), 109–118 (1982)
8. Moroney, M.: Facts from Figures (1956)
9. Xu, M., Orwell, J., Lowey, L., Thirde, D.: Architecture and Algorithms for Tracking Football Players with Multiple Cameras. In: IEE Proceedings - Vision, Image and Signal Processing. vol. 152, pp. 232–241 (2005)