Lifted Relational Team Embeddings for Predictive Sport Analytics

Ondřej Hubáček, Gustav Šourek, and Filip Železný

Czech Technical University, Prague, Czech Republic
{hubacon2,souregus,zelezny}@fel.cvut.cz

Abstract. We investigate the use of relational learning in domain of predictive sports analytics, for which we propose a team embedding concept expressed in the language of Lifted relational neural networks, a framework for learning of latent relational structures. On a large dataset of soccer results, we compare different relational learners against strong current methods from the domain to show some very promising results of the relational approach when combined with embedding learning.

1 Introduction

Sport analytics is a popular multi-billion dollar world-wide industry. It is a natural application domain for mathematical modelling, yet only recently we have been seeing penetration of modern machine learning methods into the field, with standard predictive techniques still being geared towards simple statistical models [9]. We argue that incorporating relational learning techniques might benefit the field considerably. It only seems natural as the data arising from sport records possess interesting relational characteristics on many levels of abstraction, from the matches themselves forming relations between teams, players and seasons, to the course of the individual matches being driven by the rules of each sport with game-play patterns stemming from these.

We investigate viability of the relational approach to the domain via experimental evaluation on soccer match outcome predictions based solely on historical results. We propose simple relational representations, background knowledge and modelling concepts for which we provide some interpretable insights. Particularly, we focus on expressing a concept we called "Lifted relational team embeddings" in the framework of Lifted relational neural networks (LRNNs) [13], combining relational fuzzy logic with gradient descend optimization. Finally, we experimentally compare different relational approaches with strong methods from the domain for their predictive performance on a large dataset of real soccer records.

1.1 Predictive Sports Analytics

In predictive sport analytics, the ultimate goal is to predict results of future matches. Given the stochastic nature of sports, the goal translates to correctly

estimating probabilities of the corresponding outcomes. Particularly for a game of soccer, the aim is to estimate the probabilities of the three possible outcomes *loss*, *draw*, *win*.

The task of predicting soccer results is well established in the literature. Typical approaches include statistical models based on Poisson distribution and its variations [5,10], as well as rating systems [2,7]. An example of relational learning approach was also introduced in [14], however the literature remains very scarce with these.

2 Predictive Models

We compare the proposed relational team embedding concept against a multitude of diverse learners. These consist of a simple prior probability baseline predictor, RDN-boost, a powerful SRL method for boosting Relational Dependency Networks [12], and an actual state-of-the-art model that won the mentioned Soccer Prediction Challenge [6]. Note that each of these learners has been actually selected for being a strong performer in the given task.

Baseline predictor is a simple model aggregating the prior probabilities of the individual home and away outcomes in each league. Being often surprisingly hard to beat, we include it as a baseline to serve as a natural lower bound for other learners' performance.

RDN boost learner follows a functional gradient boosting strategy on top of Relational Dependency Networks [12], powerful lifted graphical models designed to learn from data with relational dependencies using pseudo-likelihood estimation techniques. Similarly to LRNNs, RDN-boost learns from Herbrand interpretations for which it utilizes fragment of relational logic for representation, where the inner nodes of the individual regression trees of the resulting ensemble model represent conjunctions of the original predicates.

State-of-the-art model is the actual winning solution [6] from the mentioned 2017 Soccer Prediction Challenge. It is an ensemble, gradient boosted trees-based model utilizing expert-designed features. Some of these features are derived from other, already sophisticated, models from literature, such as the *pi-ratings* [2] or *page-rank* [8]. Other features are statistics based on expert insights incorporating the home advantage, historical strength, current form, or match importance. These are further aggregated in different ways w.r.t. seasons and leagues, to finally form an input into a carefully tuned XGBoost algorithm [1].

2.1 Lifted Relational Neural Networks

LRNNs [13] is a relational learning framework utilizing a parametrized fragment of relational fuzzy logic as a language for representation of various models and a

Table 1. Overview of predicates extracted from the data for the relational learners.

Predicate	Description
home(Tid)	Team Tid is home team w.r.t. prediction match.
away(Tid)	Team Tid is away team w.r.t. prediction match.
team($Tid, name$)	Team Tid has name $name$.
win(Mid, Tid_1, Tid_2)	Win of home team Tid_1 over away team Tid_2 in match Mid.
draw(Mid, Tid_1, Tid_2)	Draw between home team Tid_1 and Tid_2 in match Mid.
loss(Mid, Tid_1, Tid_2)	Loss of home team Tid_1 to team Tid_2 in match Mid.
scored(Mid, Tid, n)	The team Tid scored more than n goals in match Mid.
conceded(Mid, Tid, n)	The team Tid conceded more than n goals in match Mid.
goal_diff(Mid, n)	Difference in goals scored by the teams is greater than n.
recency(Mid, n)	The match Mid was played more than n rounds ago (w.r.t. prediction match).

gradient descend technique for their parameter training. The model representation can be viewed as a lifted *template* for neural networks, as it enables neural computations to be performed upon relational data by constructing a different computational graph, or neural network, for each of the differently structured relational examples.

For a regular training of an LRNN, as we do in experiments reported in this paper, one firstly needs to manually create the template, which may encode some background knowledge, or intuition, together with various modelling constructs. Secondly, one needs learning examples encoded in relational logic together with corresponding target predicate labels. Subsequently in the learning process, the LRNN engine grounds the template w.r.t. the different examples to create the corresponding neural networks, which are then jointly trained w.r.t. the labels, in a manner similar to that of standard deep learning frameworks.

2.2 Knowledge Representation

In its raw form, the match records contain merely the team names and the result, hence we tried to extract as much useful information as possible for each of the models. For the baseline this was straightforward, and for the SotA model this was already done [6]. For the approaches of RDN-boost and LRNNs we had to derive appropriate relational representation. Since they both learn from Herbrand interpretations, we encoded the records with numerical outcomes into predicates, which we describe in Table 1.

3 Lifted Relational Team Embeddings

Here we describe the proposed relational embedding model as expressed in the language of LRNNs. Firstly, we tested the hypothesis that there exists some predictive latent space embedding the teams. This is based on an intuition from various rating systems, such as the pi-ratings [2], where each team is assigned

one or more parameters denoting its particular strength, possibly within different areas, such as when playing at home stadium and when playing away. However, opposite to the existing rating systems, the idea of the embedding approach is to explore meaning of these latent parameters automatically by the means of regular learning from data. We can encode this scenario in LRNNs as follows.

$$
\begin{aligned}
&w_1^{(0)} : \mathit{type1}(T) \leftarrow \mathit{team(T,chelsea)} \\
&w_2^{(0)} : \mathit{type1}(T) \leftarrow \mathit{team(T,arsenal)} \\
&\dots \\
&w_j^{(0)} : \mathit{type3}(T) \leftarrow \mathit{team(T,everton)}
\end{aligned}
$$

where the types $\mathit{type}_1 \dots \mathit{type}_3$ denote individual embedding dimensions of the teams. We may directly use aggregation of such embeddings for prediction of outcome of *home* vs. *away* team matches using the following rules.

$$
\begin{aligned}
&w_{(1;1)}^{(1)} : \mathit{outcome} \leftarrow \mathit{home}(T1) \wedge \mathit{type}1(T1) \wedge \mathit{away}(T2) \wedge \mathit{type}1(T2) \\
&w_{(1;2)}^{(1)} : \mathit{outcome} \leftarrow \mathit{home}(T1) \wedge \mathit{type}1(T1) \wedge \mathit{away}(T2) \wedge \mathit{type}2(T2) \\
&\dots \\
&w_{(3;3)}^{(1)} : \mathit{outcome} \leftarrow \mathit{home}(T1) \wedge \mathit{type}3(T1) \wedge \mathit{away}(T2) \wedge \mathit{type}3(T2)
\end{aligned}
$$

This construct in principle creates a fully connected neural network with one hidden embedding layer, such as e.g. in the famous *word2vec* embedding architecture [11]. For all the historical matches we then jointly perform corresponding gradient updates of the weights to reflect the actual values of the *outcome* labels. We further denote this architecture as *embeddings*.

In theory, the embeddings possibly capture some information on the relational interplay between the matches as they are jointly optimized on the whole match history. However, we find this approach quite limited as it is rather naive to expect the flat, fixed-size embeddings to reflect all the possible nuances of the complex relational structure stemming from the different outcomes of different historical matches played between different teams in different orders. Fortunately with LRNNs, we can easily capture the relational structures explicitly while keeping the benefits of embedding learning. For that we first extend the template with a predicate capturing the different outcomes of *historical* matches (w.r.t. prediction match) through a learnable transformation as

$$
\begin{aligned}
w_1^{(2)} : \quad & \mathit{outcome}(M,H,A) \quad \leftarrow \quad \mathit{win}(M,H,A) \\
w_2^{(2)} : \quad & \mathit{outcome}(M,H,A) \quad \leftarrow \quad \mathit{draw}(M,H,A) \\
w_3^{(2)} : \quad & \mathit{outcome}(M,H,A) \quad \leftarrow \quad \mathit{loss}(M,H,A)
\end{aligned}
$$

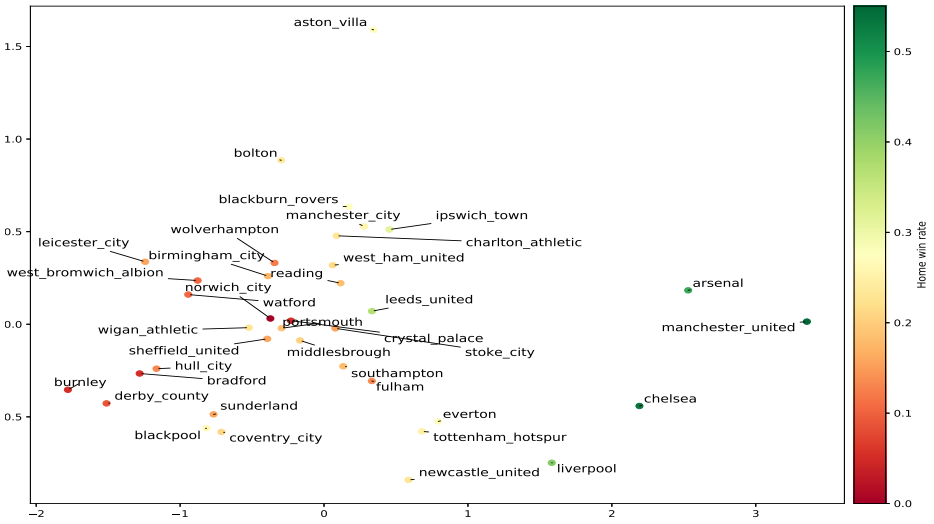

Fig. 1. Visualization of PCA projection of the learned embeddings of individual teams from the home-win model. A significant relationship between the home win rate, captured by the colorscale, and the variance captured by the main X axis can be observed.

with which we accordingly extend the predictive rules as

$$w^{(1)}_{h-h(1;1)} : outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T1, T2) \wedge type1(T2).$$
$$w^{(1)}_{h-a(1;1)} : outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T2, T1) \wedge type1(T2).$$
$$w^{(1)}_{h-h(1;2)} : outcome \leftarrow home(T1) \wedge type1(T1) \wedge outcome(M, T1, T2) \wedge type2(T2).$$
$$\ldots$$
$$w^{(1)}_{a-a(3;3)} : outcome \leftarrow away(T1) \wedge type3(T1) \wedge outcome(M, T2, T1) \wedge type3(T2).$$

reflecting the possible settings of *historical* home and away positions of the *actual* home and away teams in all historical matches played. By grounding this template, the LRNN engine assures to create the corresponding relational histories transformed into respective, differently structured, neural networks. We denote this architecture as *relational embeddings*. These embeddings of teams extracted from the model learned to predict home team win can be seen in Fig. 1.

4 Experiments

We compared approaches discussed in this paper on data from the 2017 Soccer Prediction Challenge [3], organized in conjunction with the MLJ's special issue on Machine Learning for Soccer. Particularly for this paper, we selected the

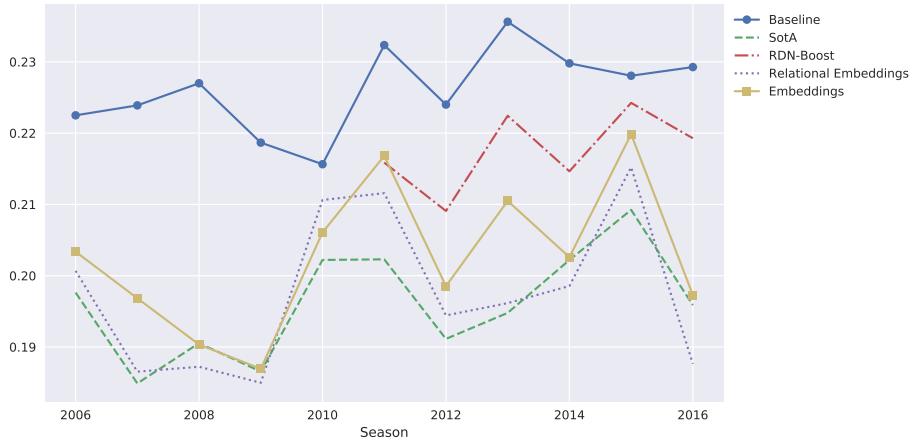

Fig. 2. Comparison of performance of the learners on English Premier League as measured by the RPS metric (lower is better) from the 2017 Soccer Prediction Challenge.

world's most prestigious English Premier League over the seasons 2006-2016. In the dataset, for each historical match there is merely a record of the team names and the resulting score. Contestants' models were evaluated using Ranked Probability Score (RPS) [4], an evaluation metric designed for the ordinal outcomes.

For each of the historical matches, we extract 3 learning examples for each respective outcome $(loss, draw, win)$, and learn one corresponding model for each of the latter. For each learner we then normalize the three outputs from the three models to obtain the final predictions that form input to the RPS metric.

Calculation of the baseline involved no setup and for setting of the SotA and RDN-boost models we refer to the Prediction Challenge submission [6]. For LRNNs we set the learning rate (0.1) and number of learning steps (50), and we utilized just a subset of the predicates (Section 3). LRNNs were trained sequentially with a history span of 5 years.

We display the final results in Fig. 2. All the learners easily pass the natural baseline (mean RPS 0.2260), with LRNNs (0.1976) performing significantly better than RDN-boost (0.2175), while trailing just closely behind the state-of-the-art model (0.1961). We also see that the relational embeddings generally dominate the standard embeddings (0.2027).

5 Conclusion

We discussed how the domain of predictive sports analytics might benefit from relational learning approaches, and experimentally proved that even simple relational templates with latent structures may lead to surprisingly strong, competitive results in predicting soccer game outcomes.

Acknowledgements Authors acknowledge support by "Deep Relational Learning" project no. 17-26999S granted by the Czech Science Foundation. Computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures".

References

1. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.
2. Anthony Costa Constantinou and Norman Elliott Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 2013.
3. Werner Dubitzky, Philippe Lopes, Jesse Davis, and Daniel Berrar. Open international soccer database, Aug 2017.
4. Edward S Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969.
5. John Goddard. Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2):331–340, 2005.
6. Ondřej Hubáček, Gustav Šourek, and Filip Železný. Learning to predict soccer results from relational data with gradient boosted trees. *Machine Learning*, 2018.
7. Lars Magnus Hvattum and Halvard Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3), 2010.
8. Verica Lazova and Lasko Basnarkov. PageRank approach to ranking national football teams. *arXiv preprint arXiv:1503.01331*, 2015.
9. Christophe Ley, Tom Van de Wiele, and Hans Van Eetvelde. Ranking soccer teams on basis of their current strength: a comparison of maximum likelihood approaches. *arXiv preprint arXiv:1705.09575*, 2017.
10. Ian McHale and Phil Scarf. Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 2007.
11. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
12. Sriraam Natarajan, Tushar Khot, Kristian Kersting, Bernd Gutmann, and Jude Shavlik. Boosting relational dependency networks. In *Online Proceedings of the International Conference on Inductive Logic Programming 2010*, pages 1–8, 2010.
13. Gustav Šourek, Vojtěch Aschenbrenner, Filip Železný, Steven Schockaert, and Ondřej Kuželka. Lifted relational neural networks: Efficient learning of latent relational structures. *Journal of Artificial Intelligence Research*, 62:69–100, 2018.
14. Jan Van Haaren and Guy Van den Broeck. Relational learning for football-related predictions. In *Latest Advances in Inductive Logic Programming*. World Scientific, 2015.