Bayesian Based Approach Learning for Outcome Prediction of Soccer Matches

Laura Hervert-Escobar[1(✉)], Neil Hernandez-Gress[1], and Timothy I. Matis[2]

[1] Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, Mexico
laura.hervert@itesm.mx
[2] Texas Tech University, Lubbock, USA

Abstract. In the current world, sports produce considerable data such as players skills, game results, season matches, leagues management, etc. The big challenge in sports science is to analyze this data to gain a competitive advantage. The analysis can be done using several techniques and statistical methods in order to produce valuable information. The problem of modeling soccer data has become increasingly popular in the last few years, with the prediction of results being the most popular topic. In this paper, we propose a Bayesian Model based on rank position and shared history that predicts the outcome of future soccer matches. The model was tested using a data set containing the results of over 200,000 soccer matches from different soccer leagues around the world.

Keywords: Machine learning · Soccer · Bayesian models
Sport matches · Prediction

1 Introduction

The sport is an activity that the human being performs mainly with recreational objectives. It has become an essential part of our lives as it encourages connivance, and when professionally engaged, it becomes a way to survive. The sport has become one of the big businesses in the world and has shown an important economic growth. Thousands of companies have their main source of income in it. The most popular sport in the world, according to Russell [1], is football soccer. Soccer detonates a great movement of money in bets, sponsorships, attendance to parties, sale of t-shirts and accessories, etc. That is why it has aroused great interest in building predictive and statistical models for it.

Professional soccer has been in the market for quite some time. The sports management of soccer is awash with data, which has allowed the generation of several metrics associated with the individual and team performance. The aim is to find mechanisms to obtain competitive advantages. Machine learning has become a useful tool to transform the data into actionable insights.

Machine Learning is a scientific discipline in the field of Artificial Intelligence that creates systems that learn automatically. Learning in this context means

identifying complex patterns in millions of data. The machine that really learns is an algorithm that reviews the data and is able to predict future behavior. It finds the sort of patterns that are often imperceptible to traditional statistical techniques because of their apparently random nature.

When the scope of data analysis techniques is complemented by the possibilities of machine learning, it is possible to see much more clearly what really matters in terms of knowledge generation, not only at a quantitative level, but also ensuring a significant qualitative improvement. Then researchers, data scientist, engineers and analysts are able to produce reliable, repeatable decisions and results [2].

With data now accessible about almost anything in soccer, machine learning can be applied in a range. However, it has been used mostly for prediction. This type of models are known as multi-class classification for prediction, an it has three classes: win, loss and draw. According to Gevaria, win and loss are comparatively easy to classify. However, the class of draw is very difficult to predict even in real world scenario. A draw is not a favored outcome for pundits as well as betting enthusiasts [3].

In this research we present a new approach for soccer match prediction based on the performance position of the team in the season and the history of matches. The model was tested using a training data set containing the results of over 200,000 soccer matches from different soccer leagues around the world. Details of data set are available at [4].

The remainder of this paper is organized as follows. Section 2 gives a summary of previous work on football prediction. A general description of how the problem is addressed is presented in Sect. 3. Section 4 describes the procedures for preprocessing data, followed by the description of the proposed model. Experiments and results are described in Sects. 6 and 7, respectively. Finally, discussion of the results are in Sect. 8.

2   Related Work

Since soccer is the most popular sport worldwide, and given the amount of data generated everyday, it is not surprising to find abundant amount of research in soccer prediction.

Most of related work is focused on developing models for a specific league or particular event such as world cup. Koning [5] used a Bayesian network approach along with a Monte-Carlo method to estimate the quality of soccer teams. The method was applied in the Dutch professional soccer league. The results were used to assess the change over the time in the balance of the competition.

Rue [6] analyzed skills of all teams and used a Bayesian dynamic generalized linear model to estimate dependency over time and to predict immediate soccer matches.

Falter [7] and Forrest [8] proposed an approach focused more on the analysis of soccer matches rather than on prediction. Falter proposed an updating process for the intra-match winning probability while Forrest computes the uncertainty

of the outcome. Both approaches are useful to identify the main decisive elements in a soccer league and use them to compute the probability of success.

Crowder [9] proposed a model using refinements of the independent Poisson model from Dixon and Coles. This model considers that each team has attack and defense strategies that evolves over time according to some unobserved bivariate stochastic process. They used the data from 92 teams in the English Football Association League to predict the probabilities of home win, draw and lost.

Anderson [10] evaluates the performance of the prediction from experts and non-experts in soccer. The procedure utilized was the application of a survey to a 250 participants with different levels of knowledge in soccer. The survey consist on predicting the outcome of the first round of the World Cup 2002. The results shows that a recognition-based strategy seems to be appropriate to use when forecasting worldwide soccer events.

Koning [11] proposed a model based on Poisson parameters that are specific for a match. The procedure combines a simulation and probability models in order to identify the team that is most likely to win a tournament. The results were effective to indicates favorites, and it has the potential to provide useful information about the tournament.

Goddard [12] proposed an ordered probit regression model for forecasting English league football results. This model is able to quantify the quality of prediction along with several explanatory variables.

Rotshtein [13] proposed a model to analyzed previous matches with fuzzy knowledge base in order to find nonlinear dependency patterns. Then, they used genetic and neural optimization techniques in order to tune the fuzzy rules and achieve a acceptable simulations.

Halicioglu [14] analyzed football matches statistically and suggested a method to predict the winner of the Euro 2000 football tournament. The method is based on the ranking of the countries combined with a coefficient of variation computed using the point obtained at the end of the season from the domestic league.

Similar approaches applied to different sports can be found in [15–17]. Their research is focused on the prediction of American football and baseball major league.

Among the existing works, the approach of [18] is most similar to ours. Their system consists of two major components: a rule-based reasoner and a Bayesian network component. This approach is a compound one in the sense that two different methods cooperate in predicting the result of a football match. Second, contrary to most previous works on football prediction they use an in-game time-series approach to predict football matches.

3   General Ideas

Factors such as morale of a team (or a player), skills, coaching strategy, equipment, etc. have a impact in the results for a sport match. So even for experts, it is very hard to predict the exact results of individual matches. It also raises very

interesting questions regarding the interaction between the rules, the strategies and the highly stochastic nature of the game itself.

How possible is to have high accuracy prediction by knowing previous results per team? How should be the selection of factors that can be measured and integrated into a prediction model? Are the rules of the league/tournament a factor to consider in the prediction model?

Consider a data set that contains the score results of over 200,000 soccer matches from different soccer leagues around the world. There is no further knowledge of other features such as: importance of the game, skills of the players or rules of the league. In this way and without experience or knowledge on soccer, our hypothesis is that soccer results are influenced by the position rank of the teams during the season as well as the shared history between matched teams.

In general, the methodology proposed decides over two approaches. The first approach consist in finding patterns in the history match of teams that indicates a trend in the results. The second approach considers the given information to rank teams in the current season. Then, based on the ranking position, a Bayesian function is used to compute the probability of win, lose or draw a match.

4   Data Pre-processing and Feature Engineering

The data set contains the results of over 200,000 soccer matches from different soccer leagues around the world. With the information of date, season, team, league, home team, away team, and the score of each game during the season. Details of data set is available at [4].

The main objective in pre-processing the data is to set the initial working parameters for the prediction methodology. Then, the metrics to obtain in this procedure are: the rank position of the teams, the start probabilities for the Bayesian function and the shared history between two teams. Preprocessing procedures were easily implemented using R.

Equations used during the pre-processing data are as follows. Index i refers to team, index t refers to the season of the team playing in the league, finally n refers to total games played by team i during season t.

$$sg_t^i = \sum_n \left(3w_{n,t}^i + d_{n,t}^i \right) \tag{1}$$

Equation (1) describes the computation of the score based on game performance sg. The score computation gives 3 points for each game won (w) during the season, 1 point for a draw (d) and zero points for a lost (l) game. This method is based on the result points from FIFA ranking method. Match status, opposition strength and regional strength are not considered due to the lack of information in the dataset.

$$sb_t^i = \sum_n \left(gf_{n,t}^i - ga_{n,t}^i \right) \tag{2}$$

Equation (2) describes the computation of the score based on the number of goals during the season sb. In this way, the score is given by the number of goals in favor gf minus the number of goals against ga.

$$gs_t^i = sg_t^i + sb_t^i \qquad (3)$$

$$score_t^i = \begin{cases} gs_t^i & t = 1 \\ 0.2\left(gs_{t-1}^i\right) + 0.8\left(gs_t^i\right) & t > 1 \end{cases} \qquad (4)$$

A partial score given in Eq. (3) is the sum of Eqs. (1) and (2). The total score for each season in given in Eq. (4).

The teams of the league in each season may vary according to promotions or descents derived from their previous performance. As shown in Eq. (4), the previous season has a weight of 20% on the total score. The current season has a weight of 80%. In this way, the ranking process takes into account a previous good/bad performance. But it also gives greater importance to the changes that the team makes in the current season. This measure was designed to have a fair comparison between veteran teams playing and rookie teams in the league. In this way, the history of each team will have an influence on their current rankings (whether positive or not) and rookie teams will have a fair comparison that alleviates league change adjustments.

The rank of the team $rank_i^t$ in Eq. (5) is given by its position according to the total score. Given a collection of M teams, the rank of a team i in season t is the number of teams that precede it.

$$rank_i^t = \left|\left\{rank_i^t \,\middle|\, rank_i^t < rank_j^t\right\}\right| \qquad \forall \qquad i \neq j, \qquad i,j \in M_t \qquad (5)$$

As expected, not all teams are participating in all seasons. Then, missing teams are not considered in the ranking of the current season.

Equations (6) and (7) are used to obtained start probabilities to be used in the Bayesian function,

$$mrank_i^t = 1 - \frac{rank_i^t}{(Max(rank^t) + 1)}; \qquad (6)$$

$$Pstart^t = \frac{mrank_i^t}{\sum\limits_i mrank_i^t} \qquad (7)$$

Finally, the shared history of the teams is a list that summarizes the number of cases that the same match has been played. The list also contains the probability of win pRw_{i-j}, lose pRl_{i-j}, and draw pRd_{i-j} a game based on the total matches tg for a given period. See Eq. (8).

$$pRw_{i-j} = \sum_n \left(\frac{w}{tg}\right)_{i-j}; \quad pRd_{i-j} = \sum_n \left(\frac{d}{tg}\right)_{i-j}; \quad pRl_{i-j} = \sum_n \left(\frac{l}{tg}\right)_{i-j} \qquad (8)$$

5   Bayesian Algorithm

A pseudo-code for the Bayesian function proposed is given in Algorithm 1. The procedure starts by computing the prior probability of the two teams in the match (step 1). The team with higher probability is labeled as a team, and the team with lower prior probability is subindex as b(step 2). Then, prior probability of the a team is used to generate 1000 random variables using a triangular distribution.

$TD[0, 1, prior_a^t]$ represents a continuous triangular statistical distribution supported over the interval $min = x = max$ and parameterized by three real numbers 0, 1, and $prior_a^t$ (where $0 < prior_a^t < 1$) that specify the lower endpoint of its support, the upper endpoint of its support, and the -coordinate of its mode, respectively. In general, the PDF of a triangular distribution is triangular (piecewise linear, concave down, and unimodal) with a single "peak", though its overall shape (its height, its spread, and the horizontal location of its maximum) is determined by the values of 0, 1, and $prior_a^t$.

Using the random variables, posterior probabilities are computed in step 5. Then, the probability corresponding to mode of posterior is used to compute and adjust measure. The adjust measure is apply to the start probabilities for the next period (step 9). Finally, the probability of win/lose the match in the period $t + 1$, knowing the probabilities of the current period t is given by equations in step 10. This equations correspond to the prior probability based on the adjusted start probability.

The procedure for the soccer prediction using Bayesian function and shared history data is given in Algorithm 2. As the pseudo-code shows. The probability taken for the prediction model is chosen between two options, shared history or ranking. Either choice allows to update results in the Bayesian function.

The procedure starts by checking the shared history of the match to predict. Based on the total matches, the next step is either use history probability or Bayesian probability. The threshold to decide is set at least 10 games of shared history.

Then, if the threshold value is greater or equal to 10, the probability lies on previous results. Otherwise, the probability is given by their rank position in the season-league along with the Bayesian function.

6   Experiments

Procedures were implemented on R statistical free license software. In order to prove the value of the methodology the training data set given by [4] was split in two parts for all leagues. First part contains the results from 2000 to 2015. Second part contains data from 2016–2017 and was used as the matches to predict.

The metric used in the challenge is the ranked probability score (RPS). The RSP helps to determine the error between the actual observed outcome of a match and the prediction. Description of the metric can be found at [4].

Algorithm 1. Pseudocode for probability of win/lose a match game

Require:
    $Pstart$: list of start probabilities (See Equation(5))
Ensure: Probability of win/lose the match game $f_m(winteam(a), loseteam(b))$
1: Compute prior probabilities for teams in the match

$$prior_i^t = \frac{Pstart_i^t}{Pstart_i^t + Pstart_j^t}; \qquad prior_j^t = \frac{Pstart_j^t}{Pstart_i^t + Pstart_j^t}$$

2: Set probable winner and loser team

$$prior_a^t = maxPstart_i^t + Pstart_j^t; \qquad prior_b^t = minPstart_i^t + Pstart_j^t$$

3: Continuous triangular prior distribution evaluated at 1000 equally spaced points
   using prior probability

$$prior^t = TD\{[0, 1, prior_a^t], x\}, \{0, 1, 0.001\}$$

4: Prior discretized into a probability mass function and discretized prior probability
   masses

$$dprior = \frac{prior}{\sum\limits_{1000} prior}; \qquad probs_i^t = \{i_k = i_{k-1} + 0.001 | i \in [0, 1)\} 1 < k < 1000]$$

5: Posterior distribution

$$posterior_i^t = \frac{probs_i^t * dprior_i^t}{\sum\limits_i probs * dprior}$$

6: Probability corresponding to mode of posterior

$$c = \underset{i}{Max}\{posterior_i^t\} \times 0.001$$

7: Adjust probabilities for current rankings

$$adjust = c \times \left(Pstart_a^t - Pstart_b^t\right)$$

8: Update start probabilities

$$Pstart_a^{t+1} = Pstart_a^t + adjust; \qquad Pstart_b^{t+1} = Pstart_a^t - adjust$$

9: Computing final win/lose probabilities

$$pwinteam(a) = \frac{Pstart_a^{t+1}}{Pstart_a^{t+1} + Pstart_b^{t+1}}; ploseteam(b) = \frac{Pstart_b^{t+1}}{Pstart_a^{t+1} + Pstart_b^{t+1}}$$

10:

Two types of outcomes were tested. In a first outcome, the variables xW, xD and xL were defined as binary numbers. In this outcome, the strategy was to check how accurate was the method in order to predict an exact result.

Algorithm 2. Pseudo-code for soccer prediction method

Require:

    $f(winteam(a), loseteam(b))$

    Shared history list

Ensure: Prediction outcome for match xW, xD, xL

 1:

 2: **if** Shared History $\in$ prediction$\{a,b\} \geq 10$ **then**

 3:

 4:      $xW = pRW; xD = pRD; xL = pRL$

 5:

 6: **else**

 7: Compute Bayesian Function $f_m(winteam(a), loseteam(b))$

 8: $\Delta = pwinteam(a) - ploseteam(b)$

 9:

10:      **if** $\Delta \leq 0.2$ **then**

11:        $xW = \frac{\Delta}{2}; xD = 1 - \Delta; xL = \frac{\Delta}{2}$

12:      **end if**

13:      $xW = Pstart_a^{t+1}; xD = 1 - Pstart_a^{t+1} + Pstart_b^{t+1}; xL = Pstart_b^{t+1}$

14: **end if**

15: NEXT MATCH

The second approach was to preserve the nature of the computation. Then, the outcome variables xW, xD and xL are in the rank of $[0, 1]$, where the sum is equal to 1.

Additionally, a real prediction was performed based on a call challenge of soccer. Detail of the call can be found at [4].

7   Results

Figure 1 shows the result obtained using both approaches using the training data set. As observed, the RSP improves when nature of the variables are continuous rather than binary. Additionally, the bars indicate the proportion of the training predictions made by history matches and for rank procedure. For the training data set, the RSP has not significant changes related to the prediction method.

As mentioned above, the methodology proposed was tested under the requirements of a call for a challenge soccer. Details results for the challenge soccer can be found at [19]. The results of the prediction for the call of the challenge soccer are shown in Fig. 2. The figure shows the proportion of the prediction defined by history match and for ranking procedure. Additionally, shows the average RSP obtained for each type of prediction. As shown, for leagues where greater proportion of prediction were made by history matches, the average RSP is around 33%, for one league it reaches a desirable 0%. On the other hand, predictions made mainly with rank procedure, the RSP average is over 40%, with one case of 0%.
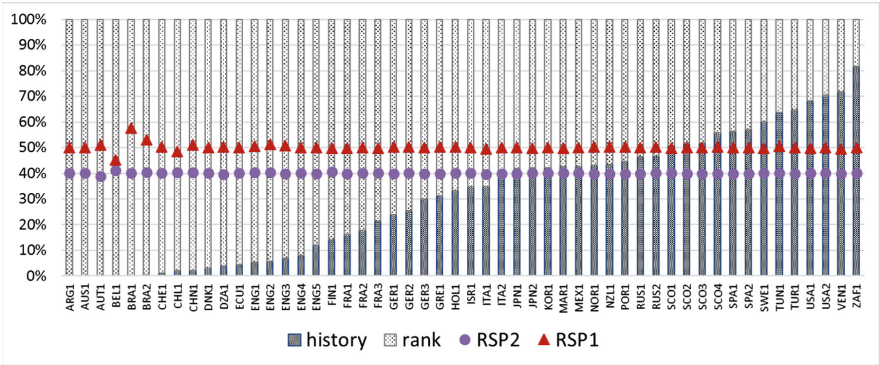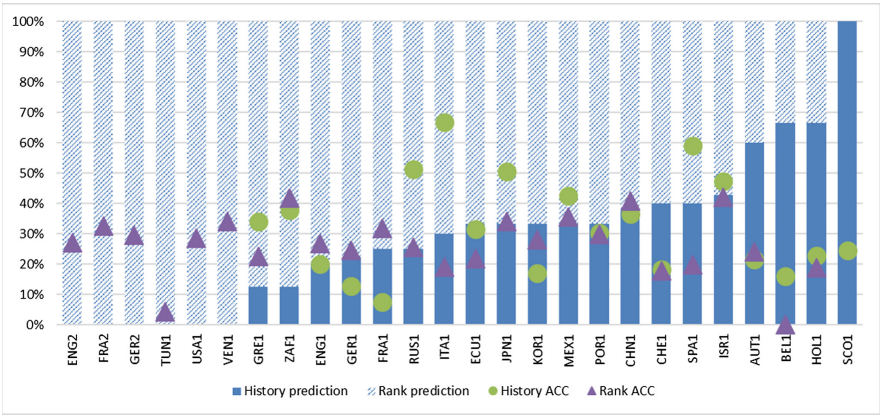
Fig. 1. Prediction results for each league

Fig. 2. Results of RSP according to prediction method

8 Conclusions

Main motivation of this work was the chance to participate in the call for the soccer challenge as a way to test a basic Bayesian model along with other techniques to predict the outcome of matches in soccer. Despite the lack of knowledge about soccer in general, we were able to first understand the challenge and then developed a prediction model that is easy to implement. From literature reviewed we learned that each league is driven by different motivations that influence the result of a match game. Then, information based only in the result of matches may no accurate allows to recognize useful patterns for prediction. Most of the time inverted in the process of defining the better way of ranking as well as programming the procedures, trying to make them as efficient as possible.

The methodology proposed is simply an instance of a more general framework, applied to soccer. It would be interesting to try other sports. In this

section, we consider the possibilities for extension. Even though the framework can in principle be adapted to a wide range of sports domains, it cannot be used in domains which have insufficient data. Another approach to explore in the future is a Knowledge-based system. This usually require knowledge of relatively good quality while most machine learning systems need a huge amount of data to get good predictions. It is important to understand that each soccer league behaves according to particular environment. Therefore, a better prediction model should include particular features of the match game, such as the importance of the game. Availability of more features that can help in solving the issue of predicting draw class would improve the accuracy.

Future work in this area includes the development of a model that attempt to predict the score of the match, along with more advance techniques and the use of different metrics for evaluating the quality of the result.

References

1. Rusell, B.: Top 10 most popular sports in the world (2013)
2. SAS: Machine learning what it is & why it matters (2016)
3. Vaidya, S., Sanghavi, H., Gevaria, K.: Football match winner prediction. Int. J. Emerg. Technol. Adv. Eng. **10**(5), 364–368 (2015)
4. Dubitzky, W., Berrar, D., Lopes, P., Davis, J.: Machine learning for soccer (2017)
5. Koning, R.H.: Balance in competition in dutch soccer. J. Roy. Stat. Soc. Ser. D (Stat.) **49**(3), 419–431 (2000)
6. Rue, H., Salvesen, O.: Prediction and retrospective analysis of soccer matches in a league. J. Roy. Stat. Soc. Ser. D (Stat.) **49**(3), 399–418 (2000)
7. Falter, J.M., Perignon, C.: Demand for football and intramatch winning probability: an essay on the glorious uncertainty of sports. Appl. Econ. **32**(13), 1757–1765 (2000)
8. Forrest, D., Simmons, R.: Outcome uncertainty and attendance demand in sport: the case of english soccer. J. Roy. Stat. Soc. Ser. D (Stat.) **51**(2), 229–241 (2002)
9. Crowder, M., Dixon, M., Ledford, A., Robinson, M.: Dynamic modelling and prediction of english football league matches for betting. J. Roy. Stat. Soc. Ser. D (Stat.) **51**(2), 157–168 (2002)
10. Andersson, P., Ekman, M., Edman, J.: Forecasting the fast and frugal way: a study of performance and information-processing strategies of experts and non-experts when predicting the world cup 2002 in soccer. In: SSE/EFI Working Paper Series in Business Administration 2003, vol. 9. Stockholm School of Economics (2003)
11. Koning, R.H., Koolhaas, M., Renes, G., Ridder, G.: A simulation model for football championships. Eur. J. Oper. Res. **148**(2), 268–276 (2003). Sport and Computers
12. Goddard, J., Asimakopoulos, I.: Forecasting football results and the efficiency of fixed-odds betting. J. Forecast. **23**(1), 51–66 (2004)
13. Rotshtein, A.P., Posner, M., Rakityanskaya, A.B.: Football predictions based on a fuzzy model with genetic and neural tuning. Cybern. Syst. Anal. **41**(4), 619–630 (2005)
14. Halicioglu, F.: Can we predict the outcome of the international football tournaments: the case of Euro 2000. Doğşu Üniversitesi Dergisi **6**(1) (2005)
15. Martinich, J.: College football rankings: do the computers know best? Interfaces. Interfaces **32**(4), 84–94 (2002)

16. Amor, M., Griffiths, W.: Modelling the behaviour and performance of australian football tipsters. Department of Economics - Working Papers Series 871, The University of Melbourne (2003)
17. Yang, T.Y., Swartz, T.: A two-stage bayesian model for predicting winners in major league baseball. J. Data Sci. **2**(1), 6173 (2004)
18. Min, B., Kim, J., Choe, C., Eom, H., (Bob) McKay, R.I.: A compound framework for sports results prediction: a football case study. Know. Based Syst. **21**(7), 551–562 (2008)
19. Hervert, L., Matis, T.: Machine learning for soccer-prediction results (2017)