Improving Predictive Accuracy Using Smart-Data rather than Big-Data: A Case Study of Soccer Teams' Evolving Performance

Anthony Constantinou
Queen Mary University of London,
London, UK, E1 4NS
a.constantinou@qmul.ac.uk

Norman Fenton
Queen Mary University of London,
London, UK, E1 4NS
n.fenton@qmul.ac.uk

EXTENDED ABSTRACT

(this paper is published as extended abstract only)

In an era of big-data the general consensus is that relationships between variables of interest surface almost by themselves. Sufficient amounts of data can nowadays reveal new insights that would otherwise have remained unknown. Inferring knowledge from data, however, imposes further challenges. For example, the 2007-08 financial crisis revealed that big-data models used by investment banks and rating agencies for decision making failed to predict real-world financial risk. This is because while such big-data models are excellent at predicting past events, they may fail to predict similar future events that are influenced by new and hence, previously unseen factors.

In many real-world domains, experts comprehend vital influential processes which data alone may fail to discover. Yet, such knowledge is normally disregarded in favor of automated learning, even when the data are limited. While automation provides major benefits, these benefits sometimes come at a cost for accuracy. This study focuses on a prediction problem that has similarities to financial risk, namely predicting evolving soccer team performance. Soccer is the world's most popular sport and constitutes an important share of the gambling market. Just like in financial risk, future team performance can be suddenly and dramatically affected by rarely seen, or previously unseen, events and so both require smarter ways of data engineering and modeling, rather than just larger amounts of data.

Most of the previous extensive work on soccer has focused on results predictions based on historical data of relevant match instances. In this study we do not consider individual match results, but rather exploit external factors which may influence the strength of a team and its resulting performance. The aim is to predict a soccer team's performance for a whole season (measured by total number of league points won) before the season starts. This is an important and enormous gambling market in itself - betters start placing bets such as which team will win the title, finish in top positions, or be relegated, as soon as the previous season ends. The need for greater accuracy in such predictions has become the subject of international interest following the 2015-16 English Premier League (EPL) season when Leicester City finished top of the league, having been priced at 5,000 to 1 to do so by many bookmakers.

We use a data and knowledge engineering approach that puts greater emphasis on applying causal knowledge and real-world 'facts' to the process of model development for real-world decision making, driven by what data are really required for inference, rather than blindly seeking 'bigger' data. We refer to this as the 'smart data' approach. We use a Bayesian network (BN) as the appropriate modelling method. Based on the soccer case study, we illustrate the reasoning towards this smart-data approach to BN modeling with two subsystems:

1. A knowledge-based intervention for informing the model about real-world time-series facts; and
2. A knowledge-based intervention for data-engineering purposes to ensure data adhere to the structure of the model.

The BN model incorporates factors such as player injuries, managerial changes, team involvement in other European competitions, and financial investments relative[1] to adversaries. The BN model is based on three distinct time components:

1. Observed events from previous season that have influenced team performance;
2. Observed events during the summer break that are expected to influence team performance;
3. Expected performance for next season, accounting for the uncertainty which arises from other unknown events which may influence team performance, such as injuries.

This process is repeated for each new season, for a total of 15 seasons. This approach enabled us to provide far more accurate predictions compared to purely data-driven standard

[1] Team A may spend £20m to improve their squad, but if the average adversary spends £30m, then the strength of Team A is expected to diminish relative to the average adversary.

non-linear regression models, which still represent the standard method for prediction in critical real-world risk assessment problems, such as in medical decision analysis (Kendrick, 2014). Specifically, we demonstrate how we managed to generate accurate predictions of the evolving performance of soccer teams based on limited data that enables us to predict, before a season starts, the total league points to be accumulated. Predictive validation over a series of 15 EPL seasons demonstrates a mean error of 4.06 points (the possible range of points a team can achieve is 0 to 114). In contrast, for two different regression based methods, the mean errors are 7.27 and 7.30.

The implications of the paper are two-fold. First, with respect to the application domain, the current state-of-the-art is extended as follows:

1. This is the first study to present a model for accurate time-series forecasting in terms of how the strength of soccer teams evolves over adjacent soccer seasons, without the need to generate predictions for individual matches.

2. Previously published match-by-match prediction models (some of them include: Karlis & Ntzoufras, 2003; Rotshtein et al., 2005; Baio & Blangiardo, 2010; Hvattum & Arntzen, 2010; Constantinou & Fenton, 2012; Constantinou & Fenton, 2013b) which fail to account for the external factors influencing team strength, are prone to an error of 8.51[2] league points accumulated per team, in terms of prior belief for team strength, and for each subsequent season. Therefore, one could improve match-by-match predictions by reducing the error in terms of prior belief.

3. Studies which assess the efficiency of the soccer gambling market (Dixon & Pope, 2004; Goddard & Asimakopoulos, 2004; Graham & Stott, 2008; Constantinou & Fenton, 2013b) may find the BN model helpful in the sense that it could help in explaining previously unexplained fluctuations in published market odds.

Second, with respect to the general strategy for learning from data, we demonstrate that seeking 'bigger' data is not always the path to follow. The model presented in this paper, for instance, is based on just 300 data instances generated over a period of 15 years. With a smart-data approach, one should aim to improve the quality, as opposed to the quantity, of a dataset which also directly influences the quality of the model. We highlight the importance of developing models based on what data we really require for inference, rather than generating a model based on what data are available which represents the conventional approach to big-data solutions. With smart-data one has to have a clear understanding of the inferences of interest. Inferring knowledge from data imposes further challenges and requires skills that merge the quantitative as well as qualitative aspects of data.

For future research, we question whether automated learning of the available data is capable of inferring real-world facts such as those incorporated into the BN model presented in this paper. It may be the case that, for many real-world problems, resulting inferences will be limited in the absence of expert intervention for data engineering as well as modeling purposes. Future research will examine the capability of causal discovery algorithms in terms of realizing various real-world facts from data, and the impact various data-engineering interventions may have on the results.

Keywords: data engineering; dynamic Bayesian networks; expert systems; football predictions; smart data; soccer predictions; temporal Bayesian networks.

[2] Note that this error assumes EPL teams, and is dependent on the size of the league. For instance, the EPL consists of 20 teams and each team has to play 38 matches. Hence, the maximum possible accumulation of points is 114.

ACKNOWLEDGEMENTS

We acknowledge the financial support by the European Research Council (ERC) for funding this research project, ERC-2013-AdG339182-BAYES_KNOWLEDGE, and Agena Ltd for software support.

REFERENCES

Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37:2, 253- 264.

Constantinou, A., Fenton, N., & Neil, M. (2012). pi-football: A Bayesian network model for forecasting Association Football match outcomes. *Knowledge-Based Systems*, 36: 322, 339.

Constantinou, A., & Fenton, N. (2013a). Profiting from an inefficient Association Football gambling market: Prediction risk and Uncertainty using Bayesian networks. *Knowledge-Based Systems*, 50: 60-86.

Constantinou, A, & Fenton, N. (2013b). Profiting from arbitrage and odds biases of the European football gambling market. *The Journal of Gambling Business and Economics*, Vol. 7, 2: 41-70.

Dixon, M., & Pope, P. (2004). The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20, 697-711.

Goddard, J., & Asimakopoulos, I. (2004). Forecasting Football Results and the Efficiency of Fixed-odds Betting. *Journal of Forecasting*, 23, 51-66

Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. *Applied Economics*, 40, 99-109.

Hvattum, L. M., & Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26, 460-470.

Karlis, D., & Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *The Statistician*, 52: 3, 381-393.

Kendrick, M. (2014). *Doctoring Data: How to sort out medical advice from medical nonsense*. UK, Columbus Publishing.

Rotshtein, A., Posner, M., & Rakytyanska, A. (2005). Football predictions based on a fuzzy model with genetic and neural tuning. *Cybernetics and Systems Analysis*, 41: 4, 619- 630.