IOP Conference Series: Materials Science and Engineering

Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)

To cite this article: Nazim Razali *et al* 2017 *IOP Conf. Ser.: Mater. Sci. Eng.* **226** 012099

Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)

Nazim Razali[1], **Aida Mustapha**[1], **Faiz Ahmad Yatim**[2], **Ruhaya Ab Aziz**[1]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
[2] Co-Operative College of Malaysia, 103 Jalan Templer, 46700 Petaling Jaya, Selangor, Malaysia

E-mail: nazim.uthm@gmail.com, aidam@uthm.edu.my, faiz@mkm.edu.my, ruhaya@uthm.edu.my

Abstract. The issues of modeling asscoiation football prediction model has become increasingly popular in the last few years and many different approaches of prediction models have been proposed with the point of evaluating the attributes that lead a football team to lose, draw or win the match. There are three types of approaches has been considered for predicting football matches results which include statistical approaches, machine learning approaches and Bayesian approaches. Lately, many studies regarding football prediction models has been produced using Bayesian approaches. This paper proposes a Bayesian Networks (BNs) to predict the results of football matches in term of home win (H), away win (A) and draw (D). The English Premier League (EPL) for three seasons of 2010-2011, 2011-2012 and 2012-2013 has been selected and reviewed. K-fold cross validation has been used for testing the accuracy of prediction model. The required information about the football data is sourced from a legitimate site at http://www.football-data.co.uk. BNs achieved predictive accuracy of 75.09% in average across three seasons. It is hoped that the results could be used as the benchmark output for future research in predicting football matches results.

1. Introduction

Association football or soccer is a famous amongst the most well known sports on the planet. Predicting the results of football matches has attract so many people who love and have so much passion on football, from the managerial football team until to the fans themselves. Prediction in football has become an intriguing research problem, partly because of the difficulty, because there are many factors can influence the football matches outcome, such as teamwork, skills, weather, home advantage and many others. The challenge is also faced by the experts in football as well because it is very difficult to anticipate the actual results of football matches. Anything can be happened within 90 minutes of football match with none or extra time such as injury or sent off of player by red card. Luck also can be a factor that influence the results of football matches as the strong team are not necessarily win the match against the weak team. Due to the different factors that influence the football match, this research resort to Bayesian Networks (BNs) due its proven applicability in predicting the weather [1], sports [2] and many others. In this paper, a Bayesian approach for the prediction of football results is presented by using the attributes provided by `http://www.football-data.co.uk` in the section of historical data for English Football Results.

The remainder of this paper is organized as follows. A summary of previous work on football predictions is presented in Section 2. In Section 3, the experiments including the dataset, the Bayesian Networks as well as the results in terms of predictive accuracies are presented. Finally, the conclusions are in provided in Section 4.

2. Related Work

Football is the world's renowned sport since 1930. There will be a FIFA World Cup organised and hosted by a nation once in every four years. Not only that, there are also popular football league and have their own fans around the world such as English Premier League (England), French Ligue 1 (France), Spanish La Liga (Spain), Italian Seria A (Italy) and German Bundesliga (Germany). In this manner, it is not suprised that there has been a generous measure of research in football prediction. The research for predicting the results of football matches outcome as started as early 1977 by [3]. [3] developed a model called least squared model that rate both strength of the home and the away team using matrix of goal scoring distribution. However, the first football prediction model incorporate with the application of Bayesian Networks only started in work of [4]. [4] proposed a Bayesian networks in order to take account the variation of time for all properties simultaneously (Dynamic) which also known as Dynamic Bayesian networks (DBNs). As the result, the offensive and defensive strength of home team and away team will be varied over time.

Next, [5] presented a complex framework for predicting football matches results. This complex framework known as FRES system consists of two main components: rules based theorem and Bayesian network component. Thus, FRES system is a compound of two techniques which cooperate together for predicting football matches results. Besides, FRES system also was implemented in-game time-series approach which make the prediction more realistic. However, FRES system need sufficient expert knowledge in order to run it well. [2] suggested Bayesian hierarchical model for the prediction of football results. The number of goals scored by the two teams in each match have been used to developed Bayesian hierarchical model. [6] developed a football prediction model called pi-rating to generate forecasts about the football matches outcome whether home win,draw or away win for English Premier League (EPL) matches during seasons 2010/2011 which incorporate objective information and subjective information such as team strength, team form, psychological impact and fatigue. [7] extend the work of [8] on Poisson distribution that expressed the attacking and defensive strength from goal scoring distribution. [7] developing a statistical model for the analysis and predicting of football match results which assumes a bivariate Poisson distribution with intensity coefficients that change randomly over time. However, [7] claimed that the work was based on classical perspective and suggested the used of Bayesian Networks for better account of parameter uncertainty. In conclusion, it is shown that Bayesian networks has significant value for predicting football matches results.

Among the related works in the literature, this paper takes off from [9], who presented an approach to forecasting results in which the Bayesian Networks provided a means for representing, displaying, and predicting the results of expert knowledge in a football game. Their results showed that the Bayesian networks is generally superior to the other techniques such as the MC4, a decision tree learner, naive Bayesian learner (NB), and k-nearest neigbhbor learner (KNN) for this domain in terms of predictive accuracy. The Bayesian networks proposed by [9] successfully gained predictive accuracy of 59.21% which outperformed other machine learning techniques by 41.72% (MC4), 47.86% (NB) and 50.58% (KNN). Note that [9] used the presence and absence of 3 key players, the home advantage, opposition team quality based on position in league table and the position of main key player named Wilson whether he played as midfield or not as the attributes for predicting the football matches results.

3. Experiments

The prediction experiments are carried out using the WEKA software. WEKA is open source software under the GNU (General Public License), where the software can provide the implementation state-of-the-art data mining and machine learning algorithms. In the experiments, each game was considered separately because each game has different values for the factors and vary based on the state of the game. Therefore, the prediction experiments are repeated for all 380 matches in three seasons separately. The results are then compared between one season to another.

3.1. Dataset

Table 1 shows the main factors and some values in the football match prediction that will be used in modeling the matches via Bayesian Networks. This research, in particular, considered the English Premier League for the seasons of 2010-2011, 2011-2012 and 2012-2013. The league is made by a number of $T = 20$ teams, and each team plays each other twice in a season (one at home and one away). The data was sourced from `http://www.football-data.co.uk`.

Table 1. Main Factors in Football Match Prediction

Attributes	Sample Values
Home Team	Manchester United
Away Team	Wigan
Home Team Shots	17
Away Team Shots	8
Home Team Shots on Target	10
Away Team Shots on Target	4
Home Team Corners	5
Away Team Corners	5
Home Team Fouls Committed	10
Away Team Fouls Committed	14
Home Team Yellow Cards	2
Away Team Yellow Cards	2
Home Team Red Cards	0
Away Team Red Cards	0
Half Time Home Team Goals	0
Half Time Away Team Goals	0
Full Time Home Team Goals	4
Full Time Away Team Goals	0

3.2. Bayesian Networks

In recent years, Bayesian networks (BNs) has become a popular way of probability models the relationship between a set of variables for a particular domain. Bayesian networks is a graphical model for reasoning under uncertainty, where the nodes represent discrete or continous variables and arcs represent direct connections between them [10]. This direct connection represents a causal connection. The strength of BNs is that it quantitatively model relationships among variables, whereby the probability of the belief will be automatically updated as new information becomes available. The main idea of Bayesian networks is actually come from Thomas Bayes works called Bayes theorem. Bayes theorem technically is stated in Equation 1:

$$\mathrm{P(X|Y)} = \frac{\mathrm{P(Y|X).P(X)}}{\mathrm{P(Y)}} \tag{1}$$

where:

- $P(X)$ is the prior probability or marginal probability of X.
- $P(X|Y)$ is the posterior probability or conditional probability of X given Y.
- $P(Y|X)$ is the conditional probability of Y given X (the likelihood of data Y).
- $P(Y)$ is the prior probability or marginal probability of data Y (the evidence).

A Bayesian network consists of two components: quantitative and qualitative. The quantitative components of BNs can be represented in form of network parameters called conditional table while the qualitative components of BNs can be represented in form of network structure. There are a set conditional probabilities for discrete data or probability density function for continuous data in conditional table. A qualitative component of Bayesian networks (BNs) is represented by its network structure called directed acyclic graph (DAG). A directed acyclic graph made up of a set of nodes that represent random variables from the domain and directed edges connecting nodes to represent the conditional dependencies between nodes. Though, the directed edges cannot form any directed cycles (no loop). When building a Bayesian network from prior knowledge alone, the probabilities will be based on Bayes' rule or Bayes' theorem. When learning these network from data, the probabilities will be physical and the values may be uncertain.

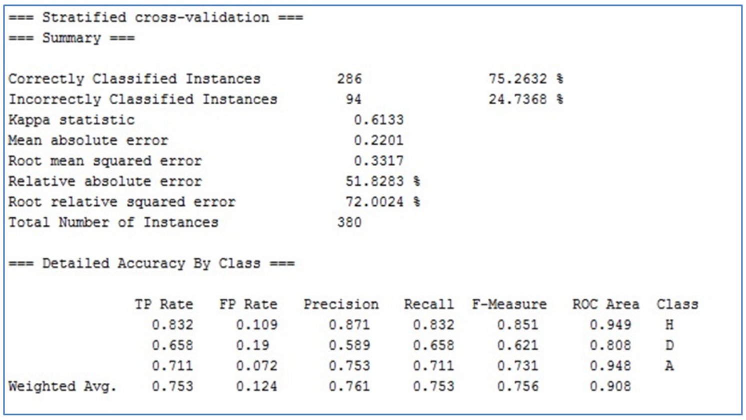
3.3. Results

Table 2 shows the overall prediction accuracies for the football matches in the English Premier League during Season 2010-2011, Season 2011-2012, and Season 2012-2013. Again, the prediction results were obtained from separate experiments for each season.

Table 2. Overall Prediction Results across Three Seasons in EPL

Season	Prediction Accuracy (%)
2010-2011	75.26
2011-2012	79.47
2012-2013	70.53

From the table 2 showed that the overall average percentage of 75.09% across three seasons are successfully achieved in this research is well above the overall average predictive accuracy percentage presented in [9] with 59.21%. Next, Figure 1, Figure 2, and Figure 3 show the detailed of the results.



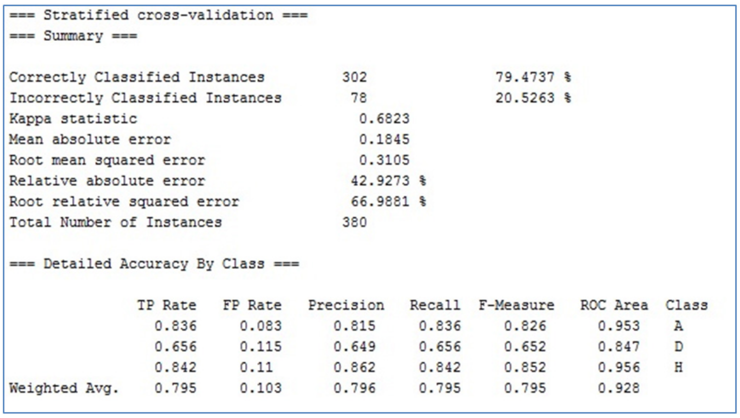
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         286               75.2632 %
Incorrectly Classified Instances        94               24.7368 %
Kappa statistic                          0.6133
Mean absolute error                      0.2201
Root mean squared error                  0.3317
Relative absolute error                 51.8283 %
Root relative squared error             72.0024 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.832     0.109     0.871      0.832     0.851      0.949    H
                 0.658     0.19      0.589      0.658     0.621      0.808    D
                 0.711     0.072     0.753      0.711     0.731      0.948    A
Weighted Avg.    0.753     0.124     0.761      0.753     0.756      0.908
```

Figure 1. Results for Season 2010-2011



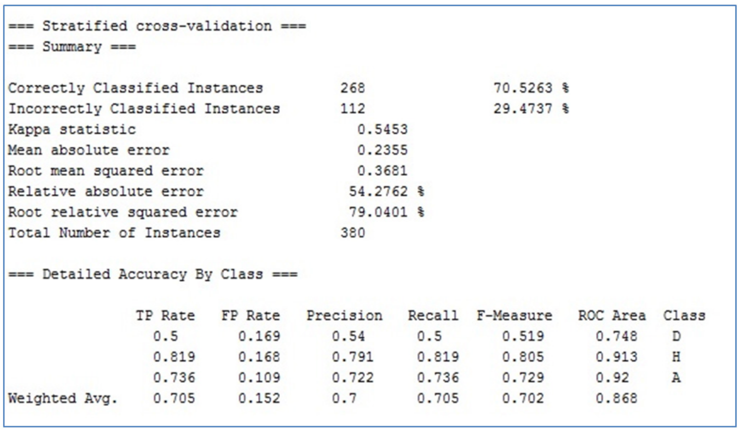
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         302               79.4737 %
Incorrectly Classified Instances        78               20.5263 %
Kappa statistic                          0.6823
Mean absolute error                      0.1845
Root mean squared error                  0.3105
Relative absolute error                 42.9273 %
Root relative squared error             66.9881 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.836     0.083     0.815      0.836     0.826      0.953    A
                 0.656     0.115     0.649      0.656     0.652      0.847    D
                 0.842     0.11      0.862      0.842     0.852      0.956    H
Weighted Avg.    0.795     0.103     0.796      0.795     0.795      0.928
```

Figure 2. Results for Season 2011-2012



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         268               70.5263 %
Incorrectly Classified Instances       112               29.4737 %
Kappa statistic                          0.5453
Mean absolute error                      0.2355
Root mean squared error                  0.3681
Relative absolute error                 54.2762 %
Root relative squared error             79.0401 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.5       0.169     0.54       0.5       0.519      0.748    D
                 0.819     0.168     0.791      0.819     0.805      0.913    H
                 0.736     0.109     0.722      0.736     0.729      0.92     A
Weighted Avg.    0.705     0.152     0.7        0.705     0.702      0.868
```

Figure 3. Results for Season 2012-2013

Figure 1, Figure 2, and Figure 3 have showed the prediction accuracy of 75.26% (286/380), 79.47% (302/380) and 70.53% (268/380) respectively. K-fold cross validation have been used to measure the performance of prediction model. According to [11], cross-validation is a powerful general technique for estimate model prediction performance. In this research, 10 fold cross validation have been used. Thus, 380 matches data of every season of English Premier League have been divided into 10 equal set of size. Each set is divided into two groups: 90% matches data are used for training and 10% matches data are used for testing. Then, 90% matches data will be trained using Bayes Classifier and applied to remaining 10% of matches data for testing in set 1. The same steps will be repeated for remaining 9 equal sized set of matches data. As the result, the averages of accuracy of the 10 Bayes classifiers are produced from 10 equal sized (90% training and 10% testing) set same like have been shown in Figure 1, Figure 2, and Figure 3 in term of correctly classified instances for prediction model accuracy.

4. Conclusions

Matches predictions in football has become a problem of interest and researchers have been trying to find solutions that outperforms one another. This paper described the use of Bayesian Networks to predict the football results in the English Premier League. The average accuracy of 75.09% across three seasons using the BN in this dataset is well above the overall average predictive accuracy percentage in [9] with 59.21%. It is hoped that the results could be used as the benchmark output for future research in predicting football matches results.

Acknowledgements

This research is funded by the Ministry of Higher Education Malaysia under the Research Acculturation Collaborative Effort (RACE) Grant Scheme Vot 1513.

References

[1] Hesar A S, Tabatabaee H and Jalali M 2012 *International Conference on Information and Knowledge Management*

[2] Baio G and Blangiardo M 2010 *Journal of Applied Statistics* 1–13

[3] Stefani R 1977 *IEEE Transactions on Systems, Man, and Cybernetics* **7** 117–121

[4] Rue H and Salvesen O 2000 *Journal of the Royal Statistical Society: Series D (The Statistician)* **49** 399–418

[5] Min B, Kim J, Choe C, Eom H and (Bob) McKay R I 2008 *Knowledge-Based Systems* **21** 551–562

[6] Constantinou A C, Fenton N E and Neil M 2012 *Knowledge-Based Systems* **36** 322–339

[7] Koopman S J and Lit R 2015 *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178** 167–186

[8] Maher M 1982 *Statistica Neerlandica* **36** 109–118

[9] Joseph A, Fenton N E and Neil M 2006 *Knowledge-Based Systems* **19** 544–553

[10] Pearl J and Russell S 2003 *The Handbook of Brain Theory and Neural Networks* **2** 157–160

[11] Seni G and Elder J 2010 *Synthesis Lectures on Data Mining and Knowledge Discovery*