International Journal of Integrated Engineering:
Special Issue 2018: Data Information Engineering, Vol. 10 No. 6 (2018) p. 155-158.
© Penerbit UTHM DOI: https://doi.org/10.30880/ijie.2018.10.06.022

Bayesian Approach to Classification of Football Match Outcome

Muhammad Haleq Azhar Abdul Rahman[1], Aida Mustapha[1*], Rahmat Fauzi[2], Nazim Razali[1]

[1]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia.
[2]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia

Received 28 June 2018; accepted 5 August 2018, available online 24 August 2018

Abstract: The football match outcome prediction particularly has gained popularity in recent years. It attract lots type of fan from the analyst expert, managerial of football team and others to predict the football match result before the match start.There are three types of approaches had been proposed to predict win, lose or draw; and evaluate the attributes of the football team. The approaches are statistical approach, machine learningapproach and Bayesian approach. This paper propose the Bayesian approaches within machine learning approaches such as Naive Bayes (NB), Tree Augmented Naive Bayes (TAN) and General Bayesian Network (K2) to predict the football match outcome. The required of football data is the English Premier League match results for three seasons; 2016 – 2017, 2015 – 2016 and 2014 – 2015 downloaded from http://www.football-data.co.uk. The experimental results showed that TAN achieved the highest predictive accuracy of 90.0 % in average across three seasons among others Bayesian approach (K2 and NB). The result from this research is hope that it can be used in future research for predicting the football match outcome.

Keywords: Football match prediction, Bayesian algorithm, data mining.

1. Introduction

The English Premier League (EPL) is the most watched football league alongside Serie A (Italy), Bundesliga (Germany) and La Liga (Spain).It is particularly popular among Asia. The football match outcome prediction particularly has gained popularity in recent years. The football match outcome predictions attract lots type of fan even from the analyst expert, managerial of the football team and others. Predicting the football match outcome become a bit difficult as there are various factors that could affect the result such as the skill and teamwork of the players in each teams, the venue of the match, the duration of the match and many others.

The Bayesian approach is very popular among prediction as it works to predict the weather, diseases, technology and sports.In association football or soccer, Bayesian approach has proven that it successfully outperforms other machine learning techniques [1].However, there are many type of Bayesian learning algorithm and three of themhas been chosen and used in this paper which is Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN) and General Bayesian Networks (K2). The EPL is chosen in this paper due to its popularity among Asia and the availability of dataset that is used in this project is source from thesource from the legitimate site at http://www.football-data.co.uk.

The raw data that is extracted is from three seasons of EPL from year 2015 to 2016.Then, the raw data has been run the process of data cleaning and data selection to remove the unnecessary attributes that is not effect the result of prediction of the match outcome. This paper is set to predict the football match outcome using three different types of Bayesian algorithms, which are Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN) and General Bayesian Networks (K2). Next, the prediction accuracy will be compared.

The remaining of this paper proceeds as follows. Previous work on football match outcome predictions is summarized in Section 2. In Section 3, the experimental setup including the dataset, the Bayesian Algorithm as well as the results in terms of predictive accuracies are presented. Finally, the conclusions are in provided in Section 5.

2. Related Work

Football is one of the foremost prevalent sports. Forecasting the results of football matches is curiously to numerous, from fans to punters and researchers. It is additionally interesting to the researchers due to its challenge since the result of a football match is dependenton numerous factors, such as a teamwork's, skills, and condition. Even for football specialists, it is exceptionally hard to anticipate the precise result about

Correspondingauthor: aidam@uthm.edu.my the football matches outcome. However, the football prediction model with the application of Bayesian approach has been carried out in [2, 3]. Both research proposed the same Bayesian algorithm as they claims that the Bayesian Networks become very popular in artificial intelligence and it is the most powerful algorithm as compared to other techniques in predicting football match outcome.

In [2], the Bayesian Network performance is tested in the area of predicting the football match result that involving the Barcelona FC in Spanish league (La Liga) for the season 2008-2009 while [3] predict the football match result for three seasons of English Premier League(EPL) and the accuracy of the prediction model were tested using the k-fold cross validation. Next, [4] developed other football predicting model that can generate forecast on the football match outcome for English Premier League (EPL) in season 2010-2011 which is called pi-rating model. The result in work of [1] showed that Bayesian network achieved higher accuracy as compared to other machine learning techniques such as Naïve Bayes, k-nearest neighbor (KNN) and decision tree. [5] proved that the general Bayesian Network (GBN) perform better than Naïve Bayes (NB) with the performance that is close to Tree Augmented Naïve Bayes (TAN) and challenge the claims of [6] which the GBN perform no better than NB on analyzed datasets.

3. Experimental Setup

The experiment is carried out using the open source software called WEKA that is under the General Public License (GNU).The WEKA software is provided with the implementation of data mining which is Bayesian algorithm such TAN, BN, and K2. The accuracy of the prediction model performance are measured by using k-fold cross validation as it is the general technique that is used to estimate the prediction accuracy.

3.1 Dataset

The dataset that is chosen for this work is the English Premier League (EPL) for three seasons which from 2014-2015, 2015-2016 and 2016-2017. The dataset isaccessible and were collected from the website at http://www.football-data.co.uk.The English Premier League (EPL) consist of 20 teams that will play twice in the season and each team will play at their home and away (double round robin league format). Thus, 1 team played 19 matches and overall 380 matches will be played in 1 complete season of EPL. As a result, 1140 data (3 complete season of EPL) have been used for this research. The 20 attributes have been chosen to follow the work in [3] exclude the attributes from bookmaker odds companies for the matches result. Table 1 shows the total of 20 attributes that were used in this experiment to predict the match outcome by using three different of the Bayesian algorithm such as Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN) and General Bayesian Network (K2).

Table 1 List of 20 attributes used in the experiment.

Label	Attribute	Sample
		Values
Home Team		Chelsea
Away Team		West Ham
FTHG	Full Time Home Goal	2
FTAG	Full Time Away Goal	1
FTR	Full Time Result	H
HTHG	Half Time Home Goal	0
HTAG	Half Time Away Goal	0
HTR	Half Time Result	D
HS	Home Shot	16
AS	Away Shot	7
HST	Home Shot Target	6
AST	Away Shot Target	3
HF	Home Foul	16
AF	Away Foul	16
HC	Home Card	7
AC	Away Card	1
HY	Home Yellow	5
AY	Away Yellow	2
HR	Home Red	0
AR	Away Red	0

3.2 Pre-processing

The raw data that had been collected might had some missing values, inconsistency and this would affect the quality and the accuracy of the football match prediction. In order to improve and maintain the quality of the data and the prediction result, the raw data undergo data preprocessing such as data cleaning and data reduction.

Data cleaning is work as to clean the raw data by filling in the missing value, smoothing the noisy data and resolving any of inconsistencies data. The missing value can be overcome by various method such asuse the attribute mean, use attribute mean for all samples that belong to the same class as the given tuple and others. The smoothing technique that could help to smooth the noisy data by using binning, clustering, regression and human inspection.

A huge amount of data often take more time to make an analysis and Data reduction is done by reducing the quantity or volume of the attributes to make the analysis be more efficient and produce almost the same of analytical result. The dataset that were extracted for this experiment undergoes both of the data preprocessing as described before to ensure only the relevant attributes that will be used to predict the football match outcome and reduce the time consume for the experiment.

3.3 Bayesian Algorithm

There are three variations of Bayesian algorithms were chosen and presented here which are Naïve Bayes (NB), Tree Augmented Naïve Bayes (TAN) and K2. K2 algorithm has been chosen for representing general Bayesian network (GBN). According to [7], K2 algorithm is the best algorithm in general Bayesian network in construction time and demonstrate good performance for structure learning which not too complex compared to other presented algorithms in their work.

The Naive Bayes (NB) classifier is based on Bayes' theorem with a strong assumption that all variables are independent between predictors.A Naive Bayesian model is straightforward and easy to build, which makes it particularly useful for very large datasets. Bayes theorem calculates the posterior probability, $P(c|x)$from$P(c), P(x)$,and$P(x|c)$.Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. The Bayesian probability is shown in Equation 1:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

where

- $P(c)$is the prior probability of c.
- $P(c|x)$is the conditional probability of c given x.
- $P(x|c)$ is the conditional probability of x given c.
- $P(x)$is the prior probability of data x.

Fig. 1 to Fig. 3 shows the variation of the Bayesian algorithms. According to [8], the term naïve because of it depends on two simple assumptions. In particular, it assumes that conditional properties are conditional on a given class, it states that no hidden or latent attributes affect the prediction process.The model of Naïve Bayes network is shown in Fig 1.

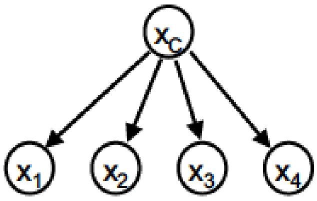


Fig. 1A naïve Bayesian classifier [5].

According to [6], the Tree –Augmented Naïve Bayes is an extension of Naïve Bayes which can have one other edge indicating to each node. The class attribute of TAN network has single class attribute have no parents while each of other attribute have a class as the parents and become a parent for the most to one other attribute. Fig. 2 shows the well-known TAN network that will be used in this research.

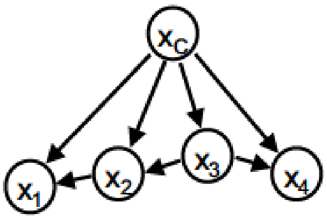


Fig. 2A tree augmented naïve Bayesian classifier [6].

According to [9], the K2 algorithm use a greedy search which there are no restriction on the number of the parents for a node has. The K2 algorithm assuming that the node has no parent and then adding gradually the parent to the node which resulting the most higher score. It only stop adding parent when the node score have not increase. Although, there are many other Bayesian learning algorithm in general Bayesian network family such as Hill Climbing and Greedy Thich Thinning, K2 algorithm is the best algorithm in general Bayesian network in construction time and demonstrate good performance for structure learning which not too complex compared to other presented algorithms. Fig. 3 shows the network model for K2 algorithm.

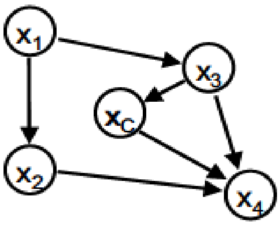


Fig. 3A general Bayesian network (K2) classifier [5].

3.4 Validation

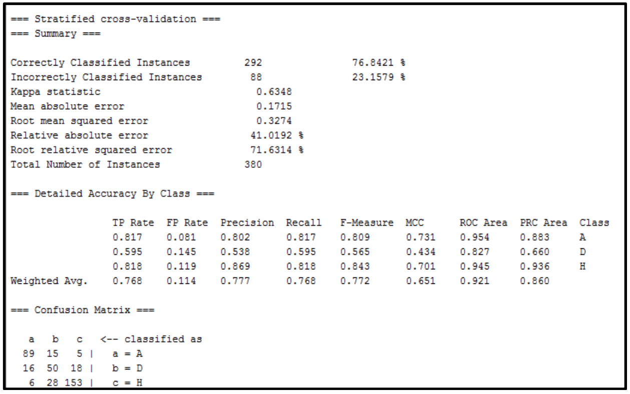
Cross-Validation is a statistical method to evaluate and comparing learning algorithm by dividing the data into two different groups, which is training set and testingset. Thus, 10 equal of set size had been divided from 380 matches data for each season of English Premier League (EPL). Then, Each of set have been separated into two type of group which is 90% data will be set for training and 10% of the data will be set as testing. The balance of 9 equal of set size will be repeated using the same process. According to [10], estimation of prediction error is important to evaluate the performance of the propose model. 10-fold cross validation is used to estimate the prediction error that occurs during the experiment of each Bayesian Algorithm due to [11] claims that 10 – fold cross validation can perform better and it can be served as standard procedure to estimate the performance and model selection.

4. Results and Discussion

The experiment is conducted using WEKA software and the dataset that is stated in section 3.1. Table 4.1 show the table of comparison result prediction performance in term of accuracy in percentage (%) among General Bayesian Network (K2), Tree Augmented Naïve Bayes (TAN) and Naïve Bayes (NB) across three seasons in English Premier League (EPL) from 2015 to 2017.

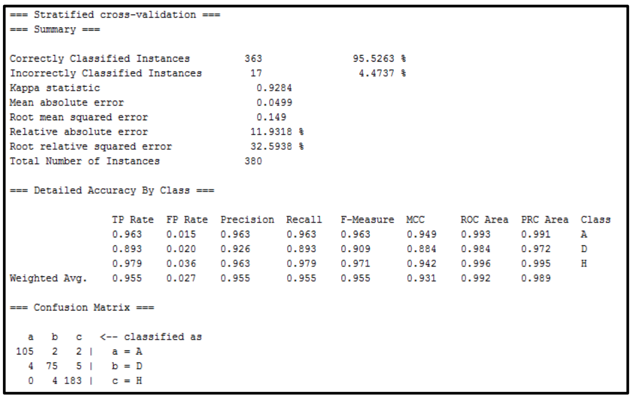
Data Sample (season)	**K2**	**TAN**	**NB**
2014-2015	76.84%	91.32%	72.89%
2015-2016	72.11%	83.16%	71.05%
2016-2017	76.84%	95.53%	78.16%
Average Accuracy Percentage (%)	**75.26%**	**90.00%**	**74.03%**

Overall, TAN successfully outperform K2 and NB algorithm in term of seasonal accuracy and average accuracy for three EPL seasons while K2 successfully outperform NB for two seasons (EPL season 2014-2015 and 2015-2016). The NB algorithm only manage to outperform K2 algorithm for season 2016-2017 by differential of 1.32% of accuracy. A sample for EPL season 2016-2017 of the output resultproduced by WEKA for each Bayesian algorithm (K2, TAN and NB) that is used to predict the football match outcome are shown in Fig. 4, Fig. 5 and Fig. 6, respectively.



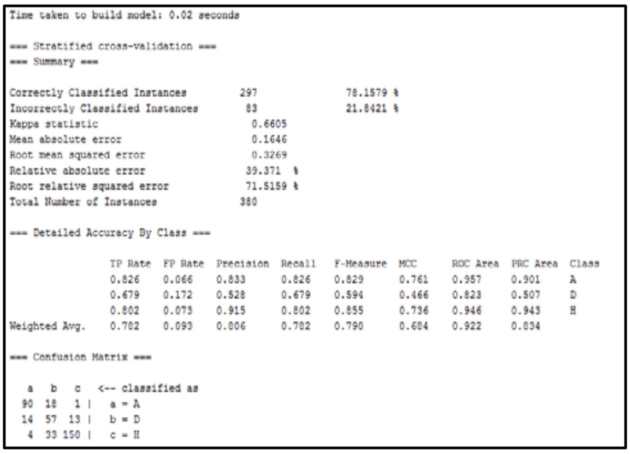
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         292               76.8421 %
Incorrectly Classified Instances        88               23.1579 %
Kappa statistic                          0.6348
Mean absolute error                      0.1715
Root mean squared error                  0.3274
Relative absolute error                 41.0192 %
Root relative squared error             71.6314 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.081    0.802      0.817    0.809      0.731    0.954     0.883     A
                 0.595    0.145    0.538      0.595    0.565      0.434    0.827     0.660     D
                 0.818    0.119    0.869      0.818    0.843      0.701    0.945     0.936     H
Weighted Avg.    0.768    0.114    0.777      0.768    0.772      0.651    0.921     0.860

=== Confusion Matrix ===

   a   b   c   <-- classified as
  89  15   5 |   a = A
  16  50  18 |   b = D
   6  28 153 |   c = H
```

Fig. 4 Results for season 2016 – 2017 for K2 algorithm.



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

Correctly Classified Instances         363               95.5263 %
Incorrectly Classified Instances        17                4.4737 %
Kappa statistic                          0.9284
Mean absolute error                      0.0499
Root mean squared error                  0.149
Relative absolute error                 11.9318 %
Root relative squared error             32.5938 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.963    0.015    0.963      0.963    0.963      0.949    0.993     0.991     A
                 0.893    0.020    0.926      0.893    0.909      0.884    0.984     0.972     D
                 0.979    0.036    0.963      0.979    0.971      0.942    0.996     0.995     H
Weighted Avg.    0.955    0.027    0.955      0.955    0.955      0.931    0.992     0.989

=== Confusion Matrix ===

   a   b   c   <-- classified as
 105   2   2 |   a = A
   4  75   5 |   b = D
   0   4 183 |   c = H
```

Fig. 5 Results for season 2016 – 2017 for TAN algorithm.



```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         297               78.1579 %
Incorrectly Classified Instances        83               21.8421 %
Kappa statistic                          0.6605
Mean absolute error                      0.1646
Root mean squared error                  0.3269
Relative absolute error                 39.371  %
Root relative squared error             71.5159 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.826    0.066    0.833      0.826    0.829      0.761    0.957     0.901     A
                 0.679    0.172    0.528      0.679    0.594      0.466    0.823     0.507     D
                 0.802    0.073    0.915      0.802    0.855      0.736    0.946     0.943     H
Weighted Avg.    0.782    0.093    0.806      0.782    0.790      0.684    0.922     0.834

=== Confusion Matrix ===

   a   b   c   <-- classified as
  90  18   1 |   a = A
  14  57  13 |   b = D
   4  33 150 |   c = H
```

Fig. 6 Results for season 2016 – 2017 for NB algorithm.

5. Conclusions

This paper was set to study and describe the use of Bayesian Approach namely TAN, NB and K2 which representing the general Bayesian network to predict the football match outcome as the prediction in football become popular among the researchers that trying to find solution and overcome the problem in predicting the football match outcome. The challenge and difficulty of predicting the football match outcome is an interest of researchers to compete and the football data accessibility via other medium besides website. The average of predictive accuracy for TAN (90.00%) hassuccessfully outperformed other Bayesian algorithm (K2 and NB) while K2 (75.26%) and NB (74.03%) sharing a close accuracy performance of predicting the football match outcome across three season of English Premier League (EPL). The result from this research is hopedto be used for further research on predicting the football match outcome.

Acknowledgement

This project is sponsored by UniversitiTun Hussein Onn Malaysia.

References

[1] Joseph, A., Fenton, N.E., Neil, M. Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), (2006), pp. 544-553.

[2] Owramipur, F., Eskandarian, P., Mozneb, F.S. Football result prediction with Bayesian network in Spanish League-Barcelona team. International Journal of Computer Theory and Engineering, 5(5),(2013).

[3] Razali, N., Mustapha, A., Yatim, F.A., Aziz, R.A.Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). In IOP Conference Series: Materials Science and Engineering, vol. 226, no. 1, (2017). p. 012099.

[4] Constantinou, A.C. Bayesian networks for prediction, risk assessment and decision making in an inefficient association football gambling market Doctoral dissertation, Queen Mary University of London, UK. (2013).

[5] Madden, M.G. On the classification performance of TAN and general Bayesian networks. Knowledge-Based Systems, 22(7), (2009), 489-495.

[6] Friedman, N., Geiger, D., Goldszmidt, M. Bayesian network classifiers. Machine learning, 29(2-3), (1997), pp. 131-163.

[7] Hesar, A.S., Tabatabaee, H., Jalali, M. Structure learning of Bayesian networks using heuristic methods. In Proc. of International Conference on Information and Knowledge Management (ICIKM 2012).

[8] John, G. H., Langley, P. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence.* Morgan Kaufmann Publishers Inc. (1995), pp. 338-345.

[9] Cooper, G.F., Herskovits, E.A. Bayesian method for the induction of probabilistic networks from data. *Machine learning*, *9*(4), (1992), pp. 309-347.

[10] Fushiki, T. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing, 21(2), (2011), pp. 137-146.

[11] Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI, v*ol. 14, No. 2, (1995), pp. 1137-1145.