A Comparative Study of Data Mining Techniques on Football Match Prediction

IOP Conf. Series: Journal of Physics: Conf. Series **1020** (2018) 012003 doi:10.1088/1742-6596/1020/1/012003

A Comparative Study of Data Mining Techniques on Football Match Prediction

Che Mohamad Firdaus Che Mohd Rosli[1], Mohd Zainuri Saringat[1], Nazim Razali[1], Aida Mustapha[1]

[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia

E-mail: chemohamadfirdaus94@yahoo.com, zainuri@uthm.edu.my, nazim.iium@gmail.com, aidam@uthm.edu.my

Abstract. Data prediction have become a trend in today's business or organization. This paper is set to predict match outcomes for association football from the perspective of football club managers and coaches. This paper explored different data mining techniques used for predicting the match outcomes where the target class is win, draw and lose. The main objective of this research is to find the most accurate data mining technique that fits the nature of football data. The techniques tested are Decision Trees, Neural Networks, Bayesian Network, and k-Nearest Neighbors. The results from the comparative experiments showed that Decision Trees produced the highest average prediction accuracy in the domain of football match prediction by 99.56%.

1. Introduction

In data generalization, there are two major topics that are concentrated on this subject; Data Mining and Statistics. According to [1], the reason that prediction approaches have been widely used in many companies around the world today is because prediction is the heart of remarkable disciplines in Science. This approach enables companies and organizations to predict and prepare reasonable future plans. One particular applications of prediction is in sport. In order to forecast the future of a particular organization, the logic of prediction is employed. In sport analytic for association football, match prediction is used to predict the matches before it started using a particular technique so that the managers and players able to forecast the possibilities of either win, draw, or loss of the game before it started. Positive outcome will definitely boost up player's spirit but if the outcome is negative, it does not mean to lowered player's sportsmanship but as a guide for them to play extra cautious and provide an appropriate counter-attack.

There are already many researches on the prediction of match outcome in sports, in basketball and football particularly. However, most of those researchers analyzed and predict using a single technique only. There are research that predicts using different techniques but the result might be biased since there are techniques that using a particular technique to increase the accuracy of the prediction. For example, [2] predicted football and basketball using Least Square while [3] predicted football outcome using Neural Tuning and Genetic Tuning fuzzy model. Both researches showed how to match using different techniques with a specific technique create by researchers.

To address this gap, a comparative analysis of several data mining techniques using only raw data is essential to find the most accurate techniques using only raw data without any elements that could affect the result like derivation of data or particular technique. This paper is set to achieved three objectives; (1) prediction of football match result (win, lose and draw) using raw data of previous match outcome, (2) model generation of those results using several techniques in data mining, and finally (3) comparative analysis to observe the result of accuracy metrics. This research is limited to previous match's data of three seasons of English Premier League from 2013-2014, 2014-2015, and 2015-2016 totaling 1,140 matches of both home and away.

The rest of the paper is organized as follows: Section II describes the related work on profiling and classification models. Section III presents the method uses in this paper to produce comparative analysis results. Section IV presents the experimental results. Finally, Section V concludes the work and highlights a direction for future research.

2. Related Work

Numerical prediction is a method where a continuous-valued function or ordered-value are predicts by the model constructed. This particular model is also known as predictor. The accuracy of the predictor depends on how well the provided predictor able to predicts the value of predicted attribute for a new data. When compared to classification methodology in data mining, predicted values in prediction methods are usually continuous whereas classifications are discreet, making a classification problem seen as a predictor of classes. Prediction are normally about predicting the future and classification are about classifying the present. Machine learning approaches have been used in predicting match outcome in sports especially for basketball and football association. This prediction is carried out by generating past match results. In this work, four standard prediction algorithms are used for comparative purposes, which are Decision Trees, Neural networks, Bayesian networks, and k-Nearest neighbors.

A decision tree is a structure that includes a leaf nodes, branch nodes, and root nodes. Each leaf node holds a class label, each internal node denotes a test on an attribute, and branch denotes the outcome of a test. The topmost node in the tree is the root node. Decision tree model was used in predicting football matches outcome by [4]. In their experiment, the decision model was developed using the Machine Learning package in order to generate MC4 decision tree. MC4 decision tree is different than the original because it used the tree pruning technique in order to reduce the size and the complexity of the tree. However, MC4 decision tree only managed to predict 45.77% and 41.72% in accuracy for both expert model data and general model data.

Neural Network is a parallel distributed processor that has a propensity for storing experiential knowledge and making it available for users [5]. Neural computing is the study of networks of adaptable nodes which through a process of learning from task. According to [6], neural network model was used to predict the match outcome between two teams during World Cup in 2006. Technically, they proposed a supervised multi-layer perceptron neural network (MLP) with error propagation learning rule (BP) to predict the winning rate for the teams. Then, they compared the result in order to determine the relationship between win and loss. The accuracy for this model was 76.9% excluding the tied games due to limitation in predicting draw outcome.

Belief networks or Bayes nets for short, were the other name for Bayesian Network. It belong to the probabilistic graphical models. These models can be used to represent knowledge about uncertain amount. In particular, the graph contains nodes and edges where nodes represent a random variables and the edges represent the relation of probabilistic dependencies among the corresponding random variables. This Bayesian Network model was used by [7] to predict football outcomes using the Skellam's distribution. This paper focuses on the goal differences instead of the number of goals like most of researchers. The research was made to a specific

team with observed number point greater than 1.95 or probability of observed outcome lower than 20%. For instance, the prediction accuracy of win for Bolton team is 53%.

The k-Nearest Neighbors method begins in the early 1950s. The method is labor-intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. Nearest Neighbors are based by comparing a given test tuple with training tuples that are similar to it. [4] used the IB classifier from the family of k-Nearest Neighbors (KNN) in the MLC++ library for predicting matches outcome for Tottenham Hotspur Football Club for 2 seasons. The KNN technique used likeness approach, where a Euclidean distance is measured and plotted on and n-dimensional graph where each dimensional is one of the supplied attributes. KNN technique produced prediction accuracy at 50.58% and 47.21% for two seasons.

3. Methodology

In this section, the dataset, the classification techniques, and the performance measures are presented and discussed. Experiments for Decision Tree, Bayesian Network, and k-Nearest Neighbors are implemented using Weka 3.8 [8] while the Neural Networks algoritm is implemented using the GMDH Shell DS (`https://www.gmdhshell.com/`). In Weka 3.8, the tool uses 10-fold cross-validation method to validate the data. As for classifier, Decision Tree uses J48 classifier while Bayesian Network uses BayesNet classifier. IBk will be the classifier appointed for k-Nearest Neighbors technique. In GMDH Shell DS, the Neural Network technique is implemented using the cross-validation criterion PRR to validate the data.

3.1. Dataset

The dataset chosen to be implement in this research was the data of the English Premier League of 2013 to 2016 season. English Premier League consisted of 20 teams played both home and away totaled 380 matches throughout the season. This is mean there are approximately 1,140 matches for three seasons of English Premier League. The data consists details of the number of full time home and away goal, full time result (win, draw or lose), the number of home and away shots, the number of home and away shots on target and the number of home and away corner. The data was retrieved from the website `http://www.football-data.co.uk`. There are eight features extracted from the datasets, which are `Full Time Home Goal`, `Full Time Away Goal`, `Home Team Shot`, `Away Team Shot`, `Home Shot On Target`, `Away Shot On Target`, `Home Corner`, and `Away Corner` while `Full Time Results` which are win, draw or lose as target class for prediction.

3.2. Evaluation metrics

The performance measures used in this paper is the accuracy. Overall, accuracy measures how often is the classifier or the classification techniques (Decision Trees, Neural Networks, Bayesian Networks, k-Nearest Neighbor) getting correct target class of classification. Equation 1 shows how accuracy is measured.

$$\text{accuracy} = \frac{TP + TN}{P + N} \tag{1}$$

From the equation, TP represents true positives or equivalent to hit, TN represents true negatives or equivalent to correct rejection, P is the total number of real positive cases in the data, and finally N is the total number of real negative cases in the data. In context of football match results, its refer to win, draw or lose target class. For example, if the predicted result of a match hit win but the actual result is lose the result will not be counted in TP and TN. However, it will counted in P which it is actually false positve. Literally, its can be explained

that the total number of correct predicted match results (win, draw or lose) are divided with total number of matches.

4. Results

The experiment generated the percentage of accuracy for each technique. All four techniques have been implemented using the specific tool to acquire prediction data. The result were recorded and analyzed. A comparison of the results collected have been made to observe the most accurate technique to predict football match outcome. In order to validate the data, 10-fold cross-validation techniques have been choosen.

4.1. Decision trees

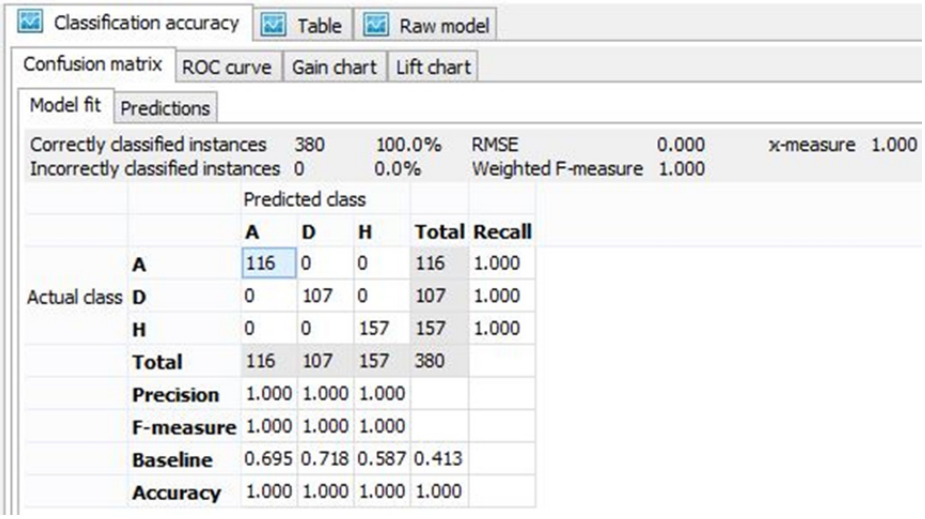
The data retrieved have been imported into Weka 3.8. Then, the data are run using Decision Tree technique. In this case, Weka 3.8 uses Decision Tree J48 Classifier which are also called C4.5 Decision Tree to generate the result. There is no need to discretize the data because C4.5 Decision Tree can handle numeric attributes. The summary of the result is shown in Figure 1. The summary report shows that for 15/16 seasons of English Premier League matches, the accuracy percentage generated was 99.47%.

```
=== Summary ===

Correctly Classified Instances         378               99.4737 %
Incorrectly Classified Instances         2                0.5263 %
Kappa statistic                          0.992
Mean absolute error                      0.0035
Root mean squared error                  0.0549
Relative absolute error                  0.8059 %
Root relative squared error             11.7266 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.008    0.983      1.000    0.991      0.988    1.000     1.000     A
               0.981    0.000    1.000      0.981    0.991      0.987    1.000     1.000     D
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     H
Weighted Avg.  0.995    0.002    0.995      0.995    0.995      0.993    1.000     1.000

=== Confusion Matrix ===

   a   b   c   <-- classified as
 116   0   0 |   a = A
   2 105   0 |   b = D
   0   0 157 |   c = H
```

Figure 1. Summary Report of J48 Decision Tree of 15/16 Seasons

4.2. Neural networks

Different tool called the Group Method of Data Handling (GMDH) has been used for running the football prediction using Neural Networks. The same validation techniques have used in order to generate prediction accuracy for Neural Networks technique to be compared to accuracy of Decision Tree, Bayesian Network, and k-Nearest Neighbors, which were all implemented in Weka 3.8. The result of Neural Networks is surprisingly accurate with 100% accuracy throughout 380 matches. Nevertheless, this result does not make Neural Networks as the best technique because the the preduction accuracy is averaged out to take into account performance across three seasons. Figure 2 shows the result for 2015-2016 season.


Classification accuracy | Table | Raw model

Confusion matrix | ROC curve | Gain chart | Lift chart

Model fit | Predictions

Correctly classified instances 380 100.0% RMSE 0.000 x-measure 1.000
Incorrectly classified instances 0 0.0% Weighted F-measure 1.000

		Predicted class				
		A	**D**	**H**	**Total**	**Recall**
Actual class	**A**	116	0	0	116	1.000
	D	0	107	0	107	1.000
	H	0	0	157	157	1.000
	Total	116	107	157	380	
	Precision	1.000	1.000	1.000		
	F-measure	1.000	1.000	1.000		
	Baseline	0.695	0.718	0.587	0.413	
	Accuracy	1.000	1.000	1.000	1.000	

Figure 2. Twice-Multilayered Neural Network of 15/16 Seasons

4.3. Bayesian networks

Weka 3.8 was used to generate model using Bayesian Network technique as well as Decision Tree. The data set of English Premier League of 13/14 seasons to 15/16 seasons were then implemented into Weka 3.8 to generate prediction data. The summary report of Bayesian Network of 15/16 Seasons in Figure 3 shows 75.53% of accuracy from 380 matches throughout the seasons.

```
=== Summary ===

Correctly Classified Instances         287               75.5263 %
Incorrectly Classified Instances        93               24.4737 %
Kappa statistic                          0.625
Mean absolute error                      0.2274
Root mean squared error                  0.3229
Relative absolute error                 51.9126 %
Root relative squared error             68.9984 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.776     0.083     0.804      0.776     0.789      0.700    0.950     0.898     A
               0.561     0.139     0.612      0.561     0.585      0.433    0.808     0.675     D
               0.873     0.148     0.806      0.873     0.838      0.718    0.959     0.945     H
Weighted Avg.  0.755     0.126     0.751      0.755     0.752      0.632    0.914     0.855

=== Confusion Matrix ===

   a   b   c   <-- classified as
  90  19   7 |   a = A
  21  60  26 |   b = D
   1  19 137 |   c = H
```

Figure 3. Summary Report of Bayesian Network of 15/16 Seasons

4.4. k*-Nearest neighbors*

The *k*-Nearest Neighbors technique was also applied to the football dataset by using Weka 3.8. The experiment used IBk classifier with 10-fold cross-validation technique similar to J48 Decision Tree and BayesNet Bayesian Network algorithms. Figure 4 shows the summary report generated by Weka 3.8 of IBk k-Nearest Neighbors technique. The result produced accuracy of 78.16% for 2015-2016 season.

```
=== Summary ===

Correctly Classified Instances         297               78.1579 %
Incorrectly Classified Instances        83               21.8421 %
Kappa statistic                          0.6687
Mean absolute error                      0.1482
Root mean squared error                  0.38
Relative absolute error                 33.8403 %
Root relative squared error             81.1951 %
Total Number of Instances              380

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.776    0.076    0.818      0.776    0.796      0.711    0.850     0.703     A
               0.692    0.165    0.622      0.692    0.655      0.511    0.763     0.517     D
               0.847    0.081    0.881      0.847    0.864      0.771    0.883     0.809     H
Weighted Avg.  0.782    0.103    0.789      0.782    0.784      0.679    0.839     0.695

=== Confusion Matrix ===

   a   b   c   <-- classified as
  90  23   3 |   a = A
  18  74  15 |   b = D
   2  22 133 |   c = H
```

Figure 4. IBk k-Nearest Neighbors of 15/16 Seasons

4.5. Comparative results

The comparative results are shown in Table 1. The results using Neural Network technique generated by GMDH is surprisingly accurate with 100% accuracy throughout 380 matches for season 2014-2015 and 2015-2016. However, the technique still does not suggested as the best since the average accuracy of three seasons been taken into account. The comparative results shown that Desicion Trees have the highest average accuracy by 99.56% follow by Neural Networks and k-Nearest Neighbors technique by 96.83% and 77.54% while Bayesian Networks have the worst average accuracy by 76.41%.

Table 1. Comparison of Accuracy Percentage

Accuracy	Decision Trees	Neural Networks	Bayesian Networks	k-Nearest Neighbors
Season 2013-2014	99.74%	90.50%	77.37%	78.68%
Season 2014-2015	99.47%	100.00%	76.32%	75.79%
Season 2015-2016	99.47%	100.00%	75.53%	78.16%
Average	99.56%	96.83%	76.41%	77.54%

5. Conclusions and Future Works

This study compared four standard data mining algorithms in football match outcome prediction. The experimental results proved that Decision Tree is the most accurate technique to predict football match outcome with 99.56% accuracy as compared to other techniques, which are Neural Network (96.83%), Bayesian Network (76.41%), and k-Nearest Neighbors (77.54%), respectively. However, the research can be improved by studying and creating the same technique for each technique to increase the accuracy of results by using other attributes that correlated to match outcome besides the attributes that have been used in this research such as the distance travelled by away team or current performance of competing teams in the league or the number of yellow or red cards fouled by player which offcourse influece the football match outcome. Moreover, the techiques used in this research is static over time because it depend on past

match data while the real time football match statistic during match also need to be measured and taking account for prediction (dynamic).

In any sports, knowing the possibilities of winning a game is essential. Result of outcome prediction can be used to elevate player's sportsmanship or to prepare a counter-attack against their opponents. Match outcome prediction should become a trend in our national sports in order to produce a profitable and quality organization of a particular sports while increases athletics' performances. Management of a particular sports also can used this prediction method to forecast the future of an organization.

Acknowledgements
This project is sponsored by the Malaysian Ministry of Higher Education (MOHE) under the Research Acculturation Grant Scheme vot 1513.

References

[1] Dhar V 2013 *Communications of the ACM* **56** 64–73

[2] Stefani R T 1977 *IEEE Transactions on systems, man, and cybernetics* **7** 117–121

[3] Rotshtein A P, Posner M and Rakityanskaya A 2005 *Cybernetics and Systems Analysis* **41** 619–630

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

[5] Rumelhart D E, Hinton G E and Williams R J 1988 *Cognitive modeling* **5** 1

[6] Huang K Y and Chang W L 2010 A neural network method for prediction of 2006 world cup football game *Neural Networks (IJCNN), The 2010 International Joint Conference on* (IEEE) pp 1–8

[7] Karlis D and Ntzoufras I 2009 *IMA Journal of Management Mathematics* **20** 133–145

[8] Witten I H, Frank E, Hall M A and Pal C J 2016 *Data Mining: Practical machine learning tools and techniques* (Morgan Kaufmann)