Premier League Matchs Forecasting Using Multi Time-Frame Approach

Final project report

[redacted]

[redacted]

Abstract — **Predicting soccer matches' outcome poses a significant challenge due to the fact that a sport event is stochastic and involves complicated correlation between its relevant features. In order to reach an effective prediction framework, the factors of sports events and the dynamic correlations between them should be taken into account. As part of this work, a multi-time frame approach has been introduced to predict the match outcome of Manchester City Football Club (MCFC) analytical community dataset. The framework simulates the dynamic nature of sport events using both the objective analysis for sports historical datasets supported with the subjective information from experts. The framework integrates K-nearest neighbor classifier (KNN), Naïve Byes classifier (NBC) and rule-based reasoner into one system. Based on the adopted framework, the prediction accuracy can reach up to 60%.**

Keywords—Bayesiean framework; Match prediction; Rule based reasoner; Sports data mining

I. Introduction

Due to the extremely competitive environment in sports field, there is a huge demand for data mining technologies to find, extract and learn about hidden patterns in sports statistical datasets. These knowledge discovery practices are intensively applied in sports field to perform statistical analysis, pattern discovery as well as outcome prediction processes in order to support decision making [1]. Predicting the match outcome holds the interest of sports decision makers, experts, coaches, managers, scouts, and sports fans as well. As a result, for the last couple of years, researchers proposed different methodologies to provide the best possible approach to predict tournaments/match outcomes, match results, or team performance.

The nature of sports events makes the prediction process more challenging due to the correlation between different factors. In any sport event, the performance of competitors depends on physical, psychological, strategic, and random factors. Physical features are mainly skills, and strength of an athlete as stamina, shooting, and passing accuracy in football. Psychological conditions have a huge influence that can impact the physical attributes. These conditions depend on venue, fans, rivalries, motivation, and the morale of the team [2-3]. Strategy can be a key factor in sport events; strategic choices of formation, offense/defense plan, and opponent's reactions can highly influence the performance of a team. These factors can influence the prediction process, and the correlation between them makes the prediction more difficult. In addition, sport events have a room of randomness due to the stochastic nature of them. Thus, predicting the outcome of sports match becomes a challenging and sensitive process. As a result, match outcomes are hard to be predicted by the traditional way of seeking expert knowledge only. For this reason, it is better to combine expert knowledge with data mining tools.

In this paper, a dynamic framework will be designed for predicting the match outcome. The proposed framework utilizes both experts' knowledge and data mining techniques in a dynamic and compound perspective while considering the stochastic nature of sports events. The principal idea is to use a rule-based reasoner (RBR) in a multi-time frame approach. The RBR will be built based on experts' experience and intuitions and will be used to tune the data mining tools in order to get better classification and prediction results.

The paper is organized in the following manner: the second section presents the initial proposal, by describing the approach of the first project and describing clearly the key differences between Football Result Expert System (FRES) [5] and proposed idea. The third section consists of the changes and justifications applied over the initial proposal. The fourth section describes the working system as a whole in a more detailed manner, while the fifth section discusses results followed by a conclusion.

II. Initial Proposal

This section describes the proposed project with a discussion on the proposed approach. The project will be based on Manchester City Football Club (MCFC) analytical community data set. This dataset was released to inspire the next level of analytics in the hope of new performance measures, tools for player/team comparison and profiling [4]. MCFC released 10369 rows and 236 columns; each row presents a player in a certain match while each column starting from column 6 can be considered as feature which contains every 'on the ball' event for every player in every match. It was designed and prepared to suit data mining studies and analysis.

Most revised papers suffer from lack of detailed data [1-3]. Studies generally use team-based information such as current team ranking, formation, outcome of selected matches and so on. However, these features are not enough to achieve accurate results since players have different performances in different occasions which affect the overall team performance.

Among all revised papers, it seems that FRES approach is more realistic in terms of dealing with football matches prediction and its effectiveness was demonstrated in [5]. One of the advantages of FRES is that it overcomes the stochastic nature of sport events by introducing time frame analysis technique.

This proposal will utilize this idea from FRES but with some crucial differences. Even though dataset does not contain data in time domain, time frames will be simulated through an iterative process described in more detail below. In [5], FRES was used to predict the outcome of a tournament, while the proposed method will be used for predicting the outcome of the next match. The proposed model will be composed of two main components; one component is simulating the stochastic characteristics of soccer using four Bayesian networks referring to offense, defense, fatigue, and possession indices while the second acts as a simplified coach, just like in FRES, coach will be based on rule-based reasoner.

The proposed approach works as depicted in Fig. 1 with the following steps:

Step 1. Select one key feature for each Bayesian network.

Step 2. Use gain ratio to select the most influential features regarding the key features. The features selected in this step will constitute our dataset used during match prediction.

Step 3. For a given match, input the features selected in the previous step into the four Bayesian networks accordingly (for example input the offensive key features and most related features selected by gain ratio to the Offense Bayesian network and so on). The four Bayesian networks will calculate the "state" of the two competing teams according to the given four indices.

Step 4. Input the previous step output to association rules to determine the outcome score of that time-frame.

Step 5. Using the results of Step3 and Step4 the inference rules, use forward chaining to define the changes in the strategies of the team.

Step 6. Check if the current time frame is the last time frame in the match; if not go to Step3 else go to Step7.

Step 7. Provide the end results and end simulation.

The proposed method contains four main components that are discussed in more details below:

Feature selection: The idea behind using feature selection approach is that most of researchers depend on experts' knowledge for feature selection. However, expert's knowledge can differ significantly from an expert to another, meanwhile; there are data mining techniques that can be used for feature selection based on the importance of that feature. Thus, Information gain and gain ratio will be applied to evaluate available database accordingly, valuable features that can assist match outcome prediction will be selected based on their relevance to the manually chosen key features as given in step 2. Although, features are player dependent and the Bayesian network deals with team state, each team index state will be updated using an aggregation of the players' individual offence, defense, fatigue, and ball possession indices.

Bayesian Networks: Four Bayesian networks will be designed according to the following categories that are; offence, defense, ball possession, and fatigue as given in [5]. Each network will be assigned five features (or maybe more) to evaluate the performance of the team in that category.

Association rules: will be trained using previous matches as reference and the features selected in Step 2. The training is not done in time-frame basis but the features used are the same so the results are expected to be accurate. During program execution, the association rules will be driven by results of the four Bayesian networks, in order to estimate the score.

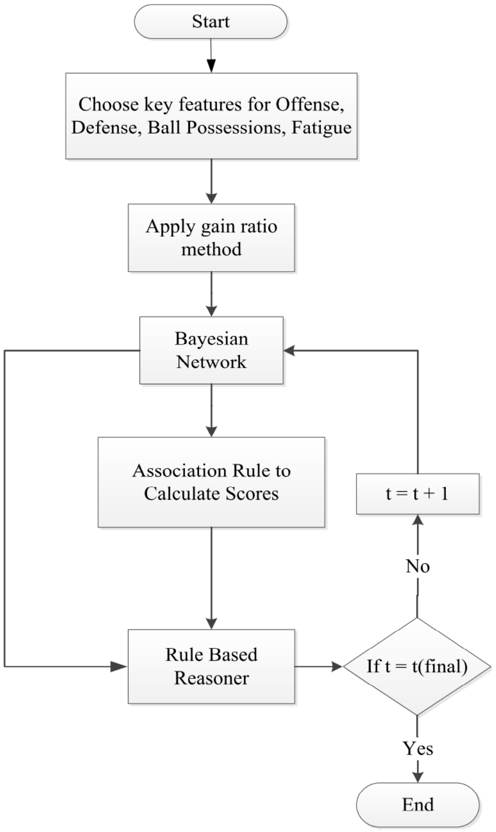
Start
Choose key features for Offense, Defense, Ball Possessions, Fatigue
Apply gain ratio method
Bayesian Network
Association Rule to Calculate Scores
t = t + 1
No
Rule Based Reasoner
If t = t(final)
Yes
End

Fig. 1 First proposal flow chart

Rule-based reasoner: will be designed using forward chaining method to determine coaches decisions on each time-frame based on outputs of association rules and Bayesian networks.

FRES's main workflow (as in [5]) is the following: the four Bayesian networks all have a desired output. They calculate the probabilities of their output based on some fixed and some probabilistic values. The desired output specifies some fixed inputs, for example if the attack network has as desired output aggressive, it sets the formation accordingly. This results in almost every case in some conflicting attribute values. At this point the rule-based reasoner takes over, modifying these values based on rules defined. These attributes then are returned to the Bayesian networks as fixed values, which are run again. The results of this run are then fed back to the rule-based reasoner, which infers the score for the time frame and also updates the state of the world, which does involve attributes like morale, which is over the influence of the coach. In contrast, our approach does not set any goals for the teams in advance. We assume, that the coach can perceive correctly the other team's attributes. When we calculate the probabilities using our Bayesian networks, based on these probabilities we provide a "vision" for the coach, where he can see the outcome of the current time frame if none of the teams change anything. This "vision" is provided by the association rules, so it can also be perceived as the experience of the coach. The coach (rule-based reasoner) gives the best advice he can, by modifying some attributes of the teams taking into account some constraints - this is done for both of the sides – and we run the Bayesian networks again, this time the association rules actually calculate the result and updates the state of the world. As it can be seen there are key differences between the two approaches regarding both assumptions and the mode the parts interact.

The project will be carried on following a sequence of steps where methods will be coded and tested individually before producing the final system code.

Firstly, the proposal will be done for one time frame taking into account that association rules will be coded and tested separately. Bayesian networks will also be designed individually with small number of training data. Once these steps are done, the match prediction will be found by combining these methods as shown in Fig. 1, yet without a rule based reasoner since association rule can give the match score. Second step will be to define the rule-based reasoner according to common knowledge of football and reasonable assumptions as given in [22]. It shall be designed by forward chaining method where it will also be tested in an independent environment. Finally, after finishing the test procedures, previous systems will be integrated as shown in Fig. 1. The proposed multi time-frame system is expected to produce better results in premier league matches prediction compared with single time-frame prediction.

III. REVIWED PROPOSAL

During the implementation different unforeseen challenges emerged that lead to changes in the plan. All modifications that were carried out are enlisted, while the next section describes the system as a whole in more detail.

The first challenge was regarding selecting features for fatigue. The dataset provided is not detailed enough to make relevant connections between the players' attributes and their level of fatigue, which lead to the decision of discarding this element from the project. The dataset also failed to provide detailed and relevant information regarding ball possession. Most of the passing attributes are correlated, but none of them was specific enough to draw conclusions about ball possession. Also, the contribution of passing features was not significant when applying feature selection. Because of the time constraint and complexity, Naïve Bayes Classifier and KNN were chosen instead of Bayesian networks. After implementing the classifiers, they were tried out as match predictors in order to have their performance as a reference for the whole system.

Association rules were not used because the match outcome can be predicted with the KNN or NBC classifiers, which lead to a simpler structure.

In Step 5, the rule-based reasoner is defined in the following manner: rules are defined in sequence, where some rules, if apply, might not let other rules apply after them.

Provided the rules, a 5-20% increase in accuracy was achieved, which justified the effectiveness of the chosen approach.

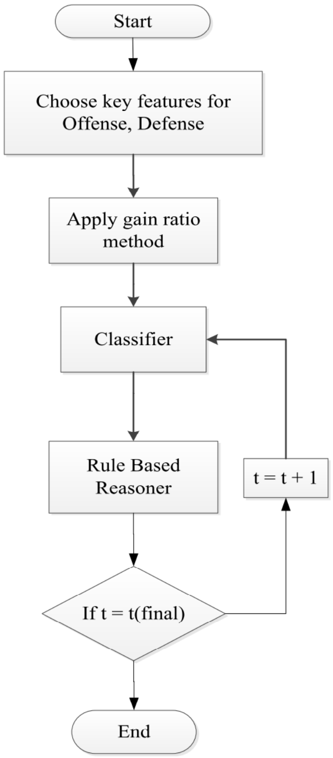

Fig. 2 Reviewed proposal flow chart

IV. FRAMEWORK IMPLEMENTAION

As shown in the reviewed proposal flowchart, the current system is built up of few different sections, where each section was designed to produce suitable outcome in order to be used as an input to the following section. In this portion of the report, an overview of the framework implementation will be presented together with justification of some decisions made towards the final framework.

The framework consists of four main parts; dataset preprocessing, random selection of test data, classification, and RBR.

A. *Database processing*

Initially, MCFC database was converted to represent team records rather than per player records. Since, the proposal targeted to predict match's outcome, all players' data should be combined and averaged to have per team data. Outcome labels (Win, Draw, and Lose) were inserted in the database. An overview of all steps involved in preprocessing database is shown in fig. 3.

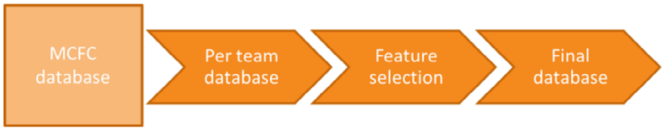

Fig. 3 Data extraction

RapidMiner software was used to select two groups of features (Offensive & Defensive) based on Gain Ratio. Thus, the output features after applying gain ratio on all features of per team database were as given in Table I.

TABLE I
SELECTED FEATURES

Offensive features	Shots On Target	Attempts Open Play on target	Unsuccessful Passes Middle third	Successful Dribbles	Total Fouls Conceded
Defensive features	Challenge Lost	Interceptions	Unsuccessful Ball Touch	Shots On Conceded	Unsuccessful Dribbles

Number of goals and conceded goals were used as classes for selecting offensive and defensive features respectively. There are some correlations between these two combinations of features. For example, "Unsuccessful Passes Middle third" correlates to "Interceptions". These correlations will be very helpful to setup some rules in RBR later on.

B. *Random selection of test data*

All steps, hereafter, were coded using octave and can be automatically executed after setting up all tunable parameters in order to optimize the overall solution output. Figures 4 and 5 show briefly the steps involved in creating two configured training and testing database.

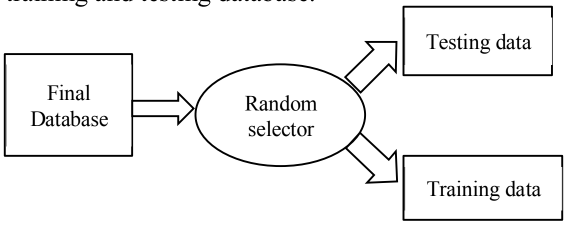

Fig. 4 Data generation scheme

After setting up the number of testing samples N, random selector is used to select N matches from the final database that contains total of 760 matches from premier league. Test dataset with 'N' samples and Training data with '760 – N' records were generated. One point worthwhile to be emphasis here, in premier league every team meet every other team twice during whole season. Therefore, if random selector chooses one match, it will not take the second meeting which will be used in the prediction system later on.

For predicting any match, we suggested to take both overall performances of both teams playing together with the previous meeting between them. Therefore, overall performance of a team is to be calculated by averaging all its records in the training dataset, assuming that test matches are unknown. Additionally, previous match is obtained from training dataset as well. Herein, a weighting parameter is used to control the contribution of seasonal performance and the teams' last match performance as shown in equation (1) which is applied for each team individually.

$$Team\ features = \omega \times team_{season} + (1-\omega) \times team_{previous\ match} \quad (1)$$

Where, ω is a weighting coefficient,
$team_{season}$ is the seasonal performance features,
$team_{previous\ match}$ is the previous match performance features.

By having a certain weight for seasonal performance from the training data (as shown in Fig. 5), some kind of static data about the team, related to its ranking, is added to testing data sample, while the dynamic side of soccer is added by considered previous match, where a team A, e.g., performs better against team B in particular even though team A is weaker or lower than team B in the ranking table.

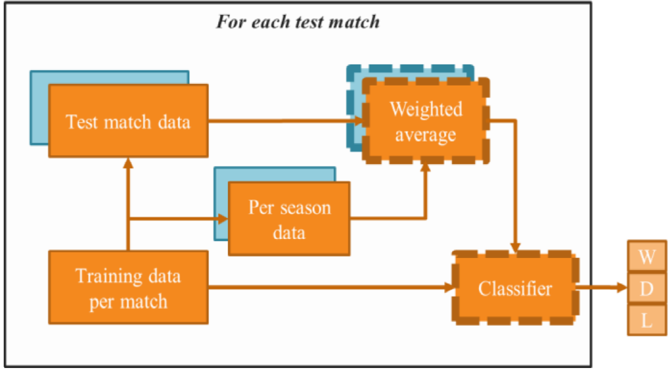

Fig. 5 Proposed frame without RBR

C. *Classification*

Both training and testing (Weighted average) datasets are ready to be classified based on the 20 features; 10 for each team. Two classifiers were coded and employed to do the classification; weighted-KNN and NBC. These two classifiers were chosen due to nature of features property of being numerical type. Both classifiers perform pretty well with this type of data as well as they are based on Bayesian framework, which is well known to such uncertainty and stochastic environment such as soccer world. For the KNN, both Inverse weighting scheme and Gaussian weighting scheme have been implemented. Employing the weighted distances among the instances was an effective approach to get the most accurate prediction outcomes.

D. *Rule based reasoner*

The key of success in the proposed framework is the correct functioning of the rule based reasoner. As it simulates the decisions carried out by the team's coach, it could introduce to the classifier some aspects which are unattainable without this approach. In order to apply the rules, the feature values are normalized with respect to the whole dataset by the following rule:

$$feature_{normalized} = \frac{feature-\mu(feature)}{\sigma(feature)} \qquad (2)$$

The rules are applied for both teams simultaneously, as a further improvement, different rule sets can be defined for each team. A rule is defined in if – then –else manner, which makes it possible to constraint each rule according to a predefined conditions.

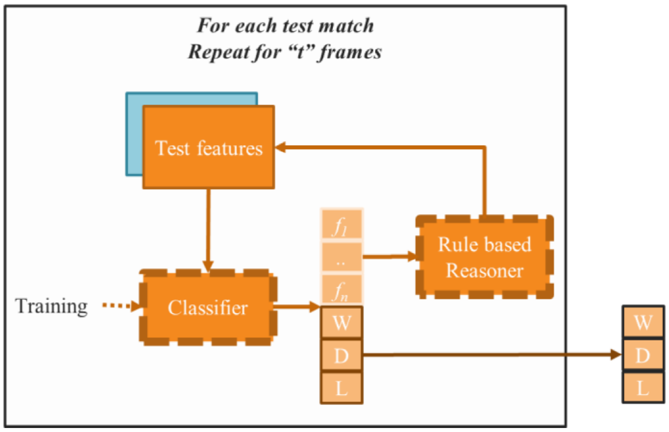

Fig. 6 Proposed frame with RBR

The RBR takes the features in every time frame and applies the predefined rules to update them for the next time frame. These rules were defined using both attack and defense features were each rule update its relevant feature as the following example:

```
If teamA (fSD) < teamB (fIN)
    Then
            TeamA attack should decrease
End
```

Where, SD is successful dribbles,
IN represents interceptions.

As in the provided example, equal number of rules were set for attach and defense were each rule is applied probabilistically to handle factors like randomness, morale, and team condition.

V. Results

A. Single-Time-Frame results

Results were carried on in two stages. First stage was done without rule based reasoner were the code was executed for one time frame. Two tests were evaluated for three classifiers; inversely weighted KNN (KNN_1), Gaussian weighted KNN (KNN_2), and Naïve Bayes Classifier (NBC) which were placed as shown in Fig. 2. On each test case, the total number of features was switched from 10 to 20.

As given in table II, the lower the number of features the higher the accuracy in all classifiers which is expected due to the effect of curse of dimensionality.

TABLE II
Single Time Frame Results

# features	KNN_1	KNN_2	NBC
10	43.3%	50%	46.7%
20	32%	37.5%	50%

B. Multi-Time-Frame results

Multi time frame testing was done by including the rule based reasoner for three iterations as shown in Fig. 2, RBR improved the classifiers accuracy as given in Table III and Table IV. The improvement was significant for the KNN classifiers where it has reached to 63.3% from 43.3% (without RBR) for KNN_1 in first test case while the improvement was lower for the NBC classifier since its accuracy increased from 46.7% to 53.3%.

TABLE III
Multi Time Frame Results for 10 Features

Iterations	KNN_1	KNN_2	NBC
1	53.3%	46.7%	50%
2	63.3%	56.7%	53.3%
3	63.3%	56.7%	53.3%

TABLE IV
Multi Time Frame Results for 20 Features

Iterations	KNN_1	KNN_2	NBC
1	53.3%	36.7%	46.7%
2	56.7%	46.7%	56.7%
3	60%	50%	56.7%

The overall results demonstrate an improvement in the classification accuracy by using RBR for all test cases and classifiers. It was found that three time frames were sufficient in providing an improvement in the classification.

VI. Conclusion

This paper showed some practical challenges faced in the complex and highly dynamic field of sports data mining. In this paper, a multi-time frame approach was combined with a rule based reasoner to produce a premier league match prediction toolbox. In all tests executed, results were improved with more or less significance which provides a statistical evidence to support the proposed hypothesis. The proposed framework has a wide range of flexibility by considering previous team's historical data and availability to define and test rules easily.

In conclusion, it is not enough to take into account the statistical data in such complex environments, but a form of reasoning and the, sometimes decisive, factor of randomness is also required which results in a more realistic model.

VII. Group Contribution

Each member of the group was involved in all meetings and discussions and the framework was design, coded, and tested by distributing the tasks among team members. Each member contributed also in the presentation and writing.

REFERENCS

[1] O. K. Solieman, "Data Mining in Sports: A Research Overview." 2006, [Online]. Available: http://ai.arizona.edu/mis480/syllabus/6_Osama-DM_in_Sports.pdf.

[2] B. Min, J. Kim, C. Choe, H. Eom, and R. I. McKay, "A compound framework for sports results prediction: A football case study," *Knowledge-Based Systems,* vol. 21(7), p. 551-562, 2008.

[3] J. Hucaljuk, and A. Hucaljuk, “Predicting football scores using machine learning techniques” *MIPRO, 2011 Proceedings of the 34th International Convention,* p. 1623, 2011.

[4] MCFC Analytics society, http://www.mcfc.co.uk/Home/The%20Club/MCFC%20Analytics.

[5] B. Min, J. Kim, C. Choe, H. Eom, and R. I. McKay, "A compound framework for sports results prediction: A football case study," *Knowledge-Based Systems,* vol. 21(7), p. 551-562, 2008.