CIS602 - Data Mining Project Proposal

English Premier League Soccer Results Prediction

Minh Nguyen - Hai Ha - Tu Luong

---

1. Overview

In this project, we will use data mining algorithms in combination with machine learning to predict results of soccer matches in English Premier League (EPL). Although it is difficult to consider all features that effect the results of the matches, we will find the most significant features which are the most effective to the results (e.g. team current form, team players index, head-to-head history, etc.) and different classifiers are tested to attempt solving the problem.

Our basic framework should be as following steps:

1. Choosing matches set data, normalizing and plotting data points;
2. Analyzing and selecting key features;
3. Data and features extraction;
4. Testing on various mining algorithms;
5. Implementing the most accurate algorithm;
6. Improving implemented algorithm.

Datasets source:

1. https://www.premierleague.com/
2. https://www.fifaindex.com/
3. https://github.com/footballcsv/eng-england

For experiments, we would like to learn data from previous 5 seasons before the current season we want to predict. We will get as much information from each season as possible, and select key features based on the results of that season. Then we predict the each match results, knowing that we already have season schedule and exclude all unexpected events such as match delay, player injury, etc. Finally, we compare the predicted results to the results we already know, with various algorithms, improving the algorithm by add more key features if possible.

2. Intelligent Merit

The aim of this project is to find an algorithm based on learning past data, to predict future matches in EPL with a reasonable precision. As the problem is one of classification problems, instead of the Bayesian dynamic model is built [1], but there are more than its parameters we think need to be considered in order to get better results and furthermore we can apply the model in others data set. Therefore, selecting features to analyze the data is the key point. Firstly, we must choose features which have the most significant in the results of the match, for example, in a particular match, features such as team's players' index, team's current formation, playing in home or away stadium will be significant impact to the final results. Secondly, extracting from data all the features we chose, then

using machine learning methods, such as KNN, SVM, and others to produce the output representative of the probable outcome of the match.
Base on testing results, we can compare and probably improve implemented algorithms by changing key features..

3. Broader Impacts

We apply this approach to the prediction of soccer matches but also expect that it may have a broader applicability for many different kinds of sports. The latter refers to the potential of the proposed project to benefit entertainment area. Some people aim to give the public investors access to the traditional private world of sports prediction, since there is a fund that concentrates on sports betting should therefore be an attractive investment as some investors like sport, and like to have a bet. Gambling is illigal but it could generate more tax. If the bets were legal, the tax revenue would be enormous.

By using the prediction, we could turn soccer community into more interesting field. Including players who can look at the estimated result for a better preparation as well as their leader or managers to recognize the power of sport, not only by their knowledge, but by their ability to bring out the best in every member of their teams. Moreover, critic can have a better review before and after the match. For the recruitment, the project can be used as a simulation to check if the player is suitable for the current team. Cheating in a game is also detected if a player play too good as it could be a sign of drug usage.

4. References

[1] Rue, Havard, and Oyvind Salvesen, “Prediction and retrospective analysis of soccer matches in a league” Journal of the Royal Statistical Society: Series D (The Statistician) 49.3 (2000): 399-418.