Saint Petersburg State University

Mathematics and Mechanics Faculty

Department of Analytical Information Systems

Malysh Konstantin

Data mining and football results prediction

Term paper

Scientific Adviser: Alexander Igoshkin, Yandex Mobile Department

Saint Petersburg, 2014

1 INTRODUCTION

Main goal of this research is a highly precise prediction of football matches' results. It is to be done by choosing the most influential data about playing teams and looking through a big number of machine learning methods to produce as high prediction rate as possible.

2 PROBLEM STATEMENT

The problem is to create a model that would be able to predict football matches' results with not less than 70% precision and finding out which features do really affect an outcome.

3 APPROACH

The current work is divided into two parts: data extracting and applying machine learning algorithms to it.

3.1 DATA MINING

3.1.1 Choosing a tournament

There are many issues connected with different leagues.

Firstly, some of them have problems with match fixing. For example, in 2006 Italian side Juventus from Turin was relegated to Serie B (Italian second tier) straight after gaining gold medals of 2005-06; Milan, Reggina, Fiorentina and Lazio were deducted from 3 to 15 points due to a huge corruption scandal.

Secondly, as the author of this work can speak only English and Russian, there can be some problems connected with extracting data in French, German etc.

Finally, there can be situations when looking at some number of matches is necessary to get more information, so translations must be available and not so boring.

Finally, the choice fell on the Premier League, top English professional league.

3.1.2 Choosing data set

In this section, features of our data set will be described, and formulas will be given.

Let us assume that 2 points are given for a victory, 1 for a draw and 0 for a defeat.

- Form of the team (one for each of two playing teams): average team's result in last five games: $Form = \frac{1}{10}\sum_{k=1}^{5} result(5)$, where $result(i)$ is a result of i-th last game.
- Team's concentration(one for each of two playing teams), showing its ability to concentrate in matches against weaker teams: $Concentration = \frac{x-1}{5}$, where x equals the number of last game with a loss against a team, which is placed 7 or more places lower (if $x \geq 7$, then $Concentration = 1$)

- Team's motivation (one for each of two playing teams), showing its passion for winning this particular match:

  Assume $Dist$ is a number of points behind the closest "critical point", where the "critical points" are:

  - 1st place
  - 2nd place
  - 3rd place
  - 4th place (making a team qualified for the UEFA Champions League)
  - 5th place (making a team qualified for the UEFA Europa League)
  - 6th place (not giving a certainty in qualifying for any of UEFA competitions)
  - 17th place (being not relegated)
  - 18th place (being relegated);

  and $ToursLeft$ as a number of matches left for a team to play in a current season.

  $Motivation = 1$, if the match is a derby (for example, Liverpool vs Everton), else

  $Motivation = 1$, if a number of matches played is more than 32 and $Dist \leq 3$, else

  if $1 - \frac{Dist}{3*ToursLeft} < 0$ or $1 - \frac{Dist}{3*ToursLeft} > 1$, then $Motivation = 0$, else

  $Motivation = 1 - \frac{Dist}{3*ToursLeft}$;

- Goals difference (one for each match): $GoalsDif = \frac{1}{2} + \frac{goalsdifference}{2*maxgoalsdifference}$, Where $goalsdifference$ is a difference of differences of scored and conceded goals and $maxgoalsdifference$ is a maximum of such value from all of competing teams.
- Points difference (one for each match): $PointsDif = \frac{1}{2} + \frac{pointsdifference}{2*maxpointdifference}$, where $pointsdifference$ is a difference of points between two playing teams and $maxpointdifference$ is a difference of points between 1st and last placed teams.
- History of this meeting: $History = \frac{p_1+p_2}{4}$, where p_i is a result of i-th last meeting of these teams(for a home side)

3.1.3 Extracting data

After looking through many of British and Russian football-associated websites, two of them were chosen as the most informative and simple to parse:

a) www.championat.com (in Russian)

b) www.statto.com (in English)

After working with these websites, scripts were written, which gave us a full data set for seasons 2012/13 and 2013/14.

3.2 MACHINE LEARNING

After getting a data set, different machine learning algorithms were applied:

- SVM with linear kernel;
- SVM with RBF kernel;
- Logistic regression.

All these methods gave nearly the same precision, which equals 0.54.

4 FURTHER RESEARCH

As it can be seen, the precision is not high enough to satisfy with itself. The directions of further work and investigations are:

- Changing and improving data set.

  It is obviously necessary to add more features, talking about players inside of teams; also some more psychological values (tiredness, changes in squads);

- Trying more machine learning algorithms;
- Going deeper through the years: extracting more data from more seasons;
- Going wider: looking at 5 best European leagues (England, Italy, Spain, Germany, France), applying same methods, looking at precisions for each team depending on its position/name/etc.
- Going to wealth: adding bookmakers odds, trying to increase the amount of won money, punishing machine for predicting a result wrong.

5 RELATED WORK

Some other attempts of coming close to a highly precisive outcome were made by different researchers.

2006 reserarch by Babak Hamadani of Stanford University was based on three seasons between 2003 and 2005. As a result, it was shown that each of the seasons had its own set, which leads us to a question of getting a right data set or productivity of our research.

2012 study by Jack David Blundell of University of Leeds has shown that there is a possible accurate model: a regression one. Also this research has given an interesting result of simple models being nearly as accurate as complex ones.

However, these and other studies were all about American football, talking about only major competition: NFL regular season. In future this study will have to take a look not only at the Premier League season, but national cup(FA cup), League cup, and last, but not least, European competitions such as UEFA Champions League and UEFA Europa League, so an outcome can be either more accurate or less accurate.

6 CONCLUSION

This research shows that machine learning also can be used in sports predictions: as it was shown, with a small data set we got a precision of 54%, which is not that bad at this point.

7 REFERENCES

[1] A. Joseph, N.E. Fenton, M. Neil. “Predicting football results using Bayesian nets and other machine learning techniques”, *Computer Science Department, Queen Mary, University of London, UK,* 2006

[2] J. Warner “Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line”, 2010

[3] J.D. Blundell “Numerical Algorithms for Predicting Sports Results”, *School of Computing, Faculty of Engineering, University of Leeds,* 2006

[4] B. Hamadani “Predicting the outcome of NFL games using machine learning”, *Stanford University,* 2005