Predicting Soccer Match Outcomes Using a Naive Bayesian Network

Owen Chapman **Patrick Shao** and **Huu Le**
Pomona College

Abstract

Predicting the outcomes of sports games is an absorbing interest for many people, from casual fans to analysts and statisticians to the sports franchises themselves. Accurate prediction of results involves a good model of sports teams relative strength, which can then be used to predict which team will win a given match with what probability. This model can be used by organizers to create fair tournament brackets, by analysts to determine rankings for a league, and by betting markets to determine the odds for a game. Applications such as these make sports modeling and prediction a multimillion-dollar industry.

Introduction

We will implement a machine that estimates the probable outcomes of soccer matches when given historical data on the two playing teams past results. The input data consists of a large database of match information, with each match containing the two teams names, which team played at home, and the score of the game. Because data more detailed than this is difficult to obtain for free, our machine will make predictions based only on these relatively easy-to-find data. The machine will then predict which team will win with what probability. We use match data from the English Premier League over the past fifteen years as our dataset.

Systems Description

To generate an accurate model of a soccer team, we first needed enough data for a machine to plausibly construct an accurate model of a teams relative strength. To do this, we turned to the Internet. The English Premier League has well-documented and extensive English-language match data available for free from Wikipedia.org and rsssf.com, which we downloaded and processed to create a uniform database of EPL match results from the past fifteen years. To read this database, we also wrote a basic API to compile match data according to team and to enable easy access to useful subsets of the data using function calls. Using this API, we are running several different algorithms to evaluate matches. We are starting with a simple Nave Bayes classifier, which uses calculated feature inputs from the previous season to predict match outcomes in the next season. We estimate the probability of the label as the number of results with that label in the previous season out of the total games played during that season. We likewise calculate features such as the probability of playing a given opposing team given the result using the previous years data. We plan to evaluate other models of increasing complexity using a validation set to test their relative accuracies. We hope to implement a series of perceptron classifiers to generate match predictions, as well as a hidden Markov model that reflects the results of recent matches when predicting the next match.

Results

When run on a testing data set, our models will generate match predictions that can be compared to the actual match results. Our results for different algorithms can be compared with each other to determine which approach classifies the most labels correctly. We can also compare these results to external benchmarks, such as a random classifier or the predictions of human sports analysts. A low bar for success would be improvement upon the random classifier, and a highly successful implementation would be competitive with pundit predictions.

References

Constantinou, A. C.; Fenton, N. E.; and Neil, M. 2013. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using bayesian networks. *Knowledge-Based Systems* 50:60–86.

McCabe, A., and Trevathan, J. 2008. Artificial intelligence in sports prediction. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on*, 1194–1197. IEEE.

Rue, H., and Salvesen, O. 2000. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)* 49(3):399–418.