

1. Introduction

2. Data Collection and Processing

3. Data Analysis Approach

The first method that we used to analyze the dataset obtained by above procedure was Weka. We used various algorithms that were identified to be promising during our preliminary investigations. Specifically, those included three tree algorithms, three Bayesian algorithms, and two logistic regression algorithms.

Complementing the Weka, we decided to create our own machine learning implementation. Specifically, we wrote a [Python script](#) implementing a neural network with one hidden layer and three softmax output units each corresponding to the probability of the input data belonging to each class. The algorithm also uses 10-fold cross-validation with random partitioning of the dataset into training/validation sets, intentionally mirroring the Weka's approach.

Finally, we sought to investigate the relative predictive power of each feature. This is not only interesting in its own right (we might be able to determine which skills are most relevant for soccer), it can also be a helpful aid for guiding machine learning implementation, in particular serving as a sanity-check for the approach. In order to measure this quantitatively, we wrote a [Python script](#) that calculates, for each attribute, the maximum information gain that can be achieved by making a split along that attribute (cf. decision trees).

4. Results

Position-averaged, categories.

Algorithm	Premier League	Bundesliga	La Liga
ZeroR	0.4801	0.4896	0.4878
J48	0.4709	0.3889	0.4177
REPTree	0.4618	0.4514	0.5213
RandomForest	0.5505	0.4896	0.4787
NaiveBayes	0.3700	0.3299	0.2927
BayesNet	0.4954	0.4896	0.4360
NaiveBayesMultinomial	0.4037	0.3125	0.4024
Logistic	0.4771	0.4479	0.4390
SimpleLogistic	0.5260	0.5000	0.5183

Table 1: 10-fold cross-validation accuracy for 2016-2017 season for three different leagues. The highest accuracy for each league has been shaded in light red.

5. Discussion

6. Conclusion