Machine Learning with Soccer Game Prediction

Jiani Lin Yi Ou

October 28, 2015

1 We have done

1. **Collecting dataset**
   We found a data bataset including fulltime and halftime results for most European league from 1993/94, which contains match detail such as the number of shots, shots on target, the number of corners, fouls, yellow and red cards referees and betting odds for each matches. We have collected match data from five seasons in the England Premier League.

2. **Clustering the teams**
   There are varieties of style of those teams, in soccer games and one certain type of teams may have more chance to win times in another types. So we adopt **K-means** algorithm and utilize number of shots(A), shots on target(B), the number of corners(C), fouls(D) and yellow cards(E) as feature to cluster all the teams.
   Depend on the number of teams, which is 20. We initially set the $k = 5$

	C1	C2	C3	C4	C5
A	465	0	608	412	495
B	137	0	215	133	163
C	490	0	397	414	433
D	204	0	237	169	217
E	26	0	31	33	30

Graph.1 the center of each cluster

C1	C2	C3	C4	C5
Stoke Crystal-Palace Leicester	Norwich Watford Bournemouth	Man-City Chelsea Arsenal Liverpool	Sunderland Swansea West-Brom Aston-Villa	Tottenham Everton Southampton Newcastle West-Ham Man-United

Graph.2 the teams of each cluster

As the algorithm indicates, Cluster 3 shows the top teams, highest shots and the number of corner indicates their attack power, Cluster 2 are all 0, because they just be upgrade to this level, Cluster 1 shows the tough teams, because they have committed the highest number of fouls. Cluster 4 have the lowest number of every feature, which means the are the weak teams. The rest cluster 5 indicates the teams are average teams, because they have the average offense and defense.

3. **Predicting according to betting odds**
   As the gambling company will provide betting odds before the match, we expect to take betting odds as a sort of feature. We have tried to predict the result only by betting odds from different gambling company, using logistic regression:

   $$h_\theta(z) = \frac{1}{1+e^{-}z} \qquad \text{where} \quad z = \theta^T x \tag{1}$$

   if $h(\theta) \geq 0.5$, predict true, else predict false. We can find optimal theta by minimize the cost function:

   $$\mathrm{J}(h_\theta, y) = -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x)) \tag{2}$$

   where, y means the label of data. We train this algorithm on season 2014, and test season 2015. The accuracy of correctly predicting if the home team can win is 67.63%. Then by using multiclass classification, we got the accuracy of predicting result (home win, away win and draw) with 47.37%.

2 Plan to do

1. we expect to use neural network to do our final prediction. What we have done (clustering the team, predicting according to betting odds) can be neurons in hidden layer. Next, we will try to extract more features to form more neurons in our hidden layer.

   (a) team information
   We will consider the recently status of teams in both side by extracting features in latest k games for two teams respectively and take it into account if these two teams are derby.

   (b) external factors
   Some external factors may be considered, like referee and the weather in that day. We will looking for the features in this category carefully to find some features affect the game indeed.

   (c) others
   There are some factors which may affect the result significantly, such as illness of the players. We will identify more attributes in this type and try to form a neurons in the hidden layer.

2. We will use cross validation to avoid over fitting and bias and may use computational learning theory to proof the number of examples we need to make our learning good enough.

3. We will try more algorithm, like using SVMs instead of logic regression, to find the algorithm with best performance.

4. We may combine betting odds to find a investment policy for the England Premier League, like which kind of games you should invest, how to invest, how much money should we invest. And we will put real money in it to test the difference between err_S and err_D.

3 Reference

[1] $http://www.math.spbu.ru/SD_AIS/documents/2014-12-341/2014-12-tw-15.pdf$
[2] $http://homes.cs.washington.edu/\ jasnyder/papers/thesis.pdf$