Feature Extraction and Aggregation for Predicting the Euro 2016

Maryam Tavakol
Hamid Zafartavanaelmi, and Ulf Brefeld

Riva del Garda, Sep 19, 2016
Agenda

• Introduction
• Feature Extraction
• Prediction & Learning
• Performance Analysis
• Summary
Introduction
Feature Extraction

• Based on available data from the past tournaments

• General country data

  • FIFA ranking, FIFA points, UEFA ranking, etc.

  • Normalising features using \textit{min} and \textit{max} rescaling —keep the order
Feature Extraction

• Player specific data
  • Market value, age, num of matches/goals, etc.
  • Obtaining the current squads
  • Goal/play ratio —host advantage for France
  • Averaging for all players of a team
  • Normalising features using \textit{min} and \textit{max} rescaling
Add a New Feature
Club Division

Juventus
Club rank = 2

<table>
<thead>
<tr>
<th>Player</th>
<th>Position</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gianluigi Buffon</td>
<td>1</td>
<td>Goalkeeper, highly likely to start next game.</td>
</tr>
<tr>
<td>Andrea Barzagli</td>
<td>15</td>
<td>Defender, highly likely to start next game.</td>
</tr>
<tr>
<td>Giorgio Chiellini</td>
<td>3</td>
<td>Defender, highly likely to start next game.</td>
</tr>
<tr>
<td>Leonardo Bonucci</td>
<td>19</td>
<td>Defender, highly likely to start next game.</td>
</tr>
<tr>
<td>Stefano Sturaro</td>
<td>14</td>
<td>Midfield, highly likely to start next game.</td>
</tr>
<tr>
<td>Simone Zaza</td>
<td>7</td>
<td>Forward, highly likely to start next game.</td>
</tr>
</tbody>
</table>

Lazio
Club rank = 212

<table>
<thead>
<tr>
<th>Player</th>
<th>Position</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federico Marchetti</td>
<td>13</td>
<td>Goalkeeper, in contention to start next game.</td>
</tr>
<tr>
<td>Marco Parolo</td>
<td>18</td>
<td>Midfield, highly likely to start next game.</td>
</tr>
<tr>
<td>Antonio Candreva</td>
<td>6</td>
<td>Midfield, highly likely to start next game.</td>
</tr>
<tr>
<td>Did not train.</td>
<td></td>
<td>Unlikely to start next game.</td>
</tr>
</tbody>
</table>
Team-Club Harmony

<table>
<thead>
<tr>
<th>Country</th>
<th>Num of Players</th>
<th>Club</th>
<th>Club Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>5</td>
<td>Barcelona</td>
<td>1</td>
</tr>
<tr>
<td>Italy</td>
<td>6</td>
<td>Juventus</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>2</td>
<td>Juventus</td>
<td>2</td>
</tr>
<tr>
<td>Germany</td>
<td>5</td>
<td>Bayern Munich</td>
<td>4</td>
</tr>
<tr>
<td>Belgium</td>
<td>3</td>
<td>Liverpool</td>
<td>42</td>
</tr>
<tr>
<td>Poland</td>
<td>3</td>
<td>Legia</td>
<td>52</td>
</tr>
<tr>
<td>Portugal</td>
<td>4</td>
<td>Sporting CP</td>
<td>179</td>
</tr>
<tr>
<td>Wales</td>
<td>3</td>
<td>Crystal Palace</td>
<td>0*</td>
</tr>
<tr>
<td>Iceland</td>
<td>2</td>
<td>Hammarby</td>
<td>0*</td>
</tr>
</tbody>
</table>

(Normalised Club rank) x (num of players)
Prediction

• A score per country is defined as a weighted sum of features, i.e., linear function

\[ s_i = \theta_i^\top x_i \]

• The probabilities are computed based on obtained scores
Prediction

if $s_i \geq s_j$:

$P_{wi} = \frac{s_i}{(s_i+s_j)}$

$P_{wj} = (1 - P_{wi}) \times s_j = P_{li}$

else:

$P_{wj} = \frac{s_j}{(s_i+s_j)}$

$P_{wi} = (1 - P_{wj}) \times s_i = P_{lj}$

$P_d = 1 - P_{wi} - P_{wj}$

Win probability for team $i$

Lose probability for team $j$

Probability of draw
Learning

• Capture the outcome probabilities from the head to head record of pair of countries

  • **Germany** vs. **France**: 27 times

  • 10 win for **Germany**, 12 for **France** and 5 draw

\[
p_{w_G} = \frac{10}{27}, \quad p_{w_F} = \frac{12}{27}, \quad p_d = \frac{5}{27}
\]
Learning

• Converting probabilities to scores

• Obtaining parameters from the closed form solution of ridge regression problem

\[ \hat{\theta} = (X^\top X + I)^{-1} X^\top \hat{s} \]
Performance Analysis

• Compare prediction results to actual tournament outcome
  • Until Quarter-Final (QF)

• Evaluation by multi class logarithmic loss

\[
Logloss = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \cdot \log(p_{ij})
\]
Overall Performance

• Error of prediction for 45 matches before QF
  • Average error: 1.3187
Insufficient Data

- Relation of performance with amount of historical data

![Graph showing num of historical data](image)

![Graph showing error per country](image)
Sufficient Data

- Reduction of error from \textbf{1.3187} to \textbf{1.1129} for teams with more than 4 historical records
Role of Past Euros

• Eliminating teams with less than 2 appearance in past Euro cups, error: 0.9680
Baseline

• Compare to a simple baseline (based on FIFA ranking only)
Summary

• Collecting data
• Feature extracting/cleaning
• New feature: team-club harmony
• Learn a linear model
• Effect of historical data on the performance
Thanks for your attention

Questions?

Email: tavakol@leuphana.de