Predicting the potential of professional soccer players

Ruben Vroonen
Tom Decroos
Jan Van Haaren
Jesse Davis

MLSA17 @ ECML/PKDD17
18/09/2017
Predicting the potential of professional soccer players

Ruben Vroonen
Tom Decroos
Jan Van Haaren
Jesse Davis

MLSA17 @ ECML/PKDD17
18/09/2017
Meet Bob, a young professional soccer player

Bob
Age: 19
Year: 2017
Bob has a set of skill ratings

Bob
Age: 19
Year: 2017

Attacking: 75/100
Defending: 67/100
Stamina: 50/100
Intelligence: 72/100
Meet Bob from the future

Bob
Age: 19
Year: 2017

Attack: 75/100
Defend: 67/100
Stamina: 50/100
Intelligence: 72/100

Bob
Age: 21
Year: 2019
What are his skill ratings?

Bob
Age: 19
Year: 2017

Attacking: 75/100
Defending: 67/100
Stamina: 50/100
Intelligence: 72/100

Bob
Age: 21
Year: 2019

Attacking: ?/100
Defending: ?/100
Stamina: ?/100
Intelligence: ?/100
Overview

Related Work
PECOTA and CARMELO

Data
SoFIFA.com ratings

APROPOS
Our approach for predicting players’ potential

Experiments
Evaluating the predictive accuracy
Overview

Related Work
PECOTA and CARMELO

Data
SoFIFA.com ratings

APROPOS
Our approach for predicting players’ potential

Experiments
Evaluating the predictive accuracy
Similar systems have already been deployed in baseball (1) and basketball (2)

(1) PECOTA

*Player Empirical Comparison Analysis Test Algorithm*

Nearest neighbors analysis on player statistics using Bill James’s similarity scores

(2) CARMELO

*Career-Arc Regression Model Estimator with Local Optimization*

Nearest neighbors analysis on Wins Above Replacement (WAR) using simple similarity score
Overview

Related Work
PECOTA and CARMELO

Data
SoFIFA.com ratings

APROPPOS
Our approach for predicting players’ potential

Experiments
Evaluating the predictive accuracy
A player card from SoFIFA.com contains 24 skill ratings for a specific player and age.
The data

Competitions:
England, France, Germany, Italy and Spain

Stats:
• 10,000 players
• 57,000 player cards
• Data from 2007-2017

Preprocessing challenges:
• Incorrect or missing age
• Position of substitute players
The most interesting categories (young and old players) have the least available data.
Most skill ratings follow a normal distribution...
... except goalkeeping skills
Overview

Related Work
PECOTA and CARMELO

Data
SoFIFA.com ratings

APROPOS
Our approach for predicting players’ potential

Experiments
Evaluating the predictive accuracy
Reminder: our task is to predict the skill ratings of future Bob

Bob
Age: 19
Year: 2017
Attacking: 75/100
Defending: 67/100
Stamina: 50/100
Intelligence: 72/100

Bob
Age: 21
Year: 2019
Attacking: ?/100
Defending: ?/100
Stamina: ?/100
Intelligence: ?/100
APROPOS follows a nearest neighbors approach

Given:

- a player $p$ and his current age $a_1$
- a future age $a_2$
- a database of players $D$

Then:

1. Search players in $D$ that are similar to $p$ at age $a_1$ and have data available for age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the ratings of similar players at age $a_2$. 
APROPOS follows a nearest neighbors approach

Given:

• a player $p$ and his current age $a_1$
• a future age $a_2$
• a database of players $D$

Then:

1. Search players in $D$ that are similar to $p$ at age $a_1$ and have data available for age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the ratings of similar players at age $a_2$. 
APROPOS follows a nearest neighbors approach

Given:
- a player \( p \) and his current age \( a_1 \)
- a future age \( a_2 \)
- a database of players \( D \)

Then:
1. Search players in \( D \) that are similar to \( p \) at age \( a_1 \) and have data available for age \( a_2 \).
2. Predict the rating of \( p \) at age \( a_2 \) by combining the ratings of similar players at age \( a_2 \).
APPROPOS follows a nearest neighbors approach

Given:
- a player $p$ and his current age $a_1$
- a future age $a_2$
- a database of players $D$

Then:

1. Search players in $D$ that are similar to $p$ at age $a_1$ and have data available for age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the ratings of similar players at age $a_2$. 

Similarity score
APPROPOS follows a nearest neighbors approach

Given:
- a player $p$ and his current age $a_1$
- a future age $a_2$
- a database of players $D$

Then:
1. Search players in $D$ that are similar to $p$ at age $a_1$, and have data at age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the ratings of similar players at age $a_2$. 
Given:

- a player $p$ and his current age $a_1$
- a future age $a_2$
- a database of players $D$

Then:

1. Search players in $D$ that are similar to $p$ at age $a_1$ and have data available for age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the ratings of similar players at age $a_2$.

APROPOS follows a nearest neighbors approach.
The **absolute** similarity score expresses the difference between skill ratings.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>81</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>85</td>
</tr>
</tbody>
</table>
The **absolute** similarity score expresses the difference between skill **ratings**

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>72</td>
</tr>
</tbody>
</table>

\[
f(Bob, Alice) = \sqrt{ (68 - 81)^2 + (72 - 81)^2 + (78 - 85)^2 }\]
The **absolute** similarity score expresses the difference between skill **ratings**

<table>
<thead>
<tr>
<th>Age</th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>72</td>
</tr>
</tbody>
</table>

\[
f(\text{Bob}, \text{Alice}) = \sqrt{(68 - 81)^2}
\]
The **absolute** similarity score expresses the difference between skill **ratings**

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Age</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Age</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>81</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>72</td>
<td>82</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>78</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
f(\text{Bob, Alice}) = \sqrt{(68 - 81)^2 + (72 - 81)^2}\]
The **absolute** similarity score expresses the difference between skill **ratings**

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th></th>
<th>Alice</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>72</td>
<td>78</td>
<td>81</td>
<td>82</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
f(\text{Bob, Alice}) = \sqrt{(68 - 81)^2 + (72 - 81)^2 + (78 - 85)^2}
\]
The **evolutional** similarity score expresses the difference between skill **evolution**

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th></th>
<th></th>
<th></th>
<th>Bob</th>
<th></th>
<th></th>
<th></th>
<th>Bob</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td></td>
<td>17</td>
<td>18</td>
<td>19</td>
<td></td>
<td>17</td>
<td>18</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>72</td>
<td>78</td>
<td></td>
<td>81</td>
<td>82</td>
<td>85</td>
<td></td>
<td>81</td>
<td>82</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>
The **evolutional** similarity score expresses the difference between skill evolution.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>+4</td>
<td>+6</td>
</tr>
</tbody>
</table>
The **evolutional** similarity score expresses the difference between skill evolution

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Age</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling score</td>
<td>68 $\rightarrow$ 72 $\rightarrow$ 78</td>
<td>81 $\rightarrow$ 82 $\rightarrow$ 85</td>
</tr>
</tbody>
</table>

The evolutional similarity score is calculated as:

$$f(Bob, Alice) = \sqrt{(4 - 1)^2}$$
The **evolutional** similarity score expresses the difference between skill evolution.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th></th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td><strong>Dribbling score</strong></td>
<td>68</td>
<td>+4</td>
<td>72</td>
</tr>
</tbody>
</table>

$$f(\text{Bob}, \text{Alice}) = \sqrt{(4 - 1)^2 + (1 - 3)^2}$$
The similarity score between players is computed as the average over all skills

\[ sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|} \]
The similarity score between players is computed as the average over all skills.

$$ sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|} $$

Total similarity between 2 players
The similarity score between players is computed as the average over all skills.

\[
sim(p, p') = \frac{\sum_{v \in V} \text{sim}_v(p, p')}{|V|}
\]

**Total similarity between 2 players**

**Normalized similarity per skill (e.g. dribbling)**
The similarity score between players is computed as the average over all skills.

Total similarity between 2 players

\[ sim(p, p') = \frac{\sum_{v \in V} sim_v(p, p')}{|V|} \]

Normalized similarity per skill (e.g. dribbling)

Total number of skills (\(=24\))
APROPOS follows a nearest neighbors approach

Given:
- a player $p$ and his current age $a_1$
- a future age $a_2$
- a database of players $D$

Then:
1. Search players in $D$ that are similar to $p$ at age $a_1$ and have data available for age $a_2$.
2. Predict the rating of $p$ at age $a_2$ by combining the player ratings at age $a_2$.

**Similarity score**
- Absolute
- Evolutinal

**Prediction method**
- Absolute
- Evolutinal
We want to predict Bob’s dribbling rating at age 21

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
</tr>
</tbody>
</table>
Alice is a similar player to Bob for whom we have historical data

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>?</td>
<td>86</td>
</tr>
</tbody>
</table>

\[
\text{Sim}(Bob, Alice) = 0.7
\]
Eve is also a similar player to Bob for whom we have historical data

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
\text{Sim}(Bob, Alice) = 0.7 \\
\text{Sim}(Bob, Eve) = 0.8
\]
The absolute prediction method uses the skill ratings of similar players.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>64</td>
<td>75</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Sim}(Bob, Alice) = 0.7, \quad \text{Sim}(Bob, Eve) = 0.8
\]
The absolute prediction method uses the skill ratings of similar players

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
Sim(Bob, Alice) = 0.7 \\
Sim(Bob, Eve) = 0.8
\]

Dribbling prediction = ____________________
The **absolute** prediction method uses the skill **ratings** of similar players

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
Sim(Bob, Alice) = 0.7 \quad Sim(Bob, Eve) = 0.8
\]

\[
Dribbling\ prediction = \frac{0.7 \times 86 + 0.8 \times 75}{0.7 + 0.8} = 80
\]
The **evolutional** prediction method uses the skill **evolutions** of similar players

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>64</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>75</td>
</tr>
</tbody>
</table>

\[
Sim(Bob, Alice) = 0.7 \\
Sim(Bob, Eve) = 0.8
\]
The **evolutional** prediction method uses the skill **evolutions** of similar players.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
</tbody>
</table>

$Sim(Bob, Alice) = 0.7$

$Sim(Bob, Eve) = 0.8$
The **evolutional** prediction method uses the skill **evolutions** of similar players.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age</strong></td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>19</td>
<td>21</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td><strong>Dribbling</strong></td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
<tr>
<td>85</td>
<td>86</td>
<td>64</td>
<td>75</td>
</tr>
</tbody>
</table>

\[
Sim(Bob, Alice) = 0.7 \quad Sim(Bob, Eve) = 0.8
\]

**Dribbling prediction** = 78 + ____________
The **evolutional** prediction method uses the skill **evolutions** of similar players.

<table>
<thead>
<tr>
<th></th>
<th>Bob</th>
<th>Alice</th>
<th>Eve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>19</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>Dribbling</td>
<td>78</td>
<td>?</td>
<td>85</td>
</tr>
</tbody>
</table>

\[
\text{Sim}(Bob, Alice) = 0.7 \\
\text{Sim}(Bob, Eve) = 0.8
\]

\[
\text{Dribbling prediction} = 78 + \frac{0.7 \times 1 + 0.8 \times 11}{0.7 + 0.8} = 84
\]
Overview

Related Work
PECOTA and CARMELO

Data
SoFIFA.com ratings

APROPOS
Our approach for predicting players’ potential

Experiments
Evaluating the predictive accuracy
We predict skill ratings for 1000 players in 2012

Similarity period = 3 years
Prediction period = 1 year
We compare 2 baseline models against 3 instances of APROPOS

1. Baseline 1: average skill rating of age group
2. Baseline 2: current skill rating as prediction
3. ABS-ABS
4. ABS-EVO
5. EVO-EVO
APROPOS performs better than baseline 1 and roughly equal to baseline 2.
APROPOS beats Baseline 2 when predicting farther in the future
The number of years used to compute player similarity has little effect on performance.
Conclusion

Predicting the potential of professional soccer players is an interesting task.

APROPOS solves this task using a nearest neighbors approach.

The best results are obtained by combining player-specific info with population-based info.