Effects Of Country Specific Features On Summer Olympics Swimming Games Medal Results

Tunç GÜLTEKİN
Türkiye olimpiyatlarda neden madalya kazanamıyor?

Türk Eğitim Sen ve Kamu Sen Genel Başkanı İsmail KONCUK:

Ülkemizi uluslararası alanda temsil edecek yüzücü çok yetişmiyor. Oysa üç yan denizlerle çevrili bir ülkeyiz. Bu ironinin nedeni nedir?

Derya Büyükuncu: Mantıksen dediğiniz gibi olması lazım. Her tarafımız deniz, suyu sevmeliyz. Ama maalesef biz sporu sevmiyoruz.

"Üç tarafı denizlerle çevrili bir ülkenin çocukları Derya Büyükuncu dışında bir isim çıkaramyorsa bu büyük bir kayıptır”

Spor Bakanı Suat Kılıç ilk röportajını VATAN’a verdi:

Özgür BOLAT
ozgurbolat@hurriyet.com.tr
2 Ağustos 2012

Türkiye neden olimpiyatlarda başarısız?
Outline

- Some Statistics about Swimming on Olympics.
- Data Collection
- Data Integration
- Data Preparation
- Data Analysis
  - Simple Visualization
  - Clustering
  - Classification
- Conclusion
Some Information

- Michael Phelps is the most decorated Olympian of all time, with a total of 22 medals.
  (18 gold, 2 silver, 2 bronze)
• Derya Büyükuncu is the first ever Turkish swimmer to win an international competition. (1997 Mediterranean Games 3rd)

• Also he has been a Turkish national team member for more than 25 years
Some Statistics

What about countries?
### Some Statistics

<table>
<thead>
<tr>
<th>Rank</th>
<th>Nation</th>
<th>Gold</th>
<th>Silver</th>
<th>Bronze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>United States (USA)</td>
<td>230</td>
<td>164</td>
<td>126</td>
<td>520</td>
</tr>
<tr>
<td>2</td>
<td>Australia (AUS)</td>
<td>57</td>
<td>60</td>
<td>61</td>
<td>178</td>
</tr>
<tr>
<td>3</td>
<td>East Germany (GDR)</td>
<td>38</td>
<td>32</td>
<td>22</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>Hungary (HUN)</td>
<td>25</td>
<td>23</td>
<td>18</td>
<td>66</td>
</tr>
<tr>
<td>5</td>
<td>Japan (JPN)</td>
<td>20</td>
<td>24</td>
<td>29</td>
<td>73</td>
</tr>
<tr>
<td>6</td>
<td>Netherlands (NED)</td>
<td>19</td>
<td>18</td>
<td>19</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>Great Britain (GBR)</td>
<td>15</td>
<td>22</td>
<td>30</td>
<td>67</td>
</tr>
<tr>
<td>8</td>
<td>Germany (GER)</td>
<td>13</td>
<td>18</td>
<td>28</td>
<td>59</td>
</tr>
<tr>
<td>9</td>
<td>Soviet Union (URS)</td>
<td>12</td>
<td>21</td>
<td>26</td>
<td>59</td>
</tr>
<tr>
<td>10</td>
<td>China (CHN)</td>
<td>12</td>
<td>17</td>
<td>8</td>
<td>37</td>
</tr>
</tbody>
</table>
Study

- Two different data was used for this;
  - Athlete medal information related to specific olympic swimming event.
  - Country specific features (Coastline Length GNI…)

• Two different data was used for this;
Athlete Medal Information

• Contains these features:
  – Athlete Name
  – Country Name
  – Sex
  – Race Course Length
  – Race Type
  – Race Time
  – Year,
  – Medal Type

• Example: *Larsen Jensen, USA, Men, 1500, Freestyle, 2004, 14:45.3, SILVER*
Country Specific Features

- **CoastLine**: Numeric, Defines the coastline length of that country.

- **EducationIndex**: Numeric, Defines the adult literacy rate of that country. It is a value which between 0 and 1.

- **GNI**: Numeric, Defines the Gross National Income of that country.

- **Population**: Numeric, Defines the Population of that country.
Country Specific Features

- **ForestArea**: Numeric, Defines Forest Area of that country.

- **Fishing**: Numeric, Defines Fisheries Production of that country.

- **BirthRate**: Numeric, Defines Birth Rate of that country. It is a value which between 0 and 1.

- **GDP**: Numeric, Defines Gross Domestic Product of that country. It is often considered an indicator of a country's standard of living.
Country Specific Features

- **CO2Rate**: Numeric, Defines CO2 Emission (metric tones) of that country.

- **DeathRate**: Numeric, Defines Death Rate of that country. It is a value which between 0 and 1.

- **HDI**: Numeric, Defines the Human Development Index of that country.
Data Integration

• Country medal information was taken from www.databaseolympics.com

• Country specific features were taken from different pages of wikipedia.

• This method created many different tables
Data Integration

Horizontal Data Integration is Required
To achieve this goal, a .net application was developed.
Data Integration

Country & Medal Information

- Country & Birth Rate
- Country & Coastline
- Country & HDI
- Country & GDP
- Country & Population
- Country & Education
- Country & GNI
- Country & Fishing
- Country & Death Rate
Data Integration

- Data was integrated over country name

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AthletName</td>
<td>CountryName</td>
<td>Coastline</td>
<td>EducationRate</td>
<td>GNI</td>
<td>Population</td>
<td>HDI</td>
<td>GDP</td>
</tr>
<tr>
<td>2</td>
<td>Larsen Jensen</td>
<td>United States</td>
<td>133312</td>
<td>0.968</td>
<td>48890</td>
<td>314419000</td>
<td>91</td>
<td>14447100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fishing</td>
<td>BirthRate</td>
<td>CO2</td>
<td>DeathRate</td>
<td>Gender</td>
<td>RaceCourseLength</td>
<td>RaceType</td>
<td>Time</td>
<td>Year</td>
<td>MedalType</td>
</tr>
<tr>
<td>100000</td>
<td>13.7</td>
<td>5461014</td>
<td>8.2</td>
<td>Men</td>
<td>1500</td>
<td>Freestyle</td>
<td>14:45.3</td>
<td>2004</td>
<td>SILVER</td>
</tr>
</tbody>
</table>
Data Preparation & Cleaning

- Country name for some feature sets are inconsistent or misspelled.
  - Ex: USA → United States → America
    China, Ch, PRC …
  - Solved programmatically.

- Some countries do not exist anymore.
  - Ex: Soviet Union, West Germany…
  - These ones are eliminated.
Data Preparation & Cleaning

• Some country specific feature information is incomplete for some countries.
  – Ex: Fishery information below 100K Tones.
  – Assigned constant value (100K), and deleted

• Country specific informations belongs to 2000’s but medal informations cover all olympics of last century.
  – Solved by using recent years’ results (such that 1996 and later.)
Some other data preparation techniques were used before performing different analysis methods.
Three different approaches were followed to analyse
- Simple visualization of data
- Clustering
- Classification
To analyse distribution of data, Weka’s Visualization tool was used without performing any special operation.

Some results were obtained.
Simple Visualization

- Probability of getting Gold Medal is higher than other medals for the countries which have very high forest area and low CO2 emission.

x: Gold
x: Silver
x: Bronze

X-Axis: Forest Area
Y-Axis: CO2
Simple Visualization

- Probability of getting a Medal is high for the countries which have very high CO2 emission and races with women athletes.

  - x: Gold
  - x: Silver
  - x: Bronze

X-Axis: CO2
Y-Axis: Gender
Simple Visualization

- Probability of getting a Medal (especially gold) is higher than other medals for the countries which have very high forest area and races with male athletes.
  
  - **x**: Gold
  - **x**: Silver
  - **x**: Bronze

X-Axis: Forest Area
Y-Axis: Gender
Simple Visualization

- Probability of getting Silver Medal is lower than other medals for the countries which have medium HDI and medium-low population.

- **x**: Gold
- **x**: Silver
- **x**: Bronze

_X-Axis: Population_
_Y-Axis: HDI_
• Probability of getting Bronze Medal is lower than other medals for the countries which have very high population and mid-low education index.

- Gold
- Silver
- Bronze

X-Axis: Education
Y-Axis: Population
Simple Visualization

What about more general results?
Clustering

- The goal is finding natural groups, which have similar features and medal results, on the data.

- Some of the attributes are removed;
  - CountryName, RaceTiming, Year

- All data belongs to 1996 and later Olympics
• K-Means and EM algorithms were run with different cluster number parameters.
  – With 3, 5 and 10

• K-Means algorithm with 5 clusters exhibited the best performance.
%66 of 304 instances were used for training and leftovers were used for testing.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Full Data</th>
<th>Cluster #0 (48)</th>
<th>Cluster #1 (14)</th>
<th>Cluster #2 (36)</th>
<th>Cluster #3 (40)</th>
<th>Cluster #4 (62)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coastline</td>
<td>66552.555</td>
<td>63901.2708</td>
<td>63334.9286</td>
<td>133312</td>
<td>29955.8</td>
<td>54178.9839</td>
</tr>
<tr>
<td>EducationRate</td>
<td>0.9583</td>
<td>0.96</td>
<td>0.9016</td>
<td>0.968</td>
<td>0.9401</td>
<td>0.9761</td>
</tr>
<tr>
<td>GNI</td>
<td>35757.45</td>
<td>37713.3333</td>
<td>15185.7143</td>
<td>18890</td>
<td>26199.25</td>
<td>37429.6774</td>
</tr>
<tr>
<td>Population</td>
<td>182442969.66</td>
<td>243053645.0625</td>
<td>124374568.6429</td>
<td>314419000</td>
<td>195846751.35</td>
<td>63351950.871</td>
</tr>
<tr>
<td>HDI</td>
<td>0.8705</td>
<td>0.8832</td>
<td>0.7217</td>
<td>0.91</td>
<td>0.8266</td>
<td>0.9022</td>
</tr>
<tr>
<td>GDP</td>
<td>5664302.195</td>
<td>6377401.2917</td>
<td>1252518.2143</td>
<td>14447100</td>
<td>2923731.425</td>
<td>2776855.8871</td>
</tr>
<tr>
<td>ForestArea</td>
<td>1688294.975</td>
<td>1518583.375</td>
<td>492660.6429</td>
<td>3030890</td>
<td>748929.725</td>
<td>914913.7903</td>
</tr>
<tr>
<td>Fishing</td>
<td>2559371.135</td>
<td>4569233.6875</td>
<td>410797.7143</td>
<td>100000</td>
<td>5616693.2</td>
<td>944072.8065</td>
</tr>
<tr>
<td>BirthRate</td>
<td>12.508</td>
<td>12.2896</td>
<td>14</td>
<td>13.7</td>
<td>12.4275</td>
<td>11.7</td>
</tr>
<tr>
<td>CO2</td>
<td>2245048.615</td>
<td>2668110.0417</td>
<td>1025752.2857</td>
<td>5461014</td>
<td>1519502.15</td>
<td>803279.9839</td>
</tr>
<tr>
<td>DeathRate</td>
<td>9.105</td>
<td>0.3646</td>
<td>13.45</td>
<td>0.2</td>
<td>9.9025</td>
<td>0.6565</td>
</tr>
<tr>
<td>Gender</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td></td>
</tr>
<tr>
<td>RaceCourse</td>
<td>267.5</td>
<td>229.1667</td>
<td>221.4286</td>
<td>181.9444</td>
<td>220</td>
<td>387.9032</td>
</tr>
<tr>
<td>RaceType</td>
<td>Freestyle</td>
<td>Freestyle</td>
<td>Freestyle</td>
<td>Backstroke</td>
<td>Medley</td>
<td>Freestyle</td>
</tr>
<tr>
<td>MedalType</td>
<td>BRONZE</td>
<td>SILVER</td>
<td>BRONZE</td>
<td>GOLD</td>
<td>BRONZE</td>
<td>BRONZE</td>
</tr>
</tbody>
</table>
Clustering

- Are there similar countries, which have not earn any silver medal, on some events?

Almost No Cluster – 3 instance for Silver Medal
Clustering

- Are there similar countries, which have not earn any silver medal, on some events?
  - 17 instances were found over 200 instance (%16)
    - **Coastline**: [Medium – Low] Scale (Mean: 29995 Km)
    - **Education Rate**: High Scale (Mean: 0.9401)
    - **GNI**: Medium Scale (Mean: 26199 Dollars)
    - **Population**: Low Scale (Mean: 135M)
    - **GDP**: [Medium – Low] Scale (Mean: 2723931 Dollars)
    - **Forest Area**: [Medium – Low] Scale (Mean: 748929 Km²)
    - **CO2**: [Medium – Low] Scale (Mean: 1519502 Tones)
    - **HDI**: High Scale (Mean: 0.8226)
    - **Category**: Women
    - **Race Type**: All types except freestyle

South Africa, Australia Poland ...
Combining some attributes can be useful.
Clustering

- K-Means algorithm result with 5 clusters.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Full Data</th>
<th>Cluster#</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(200)</td>
<td>(54)</td>
<td>(35)</td>
<td>(58)</td>
<td>(24)</td>
<td>(29)</td>
<td></td>
</tr>
<tr>
<td>Coastline</td>
<td>66552.555</td>
<td>32609.4259</td>
<td>47267.6</td>
<td>133312</td>
<td>10217.4167</td>
<td>66135.2414</td>
<td></td>
</tr>
<tr>
<td>EducationRate</td>
<td>0.9583</td>
<td>0.9625</td>
<td>0.9573</td>
<td>0.968</td>
<td>0.9293</td>
<td>0.9566</td>
<td></td>
</tr>
<tr>
<td>GNI</td>
<td>35757.45</td>
<td>35378.3333</td>
<td>30997.1429</td>
<td>48890</td>
<td>18000.4167</td>
<td>30638.9655</td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>182442969.66</td>
<td>154210489.2778</td>
<td>86905868.6857</td>
<td>314419000</td>
<td>193982076.7917</td>
<td>76815526.3448</td>
<td></td>
</tr>
<tr>
<td>HDI</td>
<td>0.8705</td>
<td>0.8846</td>
<td>0.8601</td>
<td>0.91</td>
<td>0.7835</td>
<td>0.8497</td>
<td></td>
</tr>
<tr>
<td>GDP</td>
<td>5664302.195</td>
<td>2776866.1667</td>
<td>1512588.1143</td>
<td>14447100</td>
<td>1204850.5417</td>
<td>2176581.6897</td>
<td></td>
</tr>
<tr>
<td>ForestArea</td>
<td>1688294.975</td>
<td>737830.0926</td>
<td>1433430.8857</td>
<td>3030890</td>
<td>379717.0417</td>
<td>2163491.7241</td>
<td></td>
</tr>
<tr>
<td>Fishing</td>
<td>2559371.135</td>
<td>4480266.5185</td>
<td>1858177.3714</td>
<td>100000</td>
<td>6603660.2917</td>
<td>1400544.1379</td>
<td></td>
</tr>
<tr>
<td>BirthRate</td>
<td>12.508</td>
<td>11.5537</td>
<td>13.5</td>
<td>13.7</td>
<td>12.0875</td>
<td>11.0517</td>
<td></td>
</tr>
<tr>
<td>CO2</td>
<td>2248048.815</td>
<td>1151855.111</td>
<td>601530.8286</td>
<td>5461014</td>
<td>1075534.2083</td>
<td>820840.5172</td>
<td></td>
</tr>
<tr>
<td>DeathRate</td>
<td>9.105</td>
<td>8.4444</td>
<td>8.4286</td>
<td>8.2</td>
<td>11.1167</td>
<td>11.2966</td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td>Men</td>
<td>Women</td>
<td>Men</td>
<td></td>
</tr>
<tr>
<td>RaceType</td>
<td>Freestyle</td>
<td>Freestyle</td>
<td>Freestyle</td>
<td>Freestyle</td>
<td>Medley</td>
<td>Breaststroke</td>
<td></td>
</tr>
<tr>
<td>RaceAndMedalType</td>
<td>200_BRONZE</td>
<td>200_SILVER</td>
<td>1500_GOLD</td>
<td>200_GOLD</td>
<td>200_GOLD</td>
<td>100_BRONZE</td>
<td></td>
</tr>
</tbody>
</table>
Clustering

• Are there similar countries, which have earned medal, only on 100m and 200m events?

All instances of Cluster – 4
With in range
Clustering

- Are there similar countries, which have earned medal, only on 100m and 200m events?
  - 14 instances were found over 200 instances (%13)

  - **Coastline:** Medium Scale (Mean: 66135 Km)
  - **Education Rate:** High Scale (Mean: 0.9566)
  - **GNI:** [Medium – High] - Scale (Mean: 30638 Dolars)
  - **Population:** Low Scale (Mean: 76M)
  - **GDP:** [Medium – Low] Scale (Mean: 2176581 Dolars)
  - **Forest Area:** [Medium – Low] Scale (Mean: 2163491 Km2)
  - **CO2:** [Low] Scale (Mean: 820840Tones)
  - **HDI:** High Scale (Mean: 0.8497)
  - **Category:** Men
  - **Race Type:** All types except freestyle and medley
Clustering

Can be a model built for medal type prediction?
Preparation for Classification

• There were not any correlation with medal type attribute and other features.

• Some preprocessing operations were required.
• Z Score Standardization was used to standardize all numeric attributes to have zero mean and unit variance.
Preparation for Classification

• For better classification performance, discretization method was used.
  – Sturges' Rule
    \[ k = \lceil 1 + \log_2 n \rceil \]
  – 10 Bin is required for 304 instance.
  – 5 Bin discretization showed better performance on classification.
Preparation for Classification

- After the data preparation operations;

```
Ranked attributes:
0.0271458  1 Coastline
0.0242433  2 EducationRate
0.0218696  9 BirthRate
0.0136533  3 GNI
0.0123667  11 DeathRate
0.0092241  10 CO2
0.0075056  7 ForestArea
0.0072773  5 HDI
0.0071924  4 Population
0.0071838  6 GDP
0.002023   8 Fishing
0.0014055  15 Year
0.0006656  14 RaceType
0.0005619  13 RaceCource
0.0000259  12 Gender
```

InfoGainAttrEval With Ranker
Different classification algorithms were tested with ten folds cross validation.

- J48, Decision Stump, Random Forest, RBF Network...

The most successful one is J48
Classification

- J48 classification results:

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>120</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>184</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.0916</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.4287</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.5089</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>96.4949 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>107.9713 %</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>304</td>
</tr>
</tbody>
</table>

Success rate is too low.
Random classification rate would be 33.3333 %
Classification

- J48 classification results:

```markdown
<table>
<thead>
<tr>
<th></th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>ROC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.417</td>
<td>0.328</td>
<td>0.394</td>
<td>0.417</td>
<td>0.406</td>
<td>0.534</td>
<td>GOLD</td>
</tr>
<tr>
<td></td>
<td>0.289</td>
<td>0.324</td>
<td>0.295</td>
<td>0.289</td>
<td>0.292</td>
<td>0.46</td>
<td>SILVER</td>
</tr>
<tr>
<td></td>
<td>0.471</td>
<td>0.255</td>
<td>0.49</td>
<td>0.471</td>
<td>0.48</td>
<td>0.625</td>
<td>BRONZE</td>
</tr>
<tr>
<td>Weighted Avg.</td>
<td>0.395</td>
<td>0.302</td>
<td>0.395</td>
<td>0.395</td>
<td>0.395</td>
<td>0.542</td>
<td></td>
</tr>
</tbody>
</table>

--- Confusion Matrix ---

```
```
a  b  c  <-- classified as
43 33 27 | a = GOLD
45 28 24 | b = SILVER
21 34 49 | c = BRONZE
```
• Classification by using combined attribute; RaceLengthAndMedalType

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>67</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>237</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.1548</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.1105</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.2507</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>90.3079%</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>101.3435%</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>304</td>
</tr>
</tbody>
</table>

Success rate is too low. Random classification rate would be 6.6667%.
• Knime’s decision tree view; (first three node)
Classification

- Rapid Miner’s whole tree view (without pruning);
Conclusions

• Coastline Length is the most effective feature to decide medal type but not good enough.

• Some patterns like clusters, exist in the data.

• But they are not sufficient to create a accurate model.
Thanks!

Questions

Comments