The European Commission’s science and knowledge service

Joint Research Centre
From big data to smart data

Álvaro Gómez Losada and Néstor Duch Brown
Joint Research Centre (Seville, Spain)

NAEC Conference, 16 April 2019
Background – what do we already know?

• Knowledge Discovery in Data Bases (KDD) – Fayyad et al. (1996):
• -> **Domain knowledge** + application of empirical discovery algorithms
• Volumen of data is not equivalent to knowledge
• Expensive resources to store and analyze big amount of data
• Scientific production related to new machine learning algorithms and artificial intelligence is vast (and growing)
Three empirical use cases: machine learning for small data sets

- Three examples in which the application of machine learning is possible in small data sets:

  1. a price recommendation system.
  2. a forecasting system for ozone in cities.
  3. a forecasting system for demographic population.
Application 1: a price recommendation system

- Source of data: Amazon marketplace
- Objective: forecasting product prices
- Size of data: $5 \text{ MB}$
- Methodology: collaborative filtering for recommendation systems
- Technology: python language
- Characteristics: unequally spaced time series (events)
Application 1: a price recommendation system

Transforming a TS in the (user, item, frequency) triple emulating a Recommendation scheme

**Table 1.** Some equivalences used in this study to build the forecasting recommender system (RS) from a given time series (TS).

<table>
<thead>
<tr>
<th>Concept</th>
<th>Symbol</th>
<th>RS equivalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set of distinct values in TS</td>
<td>$\mathcal{U}$</td>
<td>Set of users with cardinality $m$</td>
</tr>
<tr>
<td>Set of distinct values in TS shifted</td>
<td>$\mathcal{I}$</td>
<td>Set of items with cardinality $n$</td>
</tr>
<tr>
<td>Two distinct values in the TS</td>
<td>$u_j, u_i$</td>
<td>A pair of users</td>
</tr>
<tr>
<td>Two distinct values of the shifted TS</td>
<td>$i_i, i_j$</td>
<td>A pair of items</td>
</tr>
<tr>
<td>Number of times a distinct value in the TS and its shifted version co-occurs</td>
<td>$r_{jk}$</td>
<td>Rating of user $u_j$ on item $i_k$</td>
</tr>
<tr>
<td>TS value to which perform a forecasting</td>
<td>$u_0$</td>
<td>Active user to which recommend an item</td>
</tr>
</tbody>
</table>
Application 1: a price recommendation system

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Length</th>
<th>min</th>
<th>P50</th>
<th>max</th>
<th>P75-P25</th>
<th>Distinct values (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS-1</td>
<td>2169</td>
<td>13</td>
<td>84</td>
<td>142</td>
<td>14</td>
<td>82</td>
</tr>
<tr>
<td>TS-2</td>
<td>1224</td>
<td>7</td>
<td>17</td>
<td>36</td>
<td>8</td>
<td>22</td>
</tr>
</tbody>
</table>

Table 2. TS characteristics used in the methodology testing (P: percentile; min: minimum value, max: maximum value; in €).

Table 3. MAE performance on both TS and the different similarity measures ($s_1$, $s_2$ and $s_3$: Pearson correlation, cosine and Otsuka-Ochiai similarities, respectively), in €.

<table>
<thead>
<tr>
<th></th>
<th>TS-1</th>
<th></th>
<th>TS-2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$s_3$</td>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$s_3$</td>
<td></td>
</tr>
<tr>
<td>MAE</td>
<td>6.2</td>
<td>6.5</td>
<td>5.5</td>
<td>6.0</td>
<td>3.2</td>
<td>3.0</td>
<td></td>
</tr>
</tbody>
</table>

Evaluation
Application 2: a forecasting system for ozone in cities

- Source of data: air quality monitoring network.
- Objective: forecasting high levels of ozone to prevent population
- Size of data: ~ 100 MB
- Methodology: supervised classification + stacking of 10 classifiers
- Technology: R language
- Characteristics: time series with marked patterns
Application 2: a forecasting system for ozone in cities

Fraction of the training data

Accuracy

7/12
Application 2: a forecasting system for ozone in cities

Estimation of Ozone (O3) using the original data or different fractions produce reliable estimations.

Madrid (Spain)
Application 3: a demographic forecasting system

- Source of data: Eurostat.
- Objective: forecasting Spanish population pyramids from 2005 to 2016
- Size of data: \(~ 10 \text{ MB}\)
- Methodology: supervised learning + random forest algorithm
- Technology: \textit{R language}
- Characteristics: marked patterns, little quantitative variations
- Dissemination: European Conference on Quality in Official Statistics (Krakow, 2018)
Application 3: a demographic forecasting system
Application 3: a demographic forecasting system

Comparison with classical techniques:
Machine learning (ML) vs Arima & exponential smoothing (ES)

Units: population (individuals)

<table>
<thead>
<tr>
<th></th>
<th>ML</th>
<th>Arima</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>20318</td>
<td>81010</td>
<td>81247</td>
</tr>
<tr>
<td>MAE</td>
<td>14353</td>
<td>60596</td>
<td>62269</td>
</tr>
</tbody>
</table>

RMSE: root mean square deviation
MAE: mean absolute error
What we have learnt

In this time of Artificial Intelligence and pervasive data:
(1) These modest examples could suggest to look back to those small data size present in organizations: value in small (high quality) data sets
(2) Research is needed to adapt machine learning algorithms to small/medium data set sizes. What amount of data do we really need?
(3) We should ascertain the knowledge data sets can provide to us
(4) Democratization of technology
(5) Probably many information still remains hidden which is not analyzed due to the size of the data is not considered suitable.
(6) Lack of data analysis / machine learning skills in public administrations?
Thanks

Questions?

You can find us at:

alvaro.gomez-losada@ec.europa.eu
nestor.duch-brown@ec.europa.eu