BigData IoT in environmental modelling

Andreja Naumoski and Miroslav Mirchev

E-mail: andreja.naumoski@finki.ukim.mk
Faculty of Computer Science and Engineering – Ss. Cyril and Methodius University in Skopje, Republic of Macedonia
Data-driven environmental modelling

- Laboratory of Eco-Informatics at FINKI, part of it does, focus on applying computer science method, namely, machine learning algorithms to solve environmental problems with the power of data computer analytics.

- Models are data-driven based on data from various sources.
  - Habitat suitability modelling of organisms in water ecosystems
  - Bio-diversity modeling.
  - Early warning flood prediction.
  - Ambient air pollution with dispersion modelling.
  - Car traffic pollution modelling.
  - Predicting the trend and factors in air pollution gases.
  - Noise pollution in cities and etc.

- Modelling only with data have to always taken with caution, because the world is presented to the method only through the data point of view.
Data-driven environmental modelling

- Most of the environmental modelling is done by using supervised (white box) machine learning algorithms that are easy interpretable.
- Example of data driven models based on single target classical decision tree for habitat suitability modelling.

Single vs Multi target modelling

- **Single target models** - Model for each variable

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>245</td>
<td>5</td>
<td>0.5</td>
<td>6.5</td>
<td>12</td>
<td>2</td>
<td>C3</td>
<td>C1</td>
</tr>
<tr>
<td>42</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>65</td>
<td>12</td>
<td>6</td>
<td>15</td>
<td>1.2</td>
<td>51</td>
<td>5.9</td>
<td>1.1</td>
<td>C2</td>
<td>C1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>12</td>
<td>2</td>
<td>5</td>
<td>22</td>
<td>65</td>
<td>5.1</td>
<td>25</td>
<td>C1</td>
<td>C2</td>
</tr>
</tbody>
</table>

- **Multi target models** - Several outputs in one model

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
<th>P6</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>9</td>
<td>245</td>
<td>5</td>
<td>0.5</td>
<td>6.5</td>
<td>12</td>
<td>2</td>
<td>C3</td>
<td>C1</td>
</tr>
<tr>
<td>42</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>65</td>
<td>12</td>
<td>6</td>
<td>15</td>
<td>1.2</td>
<td>51</td>
<td>5.9</td>
<td>1.1</td>
<td>C2</td>
<td>C1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>12</td>
<td>2</td>
<td>5</td>
<td>22</td>
<td>65</td>
<td>5.1</td>
<td>25</td>
<td>C1</td>
<td>C2</td>
</tr>
</tbody>
</table>

BigSkyEarth workshop – Novi Sad, Serbia, 2018
Predicting environmental conditions

Multi-target regression tree for all physico-chemical parameters

* These are unpublished results.

BigSkyEarth workshop – Novi Sad, Serbia, 2018
Predicting environmental conditions

Multi–target regression tree for predicting trophic parameters

Multi–target regression tree for predicting concentration of metals

BigSkyEarth workshop – Novi Sad, Serbia, 2018
A multi-target regression tree predicting the water quality class from the entire lake diatom community.
Predicting the entire community dynamic

- A multi-target regression tree predicting the structure of the entire lake diatom community


BigSkyEarth workshop – Novi Sad, Serbia, 2018
Predicting TOP10 diatom dynamic

A multi-target regression tree predicting the relative abundance of the 10 most abundant lake diatoms


BigSkyEarth workshop – Novi Sad, Serbia, 2018
The most interesting from the model is the separation of two groups (clusters) of diatom species influenced by several different parameters. Actually, the OSMER irradiance is the most influencing factor on the abundance of the this diatom community according to the model, while Kpar10m and PO$_3^-$ are the less influence factors on the distribution of the diatoms.

Fururemore, the *Paraliasulcata* species are mostly influenced by the NO$_2$, while the *Naviculasp.sp.1, Naviculasp.sp.2* and *Nitzschiasp* diatoms are influences by the Oxygen demand parameter.

* These are unpublished results. (Data from Venice, Italy)
Rule induction models

Rule 1: IF NROT <= 19 AND CSCU > 19.0 AND CPLA <= 0
THEN [oligosaprobous] [196/16: 22/3] [*]

Rule 2: IF (Navicula krsticii (NKRS)) <= 0 THEN

Rule 3: IF (Sellaphora pupula (SPUP) > 4 AND Cyclotella juriljii nom. nud.(CJUR) <= 2) THEN

Fuzzy based decision tree models

- (A) PT model for the *mesotrophic* class based on TSI_TP
- (B) PT model for the *eutrophic* class based on TSI_TP


BigSkyEarth workshop – Novi Sad, Serbia, 2018
Fuzzy based decision tree models

- (A) PT model for the *brackish* class based on Conductivity parameter.
- (B) PT model for the *alkalibiotic* class based on pH parameter.

- **Fuzzy Pattern Trees are more accurate, but they demand more processing power.**
Dataset description so far..., However!!

• The dataset used in these experiments is collected during 16 month period on 14 lake measurement stations (freq = 1 month) in total 224 collected samples.

• Physico–chemical parameter set includes (temperature, dissolved oxygen, Secchi depth, conductivity, alkalinity (pH), nitrogen compounds (NO2, NO3, NH4, inorganic nitrogen), Sulphur oxide ions SO4, and Sodium (Na), Potassium (K), Magnesium (Mg), Copper (Cu), Manganese (Mn) and Zinc (Zn)).

• Biological information of the ecological system was represented with 116 different diatoms and their relative abundance.

• Today environmental monitoring and thus more precise modelling requires data with high frequency sampled, high in volume and variety of different type of data.

• System based on the IoT.
## Environmental monitoring using IoT

<table>
<thead>
<tr>
<th>Last Update</th>
<th>Mon 23:56</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>13.10 °C</td>
</tr>
<tr>
<td>Humidity</td>
<td>38.20 %</td>
</tr>
<tr>
<td>Abs Humidity</td>
<td>4.36 g/m³</td>
</tr>
<tr>
<td>Bar Pressure</td>
<td>970.00 mb</td>
</tr>
<tr>
<td>BarPres Temp</td>
<td>12.38 °C</td>
</tr>
<tr>
<td>Humindex</td>
<td>No significant</td>
</tr>
<tr>
<td>HeatIndex</td>
<td>10.74 °C</td>
</tr>
<tr>
<td>HeatIndex</td>
<td>No Heat Risk</td>
</tr>
<tr>
<td>Kitchen DewPoint</td>
<td>-0.84 °C</td>
</tr>
<tr>
<td>AirQualityIndex</td>
<td>Moderate Pollution</td>
</tr>
<tr>
<td>Dust Density</td>
<td>122.73 ug/m³</td>
</tr>
<tr>
<td>CO₂</td>
<td>53.93 ppm</td>
</tr>
<tr>
<td>MQ-2</td>
<td>379.6 mV</td>
</tr>
<tr>
<td>MQ-7</td>
<td>261.8 mV</td>
</tr>
<tr>
<td>MQ-135</td>
<td>310.3 mV</td>
</tr>
</tbody>
</table>

- Working live prototype of multi sensor environmental monitoring suite

---

BigSkyEarth workshop – Novi Sad, Serbia, 2018
Sensors generates a lot of data - BigData

- Traditional environmental monitoring usually measures every 8 hours till 2 weeks, depending on what you measure.

- Frequency of measurement ranges from every 20s, 1 min, 5 min or 20 min, depending on the nature of the measuring parameter.
Automated environmental monitoring station

Typical monitoring station will have:

- Weather station
- Chain of temperature sensors
- Dissolved oxygen
- pH
- Turbidity
- Chlorophyll fluorescence
- All every 2-5 minutes.

Credit to NETLAKE Project.
Pre-processing of the sensor generated data

• In order to process the BigData we need more processing power, and to build better models we need more data and more accurate machine learning algorithms (Fuzzy decision trees, but more power hungry).

• One way is to tamper the sampling frequency, but this is strongly depends from the nature of the measured parameter and the purpose of the measurement itself.

• Another way is to overcome the problem of volume is to remove the unnecessary attributes and reduce the size of the dataset (dimensional reduction) using feature selection.

• Methods like PCA, CCA, DCA, Attribute elevator methods (CorrelationAttributeEval technique, InfoGainAttributeEval Attribute Evaluator, Genetic algorithm, ant optimization and etc.), wrapper subset eval methods based on particular machine learning method.
Pre-processing of the sensor generated data

• **OR**, we can use more advance methods like fuzzy set theory concepts and fuzzy-rough redact concept to further improve the feature selection process.

• We have to be careful not to remove the features that describes the target attribute.

• And if this fails, there is always the way...
Massive data environmental analysis. Future work.

- Brute Force with pure CPU processing power, and hope for the best.

- Incremental learning of decision trees from streams of data.
  - The algorithm performs online and in real-time, observes each example only once at the speed of arrival, and maintains at any-time a ready-to-use model tree.

- Using ensemble's of decision trees for data streams.
  - Online Bagging*
  - Online Boosting*
  - Online RandomForest*

- Any many other machine learning methods for classification, regression, clustering, outlier detection, concept drift detection and recommender systems, which are not mention in this presentation, but deserve their place in this research area!!

Conclusion

• The presented research mainly focuses on water ecosystems, however air, soil or any other type of environmental system can be similar modelled with these machine learning algorithms.

• Many of the classical decision tree methods are easy and fast to be used in rather small datasets, but for more accurate modelling fuzzy methods provide excellent platform with a cost of performance price attach to them.

• Feature selection methods and methods for incremental learning of data streams provide interesting option to tackle some of the properties behind the concept of Bigdata in environmental modelling.

• There is no easy way to tackle the BigData when it comes to environmental modelling.
Q&A Section

Thank you for your attention

Any Questions?

Contact E-mail: andreja.naumoski@finki.ukim.mk