Machine learning approach for groundwater contaminant prediction: a quick review for model selection

Speaker: Huynh, T.M. Trang
Advisor: Ni, Chuen Fa
Date: May 10th 2018
Is groundwater important?

• Supply drinking water for 1/3 world population
• Contribute to 43% of water irrigation worldwide
• Purify and store water in good quality for centuries
• Eliminate pathogens
• Mitigate floods and droughts
• Maintaining consistent water levels in surface water bodies,
Sources of GW contamination

- Industry: Storage Tanks, Uncontrolled Hazardous Waste
- Household: Septic Systems, Landfills
- Agriculture: pesticide, fertilizers
- Atmospheric Contaminants

http://www.groundwater.org/get-informed/groundwater/contamination.html
• How to predict groundwater quality in a certain area with the limited input data and time in order to have better management practice
Outline

1. Machine learning at a glance
2. Case studies of using machine learning for nitrate prediction in groundwater
3. Discussion
4. Future work
Machine learning: Machines imitating and adapting human like behavior.

Example:

2-4
3-9
5-25
8-?
Đại Loan biến nhựa phê thai thành vàng “xanh”

Những núi chai nhựa đã sử dụng tại nhà máy tái chế Thái ở Đài Bắc bốc lên một mùi độc trung cấp rác thái khiến người ta ợ nóng, những chàng bảo lâu chứng sẽ biến thành những bộ tóc giả hay quần áo phục vụ con người.

73% lượng nhựa phê thai ở Đài Loan được tái chế.
Machine learning process

**INPUT TERMS**
- Features
- Predictors
- Attributes
- Independent variables

**OUTPUT TERMS**
- Classes
- Responses
- Targets
- Dependent variables

**Machine**
- Algorithms
- Techniques
- Models
Pre-Processing

Iterate till data is prepared

Structured Data

Learning Algorithm

Iterate to get best model

Candidate Model

Data Providers

Select Data

Raw Data

Data Processing Tools

Applications

Golden Model

Deploy Selected Model

Types of machine learning based on algorithms

**supervised**
- Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Support Vector Machines (SVM)
- Neural Networks

**unsupervised**
- K-mean clustering
- Hierarchical clustering
- Self-organizing maps
- Gaussian mixture models
- Hidden Markov models
- ..

**Semi-supervised**
- K-mean clustering
- Linear Regression
- ..

**Reinforcement**
- Q-learning
- Temporal Difference
- Deep Adversarial Networks
- ..
Languages

- Python
- R
- C++
- Java
- Scala
- Clojure
- Matlab
- ...

Softwares

- Caffe
- Tensorflow
- Deeplearning4j
- Pytorch
- Torch
- Apache SINGA
- MATLAB Neural Network...
How to distinguish machine learning and other terms?

Artificial Intelligence

Machine learning

Deep learning

Deep neural networks
## Why should we choose machine learning?

| ✔ | Wide application |
| ✔ | Multi dimension |
| ✔ | Mimic the human brain – learning by observing. |
| ✔ | Improve themselves by studying high volumes of available data |
| ✔ | Work best with Big Data |
| ✔ | High accuracy |
| ✔ | Real time simulation |
| ✔ | Can be transferred and scaled across multiple applications and millions of users |
| ✔ | Abundant and cheap |

| ✗ | Requires some understanding of the problem at hand to apply the right algorithm. |
| ✗ | Getting relevant data |
| ✗ | Choose best-fit machine learning algorithms |
2. Case study 1: Arak Aquifer (Iran)

- semi-arid region
- 1750m
- Surrounded by industrial zones and farms

http://www.achmaz.ir/News/print/739/
2. Case study 1: Arak Aquifer (Iran)

Using artificial neural network (ANN) and linear regression (LR) to model groundwater nitrate

<table>
<thead>
<tr>
<th>input</th>
<th>output</th>
<th>Study site</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH, EC, Mg, Na, K+, HCO₃⁻, SO₄⁻, Ca, TDS, TH, NO₃</td>
<td>NO₃</td>
<td>53 wells, 818 sample. 70% sample used to train models</td>
<td>-Sensitivity analysis-&gt; reduce input variables, -compare performance of ANN vs LR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No.</th>
<th>Input parameters</th>
<th>r</th>
<th>RMSE</th>
<th>MAE</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Na⁺, Mg²⁺, T, Ca²⁺, HCO₃⁻, SO₄²⁻, Cl⁻, TH, TDS, EC, pH</td>
<td>0.84</td>
<td>14.78</td>
<td>9.84</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>Na⁺, Mg²⁺, Ca²⁺, HCO₃⁻, Cl⁻, TH, TDS, EC, pH</td>
<td>0.84</td>
<td>12.68</td>
<td>8.02</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>Na⁺, Mg²⁺, Ca²⁺, HCO₃⁻, Cl⁻, TH, EC, pH</td>
<td>0.87</td>
<td>10.46</td>
<td>7.77</td>
<td>0.63</td>
</tr>
<tr>
<td>4</td>
<td>Na⁺, Mg²⁺, Ca²⁺, HCO₃⁻, Cl⁻, EC, pH</td>
<td>0.81</td>
<td>13.51</td>
<td>8.68</td>
<td>0.63</td>
</tr>
<tr>
<td>5</td>
<td>Na⁺, Mg²⁺, Ca²⁺, HCO₃⁻, Cl⁻, EC</td>
<td>0.79</td>
<td>13.56</td>
<td>9.28</td>
<td>0.63</td>
</tr>
<tr>
<td>6</td>
<td>Na⁺, Mg²⁺, Ca²⁺, HCO₃⁻, EC</td>
<td>0.80</td>
<td>14.25</td>
<td>10.47</td>
<td>0.63</td>
</tr>
</tbody>
</table>

(Zare et al., 2011)
Case study 2: Babol Iran

- situated in Mazandaran province.
- 14,301 km²
- shallow wells: water table depth is about 2.5 m, depth of 5.5 m
- Urea is the most common fertilizer used: urea, phosphate
- input layers relied on subjective judgments

https://www.worldatlas.com/as/ir/21/where-is-babol.html

(Ehteshami et al., 2016)
### Case study 2: Babol Iran

- Modeling groundwater nitrate (Ehteshami et al., 2016)

<table>
<thead>
<tr>
<th>Input</th>
<th>output</th>
<th>Study site</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>- NO2 in ground water (mg/l)</td>
<td>N-NO3</td>
<td>14,301 km², sampling 50/144 wells, 2011-2014</td>
<td>- Using Pearson method to select input data,</td>
</tr>
<tr>
<td>- Saturation percentage</td>
<td></td>
<td></td>
<td>- Compare performance of ANN feed-forward (BP) and the radial basis function (RBF)</td>
</tr>
<tr>
<td>- EC x 10^3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- pH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Organic matter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Organic carbon</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Total nitrogen</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Graphs showing comparison of measured and modeled nitrate levels](image)

- Back Propagation: R=0.77
- Radial Base: R=0.83
Case study 3: Maku, Iran

- Predicting Fluoride in groundwater (Bazerga et al., 2016)

- north of West Azerbaijan province in the northwest of Iran
- 400km²
- Cold, arid
- -7.4°C->17.2°C
- Basaltic rock, silicate minerals, apatite and fluorapatite
- Flour contamination
### Case study 3: Maku, Iran

- Predicting Fluoride in groundwater (Bazerga et al., 2016)

<table>
<thead>
<tr>
<th>Input</th>
<th>output</th>
<th>Study site</th>
<th>method</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC, pH, ion(Ca2+, Na+, K+, HCO3-)</td>
<td>FI</td>
<td>400km2, 39 sampling site (wells, springs,..) 143 samples, 2004-2008</td>
<td>-principal component analysis (PCA) used to select input</td>
</tr>
<tr>
<td>sampling</td>
<td></td>
<td>80% training, 20% testing</td>
<td>-ELM (extreme learning machine) perform better, faster than MLP (MLP)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SVM (support vector machine)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model performance</th>
<th>R2</th>
<th>NSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>0.8152</td>
<td>0.8019</td>
</tr>
<tr>
<td>SVM</td>
<td>0.8833</td>
<td>0.8658</td>
</tr>
<tr>
<td>ELM</td>
<td><strong>0.921</strong></td>
<td><strong>0.907</strong></td>
</tr>
<tr>
<td>Algorithm</td>
<td>Accuracy (R)</td>
<td>Training time</td>
</tr>
<tr>
<td>----------------------------------------------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Neural Network (NN)-Levenberg-Marquardt algorithm</td>
<td>0.8</td>
<td>NA</td>
</tr>
<tr>
<td>NN-BP (Back propagation)</td>
<td>0.77</td>
<td>NA</td>
</tr>
<tr>
<td>NN-RBF (Radial Based function)</td>
<td>0.83</td>
<td>NA</td>
</tr>
<tr>
<td>NN-MLP (Multi Layer Perceptron)</td>
<td>0.81</td>
<td>NA</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.88</td>
<td>NA</td>
</tr>
<tr>
<td>Linear</td>
<td>0.6</td>
<td>NA</td>
</tr>
<tr>
<td>Extreme learning Machine (ELM)</td>
<td>0.92</td>
<td>NA</td>
</tr>
</tbody>
</table>
3. Discussion

scheme to choose a suitable model

Type of machine learning

- Supervised/ unsupervised/ semi-supervised/reinforcement

Choose algorithm

Find predictors
3. Discussion

scheme to choose a suitable model

- Type of machine learning
  - Supervised/ unsupervised/ semi-supervised/reinforcement

- Choose algorithm
  - Regression/decision tree/ neural networks,...

- Find predictors
3. Discussion

scheme to choose a suitable model

- Supervised/unsupervised/semi-supervised/reinforcement

Choose algorithm

Find predictors

- Based on subjective adjustment?

Type of machine learning

Regression/decision tree/neural networks,…
3. Discussion

• Neural Networks or Deep Learning are most used model, while Deep learning machine is the most accurate prediction.

• There are no need for aquifer characteristic or soil parameters.

• Select input based on personal experience and sampling is needed.

• However, can machine learning predict the problem with no available target output?
4. Future work

Find more materials, decide a target output
Find indicator input
Select suitable algorithms and models
Develop methodology
Practice on a case study
Main references


Thank you