Title: Machine learning in earth sciences
URL: https://en.wikipedia.org/wiki/Machine_learning_in_earth_sciences
PageID: 68735447
Categories: Category:Geological techniques, Category:Machine learning

---
Artificial general intelligence
Intelligent agent
Recursive self-improvement
Planning
Computer vision
General game playing
Knowledge representation
Natural language processing
Robotics
AI safety
Machine learning
Symbolic
Deep learning
Bayesian networks
Evolutionary algorithms
Hybrid intelligent systems
Systems integration
Open-source
Bioinformatics
Deepfake
Earth sciences
Finance
Generative AI Art Audio Music
Art
Audio
Music
Government
Healthcare Mental health
Mental health
Industry
Software development

Translation

Military

Physics

Projects

AI alignment

Artificial consciousness

The bitter lesson

Chinese room

Friendly AI

Ethics

Existential risk

Turing test

Uncanny valley

Timeline

Progress

AI winter

AI boom

AI bubble

Glossary

v

t

e

Applications of machine learning (ML) in earth sciences include geological mapping , gas leakage detection and geological feature identification . Machine learning is a subdiscipline of artificial intelligence aimed at developing programs that are able to classify, cluster, identify, and analyze vast and complex data sets without the need for explicit programming to do so. [1] Earth science is the study of the origin, evolution, and future [2] of the Earth . The earth's system can be subdivided into four major components including the solid earth , atmosphere , hydrosphere , and biosphere . [3]

A variety of algorithms may be applied depending on the nature of the task. Some algorithms may perform significantly better than others for particular objectives. For example, convolutional neural networks (CNNs) are good at interpreting images, whilst more general neural networks may be used for soil classification , [4] but can be more computationally expensive to train than alternatives such as support vector machines . The range of tasks to which ML (including deep learning) is applied has been ever-growing in recent decades, as has the development of other technologies such as unmanned aerial vehicles (UAVs), [5] ultra-high resolution remote sensing technology, and high-performance computing . [6] This has led to the availability of large high-quality datasets and more advanced algorithms.

Significance

Complexity of earth science

Problems in earth science are often complex. [7] It is difficult to apply well-known and described mathematical models to the natural environment, therefore machine learning is commonly a better alternative for such non-linear problems. [8] Ecological data are commonly non-linear and consist of higher-order interactions, and together with missing data, traditional statistics may underperform

as unrealistic assumptions such as linearity are applied to the model. [9] [10] A number of researchers found that machine learning outperforms traditional statistical models in earth science, such as in characterizing forest canopy structure, [11] predicting climate-induced range shifts , [12] and delineating geologic facies. [13] Characterizing forest canopy structure enables scientists to study vegetation response to climate change. [14] Predicting climate-induced range shifts enable policy makers to adopt suitable conversation method to overcome the consequences of climate change. [12] Delineating geologic facies helps geologists to understand the geology of an area, which is essential for the development and management of an area. [15]

Inaccessible data

In Earth Sciences, some data are often difficult to access or collect, therefore inferring data from data that are easily available by machine learning method is desirable. [10] For example, geological mapping in tropical rainforests is challenging because the thick vegetation cover and rock outcrops are poorly exposed. [16] Applying remote sensing with machine learning approaches provides an alternative way for rapid mapping without the need of manually mapping in the unreachable areas. [16]

Reduce time costs

Machine learning can also reduce the efforts done by experts, as manual tasks of classification and annotation etc. are the bottlenecks in the workflow of the research of earth science. [10] Geological mapping, especially in a vast, remote area is labour, cost and time-intensive with traditional methods. [17] Incorporation of remote sensing and machine learning approaches can provide an alternative solution to eliminate some field mapping needs. [17]

Consistent and bias-free

Consistency and bias-free is also an advantage of machine learning compared to manual works by humans. In research comparing the performance of human and machine learning in the identification of dinoflagellates , machine learning is found to be not as prone to systematic bias as humans. [18] A recency effect that is present in humans is that the classification often biases towards the most recently recalled classes. [18] In a labelling task of the research, if one kind of dinoflagellates occurs rarely in the samples, then expert ecologists commonly will not classify it correctly. [18] The systematic bias strongly deteriorate the classification accuracies of humans. [18]

Optimal machine learning algorithm

The extensive usage of machine learning in various fields has led to a wide range of algorithms of learning methods being applied. Choosing the optimal algorithm for a specific purpose can lead to a significant boost in accuracy: [19] for example, the lithological mapping of gold-bearing granite-greenstone rocks in Hutti, India with AVIRIS-NG hyperspectral data, shows more than 10% difference in overall accuracy between using support vector machines (SVMs) and random forest . [20]

Some algorithms can also reveal hidden important information: white box models are transparent models , the outputs of which can be easily explained, while black box models are the opposite. [19] For example, although an SVM yielded the best result in landslide susceptibility assessment accuracy, the result cannot be rewritten in the form of expert rules that explain how and why an area was classified as that specific class. [7] In contrast, decision trees are transparent and easily understood, and the user can observe and fix the bias if any is present in such models. [7]

If computational resource is a concern, more computationally demanding learning methods such as deep neural networks are less preferred, despite the fact that they may outperform other algorithms, such as in soil classification. [4]

Usage

Mapping

Geological or lithological mapping and mineral prospectivity mapping

Geological or lithological mapping produces maps showing geological features and geological units. Mineral prospectivity mapping utilizes a variety of datasets such as geological maps and aeromagnetic imagery to produce maps that are specialized for mineral exploration. [21] Geological, lithological, and mineral prospectivity mapping can be carried out by processing data with ML techniques, with the input of spectral imagery obtained from remote sensing and geophysical data. [22] Spectral imaging is also used – the imaging of wavelength bands in the electromagnetic spectrum, while conventional imaging captures three wavelength bands (red, green, blue) in the electromagnetic spectrum. [23]

Random forests and SVMs are some algorithms commonly used with remotely-sensed geophysical data, while Simple Linear Iterative Clustering-Convolutional Neural Network (SLIC-CNN) [5] and Convolutional Neural Networks (CNNs) [17] are commonly applied to aerial imagery. Large scale mapping can be carried out with geophysical data from airborne and satellite remote sensing geophysical data, [20] and smaller-scale mapping can be carried out with images from Unmanned Aerial Vehicles (UAVs) for higher resolution. [5]

Vegetation cover is one of the major obstacles for geological mapping with remote sensing, as reported in various research, both in large-scale and small-scale mapping. Vegetation affects the quality of spectral images, [22] or obscures the rock information in aerial images. [5]

Random Forest ,

Support Vector Machine (SVM)

(1) Map generated with remote sensing data only has a 52.7% accuracy when compared to the geological map, but several new possible lithological units are identified

(2) Map generated with remote sensing data and spatial constraints has a 78.7% accuracy but no new possible lithological units are identified

geophysical data

Morocco

frequency electromagnetic, radiometric measurements, ground gravity measurements

Liaoning Province, China

Remote Predictive Mapping (RPM)

Landsat Reflectance, High-Resolution Digital Elevation Data

Northwest Territories, Canada

Random Forest

Landslide susceptibility and hazard mapping

Landslide susceptibility refers to the probability of landslide of a certain geographical location, which is dependent on local terrain conditions. [26] Landslide susceptibility mapping can highlight areas prone to landslide risks, which is useful for urban planning and disaster management. [7] Such datasets for ML algorithms usually include topographic information, lithological information, satellite images, etc., and some may include land use, land cover, drainage information, and vegetation cover [7] [27] [28] [29] according to the study requirements. As usual, for training an ML model for landslide susceptibility mapping, training and testing datasets are required. [7] There are two methods of allocating datasets for training and testing: one is to randomly split the study area for the datasets; another is to split the whole study into two adjacent parts for the two datasets. To test classification models, the common practice is to split the study area randomly; [7] [30] however, it is more useful if the study area can be split into two adjacent parts so that an automation algorithm can carry out mapping of a new area with the input of expert-processed data of adjacent land. [7]

Decision Trees , Logistic Regression

Feature identification and detection

Discontinuity analyses

Discontinuities such as fault planes and bedding planes have important implications in civil engineering. [31] Rock fractures can be recognized automatically by machine learning through photogrammetric analysis, even with the presence of interfering objects such as vegetation. [32] In ML training for classifying images, data augmentation is a common practice to avoid overfitting and increase the training dataset size and variability. [32] For example, in a study of rock fracture recognition, 68 images for training and 23 images for testing were prepared via random splitting. [32] Data augmentation was performed, increasing the training dataset size to 8704 images by flipping and random cropping. [32] The approach was able to recognize rock fractures accurately in most cases. [32] Both the negative prediction value (NPV) and the specificity were over 0.99. [32] This demonstrated the robustness of discontinuity analyses with machine learning.

Seoul, Korea and Jeongseon-gun, Gangwon-do, Korea

Carbon dioxide leakage detection

Quantifying carbon dioxide leakage from a geological sequestration site has gained increased attention as the public is interested in whether carbon dioxide is stored underground safely and effectively. [33] Carbon dioxide leakage from a geological sequestration site can be detected indirectly with the aid of remote sensing and an unsupervised clustering algorithm such as Iterative Self-Organizing Data Analysis Technique (ISODATA). [34] The increase in soil CO 2 concentration causes a stress response for plants by inhibiting plant respiration, as oxygen is displaced by carbon dioxide. [35] The vegetation stress signal can be detected with the Normalized Difference Red Edge Index (NDRE). [35] The hyperspectral images are processed by the unsupervised algorithm, clustering pixels with similar plant responses. [35] The hyperspectral information in areas with known CO 2 leakage is extracted so that areas with leakage can be matched with the clustered pixels with spectral anomalies. [35] Although the approach can identify CO 2 leakage efficiently, there are some limitations that require further study. [35] The NDRE may not be accurate due to reasons like higher chlorophyll absorption, variation in vegetation, and shadowing effects; therefore, some stressed pixels can be incorrectly classed as healthy. [35] Seasonality , groundwater table height may also affect the stress response to CO 2 of the vegetation. [35]

Quantification of water inflow

The rock mass rating (RMR) [36] system is a widely adopted rock mass classification system by geomechanical means with the input of six parameters. The amount of water inflow is one of the inputs of the classification scheme, representing the groundwater condition. Quantification of the water inflow in the faces of a rock tunnel was traditionally carried out by visual observation in the field, which is labour and time-consuming, and fraught with safety concerns. [37] Machine learning can determine water inflow by analyzing images taken on the construction site. [37] The classification of the approach mostly follows the RMR system, but combining damp and wet states, as it is difficult to distinguish only by visual inspection. [37] [36] The images were classified into the non-damaged state, wet state, dripping state, flowing state, and gushing state. [37] The accuracy of classifying the images was approximately 90%. [37]

Classification

Soil classification

The most popular cost-effective method od soil investigation method is cone penetration testing (CPT) . [38] The test is carried out by pushing a metallic cone through the soil: the force required to push at a constant rate is recorded as a quasi-continuous log. [4] Machine learning can classify soil with the input of CPT data. [4] In an attempt to classify with ML, there are two tasks required to analyze the data, namely segmentation and classification. [4] Segmentation can be carried out with the Constraint Clustering and Classification (CONCC) algorithm to split a single series data into segments. [4] Classification can then be carried out by algorithms such as decision trees, SVMs, or neural networks. [4]

Geological structure classification

Exposed geological structures such as anticlines , ripple marks , and xenoliths can be identified automatically with deep learning models. [39] Research has demonstrated that three-layer CNNs

and transfer learning have strong accuracy (about 80% and 90% respectively), while others like k -nearest neighbors (k-NN), regular neural nets, and extreme gradient boosting (XGBoost) have low accuracies (ranging from 10% - 30%). [39] The grayscale images and colour images were both tested, with the accuracy difference being little, implying that colour is not very important in identifying geological structures. [39]

Forecast and predictions

Earthquake early warning systems and forecasting

Earthquake warning systems are often vulnerable to local impulsive noise, therefore giving out false alerts. [40] False alerts can be eliminated by discriminating the earthquake waveforms from noise signals with the aid of ML methods. The method consists of two parts, the first being unsupervised learning with a generative adversarial network (GAN) to learn and extract features of first-arrival P-waves , and the second being use of a random forest to discriminate P-waves. This approach achieved 99.2% in recognizing P-waves, and can avoid false triggers by noise signals with 98.4% accuracy. [40]

Earthquakes can be produced in a laboratory settings to mimic real-world ones. With the help of machine learning, the patterns of acoustic signals as precursors for earthquakes can be identified. Predicting the time remaining before failure was demonstrated in a study with continuous acoustic time series data recorded from a fault. The algorithm applied was a random forest, trained with a set of slip events, performing strongly in predicting the time to failure. It identified acoustic signals to predict failures, with one of them being previously unidentified. Although this laboratory earthquake is not as complex as a natural one, progress was made that guides future earthquake prediction work. [41]

Streamflow discharge prediction

Real-time streamflow data is integral for decision making (e.g., evacuations, or regulation of reservoir water levels during flooding). [42] Streamflow data can be estimated by data provided by stream gauges , which measure the water level of a river. However, water and debris from flooding may damage stream gauges, resulting in lack of essential real-time data. The ability of machine learning to infer missing data [10] enables it to predict streamflow with both historical stream gauge data and real-time data.

Streamflow Hydrology Estimate using Machine Learning (SHEM) is a model that can serve this purpose. To verify its accuracies, the prediction result was compared with the actual recorded data, and the accuracies were found to be between 0.78 and 0.99.

Challenge

Inadequate training data

An adequate amount of training and validation data is required for machine learning. [10] However, some very useful products like satellite remote sensing data only have decades of data since the 1970s. If one is interested in the yearly data, then only less than 50 samples are available. [44] Such amount of data may not be adequate. In a study of automatic classification of geological structures, the weakness of the model is the small training dataset, even though with the help of data augmentation to increase the size of the dataset. [39] Another study of predicting streamflow found that the accuracies depend on the availability of sufficient historical data, therefore sufficient training data determine the performance of machine learning. [43] Inadequate training data may lead to a problem called overfitting. Overfitting causes inaccuracies in machine learning [45] as the model learns about the noise and undesired details.

Limited by data input

Machine learning cannot carry out some of the tasks as a human does easily. For example, in the quantification of water inflow in rock tunnel faces by images for Rock Mass Rating system (RMR), [37] the damp and the wet state was not classified by machine learning because discriminating the two only by visual inspection is not possible. In some tasks, machine learning may not able to fully substitute manual work by a human.

Black-box operation

In many machine learning algorithms, for example, Artificial Neural Network (ANN), it is considered as ' black box ' approach as clear relationships and descriptions of how the results are generated in the hidden layers are unknown. [46] 'White-box' approach such as decision tree can reveal the algorithm details to the users. [47] If one wants to investigate the relationships, such 'black-box' approaches are not suitable [48] . However, the performances of 'black-box' algorithms are usually better. [49]

References