D

Data Mining

Tao Wen
Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA

Synonyms

Data dredging; Data science; Knowledge discovery from data; Knowledge extraction; Machine learning; Pattern analysis

Definition

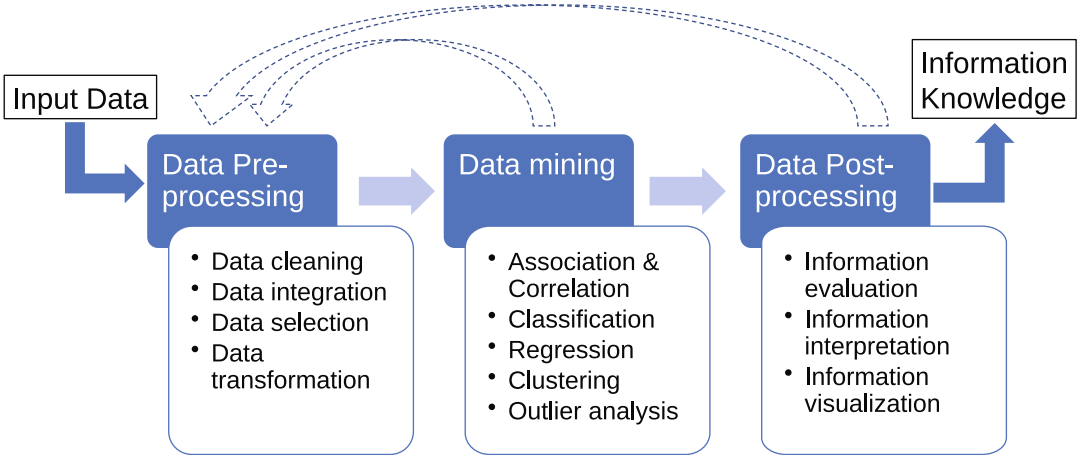
Data mining is the process of utilizing computer algorithms to discover knowledge and information from large amounts of data. Data mining, often interchangeably used with knowledge discovery from data (KDD), is sometimes viewed merely by others as one critical step of KDD (Fig. 1) (Han et al. 2011). The three essential elements of data mining are data, algorithms, and information (or knowledge). Data mining is interdisciplinary and employs many techniques from statistics, machine learning (ML), database, high-performance computing, and domain disciplines (e.g., geosciences). Here, the domain discipline refers to the field in which data mining is applied to gain knowledge and information. In geosciences, it is also not uncommon that data mining is treated as a synonym for machine learning, although some might argue that machine learning is more focused on the development of computer algorithms that are applied during the workflow of data mining. The major types of geoscience data that can be mined for information include matrix data, time-series data, network-based data, and geospatial data. Based on the type of information to be mined and the technologies to be applied, data mining tasks can be divided into five categories: association and correlation, classification, regression, clustering, and outlier analysis (Han et al. 2011). Benefiting from the ever-increasing stream of large geoscience data and the improvement of computational resources, the widespread application of data mining in geosciences becomes inevitable. The most promising near-future development of data mining in geosciences is likely the integration of data mining and physics-based modeling (Reichstein et al. 2019).

What Geoscience Data Can Be Mined?

The explosive growth and ever-increased diversity of geoscience data in the past few decades are at least partially attributed to the development of cheap automatic sensor-based monitoring devices that are deployed in the soil, water, and air. Among those large geoscience datasets, major types of data that can be mined include matrix/tabulated, time-series, geospatial, and network/graph-based. There are overlaps among these data types, e.g., all of the latter three types of data can be formatted as matrix/tabulated data.

The matrix/tabulated data is probably the most common data format in geosciences. Matrix data often has two dimensions in which each row represents one sample while each column contains a feature or characteristic of the sample. For example, Tesoriero et al. (2017) compiled a large set of groundwater quality samples including nitrate, iron, arsenic concentrations, and other bedrock and soil characteristics. In this example, each row represents one groundwater sample. Matrix data is intuitive to read. However, it becomes less convenient to use matrix data to illustrate the spatiotemporal trends of data or the intersample relationship. This is why we also need to organize geoscience data in other formats, e.g., time-series, geospatial, and network-based data. These additional data types allow the incorporation of additional spatiotemporal context and the information of the intersample relationship into the stored data.

B. S. Daya Sagar et al. (eds.), *Encyclopedia of Mathematical Geosciences*, Encyclopedia of Earth Sciences Series,
https://doi.org/10.1007/978-3-030-26050-7_77-1



Data Mining, Fig. 1 The view of knowledge discovery from data (KDD) and data mining from the perspective of machine learning and statistics communities

Time-series data refers to a set of data points indexed in time order. The time when the data point is collected is often used as the timestamp, although other types of timestamps have also been used (e.g., age of rock samples). Time-series data can be formatted as matrix data in which each row represents a sample from a time point and an additional column(s) should be added to show the timestamp of each sample. Time-series data is often used to describe the temporal trend of geoscience parameters. For example, the National Water Information System from the United States Geological Survey (https://waterdata.usgs.gov/nwis) hosts time-series data of water quantity and quality of the water body in the United States.

Geospatial data refers to the geoscience data for which we have information on the geographic location. Similar to time-series data, geospatial data can be formatted as matrix data as well when the location information is added as additional column(s) (e.g., latitude and longitude) in the matrix. In particular, an important source of geospatial data is the remote sensing data from satellites. For example, Google Maps and government agencies provide public access to low- and high-resolution satellite and multispectral imagery.

The last common data type in geoscience is network-based data. Unlike matrix data, network-based data is represented by nodes (vertices) and edges (links; connecting nodes), which can effectively illustrate the relationship between samples. For example, Agarwal et al. (2020) reported a dataset of five geochemical analytes including chloride, bromide, barium, magnesium, and sodium of >80,000 river water samples in a complex river network. Agarwal et al. used the river network-based data to inform the impact of human activities on river water chemistry. Barbier et al. (2020) visualized the compiled data from multiple experimental serpentinization studies into a network with each node representing each experiment and edges indicating the similarity between different experiments.

More and more geoscience data, in particular, the compiled datasets in large size, becomes available nowadays because more geoscience researchers are following the findability, accessibility, interoperability, and reusability (FAIR) principles when sharing data. To comply with FAIR principles, geoscience data should be transformed into a machine-readable format (Wen 2020). Geoscience data can be shared in either discipline-specific or general data repositories. A data repository is a place storing data and providing access to users (Wen 2020). For example, the HydroClient database (https://data.cuahsi.org/) is a popular database in which the hydrology community publishes time-series and geospatial data. Data in the HydroClient database can be downloaded as matrix data. In addition, the Shale Network database (https://doi.org/10.4211/his-data-shalenetwork) that is published in the HydroClient compiles water quality data for water samples collected from the regions of oil and gas production. Each of these Shale Network samples has geographic coordinates reported. The PANGAEA database (https://www.pangaea.de/) is an example of more general data repositories hosting geospatial data from many subdisciplines of geosciences.

What Methods Are Used in Data Mining?

Data mining tasks include the characterization of the data properties (descriptive task) and predictions for the past or the future by learning from existing data (predictive task). Major data mining tasks introduced below include association and correlation, clustering, outlier detection, and classification and regression. The first three tasks are descriptive while the last one is a predictive task.

The goal of association and correlation (AC) analysis, as a descriptive task, is to assess the association or correlation between parameters. In geosciences, a common correlation task is to evaluate the spatial correlation of two parameters. For example, Wen et al. (2018) reported a compiled groundwater quality dataset from the region of oil and gas production in the state of Pennsylvania in the United States. With the rapid development of oil and gas drilling in the United States, the public is getting concerned about the associated environmental impact, e.g., the groundwater contamination. Wen et al. applied a geospatial analysis tool (https://github.com/jaywt/SWGT) to assess the spatial correlation between methane (the major component of natural gas) in groundwater and the proximity to oil and gas wells. They successfully identified a few localities where some natural gas wells are potentially problematic. In paleoclimatology, scientists have striven to find geochemical proxies that correlate with the level of atmospheric oxygen, and air and sea temperatures to reconstruct the ancient climate.

Cluster analysis is the data mining task of classifying a set of samples in a way that a sample is more similar to samples in the same class than to those in other classes. Cluster analysis is one type of unsupervised learning technique. Unsupervised learning refers to those data mining techniques which discover information in a dataset with no preexisting classes. Unlike unsupervised learning, supervised learning techniques rely on datasets with preexisting classes which are often generated by humans. Cluster analysis is used to generate these previously unknown classes for a dataset. The principle of cluster analysis is to maximize intra-class similarity while minimizing inter-class similarity (Han et al. 2011). Typical clustering techniques include K-Means, spectral clustering, Gaussian mixture, self-organizing map (SOM), and affinity propagation. For example, SOM was previously used to infer distinct but spatially contiguous classes within footwall and hanging, basalts, and andesites (Cracknell et al. 2014).

Outlier analysis (OA) is the task of detecting rare or anomalous sample(s) in a dataset. These rare or anomalous samples often fail to comply with the general rules or patterns of the majority of the dataset. Such samples are called outliers. Two major applications of outlier detection in geosciences include the removal of anomalous data before performing other data mining tasks and the detection of ecosystem disturbances. Zheng et al. (2017) developed an outlier detection algorithm and applied it to water quality data along with relevant contextual features from the regions of oil and gas production to successfully detect groundwater samples likely contaminated by nearby problematic oil and gas well(s). Some examples of these contextual features include the location of water samples, the information about nearby oil and gas facilities, and geologic features (e.g., the distance to faults).

Unlike the above tasks, regression analysis (RA) and classification analysis (CA) are supervised learning techniques and predictive tasks. In RA, data mining algorithms are first used to evaluate the relationship between a target variable and often a set of predictor variables in a dataset (i.e., training data) for which values of both target variable and predictor variables are known, and then applied on another dataset (testing data) for which only values of predictor variables are known to predict the values of target variable of the testing data. Here, the target variable refers to the parameter to be predicted while predictor variables denote the set of features used in the RA/CA algorithms to make predictions for the target variable. Similar to RA, CA also requires both training data and testing data, while the goal of CA is to predict the classes of the target variable instead of its values. Typical algorithms used in RA and CA include decision trees, random forest, support vector machines (SVM), naïve Bayesian, rule-based classification, logistic/linear regression, and neural networks. For example, Wen et al. (2021) applied the logistic regression method on compiled water quality data of groundwater samples to predict the probability of dissolved methane in groundwater being originated from nearby oil and gas production activities.

Summary and Conclusions

Data mining refers to the process of discovering information from a large amount of data. Driven by the increasing volume, variety, and velocity (3Vs) of geoscience data as well as the advancement in data science techniques, data mining methods have been widely applied in geosciences. The most promising near-future development of data mining in geosciences is likely the integration of data mining and physics modeling (Reichstein et al. 2019).

Cross-References

- Geodata Science
- Machine Learning
- Machine Learning and Geosciences
- Spatial Data Sciences

Bibliography

Agarwal A, Wen T, Chen A, Zhang AY, Niu X, Zhan X, Xue L, Brantley SL (2020) Assessing contamination of stream networks near shale gas development using a new geospatial tool. Environ Sci Technol. https://doi.org/10.1021/acs.est.9b06761

Barbier S, Huang F, Andreani M, Tao R, Hao J, Eleish A, Prabhu A, Minhas O, Fontaine K, Fox P, Daniel I (2020) A review of H2, CH4, and hydrocarbon formation in experimental serpentinization using network analysis. Front Earth Sci 8:209. https://doi.org/10.3389/feart.2020.00209

Cracknell MJ, Reading AM, McNeill AW (2014) Mapping geology and volcanic-hosted massive sulfide alteration in the Hellyer-Mt Charter region, Tasmania, using Random Forests™ and Self-organising Maps. Aust J Earth Sci 61:287–304. https://doi.org/10.1080/08120099.2014.858081

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques, the Morgan Kaufmann series in data management systems. Elsevier Science, Saint Louis

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat (2019) Deep learning and process understanding for data-driven earth system science. Nature 566:195–204. https://doi.org/10.1038/s41586-019-0912-1

Tesoriero AJ, Gronberg JA, Juckem PF, Miller MP, Austin BP (2017) Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. Water Resour Res 53:7316–7331. https://doi.org/10.1002/2016WR020197

Wen T (2020) Data sharing. In: Encyclopedia of big data. Springer International Publishing, Cham, pp 1–3. https://doi.org/10.1007/978-3-319-32001-4_322-1

Wen T, Niu X, Gonzales M, Zheng G, Li Z, Brantley SL (2018) Big groundwater data sets reveal possible rare contamination amid otherwise improved water quality for some analytes in a region of marcellus shale development. Environ Sci Technol 52:7149–7159. https://doi.org/10.1021/acs.est.8b01123

Wen T, Liu M, Woda J, Zheng G, Brantley SL (2021) Detecting anomalous methane in groundwater within hydrocarbon production areas across the United States. Water Research 200:117236. https://doi.org/10.1016/j.watres.2021.117236

Zheng G, Brantley SL, Lauvaux T, Li Z (2017) Contextual spatial outlier detection with metric learning. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining – KDD '17. ACM Press, New York, pp 2161–2170. https://doi.org/10.1145/3097983.3098143