Title: Data preprocessing

URL: https://en.wikipedia.org/wiki/Data_preprocessing

PageID: 12386904

Categories: Category:Data mining, Category:Machine learning

Data preprocessing can refer to manipulation, filtration or augmentation of data before it is analyzed, and is often an important step in the data mining process. Data collection methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, and missing values , amongst other issues.

Preprocessing is the process by which unstructured data is transformed into intelligible representations suitable for machine-learning models. This phase of model deals with noise in order to arrive at better and improved results from the original data set which was noisy. This dataset also has some level of missing value present in it.

The preprocessing pipeline used can often have large effects on the conclusions drawn from the downstream analysis. Thus, representation and quality of data is necessary before running any analysis. Often, data preprocessing is the most important phase of a machine learning project, especially in computational biology . If there is a high proportion of irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase may be more difficult. Data preparation and filtering steps can take a considerable amount of processing time. Examples of methods used in data preprocessing include cleaning , instance selection , normalization , one-hot encoding , data transformation , feature extraction and feature selection .

Applications

Data mining

Data preprocessing allows for the removal of unwanted data with the use of data cleaning, this allows the user to have a dataset to contain more valuable information after the preprocessing stage for data manipulation later in the data mining process. Editing such dataset to either correct data corruption or human error is a crucial step to get accurate quantifiers like true positives, true negatives, false positives and false negatives found in a confusion matrix that are commonly used for a medical diagnosis. Users are able to join data files together and use preprocessing to filter any unnecessary noise from the data which can allow for higher accuracy. Users use Python programming scripts accompanied by the pandas library which gives them the ability to import data from a comma-separated values as a data-frame. The data-frame is then used to manipulate data that can be challenging otherwise to do in Excel. Pandas (software) which is a powerful tool that allows for data analysis and manipulation; which makes data visualizations, statistical operations and much more, a lot easier. Many also use the R programming language to do such tasks as well.

The reason why a user transforms existing files into a new one is because of many reasons. Aspects of data preprocessing may include imputing missing values, aggregating numerical quantities and transforming continuous data into categories (data binning). More advanced techniques like principal component analysis and feature selection are working with statistical formulas and are applied to complex datasets which are recorded by GPS trackers and motion capture devices.

Semantic data preprocessing

Semantic data mining is a subset of data mining that specifically seeks to incorporate domain knowledge , such as formal semantics, into the data mining process. Domain knowledge is the knowledge of the environment the data was processed in. Domain knowledge can have a positive influence on many aspects of data mining, such as filtering out redundant or inconsistent data during the preprocessing phase. Domain knowledge also works as constraint. It does this by using

working as set of prior knowledge to reduce the space required for searching and acting as a guide to the data. Simply put, semantic preprocessing seeks to filter data using the original environment of said data more correctly and efficiently.

There are increasingly complex problems which are asking to be solved by more elaborate techniques to better analyze existing information. [fact or opinion?] Instead of creating a simple script for aggregating different numerical values into a single value, it make sense to focus on semantic based data preprocessing. The idea is to build a dedicated ontology , which explains on a higher level what the problem is about. In regards to semantic data mining and semantic pre-processing, ontologies are a way to conceptualize and formally define semantic knowledge and data. The Protégé (software) is the standard tool for constructing an ontology. [citation needed] In general, the use of ontologies bridges the gaps between data, applications, algorithms, and results that occur from semantic mismatches. As a result, semantic data mining combined with ontology has many applications where semantic ambiguity can impact the usefulness and efficiency of data systems. [citation needed] Applications include the medical field, language processing, banking, and even tutoring, among many more.

There are various strengths to using a semantic data mining and ontological based approach. As previously mentioned, these tools can help during the per-processing phase by filtering out non-desirable data from the data set. Additionally, well-structured formal semantics integrated into well designed ontologies can return powerful data that can be easily read and processed by machines. A specifically useful example of this exists in the medical use of semantic data processing. As an example, a patient is having a medical emergency and is being rushed to hospital. The emergency responders are trying to figure out the best medicine to administer to help the patient. Under normal data processing, scouring all the patient's medical data to ensure they are getting the best treatment could take too long and risk the patients' health or even life. However, using semantically processed ontologies, the first responders could save the patient's life. Tools like a semantic reasoner can use ontology to infer the what best medicine to administer to the patient is based on their medical history, such as if they have a certain cancer or other conditions, simply by examining the natural language used in the patient's medical records. This would allow the first responders to quickly and efficiently search for medicine without having worry about the patient's medical history themselves, as the semantic reasoner would already have analyzed this data and found solutions. In general, this illustrates the incredible strength of using semantic data mining and ontologies. They allow for quicker and more efficient data extraction on the user side, as the user has fewer variables to account for, since the semantically pre-processed data and ontology built for the data have already accounted for many of these variables. However, there are some drawbacks to this approach. Namely, it requires a high amount of computational power and complexity, even with relatively small data sets. This could result in higher costs and increased difficulties in building and maintaining semantic data processing systems. This can be mitigated somewhat if the data set is already well organized and formatted, but even then, the complexity is still higher when compared to standard data processing. [tone]

Below is a simple a diagram combining some of the processes, in particular semantic data mining and their use in ontology.

The diagram depicts a data set being broken up into two parts: the characteristics of its domain, or domain knowledge, and then the actual acquired data. The domain characteristics are then processed to become user understood domain knowledge that can be applied to the data. Meanwhile, the data set is processed and stored so that the domain knowledge can applied to it, so that the process may continue. This application forms the ontology. From there, the ontology can be used to analyze data and process results.

Fuzzy preprocessing is another, more advanced technique for solving complex problems. Fuzzy preprocessing and fuzzy data mining make use of fuzzy sets . These data sets are composed of two elements: a set and a membership function for the set which comprises 0 and 1. Fuzzy preprocessing uses this fuzzy data set to ground numerical values with linguistic information. Raw data is then transformed into natural language . Ultimately, fuzzy data mining's goal is to help deal with inexact information, such as an incomplete database. Currently fuzzy preprocessing, as well as other fuzzy based data mining techniques see frequent use with neural networks and artificial

intelligence.

References

External links

Online Data Processing Compendium

Data preprocessing in predictive data mining. Knowledge Eng. Review 34: e1 (2019)

v

t

e

Acquisition

Augmentation

Analysis

Anonymization

Archaeology

Big

Cleansing

Collection

Compression

Corruption

Curation

Deduplication

Degradation

De-identification

Ecosystem

Editing

Engineering

Erasure

ETL / ELT Extract Transform Load

Extract

Transform

Load

Ethics

Exhaust

Exploration

Farming

Format management

Fusion

Governance Cooperatives

Cooperatives

Infrastructure
Integration
Integrity
Library
Lineage
Loss
Management
Meta
Migration
Mining
Philanthropy
Pre-processing
Preservation
Processing
Protection (privacy)
Publishing Open data
Open data
Recovery
Reduction
Redundancy
Re-identification
Remanence
Rescue
Retention
Quality
Science
Scraping
Scrubbing
Security
Sharing
Stewardship
Storage
Structure
Synchronization
Topological data analysis
Type
Validation
Warehouse

Wrangling/munging