DATA PREPARATION TECHNIQUES AND PLATFORMS IN THE CONTEXT OF MACHINE LEARNING

Genka Miteva

PhD Student, Department of Information Technologies and Communications, University of National and World Economy, *e-mail: gmiteva@unwe.bg*

Aleksandrina Murdzheva

Assoc. Prof., Department of Information Technologies and Communications, University of National and World Economy, *e-mail: amurdjeva@unwe.bg*

Abstract

As data is becoming crucial for the efficient functioning of any organization, properly preparing it for processing is also getting increasingly important. This article presents an outline of different data preparation techniques, which can be defined in the context of machine learning. Overview of the techniques in combination of algorithms and their specific requirements for the data they can work with can server as basis of the interesting research task of finding automated solutions that not only enable the use of software solutions, but also build complete solutions that support the detection of the potential of the data and application of these techniques that can be automatically identified.

Automating data preparation might be one of the steps which ensures that the machine learning process becomes quicker and more accessible. Using data preparation tools is a way to ensure more reliable and accurate data. This article aims to create an overview of existing data preparation tools and platforms. Different aspects of these platforms are considered, including data source compatibility, the data preparation techniques it includes, possibilities for integration, data security etc.

Key words: AI, data preparation, data collection, data cleaning, data transformation, feature engineeri ng, data labeling, data preparation platforms, automation

1.Data preparation

Data preparation is the process of transforming data in a way that is suitable for machine learning and other processing purposes. This is one of the main prerequisites for achieving precise results when dealing with machine learning and AI. Incredible amounts of data are being collected constantly, but a lot of it contains inaccuracies, missing values, and outliers. Furthermore, nowadays a lot of the data comes from many different sources. Finding a way to transform the data into a coherent and usable set is crucial for the machine learning process.

Some of the steps of the process usually include collection, integration, cleaning, transformation, etc. It may also include data labeling, validation, visualization, data enrichment, feature engineering.

Data collection

As previously mentioned, data is crucial to achieving precise results. Some important features of data should be taken into consideration when choosing a data set to work with:

- Size. The size of the dataset is one of the main characteristics, and in general it is considered that a simpler model which is used on a larger data set might do better than using a more complex machine learning method on a small dataset.
- Reliability. Making predictions which can later be used is impossible without data that is reliable. Unreliable data might include label errors, noise, or data that is not entirely relevant to the project at hand. Other possible problems are duplicates, omitted values etc.
- Feature representation. Some values in the data set might need to be normalized. Handling outliers might also be needed.

Possible issues that might occur with data collection also include different biases. Reporting bias, for example occurs when a data set only covers a fraction of the actual data. It can include citation bias, language bias, location bias, outcome reporting bias etc. Selection bias is especially important when talking about data collection – it occurs when proper randomization is not achieved during the data collection phase.

Data might come from many different sources. It might relate to certain events, or it might be just a snapshot of information. In general, data can be derived from sensors that collect information, tracking social media, using surveys and forms, focus groups, transactions etc.

Data integration

Data integration is the process of combining data collected through different means. It usually includes ingestion and transformation. The aim of data integration is to is to store the transformed data in a data warehouse, data lake, or data lakehouse.

According to Qlik, there are five approaches to data integration. The following illustration shows where they sit in the data management process:

The first part of the illustration shows that the data management process consists of data sources, data processing, data storage and analytics/apps. Data processing includes:

- ETL – converts raw data through three steps – extracting, transforming, and loading. The data is transformed in a staging area and then loaded, usually in a data warehouse.
- ELT – with the ELT process, data is first loaded and then transformed when it is already in the target system.
- Data streaming – the data streaming process consists of constantly moving data in real-time from source to target system.
- Application integration – application integration allows different applications to sync data between each other.
- Data virtualization – similar to streaming, in the sense that it also delivers data in real time. The difference is that data is delivered only when requested by a user or application.

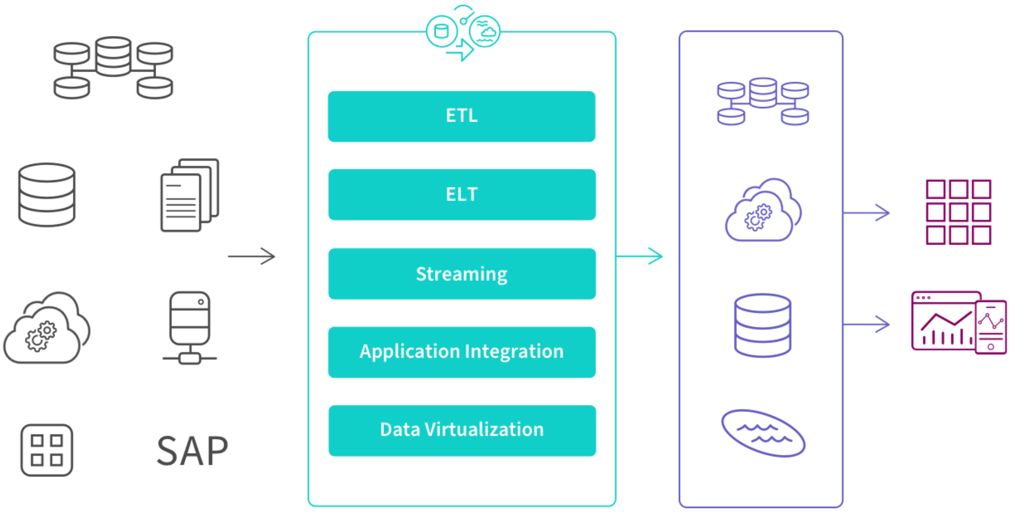

Fig. 1

Data cleaning

Data cleaning is the process of removing incorrect, duplicate or incomplete data from a data set. As mentioned before, there is no exact way of cleaning data – it depends on the needs of the organization and the specifics of the data. Nevertheless, there are a few main steps which are usually involved in the data cleaning process:

- Removing duplicate data

Removing duplicate data can begin at data collection and transformation. When combining data from different sources, duplicate data can be hard to avoid. Removing repeated values is crucial to receiving accurate results.

- Removing outliers

Removing data that does not fit within the analyzed data when it's the result of improper data entry for example, this will improve the results from the analysis. However, the existence of an outlier does not mean it is incorrect data that needs to be removed.

- Handling missing data

Many algorithms do not accept missing data, so dealing with it is crucial. Some options of handling missing data are inputting missing data based on other observations or dropping observations with missing values.

- Validation

Validating data includes making sure the data makes sense and confirming that it follows the appropriate rules for its field.

Data transformation

Data transformation is a crucial step in the data preparation process for AI and machine learning applications. It involves converting raw data into a format that is suitable for analysis, modeling, and training machine learning algorithms. The goal of data transformation is to improve the quality and relevance of the data for the specific tasks at hand. Here are some key aspects of data transformation in the context of AI:

Scaling

Scaling in AI data preparation refers to the process of adjusting the scale or range of numerical features in a dataset to make them more suitable for machine learning algorithms. Many machine learning models are sensitive to the scale of input features, and scaling is applied to ensure that all features contribute equally to the model training process. Scaling helps in improving the convergence speed and stability of certain optimization algorithms used in machine learning.

Normalization and Standardization

Normalization and Standardization is an approach to bring numerical features to a common scale. This ensures that features with different units or scales do not disproportionately influence the model. Often normalization and standardization are referred as main method for achieving data scaling.

Encoding Categorical Variables

Encoding Categorical Variables is an approach of converting categorical variables into a numerical format that can be used by machine learning algorithms. This technique is used because machine learning algorithms typically work with numerical data. In other words. encoding stands for converting categorical data into a numerical format. Common techniques include one-hot encoding, label encoding, or ordinal encoding.

Handling Outliers

An outlier is an observation in a data set that is distant from other observations. Outliers can significantly impact the performance of machine learning models. Identification of such values is very important. Various methods like trimming, winsorizing, or transformation can be applied to handle outliers.

Discretization

Discretization is an approach to preparing data for machine learning that allows transforming continuous variables, such as time, temperature, or weight, into discrete ones. Consider a dataset that contains information about people’s height. The height of each person can be measured as a continuous variable in feet or centimeters. However, for certain ML algorithms, it might be necessary to discretize this data into categories, say, “short”, “medium”, and “tall”. This is exactly what discretization does. It helps simplify the training dataset and reduce the complexity of the problem. Common approaches to discretization span clustering-based and decision-tree-based discretization.

Dimensionality reduction

Dimensionality reduction stands for limiting the number of features or variables in a dataset and only preserving the information relevant for solving the problem. This can be particularly useful when dealing with high-dimensional data.Techniques like principal component analysis (PCA), linear discriminant analysis (LDA) or t-SNE (t-distributed stochastic neighbor embedding) can be used.

Log transformation

Another way of preparing data for machine learning, log transformation, refers to applying a logarithmic function to the values of a variable in a dataset. It is often used when the training data is highly skewed or has a large range of values. Applying a logarithmic function can help make the distribution of data more symmetric.

Data Imbalance

Data imbalance refers to a situation in a classification problem where the distribution of classes in the training dataset is not equal. In other words, one or more classes have significantly fewer instances compared to other classes. The approach of nadling imbalanced classes in classification tasks is very important. Techniques such as oversampling, undersampling, or using different sampling strategies can be applied to address class imbalance.

Handling Skewed Distributions

A skewed distribution, also known as asymmetry, is a statistical term that describes the shape of a probability distribution. In a skewed distribution, the data points cluster more toward one side of the distribution than the other, creating a tail that extends in one direction. This approach refers application of techniques logarithmic or square root transformations to help make the data more symmetric and improve model performance.

Temporal Data Processing

For time-series data, features like date and time may be decomposed into components such as year, month, day, or hour. Lag features or rolling statistics can also be created to capture temporal patterns.

Feature engineering

The scope of data transformation also included feature engineering. Feature engineering is more than a transformation technique or a step in the process of preparing data for machine learning. Feature engineering involves a combination of statistical, mathematical, and computational techniques, including the use of ML models, to create features that capture the most relevant information in the data.

Model features are the inputs used by machine learning during the training process and they are significant to the accuracy of the process. Feature engineering is a key part of preparing data for machine learning. It consists of adding and constructing variables to a dataset to better understand the data and achieve a better performance of the machine learning model. It is important to have knowledge of the business problem, as well as the data source, to achieve more accurate results.

Feature engineering includes a few main processes:

- Feature creation – creating new variables.
- Transformations – transforming features from one representation to another.
- Feature extraction – selecting features from a dataset which will help find more meaningful information.
- Exploratory data analysis (EDA)– exploring the properties of data to create new hypotheses, find patterns, etc.

Feature engineering, if done correctly, can optimize a dataset to contain more important factors which affect the business model.

Data labeling

In the context of machine learning, data labeling is the process of identifying raw data and adding meaningful labels to it, in order for a machine learning algorithm to be able to learn from it. For example, labels might indicate what a photo contains, what words are used in a recording etc. Data labeling usually consists of users making judgements about unlabeled data. The algorithm then uses the labels provided by humans to learn in a process called training.

2.Data preparation platforms

According to Gartner, data preparation tools help accelerate the delivery of curated data, allowing the users to find anomalies in their data. Some popular data preparation platforms include:

Alteryx

Alteryx is a data analytics platform that provides automated data preparation and analytics with the help of machine learning and AI. The platform allows the user to integrate different data sources, and then build data pipelines to extract and load different sources into targets such as cloud data warehouses or cloud data lakes. Alteryx provides the user with the ability to automate, document, share and scale the data preparation process with low-code and no-code tools.

Talend

Talend is described as an end-to-end platform which combines data integration, data quality, and data governance in a low-code tool. It can be deployed on premises, as well as in the cloud, multi-cloud, or hybrid cloud. According to Gartner, Talend is focused on improved data quality, with the platform allowing the user to measure the validity and completeness of datasets. The platform provides a browser based self service tool which helps shorten the time it takes to clean data. It also lets the users automate data preparation and turn it into a reusable process, using data from different sources.

Power BI

Power BI provides self-service data preparation capabilities. It uses dataflows to ingest, cleanse, transform, integrate and enrich data from different sources – on-premise ones, as well as cloud based ones, which includes Dynamics 365. These dataflows store data in Azure Data Lakes, which means users can utilize Azure Machine Learning, Azure Databricks, and Azure SQL Datawarehouse for advanced analytics. Nevertheless, Power BI seems to be mainly a data visualization tool, with some users suggesting it is better to prepare data as much as possible prior to loading it into Power BI.

RapidMiner

RapidMiner is an end-to-end data science platform which allows users to create machine learning pipelines. RapidMiner allows for scaling from on-premises servers to cloud implementations. The platform allows users to extract and transform data from many sources, which include PDFs, spreadsheets and text files. RapidMiner works with both structured and unstructured data, with the ability to automate data transformation tasks.

DataRobot

DataRobot AI Platform is a full AI lifecycle platform used for predictive and generative AI. It interfaces with DataRobot Data Prep to assist with data preparation. Data Prep allows the user to explore the data, with the ability to clean, combine and shape it in a proper format for machine learning. It offers a user interface which visualizes the data in a spreadsheet style, requiring no coding.

Informatica Data Prep

Informatica Data Prep provides cloud data preparation capabilities with low-code/no-code on cloud data warehouses. It provides a data catalog with metadata which can be useful in understanding data. It also provides data profiling capabilities to find anomalies, outliers etc. The platform is also useful for data transformation and enforcing data governance and compliance policies.

3.Conclusion

Undoubtedly, the preparation of data for use in AI and Machine Learning is of utmost importance. It demonstrates the serious number of techniques that are defined and the coverage of different aspects of working with data. Clear definitions and desired end results are a good basis for seeking automation of data preparation techniques, which in the context of large volumes of data is already a serious necessity. The technological world offers a significant set of interesting solutions for automating individual data preparation tasks or of the entire processes. The wide variety of algorithms and their specific requirements for the data they can work with pose the interesting research task of finding automated solutions that not only enable the use of software solutions, but also build complete solutions that support the detection of the potential of the data and application of these techniques that can be automatically identified.

References

1. Stefanov, G. (2019). Analysis of Cloud based ETL in the Era of IoT and Big Data. In Proceedings of International Conference on Application of Information and Communication Technology and Statistics in Economy and Education. 198-202. ICAICTSEE.
2. E. Karkalikova, A. Murdjeva, Organization of Data in Data Lake – Real-Life Practice, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria.
3. Genka Miteva, Alexandrina Murdjeva CHARACTERISTICS AND PREPARATION OF DATASETS FOR MACHINE LEARNING ALGORITHMS, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria.
4. Geno Stefanov, ANALYSIS OF SERVERLESS CLOUD DATA WAREHOUSE SOLUTIONS, 11th International Conference on Application of Information and Communication Technology and Statistics in Economy and Education ICAICTSEE– 2021, Sofia, Bulgaria.
5. Delchev, D., Lazarova, V., Big Data Analysis Architecture, Economic Alternatives, 2021, Issue 2, pp. 315-328
6. Verdonck, T., Baesens, B., Óskarsdóttir, M. et al. Special issue on feature engineering editorial. Mach Learn (2021). https://doi.org/10.1007/s10994-021-06042-2
7. Edward Baumann,Charles Hsu,Hayley Buba,Taylor Cox,An Introductory Approach to Time-Series Data Preparation and Analysis, Annual Conference of the PHM Society, October 2023, DOI: 10.36001/phmconf.2023.v15i1.3561
8. Mohamad Fariq Rahmat, Zed Zulkafli, Asnor Juraiza Ishak and others, Supervised feature selection using principal component analysis, Knowledge and Information Systems, November 2023, DOI: 10.1007/s10115-023-01993-5
9. Lamia AbedNoor Muhammed, Role of data normalization in k-means algorithm results, • AL-KADHUM 2ND INTERNATIONAL CONFERENCE ON MODERN APPLICATIONS OF INFORMATION AND COMMUNICATION TECHNOLOGY, March 2023, DOI: 10.1063/5.0119267
10. What is Data Preparation? Accessed: Oct. 29, 2023. [Online]. Available: https://aws.amazon.com/what-is/data-preparation/

11. The Size and Quality of a Data Set; Google [Online]. Available: https://developers.google.com/machine-learning/data-prep/construct/collect/data-size-quality
12. What is Data Integration? Talend, [Online], Available: https://www.talend.com/resources/what-is-data-integration/
13. Data Integration, What it is, why it matters, and best practices. This guide provides definitions, examples and practical advice to help you understand the topic of data integration, QLIK, [Online], Available https://www.qlik.com/us/data-integration
14. https://www.tableau.com/learn/articles/what-is-data-cleaning
15. What is Data Labeling?; Amazon, [Online]. Available: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/
16. What Is Data Preparation? Cut through the noise with data ready for analytics, ML and more; Informatica, [Online]. Available: https://www.informatica.com/resources/articles/what-is-data-preparation.html#4
17. What is Feature Engineering?; Amazon, [Online]. Available https://aws.amazon.com/what-is/feature-engineering/
18. Feature Engineering; What is Feature Engineering for Machine Learning, Data Robot, [Online]. Available :https://www.datarobot.com/wiki/feature-engineering/
19. What is Feature Engineering — Importance, Tools and Techniques for Machine Learning; Harshil Patel, Towards Data Science, Aug 30, 2021, https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10