SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A+' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

COURSE NAME : 19CSO302 & Introduction to Data Analytics

III YEAR/ VI SEMESTER

UNIT – II PREPROCESSING AND VISUALIZATION

Topic: Data preprocessing : Error Types

P.Poonkodi

Assistant Professor

Department of Computer Science and Engineering

Data preprocessing

- Introduction
- Error Types
- Error Handling
- Filtering
- Data Transformation
- Data Merging

Introduction

- Process of transforming raw data into an understandable format
- important step in data mining as we cannot work with raw data
- quality of the data should be checked before applying machine learning or data mining algorithms

Why is Data Preprocessing Important?

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.
- **Consistency:** To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

4 major tasks in data preprocessing

Data preprocessing

- Data Cleaning
- Data cleaning is the process of removing incorrect data, incomplete data, and inaccurate data from the datasets, and it also replaces the missing values. Here are some techniques for data cleaning:
- **Handling Missing Values**
  - Ignore the tuples
  - Fill the Missing values
- **Handling Noisy Data**

Handling Missing Values

- Standard values like “Not Available” or “NA” can be used to replace the missing values.
- Missing values can also be filled manually, but it is not recommended when that dataset is big.
- The attribute’s mean value can be used to replace the missing value when the data is normally distributed
  wherein in the case of non-normal distribution median value of the attribute can be used.
- While using regression or decision tree algorithms, the missing value can be replaced by the most probable value.

Handling Noisy Data

- Noisy generally means random error or containing unnecessary data points. Handling noisy data is one of the most important steps as it leads to the optimization of the model we are using Here are some of the methods to handle noisy data.
- **Binning:** This method is to smooth or handle noisy data. First, the data is sorted then, and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin. **Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin; **Smoothing by bin median**: In this method, the values in the bin are replaced by the median value; **Smoothing by bin boundary**: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.

Handling Noisy Data

- **Regression:** This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.
- **Clustering:** This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

Data Transformation

- This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

Normalization:
It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

Attribute Selection:
In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

Discretization:
This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

Concept Hierarchy Generation:
Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

Data Transformation

- The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods for data transformation.
- **Smoothing:** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset. By smoothing, we can find even a simple change that helps in prediction.

Data Transformation

- **Aggregation:** In this method, the data is stored and presented in the form of a summary. The data set, which is from multiple sources, is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good, the results are more relevant.
- **Discretization:** The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, or 6 pm-8 pm).
- **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

Data Integration

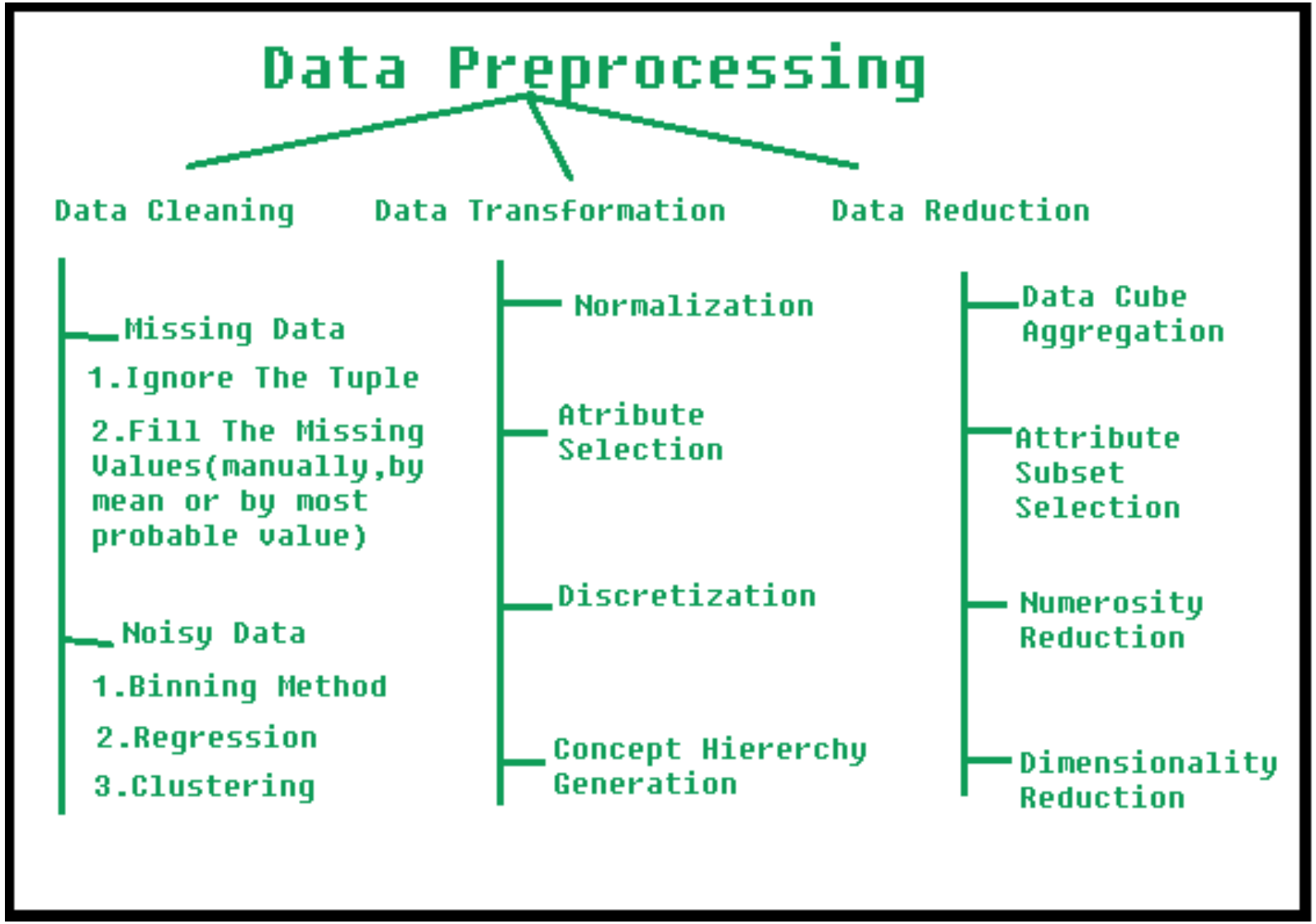
- The process of combining multiple sources into a single dataset. The Data integration process is one of the main components of data management. There are some problems to be considered during data integration.
- **Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.
- **Entity identification problem:** Identifying entities from multiple databases. For example, the system or the user should know the student *id of one database and student*name of another database belonging to the same entity.
- **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ. The attribute values from one database may differ from another database. For example, the date format may differ, like “MM/DD/YYYY” or “DD/MM/YYYY”.

Data Reduction

- This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. Some of the data reduction techniques are dimensionality reduction, numerosity reduction, and data compression.
- **Dimensionality reduction:** This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space, and computation time is reduced. When the data is highly dimensional, a problem called the “Curse of Dimensionality” occurs.

Data Reduction

- **Numerosity Reduction**: In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
- **Data compression**: The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression, it is called lossless compression. Whereas lossy compression reduces information, but it removes only the unnecessary information.


Data Preprocessing
Data Cleaning
Data Transformation
Data Reduction
Missing Data
1.Ignore The Tuple
2.Fill The Missing Values(manually,by mean or by most probable value)
Noisy Data
1.Binning Method
2.Regression
3.Clustering
Normalization
Atribute Selection
Discretization
Concept Hierarchy Generation
Data Cube Aggregation
Attribute Subset Selection
Numerosity Reduction
Dimensionality Reduction

Data Error

- Data error refers to inaccuracies or inconsistencies in the data that can occur during the collection, processing, or storage stages. Data errors can also occur due to a diverse set of issues that can arise in datasets, ranging from missing data and duplicates to outliers, inconsistencies, and inaccuracies.
- The significance of data errors lies in their potential to undermine the integrity and reliability of data-driven processes and decisions
- Data errors can lead to financial losses, legal liabilities, compromised safety, and reputational damage
- addressing data errors is crucial for ensuring data quality, trustworthiness, and the credibility of organizations and systems relying on data

Types of Data Errors

- **Missing data** refers to the absence of values in a dataset, which can hinder the training and performance of machine learning models by reducing the amount of information available for analysis.
- **Duplicate data** occurs when identical or nearly identical records exist within a dataset, potentially skewing model training and leading to redundancy in predictions.
- **Inaccurate data** encompasses information that contains errors or mistakes, undermining the reliability and precision of machine learning models.
- **Inconsistent data** refers to data that contradicts itself or exhibits variations in format or content, making it challenging for models to establish meaningful patterns.

Types of Data Errors

- **Outliers** are data points that deviate significantly from the majority of the dataset, potentially causing machine learning models to produce biased or less accurate predictions.
- **Data imbalance** indicates an unequal distribution of classes or categories within a dataset, which can result in models being biased towards the majority class and performing poorly on minority classes.
- **Bias** in data represents a systematic and non-random distortion in the dataset, introducing unfairness and prejudice into machine learning models.
- **Transformation errors** occur when data is not properly preprocessed or normalized, leading to model inefficiencies and reduced predictive accuracy.

causes data errors

- **Annotation Errors:** Mistakes made by individuals during data input, validation, or processing.
- **Incomplete Data:** Missing or incomplete information within a dataset can introduce errors and limit the validity of analyses and conclusions.
- **Inadequate Validation:** Errors may occur when data validation and quality checks are insufficiently implemented, allowing inaccurate or inconsistent data to persist.
- **Lack of Documentation:** Poorly documented data sources and procedures can lead to misunderstandings and errors in data interpretation and usage.

Prevent Data Errors

- Mitigating data errors involves a combination of strategies, including:
- **Data Validation:** Implement data validation checks, including data type verification, range constraints, and pattern matching, to ensure the accuracy and proper formatting of entered data while rejecting invalid entries.
- **Data Cleaning:** Using automated tools and manual processes to identify and rectify errors in datasets.
- **Data Quality Monitoring:** Continuously monitoring data for errors and inconsistencies using data quality frameworks and metrics.
- **Documentation:** Maintaining clear documentation of data sources, transformations, and cleaning procedures to aid in error identification and correction.

Prevent Data Errors

- **Automation:** Automate data entry processes using software tools and scripts to minimize human errors and gather data from trustworthy sources.
- **Data Governance:** Establishing data governance practices and policies to ensure data quality and accountability within organizations.

Conclusion

- Data errors are pervasive and consequential issues that can affect organizations across various industries. Addressing data errors is essential for maintaining data integrity, enabling accurate decision-making, and upholding the trustworthiness of data-driven systems. Vigilance, proactive measures, and ongoing monitoring are key to managing data errors effectively.

References

- Runkler TA, “Data Analytics: Models and algorithms for intelligent data analysis”, Springer, Third Edition 2020

- https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/#:~:text=The%20steps%20involved%20in%20data,Feature%20selection%2C%20and%20Data%20representation.


Thank You!