International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1206-1209 www.ijaem.net ISSN: 2395-5252

A review on Data Preprocessing Techniques in Data Mining

P. Swisstac Bravin
Assistant Professor
Hindustan College of Arts & Science
Padur, Chennai

Submitted: 10-05-2022 Revised: 19-05-2022 Accepted: 22-05-2022

ABSTRACT: Today 's real-world databases are highly at risk to noise, missing, and inconsistent data because of their largesize andthey are from heterogeneous sources. The dataset must be properly preprocessed before the discovery of useful information/knowledge. Data preprocessing is a crucial step in improving data efficiency. Data preprocessing is a data mining procedure that involves the preparation and manipulation of a dataset while also attempting to improve the efficiency of knowledge discovery. Cleaning, integration, transformation, and reduction are some of the techniques used in preprocessing. This study provides a comprehensive overview of data preparation strategies used in data mining.

Keywords: Data Preprocessing, Data Cleaning, DataTransformation, Data Reduction

I. INTRODUCTION

The process of extracting meaningful patterns and models from a large dataset is known as data mining. The revealed patterns and models have a big impact on how people make decisions. For data mining, data quality is critical. Raw data contains missing values, partial data, and inconsistent data. Before mining the data, it is necessary to preprocess it. Preprocessing is essential to improve data efficiency. Figure 1 shows the nine-step iterative and participative knowledge discovery method. The process of recurrence at each phase, which may require going back to prior steps. The method begins with the identification of KDD goals and ends with the application of what has been learnt [1].As indicated in Figure 1, data must be picked in order to determine the target data, and then the selected data must be preprocessed in order to improve its reliability. After the data has been preprocessed, it must be translated into a format suitable for data mining. Then, mining procedures like as clustering, classification, regression, and others will be used to extract patterns, which will be interrupted and assessed in the final stage.

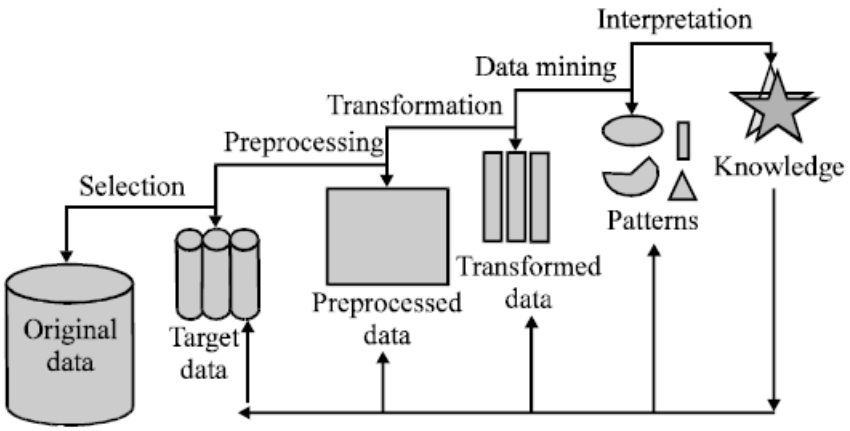

Figure 1: Steps in Knowledge Discovery

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1206-1209 www.ijaem.net ISSN: 2395-5252

II. METHODS OF DATA PREPROCESSING

The data is particularly sensitive before preprocessing, with missing values and discrepancies. Data quality has an impact on data mining outcomes. To increase data efficiency and quality, raw data is preprocessed. Data preprocessing is the most critical phase in the data mining process. Preprocessing techniques are classed as follows, as seen in figure 2.

- DataCleaning
- DataIntegration
- DataTransformation
- DataReduction

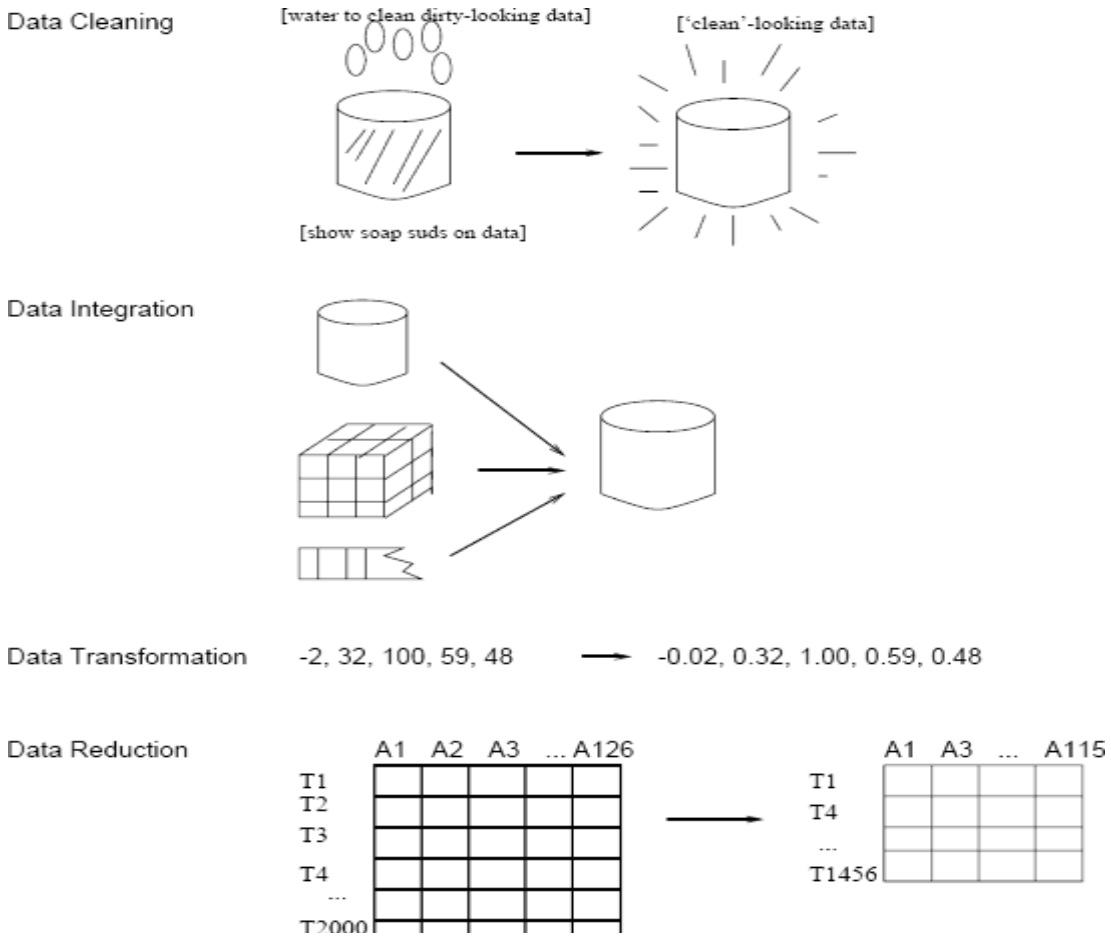

Fig 2: Different Forms of Data Preprocessing

2.1 DATA CLEANING

The data in real-world databases is incomplete (it lacks attribute values or some attributes of interest, or it only contains aggregate data), noisy (it contains errors or outlier values that diverge from the expected), and inconsistent (e.g., containing discrepancies in the department codes used to categories items). It's possible that the tuples in the databases are duplicated. As a result, data cleansing is required. Data cleansing can be done by filling in missing values, smoothing noisy data, finding or deleting outliers, and resolving inconsistencies.

MISSING VALUES: If the tuples have missing value for several attributes, then the missing values can be filled in for the attribute by different methods described below:

1. Ignore the tuple: When the class label is missing, this is done (assuming the mining task involves classification or description). This method is effective when the tuple contains several attributes with missing values.
2. Fill in the missing value manually: This method takes a lot of time and it is not feasible for a large data set with many missing values.
3. Use a global constant to fill in the missing

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1206-1209 www.ijaem.net ISSN: 2395-5252

value: In this type, the missing values are replaced by some constant such as a label like "Unknown", or -∞. This method is not recommended because the mining program may mistakenly think that they form an interesting concept sinceall have a value in common.

4. Use the attribute mean to fill inthe missing value: Using this method, the mean of the attribute can be used to fill in the missing values.
5. Use the attribute mean for all samples belonging to the same class as the given tuple: The mean of the attribute of the same class is used to fill the missing values.
6. Use the most probable value to fill in the missing value: Using a Bayesian formalism or decision treeinduction, the most probable valuemay be determined.

Method 4 is a popular strategy to fill the missing values Method 4 uses most information from the present data to predict missing values.

NOISY DATA: Noisy Data is a random error or variance in a measured variable. For example, given a numeric field say, price, how can the data be "smoothed" to remove the noise? The following techniques of data smoothing describes this.

1. **Binningmethods**:In a sorted data, this method smooths the data byconsultingtheneighborhood, or values around it. The sorted values are distributed into a number of 'buckets', or bins. Since binning methods look up the neighborhood of values, they perform local smoothing values around it.
2. **Clustering:**Clustering means grouping similar values. Using this method outliers may be detected.
3. **Combined computer and human inspection**: The combination ofcomputerandhuman inspection can identify the outliers. Forexample,to identify outlier patterns in a handwritten character database for classificationan information-theoretic measure was used.
4. **Regression:** Using regression, data can be smoothed by fitting the data to a function. Linear regression and Multiple linear regression are used to find the best fit of two variables and more.

INCONSISTENT DATA: The data recorded for some transactions may contain inconsistent data. The data inconsistencies may be corrected manually using external references. For example, data entry errors may be corrected by checking manually the paper.To detect the violation of known data constraints, knowledge engineering tools may also be used. For example, known functional dependencies between attributes can be used to find values contradicting the functional constraints.

2.2 DATA INTEGRATION

Data integration is the process of merging data from multiple sources into a single data storage, similar to data warehousing.Multiple databases, data cubes, or flat files could be used as sources. During data integration, there are numerous difficulties.How can a data analyst or a computer tell the difference between a customer_id in one database and a cust_number in another database that pertain to the same entity? Metadata exists in databases and data warehouses. Meta refers to information about information. Metadata can be utilized to help prevent data integration issues. Another key issue is redundancy. It's possible that an attribute obtained from another table is redundant. Inconsistencies in attribute or dimension name can potentially cause redundancies.

2.3 DATA TRANSFORMATION

In this method, the data are transformed or consolidated into forms which is appropriate for mining. Data transformation can be categorizedinto the following:

1. Normalization: In normalization, the attribute data are scaled to fall within a small specified range, such as -1.0 to 1.0, or 0 to1.0.
2. Smoothing:Thismethod works to remove the noise from data. Smoothing techniques include binning, clustering, andregression.
3. Aggregation: The summary or aggregation operations are applied to the data. This step is used toconstruct a data cube for analysis of the data at multiplegranularities.
4. Generalization of the data: Through the concept of hierarchies, low level or 'primitive' (raw) data are replaced by higher level. For example, categorical attributes, like street, can be generalized to higher level concepts, like city or county. Similarly, values for numeric attributes, like age, may be mapped to higher level concepts, like young, middle-aged, andsenior.

2.4 DATA REDUCTION

It takes a very long time to process

complex and large amount of data making such analysis impractical or infeasible. The integrity of the original data is not compromised in this method. Data reduction means reducing the volume or reducing the dimensions of databases.

Strategies for data reduction include the following.

1. Data cube aggregation: In the construction of a datacube, aggregation operations are applied to the data.
2. Dimension reduction: The dimensions may be detected andremoved from the data set.
3. Data compression: To reduce the data set sizeencoding mechanisms are used. Wavelet transformation and Principal Component Analysis methodsare used for data compression.
4. Numerosity reduction: The data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data e.g. regression and log-linear models), or nonparametric methods such as clustering, sampling, and the use ofhistograms.
5. Discretization and concept hierarchy generation: The raw data values for attributes are replaced by ranges or higher conceptual levels. The mining of data at multiple level of abstraction is done in concept hierarchies.

III. CONCLUSION

Data preprocessing is essential for successful data processing in data mining since real-world data is frequently incomplete, noisy, and inconsistent. Data preprocessing includes cleaning, integrating, manipulating, and decreasing data. Data cleaning processes can be used to fill in missing values, smooth noisy data, identify outliers, and remove data inconsistencies. The practice of merging data from many sources is known as data integration.The data transformation method guarantees that the data is in the proper format for mining. Data cube aggregation, dimension reduction, data compression, numerous reductions, and discretization are examples of data reduction techniques that can be used to provide a reduced representation of the data while minimizing material loss. Despite the advancement of numerous data preparation techniques, data preprocessing remains a necessity.

REFERENCES

[1]. Han J and M. Kamber, 2006, Data Mining: Concepts and Techniques. 2nd Edn, Morgan Kaufmann Publisher, San Fransisco, USA ISBN-13 978-1558609013.

[2]. Jun D., Data Preprocessing, The University of Western Ontario (2013), 1-48.

[3]. Dharmarajan R and R Vijayasanthi, 2015. An Overview on data preprocessing methods in data mining. Intl. J. Sci Res. Dev., 3: 3544-3546

[4]. Baskar S.S., Arockiam L., Charles S., A Systematic Approach on Data Pre-processing in Data Mining, An International Journal of Advanced Computer Technology 2 (11) (2013).

[5]. A.Familia,Wei-MinShenb , RichardWeberc, EvangelosSimoudisd, Data preprocessing and intelligent data analysis, Intelligent Data AnalysisVolume 1, Issues 1–4, 1997, Pages 3-23