I J R B A T, Issue (X) Vol (III) Sept 2022: 106-111
A Double-Blind Peer Reviewed & Refereed Journal
e-ISSN 2347 – 517X
Original Article

INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
 www.ijrbat.in

REVIEW OF DATA PRE-PROCESSING TECHNIQUES IN DATA MINING

[1]Karuna C. Khobragade and [2]Pankaj B. Dhumane
[1]Assistant Professor, PG Department of Computer Science,
Dr. Khatri Mahavidyalaya, Tukum, Chandrapur, Maharashtra, India.
[2]Assistant Professor, PG Department of Computer Science,
Sardar Patel Mahavidyalaya, Chandrapur, Maharashtra, India.
Correspondence Email: k12karuna@gmail.com

Communicated :16.05.2022 | Revision : 28.05.2022 & 19.06.2022 Accepted : 26.07.2022 | Published: 30.09.2022

ABSTRACT: The process of extracting relevant patterns and models from a large dataset is referred to as data mining. In a decision-making task, these models and patterns are quite useful. The quality of data is crucial for data mining. Missing values, noisy data, incomplete data, inconsistent data, and outlier data are all common characteristics of raw data. As a result, processing these data prior to mining is critical. Data pre-processing is a necessary step in improving data efficiency. Data pre-processing is a common data mining stage 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 examples of pre-processing procedures. This research provides a comprehensive overview of data preparation strategies used in data mining.
Key words : - Data Mining, Data Pre-processing, Data Cleaning, Data Integration, Data Reduction, Data Transformation..

INTRODUCTION :

In day-to-day life the amount of data being generated and stored is growing exponentially. Users who use these data are expecting more sophisticated information from them. From the ocean of these data, information is digging out with help of data mining.

"Data mining means to mine or extract "information" or "knowledge" from large amount of data", and "Data Pre-Processing" is one of the process/techniques used to carry out data mining successfully.

Data pre-processing involved major steps [1,2,3]

1) Data Cleaning
2) Data Integration
3) Data Reduction
4) Data Transformation

a) Data Cleaning:

The data cleaning routine "clean" the data by filling in missing values, smoothing noisy data, identifying and removing outlier and resolving inconsistencies. (Data Cleaning, also called data cleansing or scrubbing is the process of detecting and removing errors and inconsistencies from data in order to improve the data quality).

b) Data Integration:

This technique involves merging data from numerous heterogeneous data sources, such as multiple databases, data cubes, or files, into a cohesive data store and presenting a uniform representation of the data.

c) Data Reduction:

Techniques for data reduction can be used to provide a reduced representation of a data set that is substantially smaller in volume but produces the same analytical results [1][4][5].

d) Data Transformation:

Data transformation refers to the task of transforming data from one format to another [6]. Under this technique data is transformed in the way which is ideal and suitable for data mining.

Basic methods for data cleaning are as follows-

i) Missing Values: This occurs when a value or piece of data in the data/tuples is missing. If it is discovered that several tuples do not have any recorded value for numerous attributes, then that missing value for that attribute will be filled by using some methods like [1][2][8]-

a) Ignoring tuple: Normally this is done when the class label has missing. When the tuple that contains numerous attributes having missing values, then that approach is ineffective. This method is especially bad when an amount of missing values varies a lot by attribute. We don't use the remaining attribute values in the tuple since we are ignoring that tuple [1][2].

The method of ignoring the tuple is suitable only when the datasets are huge and numerous values inside a tuple are missing [1][2].

b) Missing values filled in by manually: This method is very time-consuming in general, and it may not be practical / feasible for huge data sets with various / multiple missing values [1][8].

This method works well with small data sets that have some missing values.

c) Using global constant for fill up the missing values: Substitute some constant for all missing attribute values such as a label like "Unknown" or $-\infty$ or like "NA" etc. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown." Hence, although this method is simple, it is not recommended [1][8].

d) Using central tendency (mean or median) for an attribute to filling in missing value: The missing values can be fill by mean/median.

Mean — We can use the attribute mean to fill in the missing value. Mean is the average value. Eg. Customer average income is 30000, then we use this value to replace missing value for income.

When the data has no outliers in that case missing value will be replaced by mean value. Mean will be affected by outliers.

Median — When the data has more outliers, it's best to replace them with the median value. Median is the middle value (50%). (If data is numerical then mean and median is used, if data is categorical the mode is used)

e) Use of most probable value to fill in missing values: This may be determined with regression, inference-based tools using **Bayesian formalism, or decision tree induction** [1][8]. Eg. In customer table, if customer_income is missing for particular customer then by calculating the average customer income we will fill the missing value.

ii) Noisy Data: "Noise" is nothing but an error (random error) or the variance (difference) in a measured value or variable. "Data that is noisy is useless. It contains any data that machines cannot understand or interpret appropriately, like an unstructured text. The binning technique, regression, and outlier data analysis can all be used to smooth out noisy data [1].

a) Binning Technique: In this approach, Data is first sorted, and then that sorted data or values can be dispersed into number of buckets called bins. This method looks at the values in the neighbourhood, or the values that are close by and conduct local smoothing.

Eg. (4,8,15,21,21,24,25,28,34)

Equal Portioned Bin	Bin Mean	Bin Boundaries
Bin1: 4,8,15	Bin1: 9,9,9	Bin1: 4,4,15
Bin2: 21,21,24	Bin2: 22,22,22	Bin2: 21,21,24
Bin3: 25,28,34	Bin3: 29,29,29	Bin3: 25,25,34

b) Regression: A data smoothing approach in which data values are conformed to a function is called regression. The regression can be linear or multiple. Linear regression has only one independent variable and multiple regression has more than one independent variables [1].

Linear regression involves finding the "best" line to fit two variables (or attributes) so that one variable can be used to predict the other.

Multiple linear regression is an extension of linear regression, where more than two attributes (variables) are involved and the data are fit to a multidimensional surface.

c) An Outlier Analysis:

Clustering can be used to identify outliers, where clusters are groupings of similar values. Outliers are values that do not fall into any of the clusters like odd man out.

a) Data Integration:

Entity Identification (Concern) Problem: During data integration, there are lot of issues to consider. Schema and Object Integration might be challenging. Entity identification problem arises when the same field with different names present in different data sources.

Eg. Identification of Emp_id in one database and Emp_no in another database which referred as same attributes, that is very difficult task for data analyst or the computer. Matching the entity identification problem is prevented by using metadata.

Redundancy and Corelation Analysis: If one attribute is derived from other attribute or from the group of attributes then it is called as redundant attribute. Redundancies may occur because of inconsistency in an attribute and/or the naming of dimension in the resulting or final data set. Many redundancies are found/detected with the help of correlation analysis and x^2 Test (Chi- Square) [9].

Eg. If we have a data set having three attributes Pizza_Name, Is_Veg, Is_Nonveg. If any Pizza_Name is not veg, it means that Pizza is surely Non_Veg. In this case one of these attributes became redundant. It means that the two attributes are very much related to each other, and one attribute can find the other. So, we can drop either the first or second attribute without any loss of information.

Pizza_Name	Is_Veg	Is_Nonveg
Veg Loaded	1	0
Chicken Fiesta	0	1

Tuple Duplication: In addition to redundancy data integration has deal with tuple duplication. For the data integration, when the table which is not in normalized form is utilised as a source then duplicate tuples may come in the resultant data. Inconsistencies always arises between various / numerous duplications, because of inaccurate or faulty data entry as well as by updating some of (but not all) the data occurrences. Eg. In Bank Database, Customer name, Address may occur in saving account and current account also.

Data Value Conflict (Detection & Resolution): Detection & resolution of the data value conflict occur when, for some entity of real-world, the values of attribute from multiple sources can differ. This may occur because of variances or the differences in scaling, encoding or representation [1].

Eg: In one system weight attribute can be saved as a metric unit and some another system we can saved it as British Imperial unit.

b) Data Reduction:

i) Dimensionality Reduction: The technique or the process in which many random variables or the attributes which are under consideration will be reduced is called as Dimensionality Reduction. It reduces an un-important or redundant attribute by applying data encoding or transformation [10][11]. The *Wavelet Transform* and the *Principal Component Analysis* (PCA) methods are available to reduce or transform an original data into a smaller / less space. With the help of an *Attribute Subset Selection* method [1], we identify (detect) and eliminate (remove) the attributes or dimensions which are redundant or irrelevant or weakly relevant.

ii) Numerosity Reduction: The data reduction approach which replace an original data into alternative smaller forms is called as Numerosity Reduction. Two methods of numerosity reduction are-

a) Parametric method **b)** Non-Parametric method

For Parametric method data is represented using some model and for non-Parametric methods we use clustering, histograms, data cube aggregation, sampling etc [12].

iii) Data Compression: In data compression approach size of the files are reduces using some encoding mechanisms like Run-Length encoding, Huffman Encoding etc. Data Compression Technique are divided as-

a) Lossless Compression and b) Lossy Compression

In *Lossless compression* large file are reduced / compressed to smaller sized file and can be decompressed again without loss of any data. In Lossy compression large files are reduced/compressed to smaller size and can be restores it to its original format with loss of some data.

iv) Attribute Subset Selection: Any data set contain huge number of attributes. Among those attributes some attributes may be redundant and some of may be irrelevant. This Attribute Subset Selection method will find the minimum set of attributes, dropping out all those irrelevant attributes which will not so affect to the data analysis cost and data utility that will be reduced [1][13].

v) Data Cube Aggregation: In this technique data is aggregated into a simpler form. That is information is gathered and expressed in summary format.

Eg. If you need to know the sale report of your company for the year 2020,2021 then summarised the data as shown below, which will give total sale per year instead six months sale.

Year 2020

H.Y.	Sale
H1	500
H2	300

Year 2021

H.Y.	Sale
H1	600
H2	100

Year	Sale
2020	800
2021	700

c) Data Transformation:

i) Smoothing: The noise can be removed from data by applying some algorithms is done with the help of smoothing technique. Smoothing can be performed by - Binning Technique, Regression Technique and the Clustering Technique.

ii) Attribute Construction: For helping the mining process it will generate new attributes and then will be added from the given attribute set. By this method, simplification of original data and efficiency of mining process will be increase.

iii) Aggregation: In this method **s**ummary or aggregate functions are applied to create data cubes. Eg. To calculate monthly and yearly amount of sale, the daily sales data will be aggregated.

iv) Normalization: When dealing with attributes on multiple scales, normalisation is usually essential; otherwise, the effectiveness of a significant, equally important attribute (on a smaller scale) may be diluted due to the values of another attribute on a bigger scale [16]. In normalization process we scale the value by giving range. It converts all data variables into given range. Like 1.0 to 2.0**.** Normalization is performed by using the techniques like Decimal Scaling, Z-score, Min-Max Normalization etc.

v) Discretization: Data discretization is a technique for transforming a large number of data values into smaller ones, making data interpretation and management easier. To put it another way, data discretization is a technique for transforming continuous data properties into a finite set of intervals with little data loss. Tin short the data is being transform into the set of small intervals. Eg. (1-10, 11-20), (Age: Young, Middle age, Senior) etc.

RESEARCH METHODOLOGY:

Research methodology is required because it allows the seamless execution of numerous procedures, results in the most efficient research feasible, giving maximum knowledge with the least amount of work, time, and money expended.

In reality, research design means to forward planning about the method which will be utilised for collecting or gathering relevant data and the relevant techniques for their analysis, while keeping the research purpose in mind.

Future Work:

a. Apart from that, it is useful for further research into emerging data pre-processing technologies.

b. Building interactive, integrated data mining environments.

c. Developing data-preparation methods and systems that are efficient and effective for single and many data sources while taking into account both internal and external data.

d. Investigating effective data preparation methods for Web intelligence.

CONCLUSION:

This paper provides an overview of data pre-processing methods with some examples.

In data mining, the data preparation is responsible for identifying quality data from the data provided by data pre-processing systems. Indeed, data preparation is very important because real-world data is impure i.e. it contain incomplete, inconsistent, noisy and missing some value. But the high-performance mining systems require quality data. Quality data yields concentrative patterns.

In this paper, we have argued for the importance of data preparation and briefly introduced the research into data preparation, where the details of each achievement can be found in this special issue. By way of summary, we now discuss the possible directions of data preparation.

The diversity of data and data-mining tasks deliver many challenging research issues for data preparation. Below we would like to suggest some future directions for data preparation.

The goal of data pre-processing is to provide quality data for any type of mining like that data mining, text mining and web mining. I conclude that data pre-processing techniques have an efficient, effective and important role in preparation, analysis, process large data-scale.

REFERENCES:

Book “DATA MINING: Concept and Techniques” 3rd edition, Jiawei Han, Micheline Kamber, Jian Pei.

Data Pre-processing in Data Mining – GeeksforGeeks, Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Data Pre processing in Data Mining -A Hands On Guide - Analytics Vidhya

data-cleaning-book-chapter.pdf (cpb-us-w2.wpmucdn.com)

Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems) (sabanciuniv.edu)

data-cleaning-book-chapter.pdf (cpb-us-w2.wpmucdn.com)

https://slideplayer.com/slide/253777/1/images/6/Figure%3A+Forms+of+data+preprocessing.jpg

Data Cleaning in data mining: https://www.slideshare.net/parvathiavc/data-mining-notes

https://www.javatpoint.com/redundancy-and-correlation-in-data-mining

https://towardsdatascience.com/dimensionality-reduction-in-data-mining-f08c734b3001

https://t4tutorials.com/dimensionality-reduction-in-data-mining/

https://www.geeksforgeeks.org/numerosity-reduction-in-data-mining/

Attribute Subset Selection in Data Mining - GeeksforGeeks

Book “Exploratory Data Mining and data Cleaning “: Tamraparni Dasu, Theodore Johnson

Book "Best Practices in Data Cleaning": Jason W OSBORNE

Data Normalization in Data Mining - GeeksforGeeks.

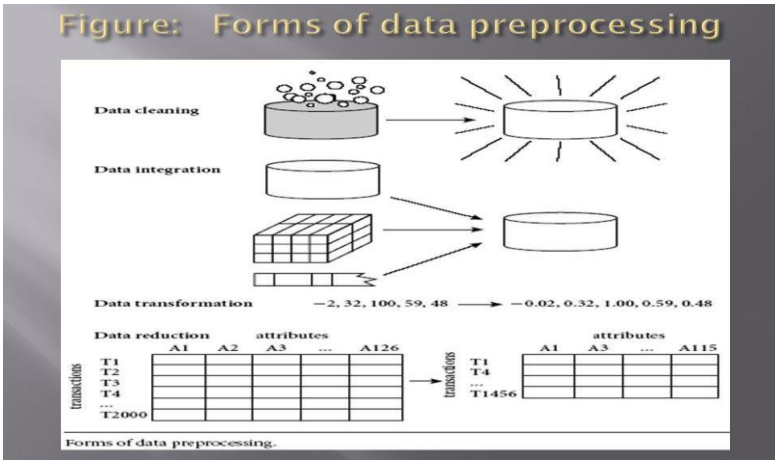

Figure: Forms of Data Pre-Processing [7]

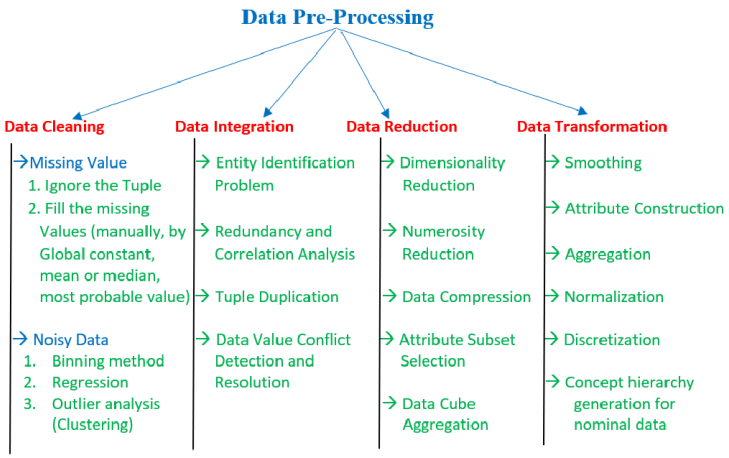

Figure1: Data Pre-Processing Techniques

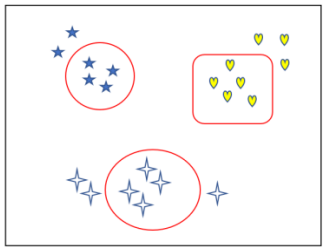

Fig: Cluster Image