Data Pre-processing Techniques in Data Mining: A Review

Pankaj Saraswat [1] and Swapnil Raj [2]

[1,2] Assistant Professor, Department of Computer Science Engineering, Sanskriti University, Mathura, Uttar Pradesh

Correspondence should be addressed to Pankaj Saraswat; pankajsaraswat.cse@sanskriti.edu.in

ABSTRACT: Data mining is the process of finding interesting patterns and models from massive datasets. In the field of natural and physical sciences, data collection, management, and analysis have evolved as the most trustworthy source of information and emergence of new findings, information, and products. The development of the most effective procedures in statistical circumstances has therefore become standard practice in the academic and industry sectors. Under actual situations, dealing with enormous datasets, there are bound to be discrepancies and abnormalities of many types that prohibit us from knowing the true results of realistic issues. These concepts and trends are helpful in decision-making situations. The quality of the data is the most important factor in data mining. For efficient information mining, computer-based data pre-processing approaches provide methods that assist the data under processing in conforming to conventional structures, hence significantly improving the efficiency of computer algorithms.

KEYWORDS: Data mining, Data preprocessing, Dataset pattern, Dataset, KDD, Knowledge Discovery.

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a method for extracting useful data from large data sets. Data mining is a phase in the KDD process that involves utilizing classification, clustering, association rules, and other methods to analyze and model large datasets [1,2]. Because of its large size, numerous resources, and collection techniques, raw data are extremely susceptible to missing, noise, outliers, and inconsistency. The outcomes of data mining will be influenced by poor data quality. As a result, preprocessing techniques must be performed to data in order to enhance its efficiency. Fig. 1 depicts the stages of the KDD process using the dataset. As illustrated in Fig. 1, data must first be chosen in order to identify the target data, and then the designated statistics must be processed in order to progress its trustworthiness. Following preprocessing, the dataset must be transformed into a structure suitable for data gathering. Patterns will subsequently be retrieved, disrupted, and analyzed in the last step utilizing mining processes like as grouping, categorization, modelling, and so on.

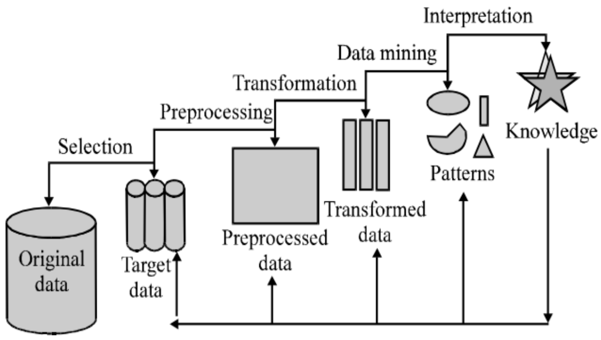


Fig.1: Illustrates the knowledge discovery steps [2]

A. Preprocessing Techniques

Data pre-processing is among few very commonly data mining jobs, and it entails making and altering data into a organization appropriate for mining. Preprocessing data attempts to decrease data size, discover data relationships, normalize data, eliminate outliers, and extract data characteristics [3]. Data cleansing, integration, transformation, and minimization are some of the methods used. Fig. 2 depicts the processes involved in data preparation.

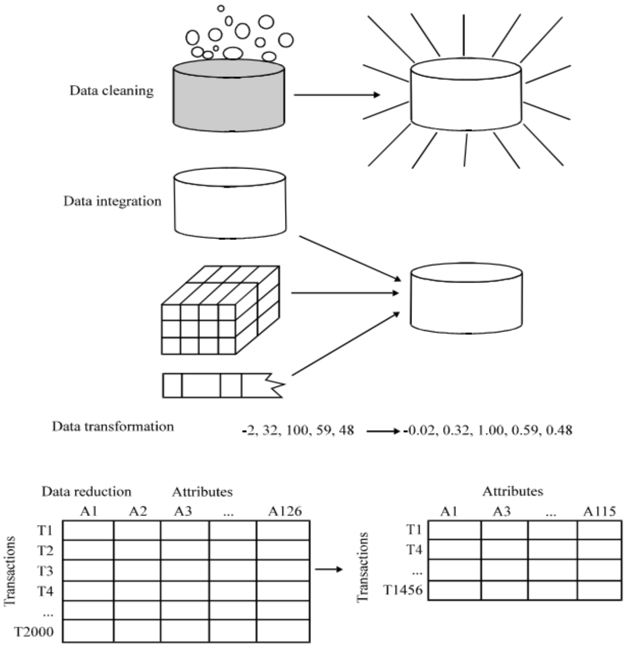
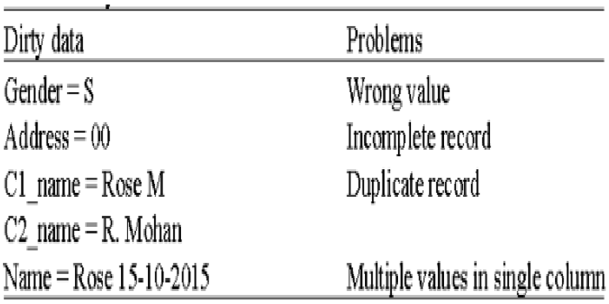


Fig. 2: Illustrates Data Pre-processing processes [4]

B. Data Cleaning

Inadequate records, level of noise in data, outliers, and unreliable data may all be found in row data. Data cleaning is the initial stage in data preparation methods, and it is used to locate missing values, smooth noisy data, identify outliers, and rectify conflicting results. These stale data will have an influence on the excavating procedure, consequential in inaccurate and unsatisfactory results[5]. As a result, it's essential to use certain procedures for cleaning of the data. Table 1 is an example of rogue data.

Table 1: Illustrates few examples of rogue data



Dirty data	Problems
Gender = S	Wrong value
Address = 00	Incomplete record
C1_name = Rose M	Duplicate record
C2_name = R. Mohan	
Name = Rose 15-10-2015	Multiple values in single column

a. Ignoring Tuple:

When the value of the class label does not exist, this option is chosen (it is used with classification mining task). This technique is futile, though it is expedient once the tuple has many vacant attributes.

b. Manually Filling the Value Which Are Missing:

In general, this method requires human work and is time intensive. It can't be utilized with a dataset that's too big.

c. Filling Missing Value by Global Constant

This technique works by substituting missing attribute values with a constant that is consistent across all entries, such as "Unknown" as a label. This technique has flaws because when missing values are substituted with a particular word, such as "Unknown," mining algorithms may think that they constitute a significant idea since they share a common value.

d. Filling Missing Value Using Attribute Mean:

This technique works by substituting a missing value for a specific property with the attribute's average value.

Use the Attribute Mean for all Samples Belonging to the d. Same Class as the Given Tuple:

For example, if we categorize users based on credit risk, the missing value for a particular tuple may be substituted by the average value of income for users in comparable credit risk classes.

e. Filling Missing Value Using Most Probable Value:

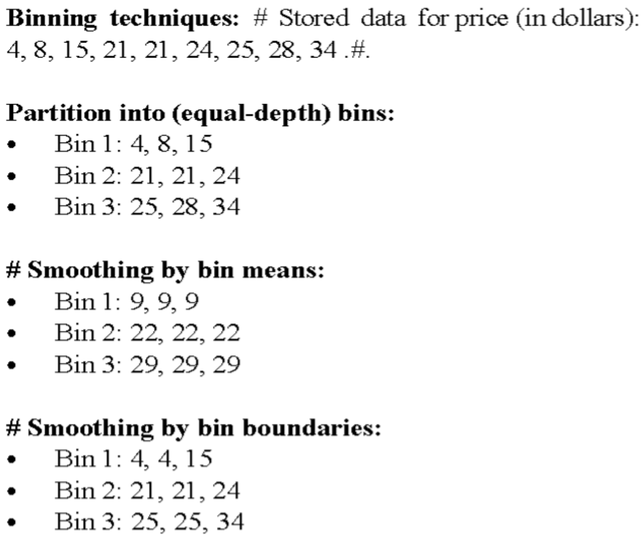
Inference-based regression utilizing a decision tree induction or Bayesian formalism are examples of methods that utilize this methodology.

f. Noise Data:

Among the greatest real challenges affecting mining activities is noise. Noise is defined as a stochastic error or change in an actual result. Dataset that involves mistakes or anomalies that deviate from the mean is referred to as "noise data." The preceding methods could be performed to repair it.

g. Binning

This technique smooths recorded data depending on its "neighborhood," or the data in its immediate vicinity. They are further split to a number of "buckets" or bins once they have been sorted. These techniques accomplish local smoothing since they rely on the data of their neighbors. The min and max values in each bin are used as bin borders in smoothing by bin boundaries. The nearest boundary value is then used to replace each value. In general, the impact of smoothing is higher when the bucket width is bigger. Alternatively, binning is employed as a discretization method in the situation of equal width buckets when the mid of values in each bucket is the same. Binning methods are shown in Fig. 3. The data for price in dollars is sorted and dispersed into equal bin widths in this graph. Every data in a basket is substituted by the average value when flattening by bin means is used. Bin 1's mean values of 4, 8, and 15 are, for example, 9. As a result, 9 is substituted for each value in the bin.



Binning techniques: # Stored data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34 .#.

Partition into (equal-depth) bins:

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

# Smoothing by bin means:

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

# Smoothing by bin boundaries:

- Bin 1: 4, 4, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 25, 34

Fig. 3: Depicts the Binning Techniques.

h. Regression

By fitting data to a function, this technique smooths it out. For example, in linear regression, the optimum line to fit two variables or characteristics is determined such that one attribute may be used to predict the other.

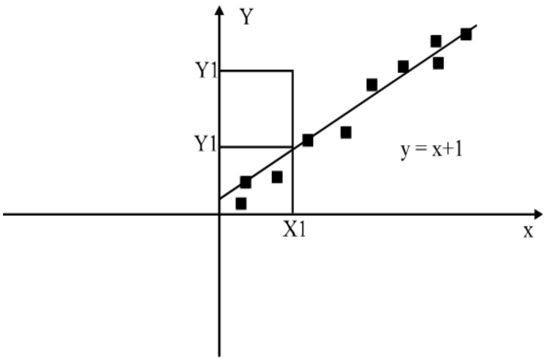


Fig. 4: Illustrates linear regression technique [6]

The term "multi-linear regression" refers to an extension of the term "linear regression." It uses two or more variables to fit data into a multi-dimensional domain. To smooth the noisy data, regression may be used to fit the data by creating a mathematical equation. Fig. 4 shows how linear regression works.

i. Clustering

Clustering is the technique of grouping together a set of points depending on their proximity. Clustering generates a group of cluster centers, each with a series of parameters that really are adjacent to one another although distant away from the others. Anomalies may be identified with this approach because it organizes similar points into groupings, while outlier points are those that fall beyond of the groupings. A clustering method in action. There are three clusters in Fig. 5, and the points that do not belong to any of them are called outliers.

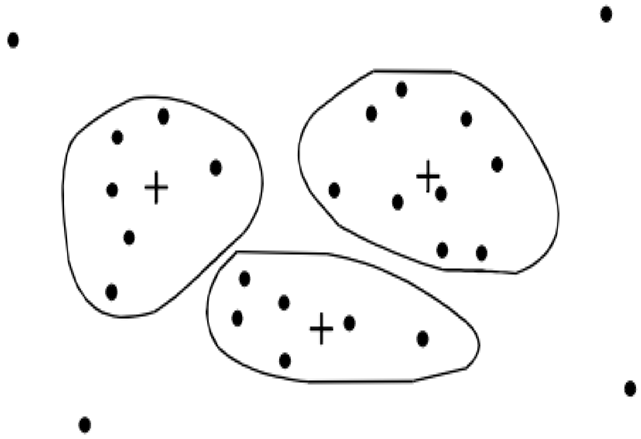


Fig. 5: Illustrates the clustering technique [7]

II. DISCUSSION

A. Data Integration

This method, similar to a data warehouse, works by integrating data from many and disparate sources into a single consistent data repository. There may be many databases, files, or data cubes in these resources. In the field of data integration, there are many problems to address, including Schema integration, object matching, and redundancy, all of which are crucial [8]. If an attribute, such as "annual revenue," is "derived" from another property or collection of characteristics, it is deemed redundant. Another kind of redundancy is attribute or dimension consistency. Some redundancies may be detected using correlation analysis. The correlation between two variables may be used to determine how strongly one characteristic can influence the other. The correlation coefficients may be used to assess the relationship between (X, Y) characteristics.

B. Data Transformation

It entails converting the data into formats that are appropriate for the mining process. It consists of the following:

a. Smoothing

It cleans up data by removing noise. Clustering, regression, and binning are some of the methods used.

b. Aggregation

It is the process of using statistical metrics such as means, medians, and variance to summarize the data. Data mining techniques utilize the resulting aggregated data. Apply aggregation to daily sales to calculate monthly and yearly sales, for example.

c. Generalization

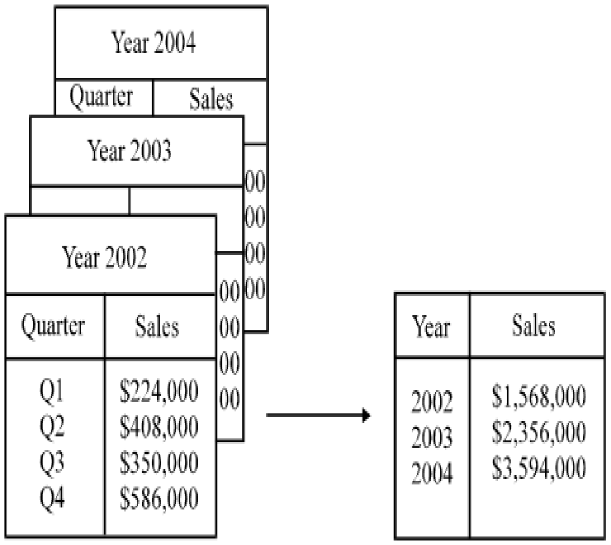
It entails utilizing hierarchical ideas to replace lower level (basic) data with higher level data. For example, a street, which is a categorical characteristic, may be substituted with city or nation, which are high-level words. Another example is that age, which is a quantitative notion, may be mapped to high-level concepts such as elder, younger, and youth.

d. Normalization

The data values are adjusted into a particular range, such as 0-1 to 1-1, using this technique. This approach is helpful for methods like as classification, artificial neural networks, and clustering algorithms in mining. To speed up the learning step, normalization may be employed to scale the data characteristics in tanning face for back propagation neural network method. Normalization techniques include minimum-maximum, z-score, and decimal scaling.

e. Data reduction

These methods may be used to decrease the size of a dataset's representation while maintaining the integrity of the original dataset. As a consequence, by using mining methods on the reduced data, improved data findings may be produced. The methods for data minimization are shown in the paragraph below.



Year 2002	
Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

Fig. 6: Illustrates Data Cube Aggregation [9]. The left side explains sales per quarter for years 2002, 2003, 2004 while the right side, data are aggregated to obtain annual sales.

f. Data cube aggregation

As illustrated in Fig. 6, this technique generates a data cube by accomplishing aggregation procedures to data without losing the essential information for data analysis.

g. Histogram

It's an unsupervised method that doesn't rely on a class label. It divides the values of characteristics into ranges (buckets). In an equal width histogram, the values are split into equal ranges, while in an equal frequency histogram, each section contains the same amount of data [10]. The method may be used to create several level hierarchies by repeating it in a closed loop. A histogram is shown in Figure 6. Each bucket contains one set of prices, as illustrated in Fig. 7.

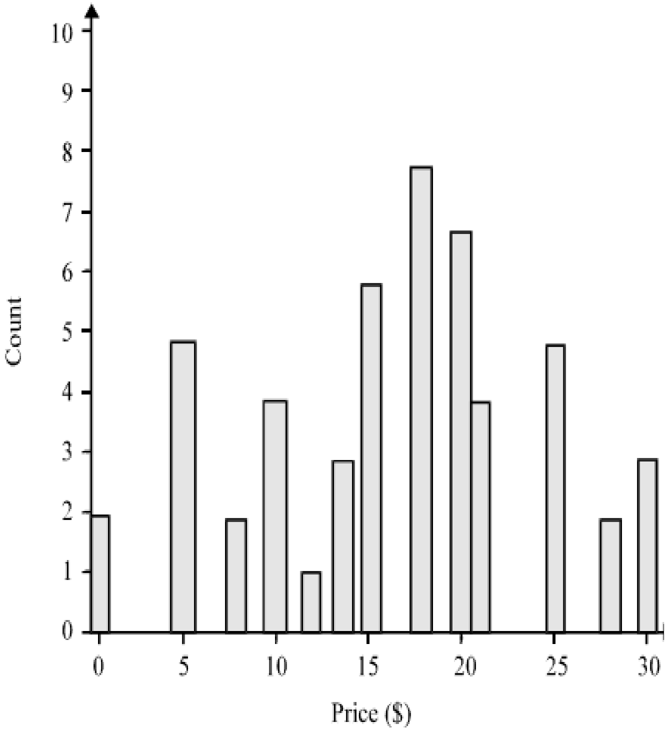


Fig. 7: Illustrates the Histogram. Hear each bucket has one pair of price [9]

III. CONCLUSION

Data preprocessing is essential for both data warehousing and data mining since real-world data is incomplete, inconsistent, noisy, and missing. Data preprocessing comprises data cleansing, data integration, data transformation, and data reduction. This research provides an overview of data preparation methods as well as some instances. Classification process is being used to organize data for various forms of extraction, such as information gathering, knowledge discovery, and web harvesting. Data screening procedures are also used to remove noisy data, fill in gaps in information, and remove extraneous data. The data integration approach links several data inputs in one area. Data transforming techniques modify the sets of data by merging schemas, whilst data mitigation strategies reduce the size of the database. We propose that data augmentation approaches are critical, economic, and successful in large-scale data collection, assessment, and treatment.

REFERENCES

[1] Storti E, Cattaneo L, Polenghi A, Fumagalli L. Customized knowledge discovery in databases methodology for the control of assembly systems. Machines. 2018;

[2] Guarascio M, Manco G, Ritacco E. Knowledge discovery in databases. In: Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. 2018.

[3] Methods DP. Data Preprocessing Techniques for Data Mining. Science (80-). 2011;

[4] Anynomous. Data Preprocessing Techniques for Data Mining. Science (80-). 2011;

[5] Mohit Sharma. What Steps should one take while doing Data Preprocessing? Hackernoon. 2018.

[6] Nguyen PM, Haghverdi A, de Pue J, Botula YD, Le K V., Waegeman W, et al. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosyst Eng. 2017;

[7] Deshmukh MA, Gulhane RA. Importance of Clustering in Data Mining. Int J Sci Eng Res. 2016;

[8] Zhang SZ, Qu XK, Sun J Bin. Data integration and mining based on web big data. Int J Multimed Ubiquitous Eng. 2015;

[9] Alasadi SA, Bhaya WS. Review of data preprocessing techniques in data mining. J Eng Appl Sci. 2017;

[10] Gama J, Pinto C. Discretization from data streams: Applications to histograms and data mining. In: Proceedings of the ACM Symposium on Applied Computing. 2006.