

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/355160516>

Data Preprocessing: The Techniques for Preparing Clean and Quality Data for Data Analytics Process

Article in *Oriental journal of computer science and technology* · January 2021

DOI: 10.13005/ojcs13.0203.03

CITATIONS

54

READS

2,657

2 authors, including:



Dr-Ashish P. Joshi

Institute of Science & Technology for Advanced Studies & Research

18 PUBLICATIONS 69 CITATIONS

SEE PROFILE



Data Preprocessing: the Techniques for Preparing Clean and Quality Data for Data Analytics Process

ASHISH P. JOSHI^{1*} and BIRAJ V. PATEL³

¹BCA Department, Vitthalbhai Patel & Rajratna P.T. Patel Science College,
Sardar Patel University, Vallabh Vidyanagar-388120, India.

³G.H.Patel Department of Computer Science and Technology, Sardar Patel University,
Vallabh Vidyanagar-388120, India.

Abstract

The model and pattern for real time data mining have an important role for decision making. The meaningful real time data mining is basically depends on the quality of data while row or rough data available at warehouse. The data available at warehouse can be in any format, it may huge or it may unstructured. These kinds of data require some process to enhance the efficiency of data analysis. The process to make it ready to use is called data preprocessing. There can be many activities for data preprocessing such as data transformation, data cleaning, data integration, data optimization and data conversion which are use to converting the rough data to quality data. The data preprocessing techniques are the vital step for the data mining. The analyzed result will be good as far as data quality is good. This paper is about the different data preprocessing techniques which can be use for preparing the quality data for the data analysis for the available rough data.



Article History

Received: 10 August 2020

Accepted: 01 September 2020

Keywords

Data Preprocessing;
Data Cleaning;
Data Transformation;
Data integration;
Data Optimization;
Data Conversion.

Introduction to Data Preprocessing

The general model for the real time data mining is as shown in fig.1. The first step is selection of the domain which determines the dataset selection.


The important thing is to select the target data and the target data must be selected from the original data set for enhance the reliability. The

data preprocessing requires after generating the target data to make it ready to use. In the next step the ready to use data works for data analysis and generating some knowledge or result by applying some mining techniques. The data preprocessing techniques includes five activities such as Data Cleaning, Data Optimization, Data Transformation, Data Integration and Data Conversion.

CONTACT Ashish P. Joshi ✉ joshiashish_mca@rediffmail.com 📍 BCA Department, Vitthalbhai Patel & Rajratna P.T. Patel Science College, Sardar Patel University, Vallabh Vidyanagar-388120, India.



© 2020 The Author(s). Published by Oriental Scientific Publishing Company

This is an  Open Access article licensed under a Creative Commons license: Attribution 4.0 International (CC-BY).

Doi: <http://dx.doi.org/10.13005/ojcs13.0203.03>

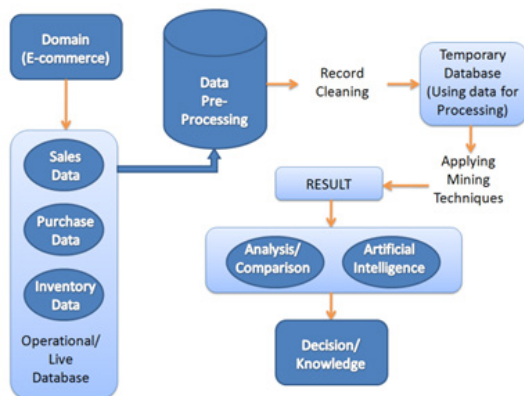


Fig.1: Model for real time data mining

Data Cleaning or Data Cleansing

Data cleaning is part of data preprocessing. Data preprocessing has many activities one of it is data cleaning. Imperfect, incorrect, Incomplete, inaccurate or irrelevant parts of the data are identified in data cleaning process. These type of dirty data can be replace, modify or delete by the specific techniques. Data cleaning is also called data cleansing. Following are the steps for the data cleaning process;

- Select datasets
- Merge datasets into one datasets (if required)
- Identify Errors.
- Standardize process
- scrub for duplicates
- Validate accuracy

Data cleaning can be achieved by many ways like adjusting the missing values, removing the duplicate row or removing the unwanted column. In the feature or any observation, values are not stored is called missing value or missing data. It must be require to filled the missing vale or remove that row. The below are some methods for handling the missing values.

- Ignore the missing values row / delete row – less preferable for small amount of data. (It is possible where percentage of missing value not more then 5%)
- Fill missing value manually – less preferable for large amount of data.
- Global constant – not much preferable (replacing global constant). Each missing value is replaced by fixed value.

- Measure of central tendency (mean, median, mode) – can be use for only numerical data. It is less preferable because its use average value of combined class.
- Measure of central tendency for each class – class wise average replacement. Most probable value (ML algorithm – linear regression, decision tree can be use to decide value).

Data Transformation

Data transformation is use for converting the structure and also use for converting the format of the attribute.

For example, if the data available in integer format and dataset requires to store it in float.

Another example is storing the 1 and 0 value by replacing the true and false value or you may say age 1-12, 13-20,21-40,41-60 can be categorized in the label like child, teen ager, young, old. The different transformation methods are given below.

Smoothing (Remove Noise from Dataset)

Data smoothing is technique which use algorithm to remove noise from a data set. This allows essential patterns to locate out. It can be used to help estimate trends.

Aggregation (Preparing Data in Abstract Format)

Data aggregation is a process which prepared summary from gathered data. It is use to get more information about class based and group based data.

Discretization (Transforming Continues Data in to Some Interval)

Discretization is a practice that transforming unintrupted data into group of fixed intervals. Majority of Data Mining activities in the real world involve unintrupted data.

There is scope of research for handle these attributes because still the existing framework of data mining are not able to do it.

Normalization (Transforming Data in to Given Range)

Basically these process includes to scaling the attribute's data. It is used to generating the data

into a smaller range, such as between 0 to 1. It is generally useful for classification algorithms.

The methods for data normalization are:

- Decimal Scaling
- Min-Max Normalization
- z-Score Normalization

Data Reduction

The data reduction is technique that compresses the data in such a way that the meaning of the data is not lost. For example, data analysis required the year wise analysis and data available quarterly, now data cube aggregation will merge the four quarter data in to year format. The following methods are use to data reduction.

Data Cube Aggregation

Merging quarterly data and make ready yearly data.

Dimension Reduction

It removes redundant features

- Step by step Forward Selection
- Step by step Backward Selection
- Combination of forwarding and Backward Selection

Example:

Initial attribute Set: {A, B,C,D,E,F}

Reduced attributes for the initial set: { }

Step1 - {A}

Step2 - {A, B}

Step3 - {A, B, E}

Reduced attributes for Final set: {A, B, E}

Data Compression

Reduce the size of files using some mechanism

- Lossless Compression
- Lossy Compression

Numerosity Reduction

The actual data is replaced with mathematical models or it may replace by smaller representation of the data instead of actual data in this reduction

technique, it is important to store the representation parameter only.

Discretization

To separate the attributes of the continuous data with a specific intervals by data discretization technique. We can replace many constant values of the attributes by marker of small intervals.

- Top-Down Discretization
- Bottom-Up Discretization

Concept Hierarchy Operation

The size of data can be reduced by collecting and then replacing the low-level concepts (such as 25 degree for tempreture) to high-level concepts (categorical variables can be as hot or cold).

- Bining
- Histogram analysis

Data Integration

Data integration means merging the two or more datasets in to one data set. Some of the application generates the database based on time interval; it requires merging if we want to process all the data at a time. For example, financial account system may generate the data yearly but if we want to perform analysis on 10 years then it requires merging 10 years dataset into one dataset that is called data integration. It also includes the process of merging data from dissimilar sources into a distinct, unified view. It integrates data at single place which are coming from multiple places. It may require to data conversion process to make unified format for each data.

Data Conversion

In the current scenario data are available in different format. The data required to conversion from the existing format to required format.

For example, python is very compatible with the csv data format bur it is not necessary that every data available in csv format. It can be in SQL data, JSON Data or XML Data. Data transformation use to converting the data into required format. The fig.2 model developed in php which is useful for converting the SQL, JSON or XML data into CSV also it is use for converting csv data to mysql.



Fig. 2: selection of json file for convert



Fig. 3: Converted csv file from json

Fig.2 shows the json file selection for convert into csv. When user select the json file and click convert the json file will be convert into csv as display in fig.3.

Conclusion

The rough data generates the errors in the data analytics process. The data analysis cannot generate the efficient result as per requirement on the basis of the rough and noisy data. The result may be varied as compare to actual result due to unprocessed data. The different techniques of the data preprocessing is useful for removing the noisy data and preparing the quality data which gives efficient result of the data analysis.

Acknowledgement

The authors acknowledge Vitthalbhai Patel and Rajratna P.T. Patel Science College managed by Charutar Vidya Mandal (Sardar Patel University) for providing us the opportunity to work in this research.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors do not have any conflict of interest.

Reference

1. Suad A. Alasadi and Wesam S. Bhaya, "Review of Data Preprocessing techniques in data mining" – *Journal of Engineering and Applied Sciences*, 1816-949X
2. Dharmarajan R and R.Vijayashanthi, "An overview on data preprocessing methods in data mining" - *International journal of Science and Research*, 3544-3546
3. Tomar D. and S. Agarwal, "A survey on preprocessing and post processing techniques in data mining" – *International Journal of database theory application*, 99-128.
4. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data Preprocessing for Supervised Learning" - *International Journal Of Computer Science Issn 1306-4428*
5. Vivek Agarwal, "Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis" - *International Journal of Computer Applications* (0975 – 8887)
6. Rima Houari, Ahcène Bounceur, Tahar Kechadi, "A New Method for Estimation of Missing Data Based on Sampling Methods for Data Mining" - https://www.researchgate.net/publication/259007815_A_New_Method_for_Estimation_of_Missing_Data_Based_on_Sampling_Methods_for_Data_Mining - (23/04/2020)
7. DeSarbo, W.S, Green, P.E, Carroll, J.D, Missing data in product-concept testing. *Decision Sciences* 17,163-185,1986
8. J.W, Some simple procedures for handling missing data in multivariate analysis. *Psychometrika* 41, 409-415,1976.