International Journal of Emerging Technology and Innovative Engineering
Volume 4, Issue 2, February 2018 (ISSN: 2394 – 6598)

Paper Received on: 16.02.2018
Paper Published on: 27.02.2018

DATA PREPROCESSING FAMILY

Jeyashree.P
Department of ECE
Kalasalingam Academy of Research and Education
Krishnan koil-626126
TamilNadu
jeyshree30@gmail.com

Mohammad Fazil.K
Department of ECE
Kalasalingam Academy of Research and Education
TamilNadu
samspop5151@gmail.com

ABSTRACT

With advancements in various digital sources, raw data generated are enormous in number. Also the massive growth in the scale of data has been observed in recent years being a key factor of big data scenario. Big data is a term which approaches in data acquiring, processing and analyzing huge amount of heterogeneous data. This article provides an overview of data pre-processing techniques, focussing on real world data problems often like incomplete and inconsistent data that contain more errors. Thus, data pre processing is a proven method of solving fore-mentioned issues. These are the primitive problems that have to be understood carefully and resolved before processing any kind of data. So, this article deals with a total of 14 data pre-processing techniques.

Keywords
Big Data, Data pre-processing, Data mining, feature extraction.

1. INTRODUCTION

Big data is a large set of data that contains voluminous, velocity, complex and a variety of data that requires a high performance processing. The life cycle of big data involves 4 stages (1) feeding data into the system (2) permanently persisting the data in storage (3) computing and analyzing data (4) finally visualizing the results. Raw data is highly susceptible to noise and missing values. It is very difficult to interpret knowledge during data mining or data analysis, if there exists much irrelevant and redundant information. So, in order to improve the efficiency and quality of data, a techniquecalled data pre-processing is done. Data pre-processing method includes datacleaning, data transformation, data integration, data reduction, feature extraction and selection.

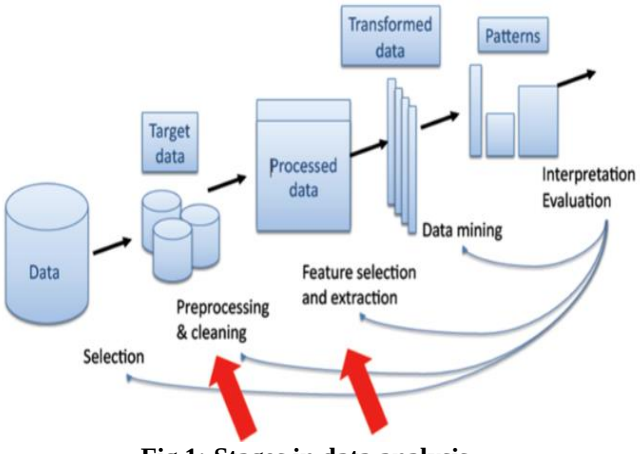

Fig 1: Stages in data analysis

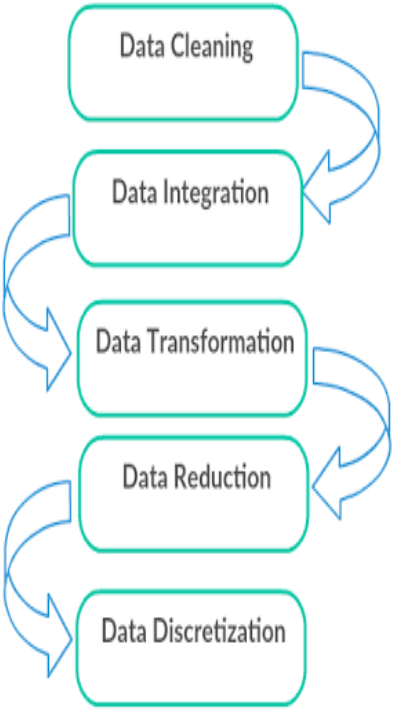

Fig 2: Data pre processing categories

DATA PREPROCESSING TECHNIQUES

1. **Data cleaning:**Repairing and filtering of dirty data is one of the most challenging tasks in data pre-processing. Data cleaning occurs in single set of data or multiple sets of data is an act of detecting and correcting corrupted srecords from a database. The term refers to identifying incomplete, incorrect, irrelevant parts of the data and then replacing, modifying or deleting those dirty data. Expecting the missing values by using learning algorithm. Filling in the missing values, correcting inconsistence in data are done in this process. Running a learning algorithm cleans the unwanted data. Following processes are the primary steps considered in data cleaning

Fig 3: The data cleaning cycle

1.1 Missing value imputation: In data acquiring process, it's pretty common to encounter the presence of missing values since raw data is often incomplete. Due to some limitations in acquisition process and data sampling process it is possible to have data missing data that cannot be avoided in data analysis such that they may create a severe difficulties.

1.2 Noisy data: Noise is generally a random variable error measured in attributes. It is either due to machine fault or man-made error. It hugely affects the data, so it has been processed through different methods often like

- Clustering
- Binning methods
- Combined computer and human inspection.

2. **Data transformation:**The data is transformed to a suitable form such that each old values could be identified with any of the old values which is appropriate for data mining.

2.1 ***Smoothing:***This technique is used to remove noisy channel from the data. It

includes binning, regression and clustering.

2.2 ***Attribute construction:***New features are constructed and added from the given set of attributes or features.

2.3 ***Normalization:*** where the attribute data are measured to fall within a smaller specified range.

2.4 ***Discretization:*** Refers to the process of diving continuous attributes to a regular intervals such that it reduces the data size. Intervals are used to replace the actual values.

3. **Data integration:** The process of combining data in different sources and providing a unified view of data is said to be data integration. Using multiple databases, data cubes or files the data is combined. The heterogeneous data of large volumes are fused. These data sources will be dynamic with extremely heterogeneous in structure.

   Steps involved in data integration:

   1. *Data* from different sources should be collected.
   2. Converting semi structured data or unstructured data into structured.
   3. Transforming structured data into data warehouse.

4. **Data reduction:**A process of obtaining a reduced representation of the data set that is much smaller in volume yet produces the same analytical results. It involves reducing the attribute number, removing irrelevant data by filtering and wrapping methods. Grouping values in clusters by aggregation and reducing the number of samples.

   4.1 ***Need for data reduction:*** Complex data mining may take long time to compute the complete data set and also a database warehouse may store only terabytes of data.

   4.2 Data reduction strategies:

   - Data cube aggregation
   - Attribute Subset Selection
   - Numerosity reduction
   - Dimensionally reduction – Data Compression.

5. **Feature extraction& selection**: A process of reducing the dimensionality by considering a subset of attributes. Feature extraction technique unites the original feature set to generate a new set of feature with less redundant information. For example, polynomial expansion expands the given set of features into polynomial spaces that new spaces are formed by the combination of all original dimensions.

n

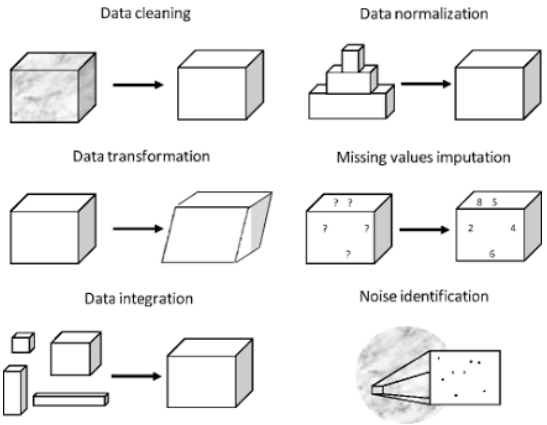

Fig 4. Data pre-processing tasks

In this article, we presented an overview of techniques involved in data pre-processing that must be followed before analysing any kind of data. Since, raw data is often incomplete,
inconsistent and irrelevant and it is likely to contain heterogeneous data. It causes a serious issue during data mining so it is important to figure out the problems and solving it. Successful data pre-processing methods are being the key factors to collectively explore new domains in Big data.

REFERENCES

[1] Garcia et al. Big Data Analytics(2016) 1:9 Data Preprocessing prospects.

[2] Data mining -101-cleaning data and sum up.

[3] Data cleaning techniques -56346092

[4] Williams D, Liao X, Xue Y, Carin Y, Krishapuram B on classification with incomplete data. IEEE trans pattern Mach intel 2017.

[5] Freney B, Classification in the presence of label noise:A survey. IEEE trans neural netw learn system 2014.

[6] R.Stein "Preprocessing Data for Neural Networks"

[7] S.P. Curram .A Comparison of Data Envelopment.

[8] Analysis and Artificial Neural Networks as Tools for Assessing the Efficiency 2014".