

## Effectiveness of Data Preprocessing for Data Mining

Sangeeta Rani<sup>Å\*</sup> and Sonika<sup>Å</sup>

<sup>Å</sup>BLS Institute of Technology Management, Bahadurgarh

Accepted 01 Oct 2014, Available online 10 Oct 2014, Vol.4, No.5 (Oct 2014)

### Abstract

Data Preprocessing is the most crucial step as the operational data is normally never captured and prepared for data mining purpose. Data in the real world is dirty because generally the data is captured from several inconsistent, poorly documented operational systems. Real world data is often incomplete and noisy say wrong values or duplicate records. This results in poor quality data which in turn results in poor quality mining results. So, many organizations or company are interested in how to transform the data into cleaned forms which can be used for high profit purposes. This goal generates an urgent need for data preprocessing. In this paper first, we show the importance of data preprocessing in data analysis, then introduce some research achievements in the area of data preprocessing. finally we suggest some future direction and development

**Keywords:** Cleaning, Reduction, Transformation, Discretization, Concept Hierarchy

### Introduction

“Data mining is the analysis of observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

Data preprocessing means perform certain tasks before the data to get process. The data preprocessing step begins with a general review of the structure of the data and quality assurance. This is performed using sampling and visualization technique. The characteristics of data to be viewed depend on the nature of data, whether the data is categorical or quantitative. For categorical variables, the visualization can be in terms of histograms, pie charts, etc. For quantitative variables, the visualization will be in terms of maxima, minima, mean, median etc. By utilizing all these methods it is possible to determine the presence of invalid and skewed data which may be incorrect. Spurious data or noise can be identified by quantitative techniques as minima and maxima analysis or by various scatter distribution parameters.

### Why we need data preprocessing

Data preprocessing is required to improve the quality of the data by removing the unwanted data from the original data.

### Data quality problems

Noise and outliers, Missing values and Duplicate data

**Noise:** Noise refers to modification of original values. Examples: distortion of a person’s voice when talking on a poor phone and snow on television screen.

\*Corresponding author: Sangeeta Rani

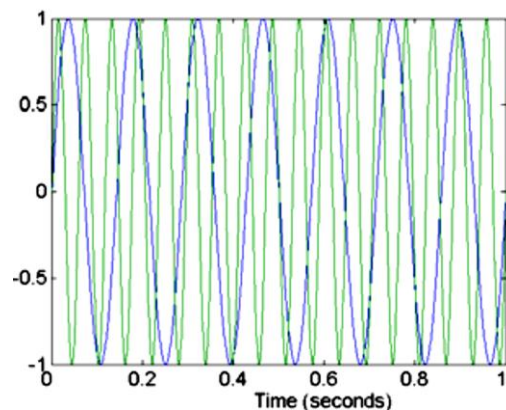


Fig.1 Two sine waves

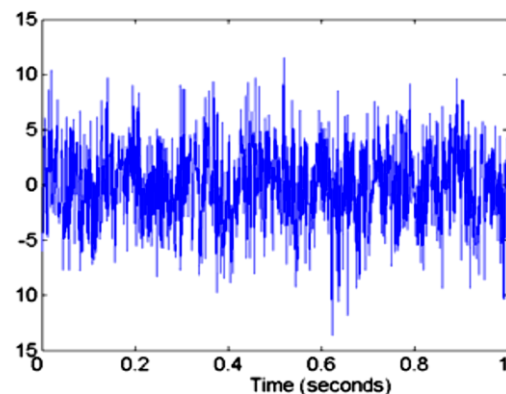
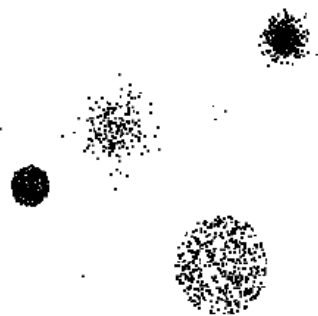


Fig.2 Two sine waves+Noise

**Outliers:** Outliers are data objects with characteristics that considers different than most of the others data objects in the data set.



**Missing Values:** When data is not complete due to some reasons. The reasons for missing values are:

- Information is not collected. e.g. people decline to give their age and weight
- Attributes may not be applicable to all cases. e.g. annual income is not applicable to children.

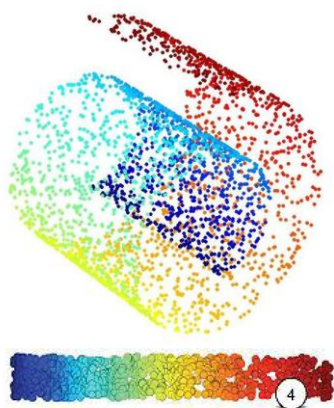
**Duplicate Data:** Data set may include data objects that are duplicates, or almost duplicates of one another. The major issue when merging data from heterogeneous sources. Examples: Same person with multiple email addresses

**Measures for data quality**

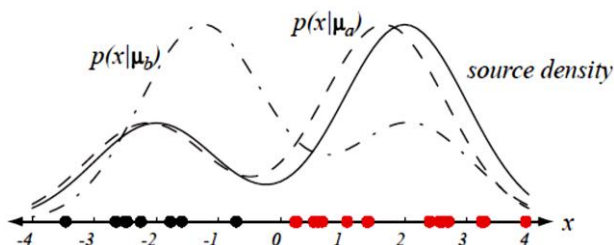
- **Accuracy** : correct or wrong, accurate or not.
- **Completeness** : not recorded, unavailable.
- **Consistency** : some modified but some not, dangling.
- **Timeliness** : timely update?
- **Believability** : how trustable the data are correct?
- **Interpretability**: how easily the data can be understood

**Data Mining Tasks:** As per the data in the data mining process we have the following types of tasks that are performed with data.

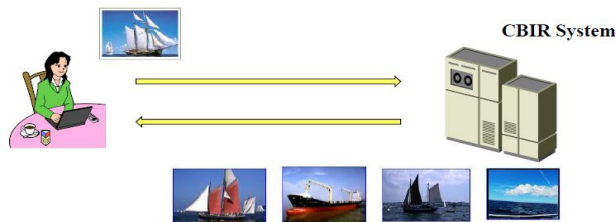
- 1) **Exploratory Data Analysis:**The task that is interactive and visual.



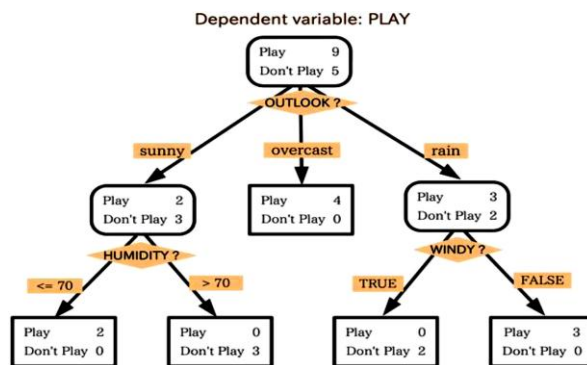
- 2) **Descriptive Modeling:** The task that describe all of the data collected.



- 3) **Discovering Patterns and Rules:** The task that detect anomalies.



- 4) **Predictive Modeling:** The task that perform inference on the collected data in order to make predictions.



**Data Preprocessing Task**

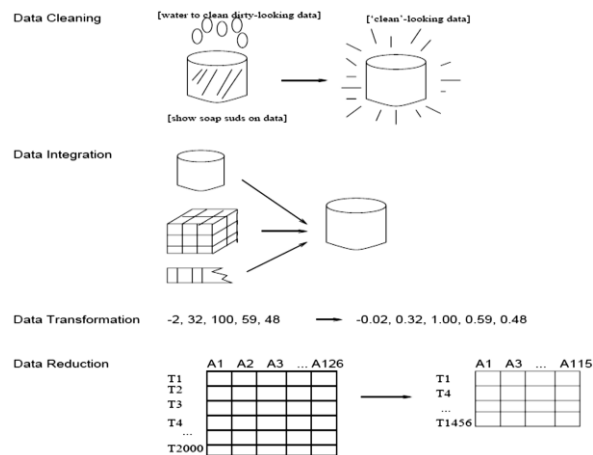
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Data Discretization and Concept Hierarchy Generation

**Data Cleaning:** Real-world data tend to be incomplete, noisy, and inconsistent. *Data cleaning* routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

**Missing Values**

1. **Ignore the Tuple:** This is usually done when the class label is missing. This method is not very effective, unless the tuple contains several attributes with missing values.
2. **Fill in the missing value manually:** In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.
3. **Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant, such as a label like "Unknown" or  $\text{---}$ . 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 foolproof.
4. **Use the attribute mean to fill in the missing value:** Use the mean value to replace the missing value for a tuple.

**5. Use the most probable value to fill in the missing value:** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction



**Noisy Data:** Noise is a random error or variance in a measured variable. To remove the noise from the data we use following methods.

- 1. Binning:** Binning methods smooth a sorted data value by consulting its “neighborhood,” that is, the values around it. The sorted values are distributed into a number of “buckets,” or *bins*. Because binning methods consult the neighborhood of values, they perform *local* smoothing
- 2. Regression:** Data can be smoothed by fitting the data to a function, such as with regression
- 3. Clustering:** Outliers may be detected by clustering, where similar values are organized into groups, or “clusters.” Intuitively, values that fall outside of the set of clusters may be considered outliers .

**Data Integration:** Data mining often requires data integration-the merging of data from multiple data stores.

- Data integration:**
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#**
  - Integrate metadata from different sources
- Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units
- Handling Redundancy in Data Integration**
  - Redundant data occur often when integration of multiple databases
  - Object identification: The same attribute or object may have different names in different databases

- Derivable data: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by correlation analysis

**Careful integration of the data from multiple sources may help reduce/ avoid redundancies and inconsistencies and improve mining speed and quality**

**Data Transformation**

- 1) Smoothing:** which works to remove noise from the data. Such techniques include binning, regression, and clustering.
- 2) Aggregation:** where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.
- 3) Generalization of the data:** where low-level data are replaced by higher-level concepts through the use of concept hierarchies.
- 4) Normalization,** where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0.0 to 1.0.
  - Min-max normalization: to  $[new\_min_A, new\_max_A]$ 
    - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to
  - Z-score normalization ( $\mu$ : mean,  $\sigma$ : standard deviation):
    - Ex. Let  $\mu = 54,000, \sigma = 16,000$ . Then
  - Normalization by decimal scaling

$$v' = \frac{v}{10^i}$$

integer such that  $Max(|v'|) < 1$

**5) Feature Construction:** Where new attributes are constructed and added from the given set of attributes to help the mining process.

**Data Reduction:** Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

- 1. Data cube aggregation:** where aggregation operations are applied to the data in the construction of a data cube.
- 2. Attribute subset selection:** where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.
- 3. Dimensionality reduction:** where encoding mechanisms are used to reduce the data set size.
- 4. Numerosity reduction:** Where the data are replaced or estimated by alternative, smaller data representations such as parametric method or nonparametric methods .

**Discretization and Concept Hierarchy Generation:** Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals.

A concept hierarchy for a given numeric attribute defines a discretization of the attribute.

**Discretization and Concept Hierarchy Generation for Numeric Data:** Concept hierarchies for numeric attributes can be constructed automatically based on data distribution analysis Methods:

**Binning:** Attribute values can be discretized by distributing the values into bin and replacing each bin by the mean bin value or bin median value. These technique can be applied recursively to the resulting partitions in order to generate concept hierarchies.

**Histogram Analysis:** Histograms can also be used for discretization. Partitioning rules can be applied to define range of values.

**Cluster Analysis:** A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms anode of a concept hierarchy, where all nodes are at the same conceptual level.

**Segmentation by natural partitioning:** Breaking up annual salaries in the range of into ranges like (\$50,000-\$100,000) are often more desirable than ranges like (\$51,263, 89-\$60,765.3) arrived at by cluster analysis. The 3-4-5 rule can be used to segment numeric data into relatively uniform "natural" intervals. In general the rule partitions a give range of data into 3,4,or 5 equinity intervals, recursively level by level based on value range at the most significant digit

**Discretization and Concept Hierarchy Generation for Categorical Data:** Categorical data are discrete data. Categorical attributes have finite number of distinct values, with no ordering among the values, examples include geographic location, item type and job category. There are several methods for generation of concept hierarchies for categorical data.

**Specification of a partial ordering of attributes explicitly at the schema level by experts:** Concept hierarchies for categorical attributes or dimensions typically involve a group of attributes. A user or an expert can easily define concept hierarchy by specifying a partial or total ordering of the attributes at a schema level. A hierarchy can be defined at the schema level such as street < city < province < state < country.

**Specification of a portion of a hierarchy by explicit data grouping:** This is identically a manual definition of a portion of a concept hierarchy. In a large database, is unrealistic to define an entire concept hierarchy by explicit value enumeration. However, it is realistic to specify explicit groupings for a small portion of the intermediate-level data.

**Specification of a set of attributes but not their partial ordering:** A user may specify a set of attributes forming a concept hierarchy, but omit to specify their partial ordering. The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept hierarchy.

### Conclusion

In this paper, we have proposed methodology for the data preprocessing. In this we categorize the data as per numeric and categorical data. Main steps in this data preprocessing is numeric and categorical data selection based on parameter determined, reduction of dimension, selection and transformation and concept hierarchy generation.

### References

- Data Mining Techniques, Arun k pujari 1st Edition
- Data warehousing, Data Mining and OLAP, Alex Berson , Smith.J. Stephen
- Data Mining Concepts and Techniques , Jiawei Han and Micheline Kambe
- Data Mining Introductory and Advanced topics, Margaret H Dunham PEA
- The Data Warehouse lifecycle toolkit , Ralph Kimball Wiley student Edition.
- Agrawal, Rakesh and Ramakrishnan Srikant, "Fast Algorithms for Mining & Preprocessing Assosiation Rules", Proceedings of the 20th VLDB Conference, Santiago, Chile (1994).
- Salleb, Ansaf and Christel Vrain, "An Application of Assosiation Knowledge Discovery and Data Mining (PKDD) 2000, LNAI 1910, pp. 613-618, Springer
- Agrawal, R., and Psaila, G. 1995. Active Data Mining. In Proceedings on Knowledge Discovery and Data Mining