

1.4.1 | Data Preprocessing Part 1



In the previous session, you have learned how to craft a data mining problem statement. In this session, you will learn how to preprocess data.

At the end of the session, you should be able to:

1. Describe the various preprocessing methods.
 2. Apply a data preprocessing method that is appropriate for a given data.
-

Real-world data are susceptible to noise and inconsistencies, incomplete and have missing components. **So how can data be preprocessed to help improve the quality of data mining results? How can the data be preprocessed to enhance the efficiency and ease of the mining process?**

Several techniques can be applied to perform data preprocessing. Some of these techniques are as follows:

1. Data cleaning can be applied to remove noise and correct inconsistencies in data.
2. Data integration merges data from multiple sources into coherent data storage such as data warehouse.
3. Data reduction can reduce the size by, for instance, aggregating, eliminating redundant features, or scaled to fall within a smaller range.

These techniques can improve the accuracy and efficiency of mining algorithms. The methods are not mutually exclusive; they can work together.

Why Preprocess Data?

Data has quality if it satisfies the requirements of the intended use. The following are the factors comprising data quality.

1. accuracy

2. completeness
 3. consistency
 4. timeliness
 5. believability
 6. interpretability
-

Among these factors, the most common that can compromise data quality in large data sets are **accuracy, completeness, and consistency**.

There are many possible reasons for **inaccurate data**. Some of these are incorrect attribute values, faulty data collection tools, human or computer errors at data entry, users may purposely submit incorrect personal information, and technology limitations. These reasons result in data inconsistencies.

Incomplete data can occur for several reasons such as unavailability of data attributes of interest, malfunctioning equipment, some data may not be recorded because they are not considered necessary at the time of entry.

e.g., Occupation=" " (missing data)

Data that were **inconsistent** with other recorded data may have been deleted. Recoding of the data history or modifications may have been overlooked.

Age="42", Birthday="03/07/2010"

Was rating "1, 2, 3", now rating "A, B, C"

discrepancy between duplicate records

Major Tasks of Data Preprocessing

1. **Data Cleaning**
 2. **Data Integration**
 3. **Data Reduction**
 4. **Data Transformation**
-



Figure 1.4.1 Forms of Data Preprocessing

Data cleaning tasks involve filling in missing values, smoothing noisy data, identifying and removing outliers, and resolving inconsistencies. Dirty data can confuse the mining procedure, resulting in unreliable output.

Data integration involves combining data residing in different sources and providing users with a unified view of them.

Data reduction obtains a reduced representation of the data set that is much smaller in volume yet produces that same analytical result. Data reduction strategies include **dimensionality reduction** and **numerosity reduction**.

In **dimensionality reduction**, data encoding schemes are applied to obtain a reduced or compressed representation of the original data. Examples include data compression techniques (e.g. wavelet transforms and principal component analysis), and attribute subset selection (e.g. removing irrelevant attributes), and attribute construction (e.g. small set of more useful attributes is derived from the original set).

In **numerosity reduction**, the data are replaced by alternative, smaller representations using parametric models(e.g. regression) or non-parametric models(e.g. histograms, clusters, sampling, or data aggregation).

Forms of **data transformation** include normalization, discretization, and concept hierarchy generation. **Normalization** is the scaling of data to a smaller range (e.g. 0.0, 1.0). **Concept hierarchy** is done by replacing low-level concepts (e.g. numeric values such as attribute age) to a higher level (e.g. young, middle-aged, senior). **Discretization** is transforming a continuous attribute to a categorical attribute.

Task 1. Data Cleaning

Real-world data tend to be dirty, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an essential step in the knowledge discovery process because quality decisions must be based on quality data. Detecting and rectifying them early, and reducing the data to be analyzed can lead to enormous payoffs for decision making.

Missing Values

The following are the methods used for cleaning data.

1. Ignore the tuple
2. Fill in the missing value manually

3. Use a global constant to fill in the missing value (e.g. replace all missing attribute values by the same constant such as label like "Unknown").
 4. Use a measure of central tendency for the attribute (e.g. mean or median) to fill in the missing value. Use mean for normal symmetric data distribution and median for skewed data distribution.
 5. Use the attribute mean or median for all samples belonging to the same class as the given tuple.
 6. Use the most probable value to fill in the missing value. This may be determined with regression or decision tree induction.
-

Noisy Data

Noise is a random error or variance in a measured variable. Use statistical description techniques (e.g. [boxplots](https://www.mathsisfun.com/data/quartiles.html) [_](https://www.mathsisfun.com/data/quartiles.html)(<https://www.mathsisfun.com/data/quartiles.html>), and [scatter plots](https://www.mathsisfun.com/data/scatter-xy-plots.html) [_](https://www.mathsisfun.com/data/scatter-xy-plots.html)(<https://www.mathsisfun.com/data/scatter-xy-plots.html>), and data visualization to identify outliers which may represent noise.

e.g., Salary="-10" (an error)

Given a numeric attribute such as price, how can you smooth out the data to remove the noise?

The following data smoothing techniques can be used to remove noise.

- **Binning.** Binning methods smooth a sorted data value by consulting its neighbourhood that is the values around it. The sorted values are distributed into several **buckets** or **bins**. Shown in Figure 1.4.2 are examples of binning methods.



Binning Methods.jpg

Figure 1.4.2 Binning Methods

In applying the binning method, data is sorted and partitioned into equal-frequency bins (e.g. three values on each bin)

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.

In smoothing by bin medians, each bin value is replaced by the bin median.

In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. The closest boundary value then replaces each bin value.

Bins may be equal width, where the interval range of values in each bin is constant.

Task 2. Data Integration

Data mining often requires data integration, the merging of data from multiple data stores. Careful integration can help avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data mining process.

The semantic heterogeneity and structure of data pose significant challenges in data integration. The following methods can be used for data integration.

1. Entity Identification Problem
2. Redundancy and Correlation Analysis
3. Tuple Duplication
4. Data Value Conflict Detection and Resolution

Entity Identification Problem is the matching up of objects in schema integration. Object identification happens when the same attribute or object may have different names in different databases. Derivable data occurs when one attribute may be a “derived” attribute in another table.

Example:

`cust_id` in one database is `cust_number` in another database
 Donald Trump = Trump Donald

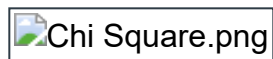
Redundancy and Correlation Analysis

An attribute may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

For nominal attributes, use the [chi-square test](http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1440#:~:text=The%20c2%20test%20is,2%20variables%20in%20the%20population) [_ \(http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1440#:~:text=The%20c2%20test%20is,2%20variables%20in%20the%20population\)](http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1440#:~:text=The%20c2%20test%20is,2%20variables%20in%20the%20population). For numeric attributes, use [correlation coefficient](http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442) [_ \(http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442\)](http://learntech.uwe.ac.uk/da/Default.aspx?pageid=1442) and [covariance](https://www.statisticshowto.com/covariance) [_ \(https://www.statisticshowto.com/covariance\)](https://www.statisticshowto.com/covariance), both of which access how one attribute's values vary from those of another.

Correlation Analysis (Nominal Data)

X^2 (chi-square) test is used to analyze the correlation nominal data. The formula is



The larger the X^2 value, the more likely the variables are related. The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count. [Correlation does not imply causality.](https://www.mathtutordvd.com/public/Why-Correlation-does-not-Imply-Causation-in-Statistics.cfm) [_ \(https://www.mathtutordvd.com/public/Why-Correlation-does-not-Imply-Causation-in-Statistics.cfm\)](https://www.mathtutordvd.com/public/Why-Correlation-does-not-Imply-Causation-in-Statistics.cfm) Causation indicates a relation between two variables in which one variable is *affected* by another. For example, there have been numerous studies that provide evidence that smoking *causes* lung cancer.

Example:

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

In X^2 (chi-square) calculation, the numbers in parenthesis are expected counts calculated based on the data distribution in the two categories.



The result of the calculation reveals that like_science_fiction and play_chess are correlated in the group.

Correlation Analysis (Numeric Data)

The correlation coefficient is also known as Pearson's Product Moment Coefficient. The formula is:

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B, σ_A and σ_B are the individual standard deviation of A and B, and $\sum(a_i b_i)$ is the sum of the AB cross-product.

If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation. If $r_{A,B} = 0$: independent or if $r_{A,B} < 0$, there is negative correlated.

Covariance (Numeric Data)

Covariance is similar to the correlation. It is mathematically represented as:

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient:
$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

where n is the number of tuples, \bar{A} and \bar{B} are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

Positive covariance: If $Cov(A, B) > 0$, then A and B both tend to be larger than their expected values.

Negative covariance: If $Cov(A, B) < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.

Independence: $Cov(A, B) = 0$ but the converse is not true:

Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence.

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

Suppose two stocks A and B have the following values in one week:
(2, 5), (3, 8), (5, 10), (4, 11), (6, 14).

Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$$




$$E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$$

$$Cov(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$$

Thus, A and B rise together since $Cov(A, B) > 0$.

Let's test your understanding of data cleaning and data integration. 

Download two raw datasets from any of the following data sources. Clean and integrate the datasets using all the data cleaning methods. Justify the data cleaning methods that you have applied in the datasets.

- [Kaggle](https://www.kaggle.com/) 
- [OpenData](https://www.gov.ph/data) 
- [OpenStat](http://openstat.psa.gov.ph/) 

Submit your output in the course bank



<https://drive.google.com/drive/folders/1WKLWnpjLQHxDyV7WdaLrb0Q6ScrEUa3u?usp=sharing>

Tap the button to assess your understanding of data preprocessing.

Assignment 

<https://tip.instructure.com/courses/9953/assignments/156976>

