Chapter 3:-   Data Preprocessing and ETL

Data preprocessing is an important task. It is a data mining technique that transforms raw data into a more understandable, useful and efficient format.

> *Data has a better idea. This idea will be clearer and understandable after performing data preprocessing*.

Why is data preprocessing required?

Real world data is generally:

Incomplete: Certain attributes or values or both are missing or only aggregate data is available.

Noisy: Data contains errors or outliers

Inconsistent: Data contains differences in codes or names etc.

Tasks in data preprocessing

1. **Data Cleaning:** It is also known as scrubbing. This task involves filling of missing values, smoothing or removing noisy data and outliers along with resolving inconsistencies.
2. **Data Integration:** This task involves integrating data from multiple sources such as databases (relational and non-relational), data cubes, files, etc. The data sources can be homogeneous or heterogeneous. The data obtained from the sources can be structured, unstructured or semi-structured in format.

3. **Data Transformation:** This involves normalisation and aggregation of data according to the needs of the data set.
4. **Data Reduction:** During this step data is reduced. The number of records or the number of attributes or dimensions can be reduced. Reduction is performed by keeping in mind that reduced data should produce the same results as original data.
5. **Data Discretization:** It is considered as a part of data reduction. The numerical attributes are replaced with nominal ones.

Data Cleaning

The data cleaning process detects and removes the errors and inconsistencies present in the data and improves its quality. Data quality problems occur due to misspellings during data entry, missing values or any other invalid data. Basically, “dirty” data is transformed into clean data. “Dirty” data does not produce the accurate and good results. Garbage data gives garbage out. So it becomes very important to handle this data. Professionals spend a lot of their time on this step.

Reasons for “dirty” or “unclean” data

1. Dummy values
2. Absence of data
3. Violation of business rules
4. Data integration problems
5. Contradicting data
6. Inappropriate use of address line

7. Reused primary keys
8. Non-unique identifiers

What to do to clean data?

1. Handle Missing Values
2. Handle Noise and Outliers
3. Remove Unwanted data

Handle Missing Values

Missing values cannot be looked over in a data set. They must be handled. Also, a lot of models do not accept missing values. There are several techniques to handle missing data, choosing the right one is of utmost importance. The choice of technique to deal with missing data depends on the problem domain and the goal of data mining process. The different ways to handle missing data are:

1. **Ignore the data row:** This method is suggested for records where maximum amount of data is missing, rendering the record meaningless. This method is usually avoided where only less attribute values are missing. If all the rows with missing values are ignored i.e. removed, it will result in poor performance.
2. **Fill the missing values manually:** This is a very time consuming method and hence infeasible for almost all scenarios.
3. **Use a global constant to fill in for missing values:** A global constant like “NA” or 0 can be used

to fill all the missing data. This method is used when missing values are difficult to be predicted.

4. **Use attribute mean or median:** Mean or median of the attribute is used to fill the missing value.
5. **Use forward fill or backward fill method:** In this, either the previous value or the next value is used to fill the missing value. A mean of the previous and succession values may also be used.
6. Use a data-mining algorithm to predict the most probable value

Handle Noise and Outliers

Noise in data may be introduced due to fault in data collection, error during data entering or due to data transmission errors, etc. Unknown encoding , out of range values (Example : Age — -10), Inconsistent Data (Example : DoB — 4th Oct 1999, Age — 50), inconsistent formats (Example : DoJ — 13th Jan 2000, DoL — 10/10/2016), etc. are different types of noise and outliers. Noise can be handled using **binning.** In this technique, sorted data is placed into bins or buckets. Bins can be created by equal-width (distance) or equal-depth (frequency) partitioning. On these bins, smoothing can be applied. Smoothing can be by bin mean, bin median or bin boundaries.

Outliers can be smoothed by using binning and then smoothing it. They can be detected using visual analysis or boxplots. Clustering can be used identify groups of outlier data.The detected outliers may be smoothed or removed.

Remove Unwanted Data

Unwanted data is duplicate or irrelevant data. Scraping data from different sources and then integrating may lead to some duplicate data if not done efficiently. This redundant data should be removed as it is of no use and will only increase the amount of data and the time to train the model. Also, due to redundant records, the model may not provide accurate results as the duplicate data interferes with analysis process, giving more importance to the repeated values.

Data Integration

In this step, a coherent data source is prepared. This is done by collecting and integrating data from multiple sources like databases, legacy systems, flat files, data cubes etc.

> *Data is like garbage.*

Issues in Data Integration

1. **Schema Integration:** Metadata (i.e. the schema) from different sources may not be compatible. This leads to *entity identification problem***.** Example : Consider two data sources R and S. Customer id in R is represented as cust_id and in S is represented is c_id. They mean the same thing, represent the same thing but have different names which leads to integration problems. Detecting and resolving them is very important to have a coherent data source.

2. **Data value conflicts:** The values or metrics or representations of the same data maybe different in for the same real world entity in different data sources. This leads to different representations of the same data, different scales etc. Example : Weight in data source R is represented in kilograms and in source S is represented in grams. To resolve this, data representations should be made consistent and conversions should be performed accordingly.
3. **Redundant data:** Duplicate attributes or tuples may occur as a result of integrating data from various sources. This may also lead to inconsistencies. These redundancies or inconsistencies may be reduced by careful integration of data from multiple sources. This will help in improving the mining speed and quality. Also, co-relational analysis can be performed to detect redundant data.

Data Reduction

If the data is very large, data reduction is performed. Sometimes, it is also performed to find the most suitable subset of attributes from a large number of attributes. This is known as dimensionality reduction. Data reduction also involves reducing the number of attribute values and/or the number of tuples. Various data reduction techniques are:

1. **Data cube aggregation:** In this technique the data is reduced by applying OLAP operations like slice, dice

or rollup. It uses the smallest level necessary to solve the problem.

2. **Dimensionality reduction:** The data attributes or dimensions are reduced. Not all attributes are required for data mining. The most suitable subset of attributes are selected by using techniques like forward selection, backward elimination, decision tree induction or a combination of forward selection and backward elimination.
3. **Data compression:** In this technique. large volumes of data is compressed i.e. the number of bits used to store data is reduced. This can be done by using lossy or lossless compression. In *loss compression,* the quality of data is compromised for more compression. In *lossless compression,* the quality of data is not compromised for higher compression level.
4. **Numerosity reduction :** This technique reduces the volume of data by choosing smaller forms for data representation. Numerosity reduction can be done using histograms, clustering or sampling of data. Numerosity reduction is necessary as processing the entire data set is expensive and time consuming.

Discretization in data mining

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

Age	1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77

Table before Discretization

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level.

concept hierarchy generation

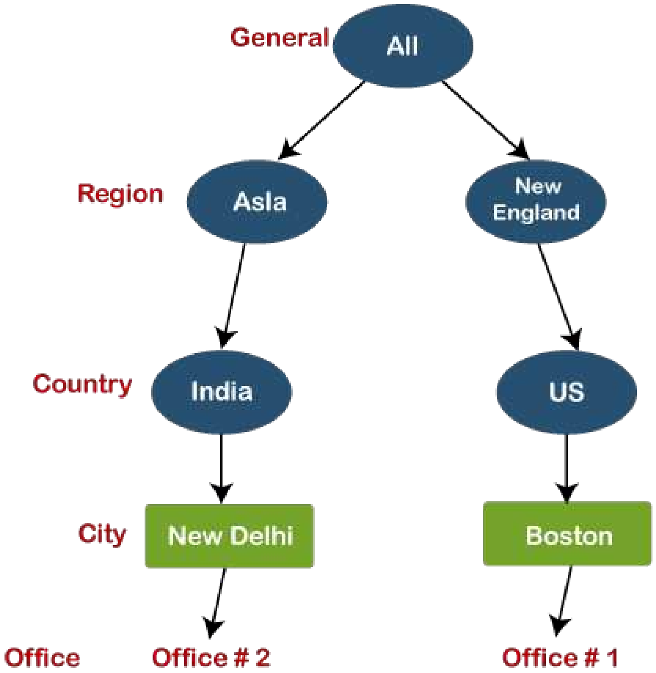
The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance. In other words, we can say that a hierarchy concept refers to a sequence of mappings with a set of more general concepts to complex concepts. It means mapping is done from low-level concepts to high-level concepts. For example, in computer science, there are different types of hierarchical systems. A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model. There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping
A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

Top-down mapping

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

Bottom-up mapping

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.



Concept Hierarchy Generation