UNIT–II: Data Preprocessing- Overview, Data Cleaning, Data Integration, Data Reduction, Data Transformation and Data Discretization.

# data preprocessing

The data present in the data warehouse may contain noisy, missing, and inconsistent data, data is to be preprocessed in order to help improve the quality of the data and data be preprocessed so as to improve the efficiency and ease of the mining process.

# steps in data preprocessing

The methods for data preprocessing are organized into the following categories

1. Data cleaning
2. Data integration
3. Data reduction
4. Data transformation

**data cleaning**

Data cleaning routine work to clean the data by filling in missing values, smoothing the noisy data, identifying or removing outliers and resolving inconsistencies.

# data integration

Combining multiple databases, data cubes, or files, data marts or other data sources under a single database or warehouse server are called data integration.

# data transformation

This is a kind of operation, such as normalization and aggregation is additional data preprocessing procedures that would contribute toward the success of the mining process.

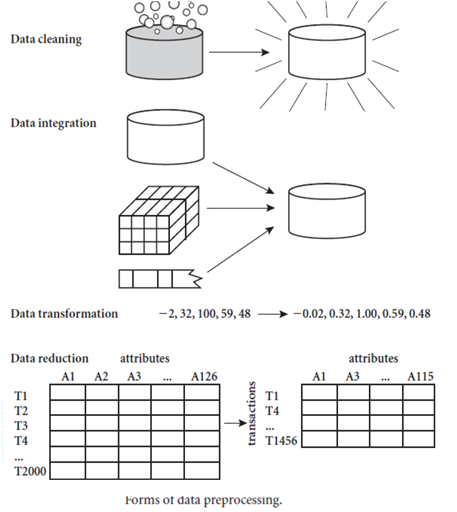
# data reduction

It obtains a reduced representation of data set that is much smaller in volume. Yet produce the same (or almost the same) analytical results.

# Forms of Data preprocessing.

The data you wish to analyze by data mining techniques are incomplete (lacking attribute values), noisy (containing errors) and inconsistent.

To overcome this problem the following data preprocessing techniques are required.



1. Data cleaning:

Data cleaning routine work to clean the data by filling in missing values,smoothing the noisy data,identifying or removing outliers and resolving inconsistencies.

1. Data integration:

Combining multiple databases, data cubes, or files, data marts or other data sources under a single database or warehouse server are called data integration.

1. Data transformation:

This is a kind of operation, such as normalization and aggregation is additional data preprocessing procedures that would contribute toward the success of the mining process.

4.Data reduction:

It obtains a reduced representation of data set that is much smaller in volume. Yet produce the same (or almost the same) analytical results. There are a number of strategies for data reduction.

This includes:

Data Aggregation.

Attribute subset selection.

Dimensionality reduction.

Numerosity reduction.

* 1. Data Aggregation:

Aggregation operation is applied to the data in the construction of the data cube.

* 1. Attribute subset selection:

Irrelevant or redundant attributes may be deleted and removed.

* 1. Dimensionality reduction:

Encoding mechanisms are used to reduce the data.

* 1. Numerosity reduction:

The data are replaced or estimated by alternative data smaller data representations.

5. Data discretization:

Data discretization is a form of data reduction that is very useful for the automatic generation of concept hierarchies from numerical data.

# data cleaning and different techniques for handling missing values

# Real-world data tend to be incomplete, noisy, and inconsistent. Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Data cleaning looks at ways of handling missing values, smoothing data to remove noisy data.

1. Missing values:

In order to handle these missing values, the following methods are employed by the data cleaning process.

* 1. Ignore the tuple:

This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute vary considerably.

* 1. Fill in the missing values manually:

In general, this approach is time-consuming and may not be feasible given a large data set with many missing values.

* 1. 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

−. 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.”

* 1. Use the attribute mean to fill in the missing value:

For example, suppose that the average income of All Electronics customers is $56,000. Use this value to replace the missing value for income.

* 1. Use the attributes mean of all samples belonging to the same class as the given tuple:

For example, if classifying customers, according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

* 1. Use the most probable value to fill in the missing value:

This may be determined by regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

# Steps for performing data cleaning

The steps for performing data cleaning as a process are,

1. Discrepancy Detection.
2. Data transformation.

# Discrepancy Detection:

The first step in data cleaning is discrepancy detection.The reason for the occurrence of discrepancies are

* 1. Poorly designed data entry forms that have many optional fields.
  2. Deliberate errors (e.g., respondents not wanting to divulge information about themselves).
  3. Data Decay (e.g., outdated address).
  4. Discrepancies may also arise from inconsistent data representations and inconsistent use of codes.
  5. Errors in instrument devices that record data, and system errors are other sources of discrepancies.
  6. Data may also be inconsistencies due to data Integrator.

Discrepancies can be avoided by using either expert knowledge or by using metadata,since it helps users to get the information about the data types,domain range and size of each individual attribute.

The data should also be examined regarding unique values,consecutive rules and null values.

Unique rule: This rule says that each value of the given attribute must be different from all other values for that attribute.

Consecutive rule: This rule says that there can be no missing value between the lowest and highest values for that attribute, and that all values must also be unique.

Null rule: A null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition.

There are a number of different commercial tools that can aid in the step of discrepancy detection.

* + 1. Data scrubbing tools: Use simple domain knowledge (e.g., knowledge of postal addresses, and spell checking) to detect errors and make corrections in the data. These tools rely on parsing and fuzzy matching techniques when cleaning data from multiple sources.
    2. Data auditing tools: find discrepancies by analyzing the data to discover rules and relationships, and detecting data that violate such conditions.

1. Data Transformation:

This is the second step in the data cleaning process. Commercial tools can assist in the data transformation step.

Data Migration tools:

It allows simple transformation to be specified, such as to replace the string “gender” by “sex”.

ETL Tools:

These tools allows user to specify transforms through a graphical user interface.

# data integration.

Data integration, which combines data from multiple sources into a data store as data warehousing.These sources may include multiple databases,data cubes or data files.

There are a number of issues to consider during data integration. They are as follows,

1. Schema integration and Objects matching.
2. Redundancy and inconsistency.
3. Detection and resolution of data value conflicts.

# Schema integration and Objects matching:

It can be tricky How can equivalent real-world entities from multiple data sources be matched up? This is referred to as the entity identification problem. For example, how can the data analyst or the computer be sure that customer\_id in one database and cust\_number in another refer to the same attribute? Examples of metadata for each attribute include the name, meaning, data type, and range of values permitted for the attribute, and null rules for handling blank, zero, or null values (Such metadata can be used to help avoid errors in schema integration. The metadata may also be used to help transform the data.

# Redundancy and inconsistency:

Redundancy is an important issue. An attribute may be redundant if it can be derived from another attribute or set of attributes.

Inconsistency occurs due to duplication, inaccuracy and partial updating.

# Detection and resolution of data value conflicts:

A third important issue in data integration is the detection and resolution of data value conflicts. For example, for the same real-world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling,or encoding. For instance, a weight attribute may be stored in metric units in one system and British Imperial units in another.

An attribute in one system may be recorded at a lower level of abstraction than the same attribute in another. For example, the total sales in one database may refer to one branch of All Electronics, while an attribute of the same name in another database may refer to the total sales for All Electronics stores in a given region.

When matching attributes from one database to another during integration, special attention must be paid to the structure of the data. This is to ensure that any attribute functional dependencies and referential constraints in the source system match those in the target system.

Careful integration of the data from multiple sources can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent mining process.

# data smoothing techniques.

Noise is a random error or variance in a measured variable.

Now we will see some of the following data smoothing techniques:

1. Binning.
2. Regression.
3. Clustering.

Binning Method:

This method removes the noisy data value from the sorted data set by communicating with its neighboring values.It is considered as local smoothing because it communicates with the values which are around it.Before initiating any of the binning methods,it is necessary to sort the given list either in increasing or decreasing order.The sorted list is then divided into a number of buckets or bins with equivalent frequency or width.

The binning process can be accomplished by using the following approaches.

a. By calculating the average mean of Bin:

This method calculates the average of data values for individual buckets and substitutes the existing values with the averaged value.

Ex:4,8,15,21,21,24,25,28,34

Partition into (equal frequency) bins:

Bin 1: 4,8,15

Bin 2: 21,21,24

Bin 3: 25,28,34

The average of bucket bin1, bin2, bin3 are 9,22,29 smoothing by means Bin 1: 9,9,9

Bin 2: 22,22,22

Bin 3: 29,29,29

b. By calculating the median:

This method is similar to that of bucket mean except that the median is calculated for individual bucket.

The median of the above list for bin1, bin2, bin3 are 8,21,28 Smoothing by bin median

Bin 1: 8,8,8

Bin 2: 21,21,21

Bin 3: 28,28,28

c. By calculating the boundaries:

In this method,the highest and lowest values in individual buckets are considered.These values are treated as boundaries of a bucket.Every data value in the bucket is substituted by the boundary value which is nearest to it.The highest and lowest values remain changed.

The highest values of bin 15,24,34 and lower values of bin 4,21,25. Smoothing by boundaries:

Bin 1: 4,4,15

Bin 2: 21,21,24

Bin 3: 25,25,34.

1. Regression:

Regression is a function that is used to remove noisy data.There are two types of regression functions:

1. Linear regression function
2. Multiple linear regression function.
3. Linear regression function:

It involves finding the “best” line to fit two attributes (or variables),so that one attribute can be used to predict the other.

1. Multiple linear regression:

It is extension of linear regression,where more than two attributes are involved and the data fit to multidimensional surface.

1. Clustering Method:

A cluster is defined as a group of data objects that have high similarity with objects belonging to same cluster and are dissimilar to objects belonging to other clusters.

Clusters can be used to delete the outliers.

# What is data reduction, data cube aggregation and selecting a subset of attributes.

Data reduction is a process of compressing massive volume of data into a limited data set.

# Data cube aggregation:

Data cube aggregation,where aggregation operations are applied to the data for construction of a data cube.Data cube store multidimensional aggregated information.Each cell holds an aggregated data value,corresponding to the data point in multidimensional space.Concept hierarchies may exist for each attribute,allowing the analyst of data at multiple levels of abstraction.Data cube provide fast access to precomputed,summarized data,there by benefiting online analytical processing as well as data mining.

The cube can be created in three ways,

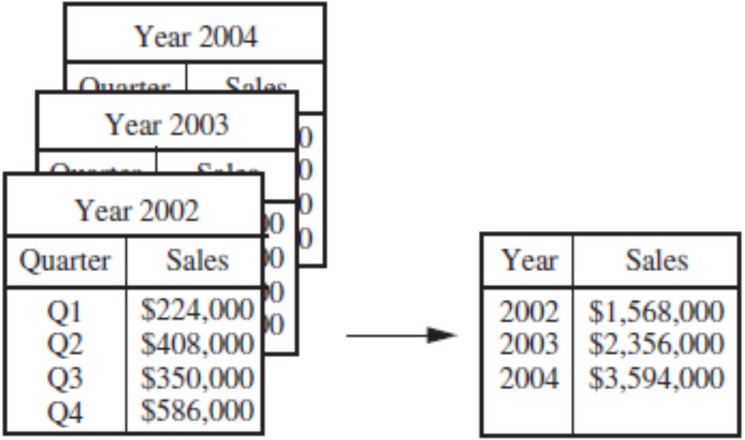
i. Base cuboid:

The cube created at the lowest level of abstraction is reffered to as base cuboid. ii.Lattice of cuboid:

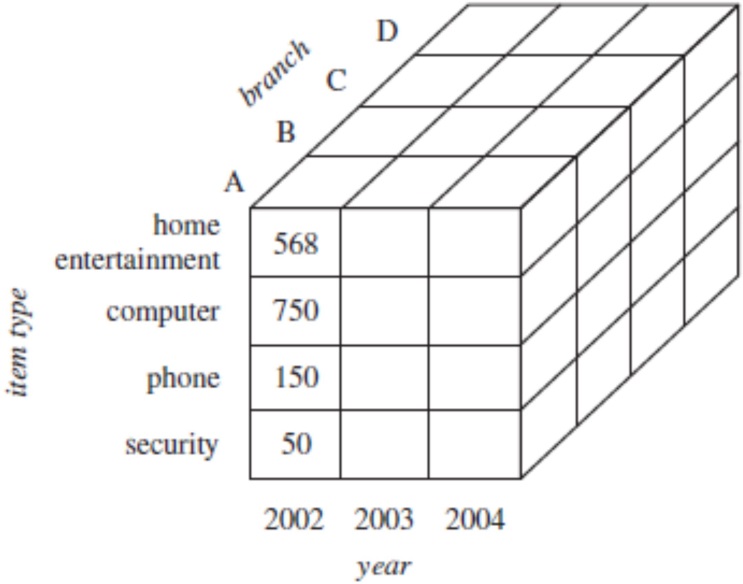
Data cubes created for varying levels of abstraction are often referred to as cuboids.

iii. Apex cuboid:

A cube at the highest level of abstraction is the apex cuboid.



On the left side,the sales are shown as per quarter.On the right,the data are aggregated to provide the annual sales.



The above figure shows a data cable for multidimensional analysis of sales data with respect to annual sales per item type for each branch.

# Attribute Subset Selection:

A database consists of the massive volume of data set which in turn are the collection of records.Each record consists of numerous attributes.Out of this attribute set many of the attributes are duplicate,inconsistent and irrelevant.It is very time consuming because data analysis is performed on all the attributes.The selection of irrelevant attributes can lead to poor quality problem,confusion and degradation in performance of the mining process.

In order to estimate the usage of irrelevant attributes a strategy called “Attribute subset selection “ is used. It reduces the data set size by removing irrelevant or redundant attributes. The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data class is as close as possible to the original distribution obtained using all attributes.

Basic heuristic methods for attribute subset selection include the following techniques, some of which are illustrated.

1. Stepwise forward selection:

The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set

Example:

Let us consider the following actual set of attributes,

{A1, A2, A3, A4, A5, A6}

Initial attributes set {A1, A2, A3, A4, A5, A6} Initial reduced set

* + { }
  + {A1}
  + {A1, A4}
  + {A1, A4, A6} Reduced attribute set.

1. Stepwise backward elimination:

The procedure starts with the full set of attributes.At each step, it removes the worst attribute remaining in the set.

Example:

Initial attributes set {A1, A2, A3, A4, A5, A6} Initial reduced set

* + {A1, A2, A3, A4, A5, A6}
  + {A1, A3, A4, A5, A6}
  + {A1,A4,A5,A6}
  + {A1, A4, A6} Reduced attribute set.

1. Combination of forward selection and backward elimination:

The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.

1. Decision tree induction:

Decision tree algorithms such as ID3,C4.5 and cart were originally intended of classification. A decision tree is a tree structure and consists of three types of nodes they are root nodes, internal nodes, and leaf or terminal nodes.

**Data Transformation**

Data transformation is a process of converting the integrated data into the correct format in order to continue the mining process.

The different techniques that are used for performing data transformation are

1. Smoothing of noisy data.
2. Aggregation of data cube.
3. Generalization of data.
4. Normalization of data.
5. Construction of attributes/features.

# Smoothing:

Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.

# Aggregation:

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.

# Generalization:

Generalization of the data, where low-level or raw data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like a street, can be generalized to higher-level concepts, like city or country.

# Normalization:

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.

An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0.

Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest-neighbor classification and clustering.

There are many methods for data normalization. Some of the methods used for normalization are

* 1. Min-max normalization,
  2. Z-score normalization,
  3. Normalization by decimal scaling.

# 5. Attribute Construction:

In attribute construction, new attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high- dimensional data.

# methods used for data discretization.

The methods used for data discretization are,

1. Supervised discretization.
2. Unsupervised discretization.
3. Top-down discretization.
4. Bottom-up discretization.

Data discretization techniques can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.

Discretization techniques can be categorized based on how the discretization is performed, such as whether it uses class information or which direction it proceeds (i.e., top-down vs. bottom-up).

The methods for data discretization are

1. Supervised dicretization:

If the discretization process uses class information, then it is supervised discretization.

1. Unsupervised discretization:

If the discretization process does not use any class information, but the data values are reduced by substituting them by limited interval descriptions, then it is unsupervised.

1. Top-down discretization (or) splitting:

If the discretization process starts by first finding one or a few points (called split points or cut points) to split the entire attribute range and then repeat the recursively on the resulting intervals, then it is called top down discretization.

1. Bottom-up discretization (or) merging:

This process starts by considering all the continuous attribute values as split points, removes some by merging neighborhood values to form intervals and then recursively apply this process to resulting intervals.