CHAPTER 1

Unit I

Introduction

Syllabus Topics

Data Mining, Data Mining Task Primitives, Data : Data, Information and Knowledge; Attribute Types : Nominal, Binary, Ordinal and Numeric attributes, Discrete versus Continuous Attributes; Introduction to Data Preprocessing, Data Cleaning : Missing values, Noisy data; Data integration : Correlation analysis; transformation: Min-max normalization, z-score normalization and decimal scaling; data reduction : Data Cube Aggregation, Attribute Subset Selection, sampling ; and Data Discretization : Binning, Histogram Analysis.

Syllabus Topic : Data Mining

1.1 Data Mining

- Data Mining is a new technology, which helps organizations to process data through algorithms to uncover meaningful patterns and correlations from large databases that otherwise may not be possible with standard analysis and reporting.
- Data mining tools can help to understand the business better and also improve future performance through predictive analytics and make them proactive and allow knowledge driven decisions.
- Issues related to information extraction from large databases, data mining field brings together methods from several domains like Machine Learning, Statistics, Pattern Recognition, Databases and Visualization.
- Data mining field finds its application in market analysis and management like for e.g. customer relationship management, cross selling, market segmentation. It can also be used in risk analysis and management for forecasting, customer retention, improved underwriting, quality control, competitive analysis and credit scoring.

Definition of Data Mining

→ (SPPU - Dec. 15)

Q. Define Data Mining. **Dec. 15, 2 Marks**

- Data mining is processing data to identify patterns and establish relationships.
- Data mining is the process of analysing large amounts of data stored in a data warehouse for useful information which makes use of artificial intelligence techniques, neural networks, and advanced statistical tools (such as cluster analysis) to reveal trends, patterns and relationships, which otherwise may be undetected.
- Data Mining is a non-trivial process of identifying :
  - Valid,
  - Novel,
  - Potentially useful, understandable patterns in data.

1.1.1 Applications of Data Mining

- Data Mining has been used in numerous areas, which include both private as well as public sectors.
- The use of Data mining in major industry areas like Banking, Retail, Medicine, insurance can help reduce costs, increase their sales and enhance research and development.

- For example in banking sector data mining can be used for customer retention, fraud prevention by credit card approval and fraud detection.
- Prediction models can be developed to help analyze data collected over years. For e.g. customer data can used to find out whether the customer can avail loan from the bank, or an accident claim is fraudulent and needs further investigation.
- Effectiveness of a medicine or certain procedure may be predicted in medical domain by using data mining.
- Data mining can be used in Pharmaceutical firms as a guide to research on new treatments for diseases, by analyzing chemical compounds and genetic materials.
- A large amount of data in retail industry like purchasing history, transportation services may be collected for analysis purpose. This data can help multidimensional analysis, sales campaign effectiveness, customer retention and recommendation of products and much more.
- Telecommunication industry also uses data mining, for e.g. they may do analysis based on the customer data which of them are likely to remain as subscribers and which one will shift to competitors.

1.1.2 Challenges to Data Mining

→ **(SPPU - Oct. 16)**

Q. Describe three challenges to data mining regarding data mining methodology. **Oct. 16, 6 Marks**

1) Mining different kinds of knowledge in databases

- Different users are interested in different kinds of knowledge and will require a wide range of data analysis and knowledge discovery tasks such as data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis.
- Each of these tasks will use the same database in different ways and will require different data mining techniques.

2) Interactive mining of knowledge at multiple levels of abstraction

- Interactive mining, with the use of OLAP operations on a data cube, allows users to focus the search for patterns, providing and refining data mining requests based on returned results.
- The user can then interactively view the data and discover patterns at multiple granularities and from different angles.

3) Incorporation of background knowledge

- Background knowledge, or information regarding the domain under study such as integrity constraints and deduction rules, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.
- This helps to focus and speed up a data mining process or judge the interestingness of discovered patterns.

1.1.3 KDD Process (Knowledge Discovery in Databases)

→ **(SPPU - Aug. 17)**

Q. Explain the knowledge discovery in database (KDD) with diagram. What is the role of data mining steps in KDD? **Aug. 17, 6 Marks**

- The process of discovering knowledge in data and application of data mining methods refers to the term **Knowledge Discovery in Databases (KDD).**
- It includes a wide variety of application domains, which include Artificial Intelligence, Pattern Recognition, Machine Learning Statistics and Data Visualisation.
- The main goal includes extracting knowledge from large databases, the goal is achieved by using various data mining algorithms to identify useful patterns according to some predefined measures and thresholds.

☞ **Outline steps of the KDD process**

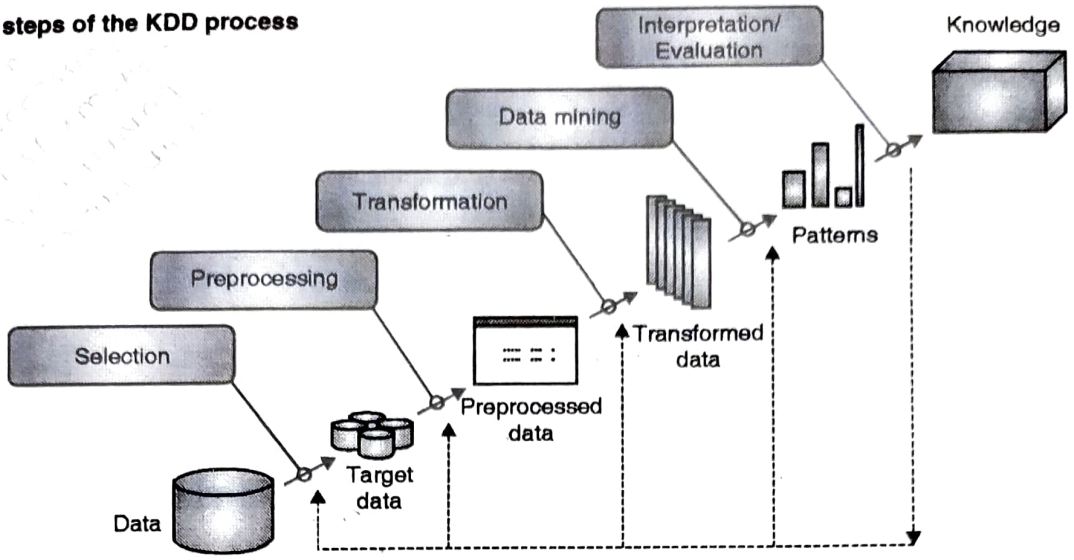


Fig. 1.1.1 : KDD Process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps :

1. Developing an understanding of

(i) The application domain

(ii) The relevant prior knowledge

(iii) The goals of the end-user.

2. Creating a target data set

Selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.

3. Data cleaning and pre-processing

(i) Noise or outliers are removed.

(ii) Essential information is collected for modeling or accounting for noise.

(iii) Missing data fields are handled by using appropriate strategies.

(iv) Time sequence information and changes are maintained.

4. Data reduction and projection

(i) Based on the goal of the task, useful features are found to represent the data.

(ii) The number of variables may be effectively reduced using methods like dimensionality reduction or transformation. Invariant representations for the data may also be found out.

5. Choosing the data mining task

Selecting the appropriate Data mining tasks like classification, clustering, regression based on the goal of the KDD process.

6. Choosing the data mining algorithm(s)

(i) Pattern search is done using the appropriate Data Mining method(s).

(ii) A decision is taken on which models and parameters may be appropriate.

(iii) Considering the overall criteria of the KDD process a match for the particular data mining method is done.

7. Data mining

Using a representational form or other representations like classification, rules or trees, regression clustering for searching patterns of interest.

8. Interpreting mined patterns

9. Consolidating discovered knowledge

The terms *knowledge discovery* and *data mining* are distinct.

KDD	Data Mining
KDD is a field of computer science, which helps humans in extracting useful, previously undiscovered knowledge from data. It makes use of tools and theories for the same.	Data Mining is one of the step in the KDD process, it applies the appropriate algorithm based on the goal of the KDD process for identifying patterns from data.

1.1.4 Architecture of a Typical Data Mining System

Architecture of a typical data mining system may have the following major components as shown in Fig. 1.1.2.

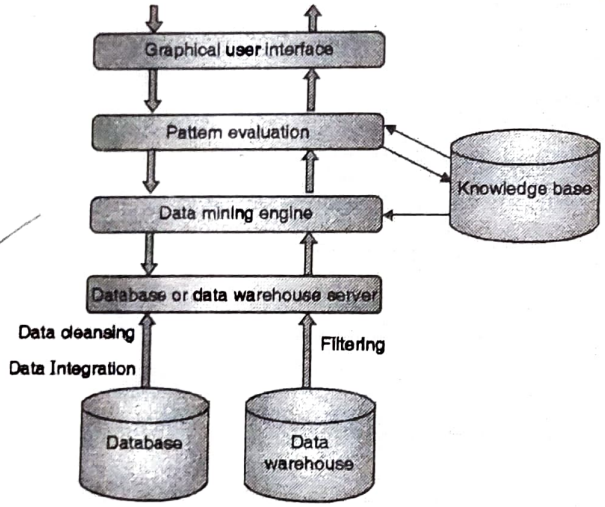


Fig. 1.1.2 : Architecture of typical data mining system

1. **Database, data warehouse, or other information repository**

   These are information repositories. Data cleaning and data integration techniques may be performed on the data.

2. **Databases or data warehouse server**

   It fetches the data as per the user's requirement which is need for data mining task.

3. **Knowledge base**

   This is used to guide the search, and gives the interesting and hidden patterns from data.

4. **Data mining engine**

   It performs the data mining task such as characterization, association, classification, cluster analysis etc.

5. **Pattern evaluation module**

   It is integrated with the mining module and it helps in searching only the interesting patterns.

6. **Graphical user interface**

   This module is used to communicate between user and the data mining system and allow users to browse database or data warehouse schemas.

Syllabus Topic : Data Mining Task Primitives

1.2 Data Mining Task Primitives

Data mining primitives define a data mining task, which can be specified in the form of a data mining query.

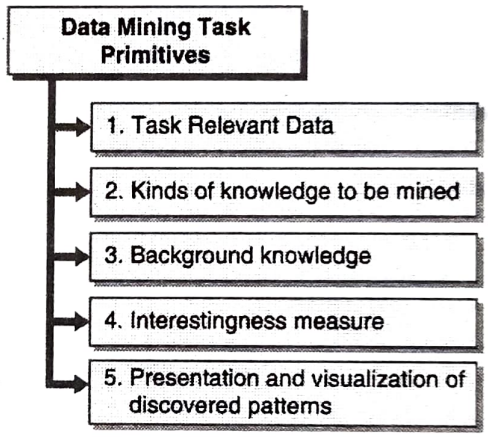


Fig. 1.2.1 : Data Mining Task Primitives

➜ **1. Task relevant data**

- Specify the data on which the data mining function to be performed.
- Using relational query, a set of task relevant data can be collected.
- Before data mining analysis, data can be cleaned or transformed.
- Minable view is created i.e. the set of task relevant data for data mining.

➜ **2. The kind of knowledge to be mined**

- Specify the knowledge to be mined.

- Kinds of knowledge include concept description, association, classification, prediction and clustering.
- User can also provide pattern templates. Also called metapatterns or metarules or metaqueries.

3. Background knowledge

- It is the information about the domain to be mined.
- Concept hierarchies is the form of background knowledge which helps to discover the knowledge at multiple levels of abstraction.

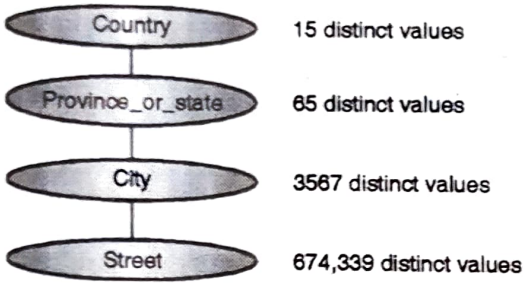


Fig. 1.2.2 : Concept hierarchy for the dimension location

Four major types of concept hierarchies

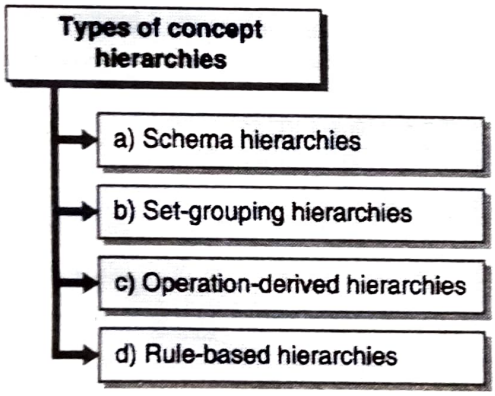


Fig. 1.2.3 : Types of hierarchies

a) Schema hierarchies

It is the total or partial order among attributes in the database schema.

Example : Location hierarchy as street < city < province/state < country

b) Set-grouping hierarchies

It organizes values into sets or groups of constants.

Example : For attribute salary, a set-grouping hierarchy can be specified in terms of ranges as in the following :

{low, avg, high} € all (salary)

{1000....5000} € low

{5001....10000} € avg

{10001....15000} € high

c) Operation-derived hierarchies

It is based on operation specified which may include decoding of information-encoded strings, information extraction from complex data objects, data clustering.

Example : URL or email address

xyz@cs.mu.in gives login name

< dept. <univ.< country

d) Rule-based hierarchies

It occurs when either whole or portion of a concept hierarchy is defined as a set of rules and is evaluated dynamically based on current database data and rule definition.

Example

Following rules are used to categorize items as *low_profit*, *medium_profit* and *high_profit_margin*.

low_profit_margin (Z) <= price (Z, A1)^cost (Z, A2)^((A1-A2)<80)

medium_profit_margin

(Z) <= price (Z, A1)^cost (Z, A2)^((A1-A2) ≥ 80)^((A1-A2)≤350)

high_profit_margin (Z) <= price (Z, A1)^cost (Z, A2)^(A1-A2)>350)

4. Interestingness measures

- It is used to confine the number of uninteresting patterns returned by the process.
- Based on the structure of patterns and statistics underlying them.
- Each measure is associated a threshold which can be controlled by the user.
- Patterns not meeting the threshold are not presented to the user.

- **Objective measures of pattern interestingness :**
  - **Simplicity :** A patterns interestingness is based on its overall simplicity for human comprehension.
  - **Example :** Rule length is a simplicity measure
  - **Certainty (confidence) :** Assesses the validity or trustworthiness of a pattern. Confidence is a certainty measure.

$$\text{Confidence } (A \Rightarrow B) = \left(\frac{\text{(Number of tuples containing both A and B)}}{\text{(Number of tuples containing A)}} \right)$$

  - **Utility (support) :** It is the usefulness of a pattern support.

$$(A \Rightarrow B) = \frac{\text{(Number of tuples containing both A and B)}}{\text{(total number of tuples)}}$$

  - **Novelty :** Patterns contributing new information to the given pattern set are called **novel patterns**. (**Example :** Data exception).

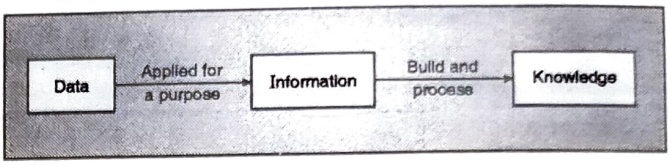
5. Presentation and visualization of discovered patterns

- Data mining systems should be able to display the discovered patterns in multiple forms, such as rules, tables, crosstabs (cross-tabulations), pie or bar charts, decision trees, cubes, or other visual representations.
- User must be able to specify the forms of presentation to be used for displaying the discovered patterns.

Syllabus Topic : Data - Data, Information and Knowledge

1.3 Data : Data, Information and Knowledge

- **Data** represents a single primary entities and the related transaction of that entity. Data are facts, which are not processed or analyzed. Example : "The price of petrol is Rs. 80 per litre".
- **Information** is obtained after processing the data and then data has been interpreted and analysed. Such information is meaningful and useful to the user. Example : "The price of petrol is increased from Rs. 80 to Rs. 85 in last 3 months". This information is useful for user who keeps a track of the petrol prices.
- **Knowledge** is useful to take decisions and actions for business. Information is transformed into knowledge. Example "When the petrol prices increases, it is likely that transportation cost also increases".
- So to get boundaries between data, information and knowledge is not easy. Sometimes data may be information for others. But finally knowledge helps to take action for business and delivers the matrix or value for decision makers to take decisions.



(a) From data to information to knowledge

(b) Knowledge leads to action and BI delivers value for decision makers

Fig 1.3.1

Syllabus Topic : Attributes Types - Nominal, Binary, Ordinal and Numeric Attributes, Discrete Versus Continuous Attributes

1.4 Attributes Types

Data Objects

- A data object is a logical cluster of all tables in the data set which contains data related to the same entity. It also represents an object view of the same.
- **Example :** In a product manufacturing company, product, customer are objects. In a retail store, employee, customer, items and sales are objects.
- Every data object is described by its properties called as attributes and it is stored in the database in the form of a row or tuple. The columns of this data tuple are known to be attributes.

Attributes types

- An attribute is a property or characteristic of a data object. For e.g. Gender is a characteristic of a data object person.

- The attributes may have values like :

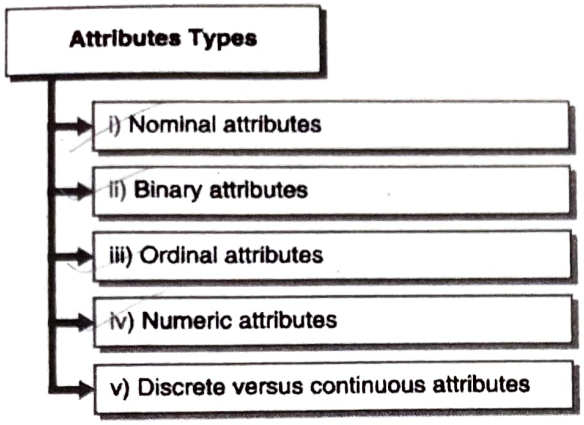


Fig. 1.4.1 : Attributes Types

➜ i) Nominal attributes

- Nominal attributes are also called as **Categorical attributes** and allow for only qualitative classification.
- Every individual item has a certain distinct categories, but quantification or ranking the order of the categories is not possible.
- The nominal attribute categories can be numbered arbitrarily.
- Arithmetic and logical operations on the nominal data cannot be performed.

Examples

- Typical examples of such attributes are :

Car owner :	1. Yes 2. No
Employment status :	1. Unemployed 2. Employed

➜ ii) Binary attributes

- A nominal attribute which has either of the two states 0 or 1 is called Binary attribute, where 0 means that the attribute is absent and 1 means that it is present.
- **Symmetric binary variable :** If both of its states i.e. 0 and 1 are equally valuable. Here we cannot decide which outcome should be 0 and which outcome should be 1. Example : Marital status of a person is "Married or Unmarried". In this case both are equally valuable and difficult to represent in terms of 0(absent) and 1(present).
- **Asymmetric binary variable :** If the outcome of the states are not equally important. An example of such a variable is the presence or absence of a relatively rare attribute. For example : Person is "handicapped or not handicapped". The most important outcome is usually coded as 1 (present) and the other is coded as 0 (absent).

➜ iii) Ordinal attributes

- A discrete ordinal attribute is a nominal attribute, which have meaningful order or rank for its different states.
- The interval between different states is uneven due to which arithmetic operations are not possible, however logical operations may be applied.

Examples

- Considering age as an ordinal attribute, it can have three different states based on an uneven range of age value. Similarly income can also be considered as an ordinal attribute, which is categorised as low, medium, high based on the income value.

Age :	1. Teenage 2. Young 3. Old
Income :	1. Low 2. Medium 3. High

➜ iv) Numeric attributes

Numeric Attributes are quantifiable. It can be measured in terms of a quantity, which can either have an integer or real value. They can be of two types,

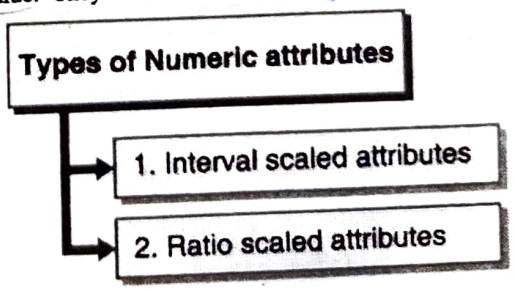


Fig. 1.4.2 : Types of Numeric attributes

➜ 1. Interval scaled attributes

- Interval scaled attributes are continuous measurement on a linear scale. Example : weight, height and weather temperature.

- These attributes allow for ordering, comparing and quantifying the difference between the values. An interval-scaled attributes has values whose differences are interpretable.

➔ **2. Ratio scaled attributes**

- Ratio scaled attributes are continuous positive measurements on a non linear scale. They are also interval scaled data but are not measured on a linear scale.
- Operations like addition, subtraction can be performed but multiplication and division are not possible.
- **Example :** For instance, if a liquid is at 40 degrees and we add 10 degrees, it will be 50 degrees. However, a liquid at 40 degrees does not have twice the temperature of a liquid at 20 degrees because 0 degrees does not represent "no temperature".
- There are three different ways to handle the ratio-scaled variables :
  - As interval scale variables. The drawback of handling them as interval scaled is that it can distort the results.
  - As continuous ordinal scale.
  - Transforming the data (for example, logarithmic transformation) and then treating the results as interval scaled variables.

➔ **v) Discrete versus continuous attributes**

- If an attribute can take any value between two specified values then it is called as continuous else it is discrete. An attribute will be continuous on one scale and discrete on another.
- ***For example :*** If we try to measure the amount of water consumed by counting the individual water molecules then it will be discrete else it will be continuous.
- ***Examples of continuous attributes*** includes time spent waiting, direction of travel, water consumed etc.
- ***Examples of discrete attributes*** includes voltage output of a digital device, a person's age in years.

Syllabus Topic : Introduction to Data Pre-processing

1.5 Introduction to Data Pre-processing

Q. What is data pre-processing ? (2 Marks)

- Process that involves transformation of data into information through classifying, sorting, merging, recording, retrieving, transmitting, or reporting is called **data processing**. Data processing can be manual or computer based.
- In Business related world, data processing refers to data processing so as to enable effective functioning of the organisations and businesses.
- Computer data processing refers to a process that takes the data input via a program and summarizes, analyse the same or convert it to useful information.
- The processing of data may also be automated.
- Data processing systems are also known as **information systems.**
- When data processing does not involve any data manipulation and only converts the data type it may be called as **data conversion.**

1.6 Different Forms of Data Pre-processing

➔ **(SPPU - Oct. 16, Dec. 16)**

Q. What are the major tasks in data preprocessing ? Explain them in brief. Oct. 16, Dec. 16, 6 Marks

Q. Explain different steps in data preprocessing. (6 Marks)

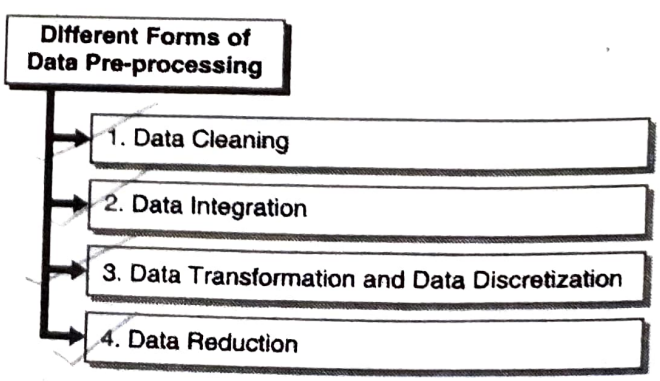


Fig. 1.6.1 : Different Forms of Data Pre-processing

Syllabus Topic : Data Cleaning

1.6.1 Data Cleaning

Data cleaning is also known as scrubbing. The data cleaning process detects and removes the errors and inconsistencies and improves the quality of the data. Data quality problems arise due to misspellings during data entry, missing values or any other invalid data.

Reasons for "Dirty" Data

- Dummy values
- Absence of data
- Multipurpose fields
- Cryptic data
- Contradicting data
- Inappropriate use of address lines
- Violation of business rules
- Reused primary keys
- Non-unique identifiers
- Data integration problems.

Why data cleaning or cleansing is required ?

- Source Systems data is not clean; it contains certain errors and inconsistencies.
- Specialised tools are available which can be used for cleaning the data.
- Some of the Leading data cleansing vendors include Validity (Integrity), Harte-Hanks (Trillium) and First logic.

1.6.1(A) Steps in Data Cleansing

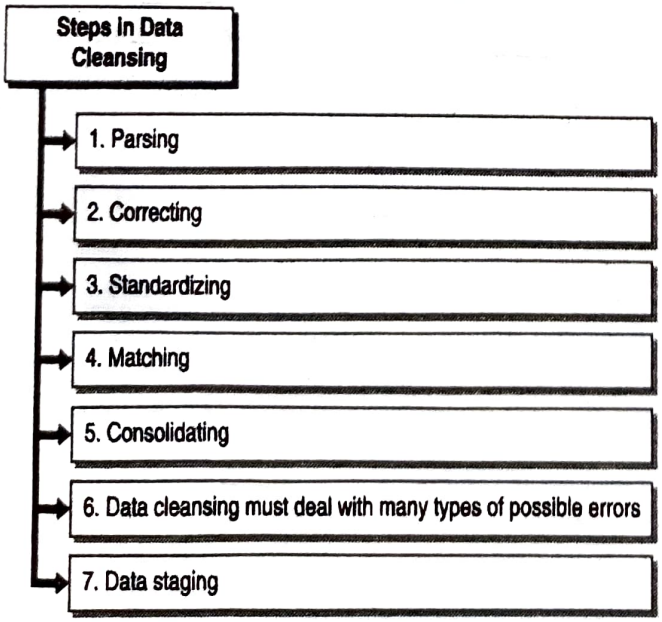


Fig. 1.6.2 : Steps in Data Cleansing

1. Parsing

- Parsing is a process in which individual data elements are located and identified in the source systems and then these elements are isolated in the target files.
- **Example :** Parsing of name into First name, Middle name and Last name or parsing the address into street name, city, state and country.

2. Correcting

- This is the next phase after parsing, in which individual data elements are corrected using data algorithm and secondary data sources.
- **Example :** In the address attribute replacing a vanity address and adding a zip code.

3. Standardizing

- In standardizing process conversion routines are used to transform data into a consistent format using both standard and custom business rules.
- **Example :** addition of a prename, replacing a nickname and using a preferred street name.

4. Matching

- Matching process involves eliminating duplications by searching and matching records with parsed, corrected and standardised data using some standard business rules.
- For example, identification of similar names and addresses.

5. Consolidating

Consolidation involves merging the records into one representation by analysing and identifying relationship between matched records.

6. Data cleansing must deal with many types of possible errors

- Data can have many errors like missing data, or incorrect data at one source.
- When more than one source is involved there is a possibility of inconsistency and conflicting data.

7. Data staging

- Data staging is an interim step between data extraction and remaining steps.
- Using different processes like native interfaces, flat files, FTP sessions, data is accumulated from asynchronous sources.
- After a certain predefined interval, data is loaded into the warehouse after the transformation process.
- No end user access is available to the staging file.
- For data staging, operational data store may be used.

Syllabus Topic : Missing Values

1.6.1(B) Missing Values

(SPPU - May 16, Dec. 16, Dec. 17)

Q. Describe the various methods for handling the missing values. **May 16, 6 Marks**

Q. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. **Dec. 16, 6 Marks**

Q. What are missing values? Explain methods to handle missing values. **Dec. 17, 6 Marks**

Missing data values

- This involves searching for empty fields where values should occur.
- Data preprocessing is one of the most important stages in data mining. Real world data is incomplete, noisy or inconsistent, this data is corrected in data preprocessing process by filling out the missing values, smoothening out the noise and correcting inconsistencies.
- There are several techniques for dealing with missing data, choosing one of them would be dependent on problems domain and the goal for data mining process.

Following are the different ways for handle missing values in databases :

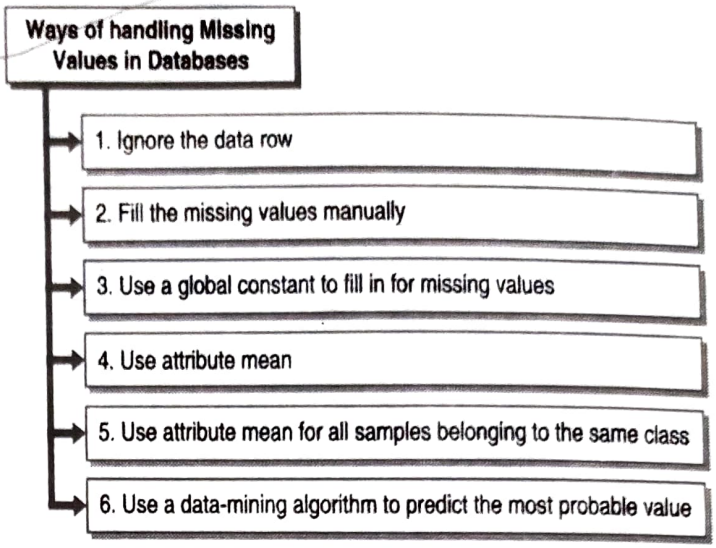


Fig. 1.6.3 : Ways of handling Missing Values in Databases

1. Ignore the data row

- In case of classification suppose a class label is missing for a row, such a data row could be ignored, or many attributes within a row are missing even in this case data row could be ignored. If the percentage of such rows is high it will result in poor performance.
- **Example :** Suppose we have to build a model for predicting student success in college. For this purpose a student's database having information about age, score, address, etc and column classifying their success in college to "LOW", "MEDIUM" and "HIGH". In this the data rows in which the success column is missing. These types of rows are of no use in the model therefore they can be ignored.

2. Fill the missing values manually

This is not feasible for large data set and also time consuming.

3. Use a global constant to fill in for missing values

- When missing values are difficult to be predicted, a global constant value like "unknown", "N/A" or "minus infinity" can be used to fill all the missing values.
- **Example :** Consider the students database, if the address attribute is missing for some students it does not makes sense in filling up these values rather a global constant can be used.

➜ **4. Use attribute mean**

- For missing values, mean or median of its discrete values may be used as a replacement.
- Example : In a database of family incomes, missing values may be replaced with the average income.

➜ **5. Use attribute mean for all samples belonging to the same class**

- Instead of replacing the missing values by mean or median of all the rows in the database, rather we could consider class wise data for missing values to be replaced by its mean or median to make it more relevant.
- **Example :** Consider a car pricing database with classes like "luxury" and "low budget" and missing values need to filled in, replacing missing cost of a luxury car with average cost of all luxury car makes the data more accurate.

➜ **6. Use a data-mining algorithm to predict the most probable value**

- Missing values may also be filled up by using techniques like regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms.
- For example, clustering method may be used to form clusters and then the mean or median of that cluster may be used for missing value. Decision tree may be used to predict the most probable value based on the other attributes.

Syllabus Topic : Noisy Data

1.6.1(C) Noisy Data

- A random error or variance in a measure variable is known as noise.
- **Noise in the data may be introduced due to :**
  - Fault in data collection instruments.
  - Error introduced at data entry by a human or a computer.
  - Data transmission errors.
- **Different types of noise in data :**
  - Unknown encoding : Gender : E
  - Out of range values : Temperature : 1004, Age : 125
  - Inconsistent entries : DoB : 10-Feb-2003; Age : 30
  - Inconsistent formats : DoB : 11-Feb-1984; DoJ : 2/11/2007

☞ **How to handle noisy data ?**

Different data smoothing techniques are given below :

1. Binning

- Considering the neighbourhood of the sorted data smoothening can be applied.
- The sorted data is placed into bins or buckets.
- Smoothing by bin means.
- Smoothing by bin medians.
- Smoothing by bin boundaries.

☞ **Different approaches of binning**

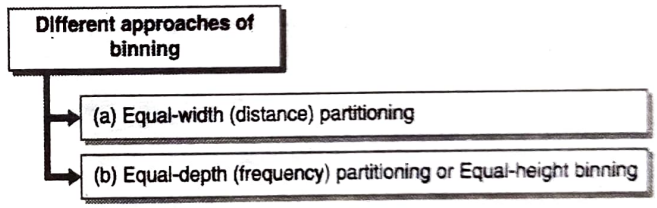


Fig. 1.6.4 : Different approaches of binning

➜ **(a) Equal-width (distance) partitioning**

- Divides the range into *N* intervals of equal size: uniform grid.

$$\text{bin width} = (\text{max value} - \text{min value}) / N$$

- **Example :** Consider a set of observed values in the range from 0 to 100.

The data could be placed into 5 bins as follows :

$$\text{width} = (100 - 0)/5 = 20$$

Bins formed are :

[0-20], (20-40], (40-60], (60-80], (80-100]

- The first and the last bin is extended to allow values outside the range : (-infinity-20], (20-40], (40-60], (60-80], (80-infinity)

Disadvantages

- Outliers in the data may be a problem.
- Skewed data cannot be held with this method.

(b) Equal-depth (frequency) partitioning or Equal-height binning

- The entire range is divided into N intervals, each containing approximately the same number of samples.
- This results in good data scaling.
- Handling categorical attributes may be a problem.
- **Example :** Let us consider sorted data for e.g. Price in INR

  4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equal-depth) bins: (N=3)

  Bin 1 : 4, 8, 9, 15

  Bin 2 : 21, 21, 24, 25

  Bin 3 : 26, 28, 29, 34
- **Smoothing by bin means**

  Replace each value of bin with its mean value.

  Bin 1 : 9, 9, 9, 9

  Bin 2 : 23, 23, 23, 23

  Bin 3 : 29, 29, 29, 29
- **Smoothing by bin boundaries**

  In this method the minimum and maximum values of the bean boundaries is found and each value is replaced with its nearest value either minimum or maximum.

  Bin 1 : 4, 4, 4, 15

  Bin 2 : 21, 21, 25, 25

  Bin 3 : 26, 26, 26, 34

2. Outlier analysis by clustering

- Partition data set into clusters and one can store cluster representation only, i.e. replace all values of the cluster by that one value representing the cluster.
- Outliers can be detected by using clustering techniques, where related values are organized into groups or clusters.

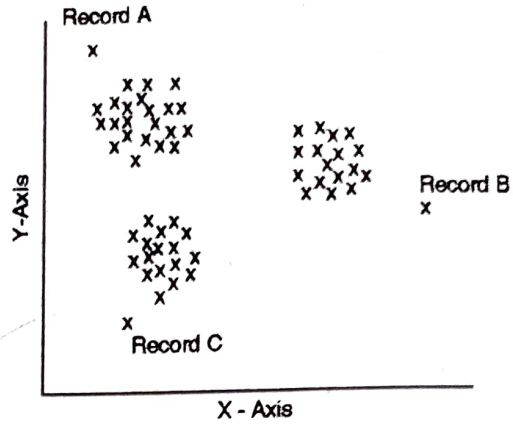


Fig. 1.6.5 : Graphical Example of Clustering

- Perform clustering on attributes values and replace all values in the cluster by a cluster representative.

3. Regression

- Regression is a statistical measure used to determine the strength of the relationship between one dependent variable denoted by Y and a series of independent changing variables.
- Smooth by fitting the data into regression functions.
- Use regression analysis on values of attributes to fill missing values.
- The two basic types of regression are linear regression and multiple regressions.
- The difference between Linear and multiple regressions is that former uses one independent variable to predict the outcome, while the later uses two or more independent variables to predict the outcome.
- The general form of each type of regression is :

Linear Regression : $Y = a + bX + u$

Multiple Regression : $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where, Y = The variable that we are trying to predict

X = The variable that we are using to predict Y

a = The intercept

b = The slope

u = The regression residual.

- In multiple regressions each variable is differentiated with subscripted numbers.

- Regression uses a group of random variables for prediction and finds a mathematical relationship between them. This relationship is depicted in the form of a straight line (Linear regression) that approximates all the points in the best way.
- Regression may be used to determine for e.g. price of a commodity, interest rates, the price movement of an asset influenced by industries or sectors.

Log linear model

- In Log linear regression a best fit between the data and a log linear model is found.
- **Major assumption :** A linear relationship exists between the log of the dependent and independent variables.
- Log linear models are models that postulate a linear relationship between the independent variables and the logarithm of the dependent variable.

  For example : $\log(y) = a_0 + a_1 x_1 + a_2 x_2 \ldots + a_N x_N$

  where y is the dependent variable; x_i, i = 1,...,N are independent variables and $\{a_i$, i = 0,...,N$\}$ are parameters (coefficients) of the model.
- Log linear models are widely used to analyze categorical data represented as a contingency table. In this case, the main reason to transform frequencies (counts) or probabilities to their log-values is that, provided the independent variables are not correlated with each other, the relationship between the new transformed dependent variable and the independent variables is a linear (additive) one.

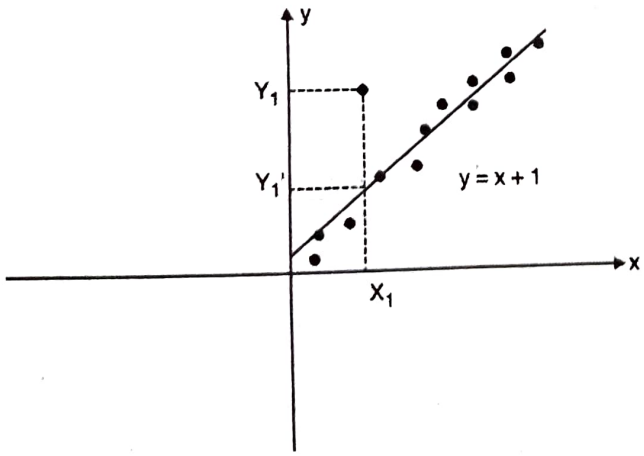


Fig. 1.6.6 : Regression example

1.6.1(D) Inconsistent Data

- The state in which the data quality of the existing data is understood and the desired quality of the data is known refers to consistent data quality.
- It is a state in which the existent data quality is been modified to meet the current and future business demands.

Syllabus Topic : Data Integration

1.6.2 Introduction to Data Integration

A coherent data store (e.g. a Data warehouse) is prepared by collecting data from multiple sources like multiple databases, data cubes or flat files.

Issues in data integration

- **Schema integration**
  - Integrate metadata from different sources.
  - Entity identification problem: identify real world entities from multiple data sources, e.g. A.cust-id ≡B.cust-#.
- **Detecting and resolving data value conflicts**
  - As the data is collected from multiple sources, attribute values are different for the same real world entity.
  - Possible reasons include different representations, different scales, e.g. metric vs. British units.
- **Redundant data occur due to integration of multiple databases**
  - Attributes may be represented in different names in different sources of data.
  - An attribute may be derived attribute in another table, e.g. yearly income.
  - With the help of co-relational analysis, detection of redundant data is possible.
  - The redundancies or inconsistencies may be reduced by careful integration of the data from multiple sources, which will help in improving mining speed and quality.

1.6.2(A) Entity Identification Problem

- Schema integration is an issue as to integrate metadata from different sources is a difficult task.
- Identify real world entities from multiple data sources and their matching is the entity identification problem.
- For example, Roll number in one database and enrolment number in another database refers to the same attribute.
- Such conflicts may create problem for schema integration.
- Detecting and resolving data value conflicts for the same real world entity, attribute values from different sources are different.

Syllabus Topic : Correlation Analysis

1.6.2(B) Redundancy and Correlation Analysis

- Data redundancy occurs when data from multiple sources is considered for integration.
- Attribute naming may be a problem as same attributes may have different names in multiple databases.
- An attribute may be derived attribute in another table e.g. "yearly income".
- Redundancy can be detected using correlation analysis.
- To reduce or avoid redundancies and inconsistencies data integration must be carried out carefully. This will also improve mining algorithm speed and quality.
- X^2 (*Chi-square*) test can be carried out on nominal data to test how strongly the two attributes are related.
- Correlation coefficient and covariance may be used with numeric data, this will give a variation between the attributes.

The X^2 (Chi-square)

- It is used to test hypotheses about the shape or proportions of a population distribution by means of sample data.
- For nominal data, a correlation relationship between two attributes, P and Q, can be discovered by an X^2 (Chi-square) test.
- These nominal variables, also called "attribute variables" or "categorical variables", classify observations into a small number of categories, which are not numbers. It doesn't work for numeric data.
- Examples of nominal variables include Gender (the possible values are male or female), Marital Status (Married, unmarried or divorced), etc.
- The Chi-square test is used to test the probability of independence of a distribution of data but does not gives you any details about the relationship between them.
- Chi-square test is defined by,

$$X^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

Where X^2 = Chi-square

E = Frequency expected which is the amount of subjects that you would *expect* to find in each category based on known information.

O = Frequency observed which is the amount of subjects you *actually* found to be in each category in the present data.

- **Degrees of freedom :** The degrees of freedom(DF) is equal to :

$$DF = (r - 1) * (c - 1)$$

where, r is the number of levels for one categorical variable and c is the number of levels for the other categorical variable.

- **Expected frequencies :** It is the count which is computed for each level of categorical attribute. The formula for expected frequency is

$$E_{r,c} = (n_r * n_c) / n$$

  - Where $Er_{,c}$ is the expected frequency count for level r of attribute X and level c of attribute Y,
  - n_r is the sum of sample observations at level r of attribute X,
  - n_c is the sum of sample observations at level c of attribute Y,
  - n is the total size of sample data.

Syllabus Topic : Data Transformation and Data Discretization - Min-max Normalization, Z-Score Normalization and Decimal Scaling

1.6.3 Data Transformation and Data Discretization

1.6.3(A) Data Transformation

Q. Explain the data transformation in detail. (4 Marks)

- Operational databases keep changing with the requirements, a data warehouse integrating data from these multiple sources typically faces the problem of inconsistency.
- To deal with these inconsistent data, transformation process may be employed.
- The most commonly used process is "Attribute Naming Inconsistency", as it is very common to use different names to the same attribute in different sources of data.
- E.g. Manager Name may be MGM_NAME in one database, MNAME in the other.
- In this one set of data names is considered and used consistently in the data warehouse.
- Once the naming consistency is done, they must be converted to a common format.
- The conversion process involves the following :
  - (i) ASCII to EBCDIC or vice versa conversion process may be used for characters.
  - (ii) To ensure consistency uppercase representation may be used for mixed case text.
  - (iii) A common format may be adopted for numerical data.
  - (iv) Standardisation must be applied for data format.
  - (v) A common representation may be used for measurement e.g. (Rs/$).
  - (vi) A common format must be used for coded data (e.g. Male/Female, M/F).
- The above conversions are automated and many tools are available for the transformation e.g. DataMapper.

☞ **Data transformation can have the following activities**

- **Smoothing :** It involves removal of noise from the data.
- **Aggregation :** It involves summarisation and data cube construction.
- **Generalization :** In generalization data is replaced by higher level concepts using concept hierarchy.
- **Normalization :** In normalization, attribute scaling is performed for a specified range.

  Example : To transform V in [min, max] to V′ in [0,1], apply

  $$V' = (V\text{-Min}) / (\text{Max-Min})$$

  Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers) :

  $$V' = (V\text{-Mean}) / \text{Std. Dev.}$$

- **Attribute/feature construction :** In this process new attributes may be constructed and used for data mining process

1.6.3(B) Data Discretization

- The range of a continuous attribute is divided into intervals.
- Categorical attributes are accepted by only a few classification algorithms.
- By Discretization the size of the data is reduced and prepared for further analysis.
- Dividing the range of attributes into intervals would reduce the number of values for a given continuous attribute.
- Actual data values may be replaced by interval labels.
- Discretization process may be applied recursively on an attribute.

1.6.3(C) Data Transformation by Normalization

➔ **(SPPU - May 17)**

Q. What are the different data normalization methods? Explain them in brief. May 17, 6 Marks

- Data Transformation by Normalization or standardization is the process of making an entire set of values have a particular property.

- Following methods may be used for normalization :

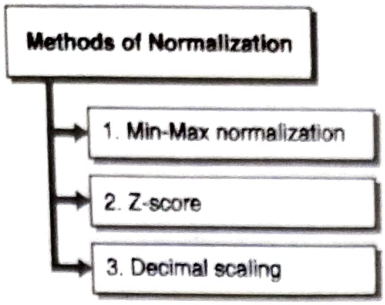


Fig. 1.6.7 : Methods of Normalization

➜ 1. Min-Max normalization

- Min-max normalization results in a linear alteration of the original data. The values are within a given range.

Following formula may be used to perform mapping a v value, of an attribute A from range [minA,maxA] to a new range [new_minA,new_maxA],

$$v' = (v - minA)/(maxA - minA) * (new\_maxA - new\_minA) + new\_minA$$

v = 73600 in [12000,98000]

v' = 0.716 in [0,1] (new range)

Ex. 1.6.1 : Consider the following group of data : 200, 300, 400, 600, 1000

(i) Use the min-max normalization to transform value 600 onto the range [0.0, 1.0]

(ii) Use the decimal scaling to transfer value 600. **SPPU - Oct. 16, 4 Marks**

Soln. :

(i) Min = Minimum value of the given data = 200

Max = Maximum value of the given data = 1000

$$V = 600 = \left(\frac{(V - min)}{(max - min)}\right) * (1 - 0) + 0$$

$$= \frac{600 - 200}{1000 - 200} = \left(\frac{400}{800}\right) * 1 = 0.5$$

(ii) Decimal scaling for 600

10^K is 10^3 = 1000

$\frac{600}{1000}$ = 0.6

➜ 2. Z-score

- In Z-score normalization, data is normalized based on the mean and standard deviation. Z-score is also known as Zero mean normalization.

$$v' = (v - meanA) / std\_devA$$

Where, MeanA = sum of the all attribute value of A

std_devA = Standard deviation of all values of A

Example

If sample data {10, 20, 30}, then

Mean = 20

std_dev = 10

So v' = (– 1, 0, 1)

➜ 3. Decimal scaling

- Based on the maximum absolute value of the attributes the decimal point is moved. This process is called as **Decimal Scale Normalization.**

$$v'(i) = v(i)/10^k \text{ for the smallest } k \text{ such that } \max(|v'(i)|) < 1.$$

Example : For the range between – 991 and 99,

10^k is 1000 (k = 3 as we have maximum 3 digit number in the range)

v'(– 991) = – 0.991 and v'(99) = 0.099

Syllabus Topic : Discretization by Binning

1.6.3(D) Discretization by Binning

- This is the data smoothing technique.
- Discretization by binning has two approaches :
  - (a) Equal-width (distance) partitioning
  - (b) Equal-depth (frequency) partitioning or Equal-height binning
- Both this binning approaches are given in section 1.6.1(C).

Syllabus Topic : Discretization by Histogram Analysis

1.6.3(E) Discretization by Histogram Analysis

In Discretization by Histogram divide the data into buckets and store average (sum) for each bucket in smaller data representation.

☞ **Different types of histogram**

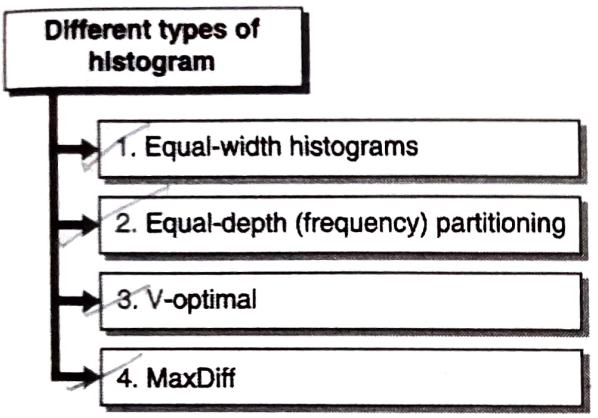


Fig. 1.6.8 : Different types of histogram

➜ **1. Equal-width histograms**

It divides the range into N intervals of equal size.

➜ **2. Equal-depth (frequency) partitioning**

It divides the range into *N* intervals, each containing approximately same number of samples.

➜ **3. V-optimal**

Different Histogram types for a given number of buckets are considered and the one with least variance is chosen.

➜ **4. MaxDiff**

After the sorting process applied to the data, borders of the buckets are defined where the adjacent values have maximum difference.

Example

1,1,5,5,5,5,5,8,8,10,10,10,10,12,14,14,14,15,15,15,

15,15,15,18,18,18,18,18,18,18,18,20,20,20,20,20,

20,20,21,21,21,21,25,25,25,25,25,28,28,30,30,30

Histogram of above data sample is,

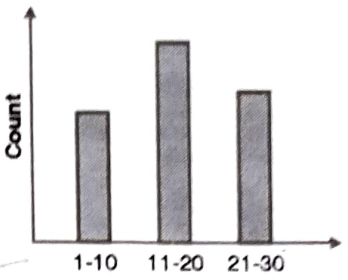


Fig. 1.6.9 : Example of histogram

Syllabus Topic : Data Reduction

1.6.4 Data Reduction

1.6.4(A) Need for Data Reduction

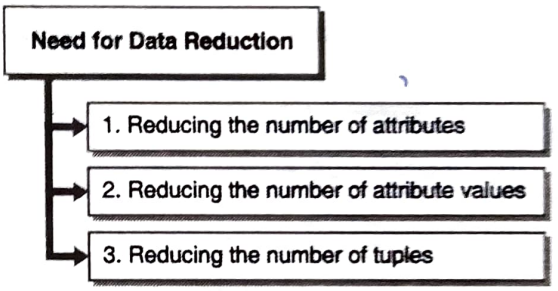


Fig. 1.6.10 : Need for Data Reduction

➜ **1. Reducing the number of attributes**

- **Data cube aggregation :** This process involves applying OLAP operations like roll-up, slice or dice operations.
- **Removing irrelevant attributes :** In this attribute selection methods like filtering and wrapper methods may be used, it also involves searching the attribute space.
- **Principle component analysis (numeric attributes only) :** This involves representing the data in a compact form by using a lower dimensional space.

➜ **2. Reducing the number of attribute values**

- **Binning (histograms) :** This involves representing the attributes into groups called as **bins,** this will result into lesser number of attributes.
- **Clustering :** Grouping the data based on their similarity into groups called as **clusters.**
- Aggregation or generalization.

➔ **3. Reducing the number of tuples**

To reduce the number of tuples, sampling may be used.

1.6.4(B) Data Reduction Technique

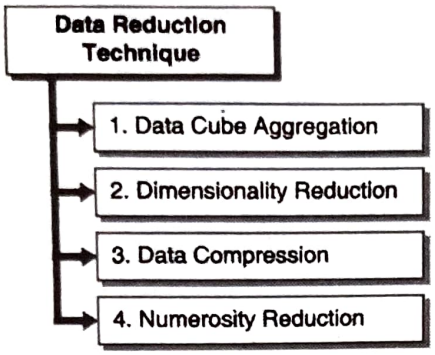


Fig. 1.6.11 : Data Reduction Technique

Syllabus Topic : Data Cube Aggregation, Attribute Subset Selection

1.6.4(B).1 Data Cube Aggregation

- It reduces the data to the concept level needed in the analysis and uses the smallest (most detailed) level necessary to solve the problem.
- Queries regarding aggregated information should be answered using data cube when possible.

☞ **Example**

Total annual sales of TV in USA is aggregated quarterly as shown in Fig. 1.6.12.

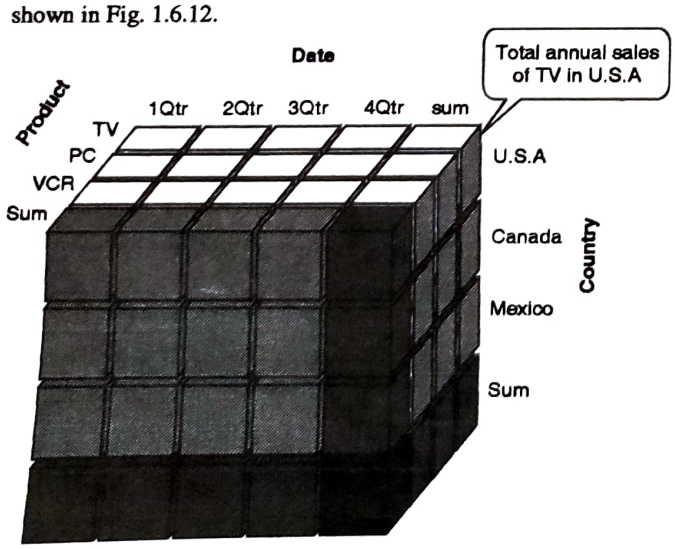


Fig. 1.6.12 : Example of data cube

1.6.4(B).2 Dimensionality Reduction

➔ **(SPPU - May 16, Dec. 17)**

Q. Enlist the dimensionality reduction techniques for text. Explain any one of them in brief.

May 16, Dec. 17, 6 Marks

- In the mining task during analysis, the data sets of information may contain large number of attributes that may be irrelevant or redundant.
- Dimensionality reduction is a process in which attributes are removed and the resulting dataset is smaller in size.
- This process helps in reducing the time and space complexity required by a data mining technique.
- Data visualization becomes an easy task.
- It also involves deleting inappropriate features or reducing the noisy data.

☞ **Attribute subset selection**

How to find a good subset of the original attributes ?

Attribute subset selection refers to a process in which minimum set of attributes are selected in such a way that their distribution represents the same as the original data set distribution considering all the attributes.

☞ **Different attribute subset selection techniques**

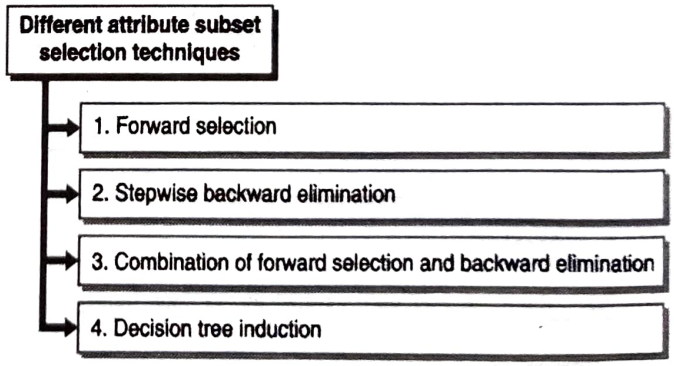


Fig. 1.6.13 : Different attribute subset selection techniques

➔ **1. Forward selection**

- Start with empty set of attributes.
- Determine the best of the original attributes and add it to the set.

- At each step, find the best of the remaining original attributes and add it to the set.

➜ **2. Stepwise backward elimination**

- Starts with the full set of attributes.
- At each step, it removes the worst attribute remaining in the set.

➜ **3. Combination of forward selection and backward elimination**

- The procedure combines and selects the best attribute and removes the worst among the remaining attributes.
- For all above method stopping criteria is different and it requires a threshold on the measure used to stop the attribute selection process.

➜ **4. Decision tree induction**

- ID3, C4.5 intended for classification.
- Construct a flow chart like structure.
- A decision tree is a tree in which :
  - Each internal node tests an attribute.
  - Each branch corresponds to attribute value.
  - Each leaf node assigns a classification.

1.6.4(B).3 Data Compression

- Data compression is the process of reducing the number of bits needed to either store or transmit the data. This data can be text, graphics, video, audio, etc. This can be usually be done with the help of encoding techniques.
- Data compression techniques can be classified into either lossy or lossless techniques. In lossy technique there is a loss of information whereas in lossless there is no loss.

☞ **Lossless compression**

- Lossless compression consists of those techniques guaranteed to generate an exact duplication of the input dataset after a compress/decompress cycle.
- Lossless compression is essentially a coding technique. There are many different kinds of coding algorithms, such as Huffman coding, run-length coding and arithmetic coding.

☞ **Lossy compression**

- In lossy compression techniques at the cost of data quality one can achieve higher compression ratio.
- These types of techniques are useful in applications where data loss is affordable. They are mostly applied to digitized representations of analog phenomenon.
- Two methods of lossy data compression :

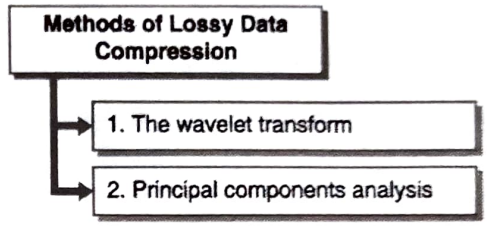


Fig. 1.6.14 : Methods of Lossy Data Compression

➜ **1. The wavelet transform**

A clustering approach which applies wavelet transform to the feature space :

- The orthogonal wavelet transform when applied over a signal results in time scale decomposition through its multi resolution aspect.
- It clusters the functional data into homogenous groups.
- Both grid-based and density-based.

☞ **Input parameters**

- Number of grid cells for each dimension.
- The wavelet and the number of applications of wavelet transform.
- Clustering approach using Wavelet transform.
- Impose a multidimensional grid like structure on to the data for summarisation.
- Use an n-dimensional feature space for representing spatial data objects.
- Dense regions may be identified by applying the wavelet transform over the feature space.
- Applying wavelet transform multiple times results in clusters of different scales.
- Clusters are identified by using hat-shape filters and also suppress weaker information in their boundary.

Major features of data compression

- It also results in Effective removal of outliers.
- The technique is Cost efficient.
- Complexity O(N).
- At different scales arbitrary shaped clusters are detected.
- The method is not sensitive to noise or input order.
- It is applicable only to low dimensional data.

➜ **2. Principal components analysis**

- Principal Component Analysis (PCA) creates a representation of the data with orthogonal basis vectors, i.e. eigenvectors of the covariance matrix of the data. This can also be derived using Singular value decomposition(SVD) method. By this projection original dataset is reduced with little loss of information.
- PCA is often presented using the eigen value/eigenvector approach of the covariance matrices. But in efficient computation related to PCA, it is the Singular Value Decomposition (SVD) of the data matrix that is used.
- A few scores of the PCA and the corresponding loading vectors can be used to estimate the contents of a large data matrix.
- The idea behind this is that by reducing the number of eigenvectors used to reconstruct the original data matrix, the amount of required storage space is reduced.

Syllabus Topic : Sampling

1.6.4(B).4 Numerosity Reduction

- Numerosity reduction technique refers to reducing the volume of data by choosing smaller forms for data representation
- Different techniques used for numerosity reduction are :

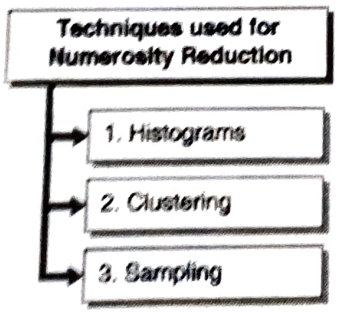


Fig. 1.6.15 : Techniques used for Numerosity Reduction

➜ **1. Histograms**

- It replaces data with an alternative, smaller data representation.
- Approximate data distributions.
- Divide data into buckets and store average (sum) for each bucket.
- A bucket represents an attribute-value/frequency pair.
- Can be constructed optimally in one dimension using dynamic programming.
- Related to quantization problems.

☞ **Different types of histogram [Refer 1.6.3(E)]**

➜ **2. Clustering**

- Clustering is a data mining technique used to group the elements based on their similarity without prior knowledge of their class labels.
- It is a technique that belongs to undirected data mining tools.
- The goal of undirected data mining is to explore structure in the data. No target variable is to be predicted, therefore there is no difference been made between independent and dependent variables.
- Categorization of clusters based on clustering techniques is given below :
  - Any example belonging to a single cluster would be termed as exclusive cluster.
  - Any example may belong to many clusters in such a case it is said to be overlapping.
  - Any example belongs to a cluster with certain probability then it is said to be probabilistic.
  - A Hierarchical representation may be used for clusters in which clusters may be at highest level of hierarchy and subsequently refined at lower levels to form sub clusters.

➜ **3. Sampling**

- Sampling is used in preliminary investigation as well as final analysis of data.
- Sampling is important in data mining as processing the entire data set is expensive and time consuming.

☞ **Types of sampling**

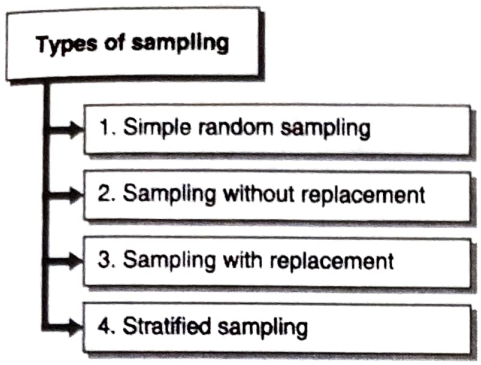


Fig. 1.6.16 : Types of sampling

➜ **1. Simple random sampling**

There is an equal probability of selecting any particular item.

➜ **2. Sampling without replacement**

As each item is selected, it is removed from the population.

➜ **3. Sampling with replacement**

The objects selected for the sample is not removed from the population. In this technique the same object may be selected multiple times.

➜ **4. Stratified sampling**

The data is split into partitions and samples are drawn from each partition randomly.

1.7 Solved University Questions and Answers

Q. 1 Discuss whether or not each of the following activities is a data mining task. **(May 16, 6 Marks)**

(i) Computing the total sales of a company.

(ii) Predicting the future stock price of a company using historical records.

(iii) Predicting the outcomes of tossing a pair of dice.

Ans. :

(i) Computing the total sales of a company

This activity is not a data mining task because the total sales can be computed by using simple calculations.

(ii) Predicting the future stock price of a company using historical records

This activity is a data mining task. Historical records of stock price can be used to create a predictive model called regression, one of the predictive modeling tasks that is used for continuous variables

(iii) Predicting the outcomes of tossing a pair of dice :

This activity is not a data mining task because predicting the outcome of tossing a fair pair of dice is a probability calculation, which doesn't have to deal with large amount of data or use complicate calculations or techniques.

Q. 2 Differentiate between Descriptive and Predictive data mining tasks **(Oct. 16, 2 Marks)**

Ans. :

(a) **Descriptive mining :** To derive patterns like correlation, trends etc. which summarizes the underlying relationship between data.

Example : Identifying items which are purchased together frequently.

Some of Descriptive mining techniques :

- Class/Concept description
- Mining of frequent patterns
- Mining of associations
- Mining of correlations
- Mining of clusters

(b) **Predictive mining :** Predict the value of a specific attribute based on the value of other attributes.

Example : Predict the next year's profit or loss.

Some of Predictive Mining techniques :

- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

❑❑❑