UNIT-II

Introduction to Data Mining

In this chapter, you will learn how data mining is part of the natural evolution of database technology.

- Why data mining is important?
- How it is defined?
- learn general architecture of data mining systems.
- Types of patterns that can be found
- How to tell which patterns represent useful knowledge?

Data Collection and Database Creation
(1960s and earlier)
- primitive file processing

↓

Database Management System
(1970s - early 1980s)
- Hierarchical and network database systems
- Relation database systems.
- Data modeling tools - entity - relational model etc.
- Query languages :- SQL
- Users interfaces : forms and reports
- Query processing & query optimization
- Transaction, concurrency control & recovery
- on-line transaction processing (OLTP)

↓

Advanced Database Systems (mid 1980s - present)

Advanced Data Analysis: Data Warehousing & Data Mining (late 1980's - present)

- Data Warehouse & OLAP
- Data Mining and knowledge discovery:
  - generalization,
  - classification,
  - association
  - clustering
  - frequent pattern,
  - outlier analysis
- Advanced data mining applications:
  - Stream data mining
  - bio-data mining
  - time-series analysis
  - text mining
  - web mining
  - Intrusion detection
- Data mining & society
  - privacy-preserving data mining.

Web-based databases (1990s - present)

↓

New generation of Integrated Data and Information Systems (Present - future)

What is Data Mining? 7 Marks Essential
4 Marks Diagram 3 Marks Explanation

Mining refers to extracting or "mining" knowledge from large amounts of data.

The term is actually a misnomer.

Many other terms carry a similar or slightly different meaning to data mining such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.

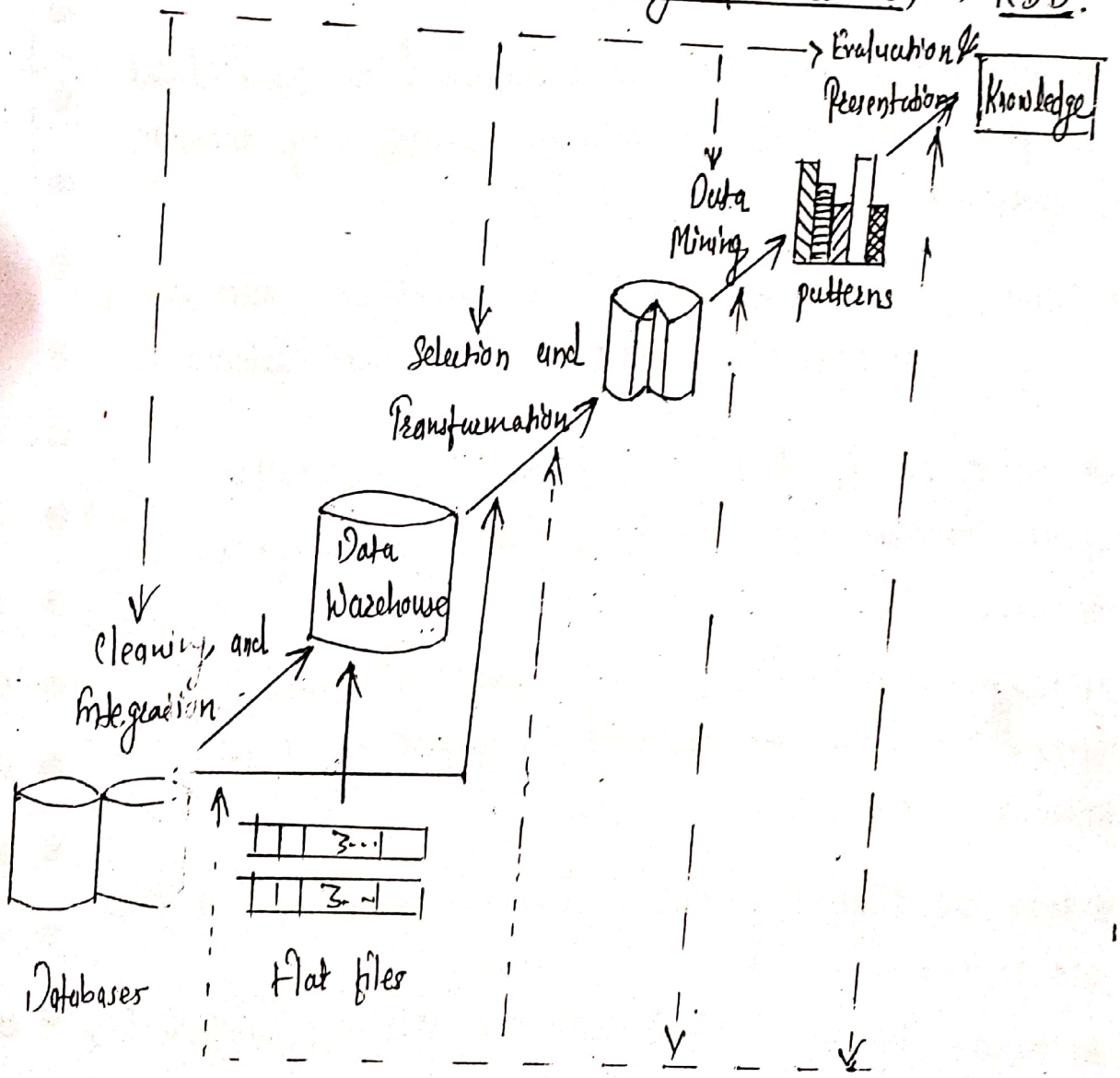


fig: Data mining as a step in the process of knowledge discovery.

- Alternatively data mining is an essential step in the process of knowledge discovery.
- Knowledge discovery as a process consists of an iterative sequence of steps:-

1) Data cleaning (to remove noise and inconsistent data)
2) Data Integration (where multiple data sources may be combined) (preprocessing step)
3) Data Selection (where data relevant to the analysis task are retrieved from the database)
4) Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5) Data Mining (an essential process where intelligent methods are applied in order to extract data patterns)
6) pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures
7) knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

- Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining.
- The data mining step may interact with the user or a knowledge base.
- The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

on IT's view, the architecture of a typical data mining may have the following major components:

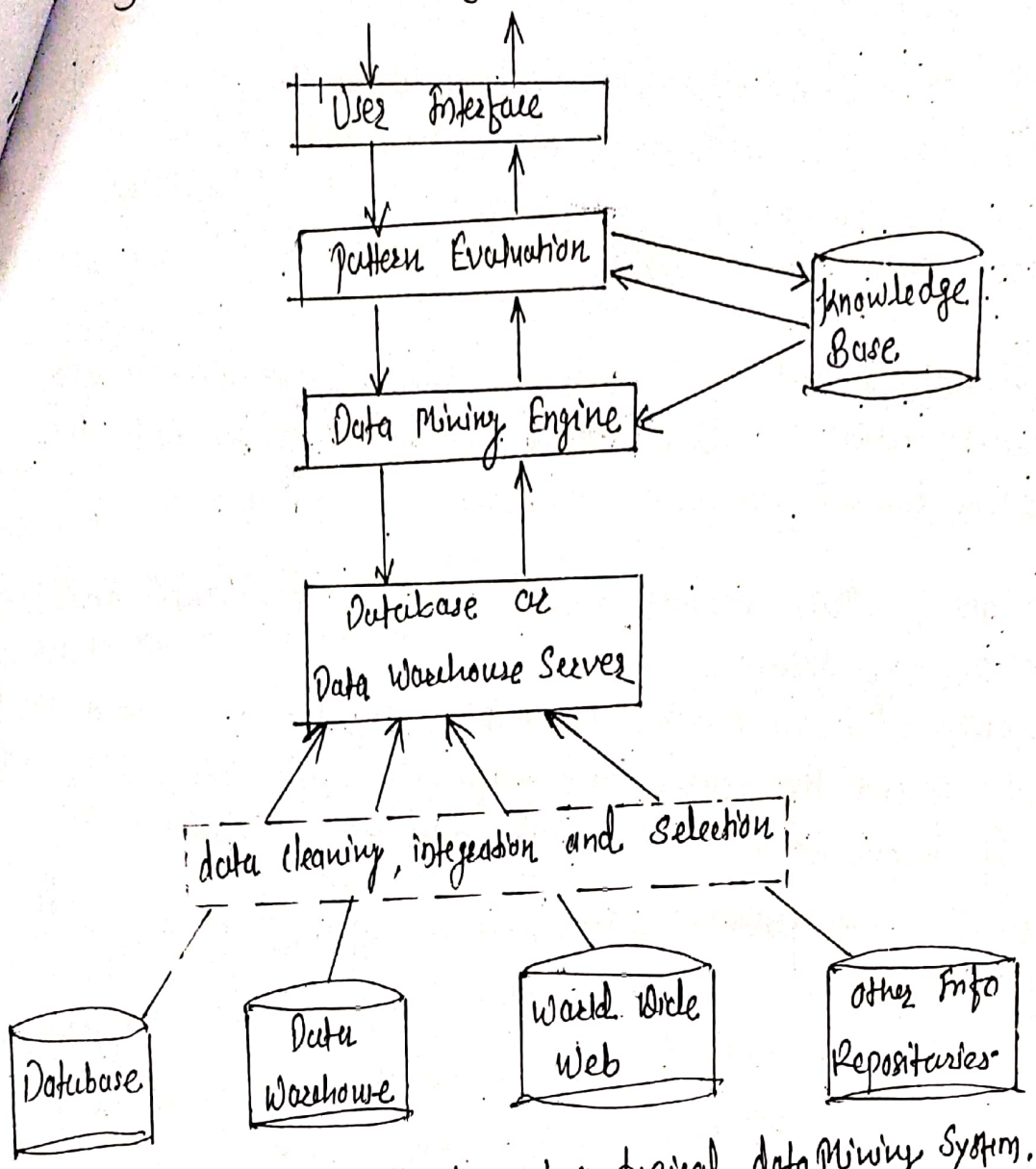


Fig: Architecture of a typical data mining system.

- Database, data warehouse, World wide web, or other info repository:
  - This is one or a set of databases, data warehouses, spreadsheets or kinds of information repositories.
  - Data cleaning and data integration techniques may be performed on the data.
- Database or data warehouse server:- responsible for fetching the relevant data, based on the user's data mining request.

Knowledge base :- used to guide the search or evaluate the interestingness of resulting patterns.

Data mining engine :- ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis and evaluation analysis.

Pattern evaluation module :- employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.

User interface :- This module communicates between users and the data mining system,

- by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
- evaluate mined patterns, and visualize the patterns in different forms.

Mining functionalities :- 7 Marks 4

What kinds of patterns can be mined? 5 points with an example.

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- Data mining tasks can be classified into two categories :
  - descriptive
  - predictive.
- Descriptive mining tasks characterize the general properties of the data in the database.
- Predictive mining tasks perform inference on the current data in order to make predictions.

⇒ Concept / class Description : characterization and Discrimination ⇒

- Data can be associated with classes or concepts.
- for example, in the AllElectronics store, classes of items for sale include computer and printers, and concepts of customers include bigspenders and budgetSpenders.
- It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms.
- Such descriptions of class or a concept are called class/concept descriptions.

1) data characterization, by summarizing the data of the class under study (often called the target class)

2) data discrimination, by comparison of the target class with one or a set of comparative classes (often ~~the~~ called the contrasting classes).

Example: Data Characterization,

- A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1000 a year at AllElectronics.
  - The result could be a general profile of the customers, such as they are 40-50 years old, employed, and have excellent credit ratings.
  - The system should allow users to drill down on any dimension, such as occupation in order to view these customers according to their type of employment.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

- The target and contrasting classes can be specified by the user, and corresponding data objects retrieved through database queries.

Example :- Data discrimination :-

- A data mining system should be able to compare two groups of AllElectronics customers, such as those who shop for computer products regularly (more than two times a month) versus those who rarely shop for such products (less than three times a year).
- The resulting description provides a general comparative profile of the customers; such as 80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education.
- Whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.
- Drilling down on a dimension, such as occupation, or adding new dimensions, such as income_level, may help in finding even more discriminative features between the two classes.

Frequent Patterns, Associations and Correlations ⇒

- Frequent patterns, are patterns that occur frequently in data.
- There are many kinds of frequent patterns, including itemsets, subsequences, and substructures.
- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

- Association analysis ⇒ suppose, as a marketing manager of AllElectronics, you would like to determine which items are frequently purchased together within the same transactions.
- An example

  buys(X, "computer") ⇒ buys(X, "software") [support = 1%, confidence = 50%]

- Where X is a variable representing a customer.
- A confidence, or certainty of 50% means that if a customer buys a computer, there is a 50% chance that she will by software as well.
- A 1% support means that 1% of all the transactions under analysis showed that computer and software were purchased together.
- above rule written simply as,

  "computer ⇒ software [1%, 50%]"

- This association rule involves single predicate are refered as single-dimensional association rules.

- Ex. age(X, "20...29") ∧ income(X, "20k...29k") ⇒ buys(X, "CD player")
  [support = 2%, confidence = 60%]

- 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player at AllElectronics.
- There is a 60% probability that a customer in this age and income group will purchase a CD player.

3) Classification and Prediction ⇒

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- The derived model is based on the analysis of a set of training data (i.e. data objects whose class label is known).
- How is the derived model presented?
  - represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks.
- Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous-valued functions. That is it is used to predict missing or unavailable numerical data values rather than class labels.

age (X, "youth") AND income (X, "high") → class (X, "A")
age (X, "youth") AND income (X, "low") → class (X, "B")
age (X, "middle_aged") → class (X, "C")
age (X, "senior") → class (X, "C")

Fig (a) IF-THEN rules

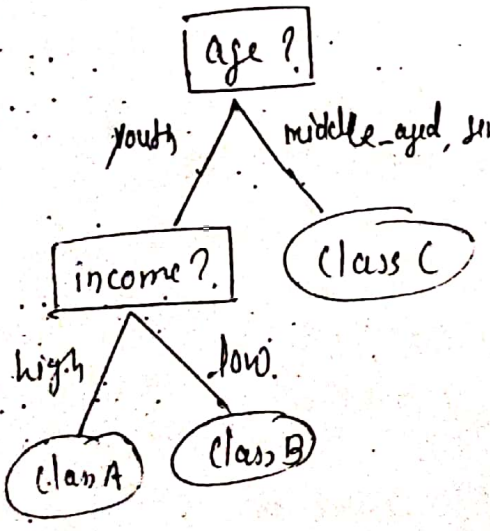


Fig (b) decision tree

Classification of Data Mining Systems ⇒

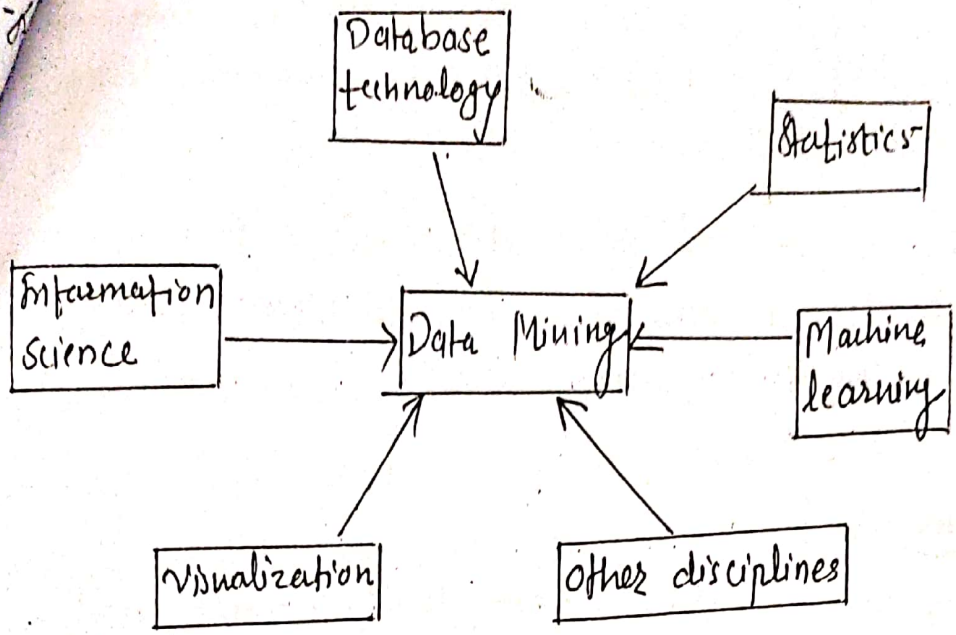


fig. Data Mining as a confluence of Multiple disciplines

- Data Mining is an interdisciplinary field
- includes database systems, statistics, machine learning, visualization and information science.
- Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, web technology, economics, business, bioinformatics, or psychology.
- because of the diversity of disciplines contributing to data mining, data mining research is expected to generate a large variety of data mining systems.
- Therefore, it is necessary to classify data mining systems according to various criteria!

4) Cluster Analysis ⇒

- Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.
- clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.

5) Outlier Analysis ⇒

- A database may contain data objects that do not comply with the general behavior or model of the data.
- These data objects are outliers.

Classification according to the kinds of databases mined :-
- A data mining system can be classified according to different criteria such as data models, or the types of data or applications involved, each of which may require its own data mining technique.
- For instance, if classifying according to data models, we may have relational, transactional, object-relational, or data warehouse system.
- We may have spatial, time-series, text, stream data, multimedia data mining system, WWW mining system.

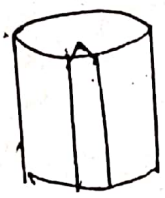
2) Classification according to kinds of knowledge mined ⇒
- that is, based on data mining functionalities, such as characterization, discrimination, association, and correlation analysis, classification, prediction, clustering, outlier analysis,

3) Classification according to the kinds of techniques utilized ⇒
- These techniques can be described according to the degree of user interaction involved (e.g. autonomous systems, interactive exploratory systems, query driven systems) or the methods of data analysis employed (e.g. database-oriented or data warehouse oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks and so on).

4) Classification according to the applications adapted ⇒
- For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, E-mail, and, so on
- Different applications often require the integration of application-specific methods.

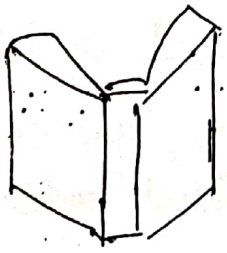
Data Mining Task Primitives ⇒

- Each user will have a data mining task in mind. That is some form of data analysis that he or she would like to have performed.
- Data mining query is used for specifying data mining tasks.
- A data mining query is defined in terms of data mining task primitives.
- The primitives allow user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths.



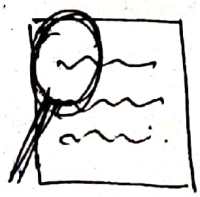
Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
conditions for data selection
Relevant attributes or dimensions
Data grouping criteria

knowledge type to be mined
Characterization, Discrimination
Association, correlation
classification, prediction,
clustering



Background knowledge
concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g. confidence)
Utility (e.g. support)



visualization of discovered patterns
Rules, tables, reports, charts, graphs decision trees, and cubes
Drill down and roll-up.

Fig. 4 Primitive for specifying a data mining task.

Concept hierarchy ⇒

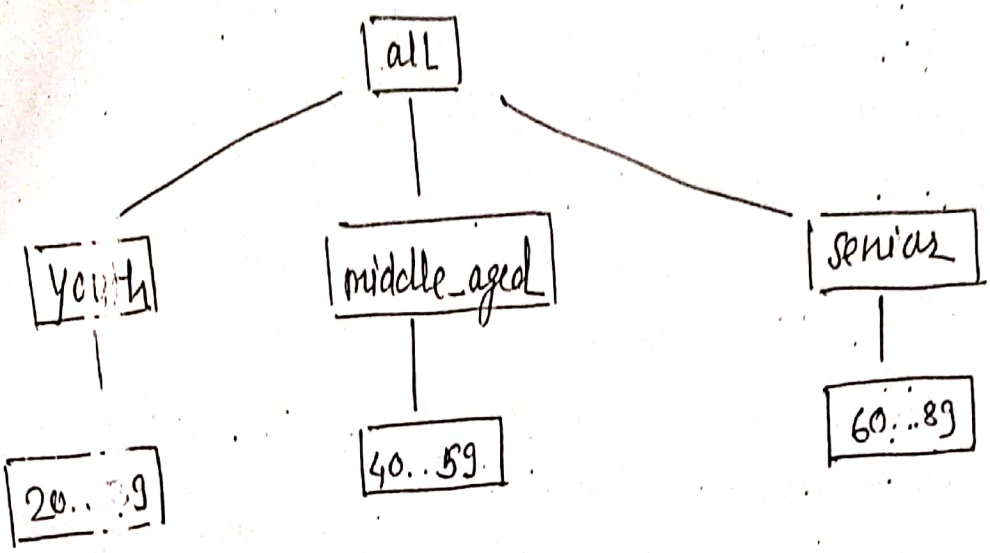


Fig. A concept hierarchy for the attribute (as dimension) age.

Integration of a Data Mining System with a DB or Data Warehouse System ⇒

- A critical question in the design of a data mining (DM) system is how to integrate or couple the DM system with a database (DB) system and/or a data warehouse (DW) system
- If a DM system works as a stand-alone system or is embedded in an application program, there are no DB or DW systems with which it has to communicate. This simple scheme is called no coupling.
- However, when DM system communicate with other information system components, such as DB and DW systems, possible schemes include no coupling, loose coupling, semitight coupling, and tight coupling.

1) No coupling ⇒ means that DM system will not utilize any function of a DB or DW system.
   - It may fetch data from a particular source (such as file system)
   - process data using some data mining algorithms, and then store the mining results in another file.
   - No coupling represents a poor design.

2) Loose coupling ⇒ means that a DM system will use some facilities of a DB or DW system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse.
   - Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing & other system facilities.

2) Performance issues ⇒
- These includes efficiency, scalability, and parallelization of data mining algorithms.

Efficiency and Scalability of data mining algorithms ⇒
- For effective data mining, data mining algorithms must be efficient and scalable.
- The running time of data mining algorithm must be predictable and acceptable in large databases.

i) Parallel, distributed, and incremental mining algorithm ⇒
- The huge size of many databases, the wide distribution of data, computational complexity of data mining methods are factors for development of parallel and distributed data mining algorithms.

3) Issues relating to the diversity of database types ⇒

i) Handling of relational and complex types of data ⇒
- However, other databases may contain complex data objects; hypertext, and multimedia data, spatial data, temporal data, or transaction data.
- It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining.
- Specific data mining systems should be constructed for mining specific kinds of data.

ii) Mining information from heterogeneous databases and global information systems ⇒
- Internet connects many sources of data.
- discovery of knowledge from different sources poses great challenges to data mining.

iii) Incorporation of background knowledge ⇒
- Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.

iv) Data mining query language and adhoc data mining ⇒
- Relational query languages (such as SQL) allow users to pose adhoc queries for data retrieval.

v) Presentation and visualization of data mining results ⇒
- Discovered knowledge should be expressed in high-level languages
- visual representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.
- such as trees, tables, rules, graphs, charts, crosstabs, matrices, or curves.

vi) Handling noisy or incomplete data ⇒
- The data stored in a database may reflect noise, exceptional cases, or incomplete data objects.

vii) Pattern evaluation ⇒ The interestingness problem ⇒
- A data mining system can uncover thousands of patterns.
- Many patterns discovered may be uninteresting to the given user.
- The use of interestingness measures or user-specified constraints to guide the discovery process and reduce the search space is another active area of research.

Pre-processing
real world data
inconsistent data
base) and their
Low-quality
How co
que

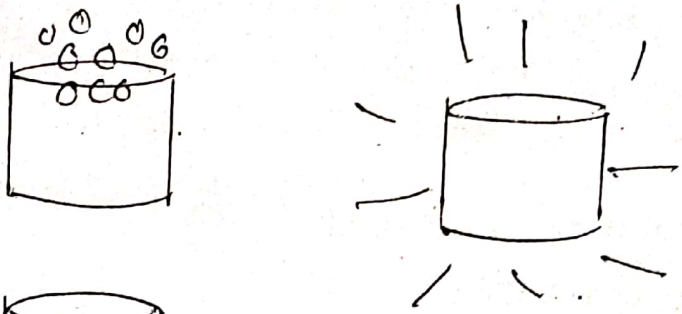
Preprocessing

- Real world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size (gigabytes or more) and their likely origin from multiple, heterogenous sources.
- Low-quality data will lead to low-quality mining results.
- How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results?
- How can the data be preprocessed so as to improve the efficiency and ease of the mining process?
- There are number of data preprocessing techniques
- i) Data cleaning ⇒ can be applied to remove noise and correct inconsistencies in the data.
- ii) Data integration ⇒ merges data from multiple sources into a coherent data store, such as data warehouse.
- iii) Data transformation ⇒ such as normalization, may be applied. for example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurement
- iv) Data reduction ⇒ can reduce the data size by aggregating, eliminating redundant features, or clustering for instance.
- These techniques are not mutually exclusive, they may work together.

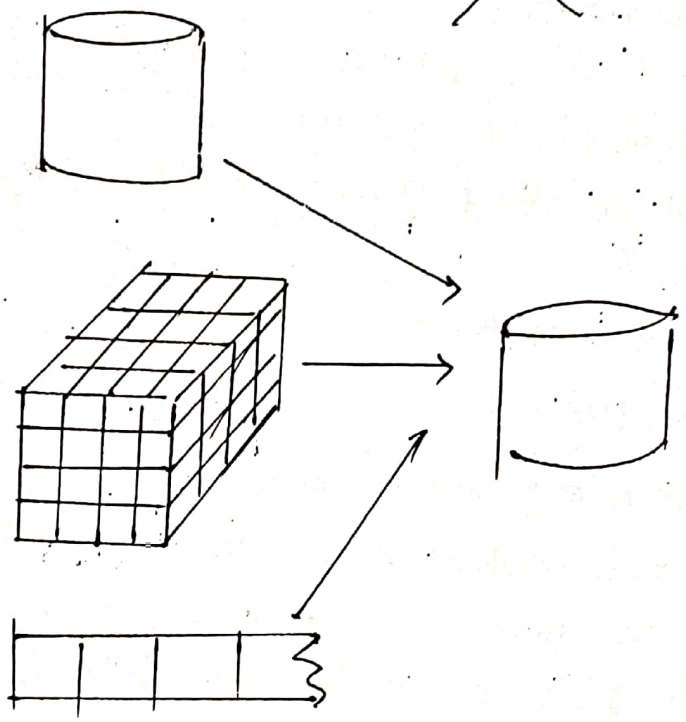
Why preprocess the Data?

- Imagine that you are a manager at ALLElectronics and have been charged with analyzing the company's data with respect to the sales at your branch.
- You carefully inspect the company's database and data warehouse, identifying & selecting the attributes or dimensions to be included in your analysis, such as item, price and units_sold.
- You notice that several of the attributes for various tuples have no recorded value.
- Furthermore, users of your database system have reported errors, unusual values, and inconsistencies in the data recorded for some transactions.
- In other words, the data you wish to analyze by data mining techniques are incomplete (lacking attribute values), noisy (containing errors, or outlier values that deviate from the expected), and inconsistent.
- Incomplete, noisy, and inconsistent data are commonplace properties of large real world databases and data warehouses.
- So to remove these consequences we use to preprocess the data with the help of data preprocessing techniques.
- Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers and resolving inconsistencies.
- Suppose that you would like to include data from multiple sources in your analysis. This would involve integrating multiple databases, data cubes, or files, that is data integration used.

Data cleaning



Data Integration



Data transformation

$-2, 32, 100, 59, 48 \rightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data reduction

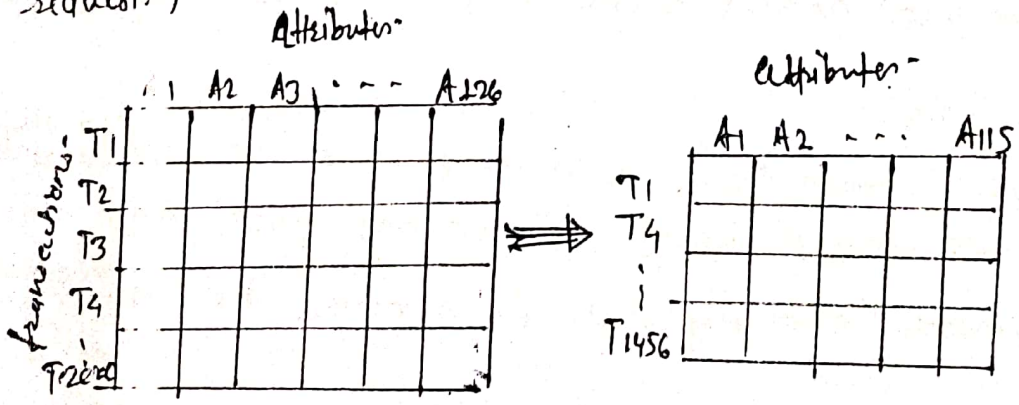


Fig. Forms of data preprocessing

Data Cleaning ⇒ Sessional 3 Marks
Explanation with an Example.

- Data cleaning (data cleansing) routines attempt to fill in miss values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

1) Missing Values ⇒ Imagine that you need to analyze AllElectronics Sales and customer data.
- you note that many tuples have no recorded val.. for several attributes, such as customer income.
- How can you go about filling in the missing values for this attribute?
- some methods we have,

i) Ignore the tuple
ii) filling the missing value manually
iii) Use a global constant to fill in the missing value ⇒
- Replace all missing attribute values by the same constant, such as label like "Unknown" .2 ...∞.
iv) Use the attribute mean to fill the missing value ⇒
v) Use the attribute mean for all samples belonging to the same class as the given tuple.
vi) Use the most probable value to fill in the missing value.

2) Noisy Data ⇒
- Noise is a random error or variance in a measured variable
- Given numerical attribute such as say, price how can we "smooth" out the data to remove the noise?

data for price (in dollars): 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

} partitioned into equal frequencies bins of size 3.

- Smoothing by bin means:-

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

} - each value in a bin is replaced by the mean value of the bin.

- Smoothing by bin boundaries:-

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

} - minimum-maximum values are identified as bin boundaries
- each bin value is then replaced by the closest boundary value.

⇒ Binning:- Binning method smooth a sorted data value by consulting its "neighborhood" that is the values around it.
- The sorted values are distributed into number of buckets or bin

⇒ Clustering:- outliers may be detected by clustering.

⇒ Regression:- Data can be smoothed by fitting the data to a function, such as with regression.

- Data Integration and Transformation ⇒

- combines data from multiple sources into a coherent store, as in data warehousing.
- These sources may include multiple databases data cubes or flat files.

Data Preprocessing

Transformation ⇒ In this the data are transformed or consolidated into forms appropriate for mining.

Strategies involve

- Smoothing: which works to remove noise from the data. Such techniques include binning, regression and clustering.
- Aggregation: where summary or aggregation operations are applied to the data.
  Eg. daily sales data may be aggregated so as to compute monthly and annual total amounts.
- Generalization ⇒ of the data, where low level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
  Eg. categorical attributes, like street, can be generalized to higher-level concepts, like city or country.
- 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.
- Attribute construction ⇒ (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

Ex. Decimal scaling ⇒ Suppose that the recorded values of A range from -986 to 917.
- The maximum absolute value of A is 986.
- To normalize by decimal scaling, we therefore divide each value by 1000 (i.e j = 3) so that,
  -98 : normalize to -0.986
  917 normalizes to 0.917.

Data Reduction

- If the data set will likely to be huge, complex and mining on huge amounts of data can take long making such analysis impractical or infeasible.
- Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume
- Strategies for data reduction include,

1) Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.
2) Attribute subset selection, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed.
3) Dimensionality reduction, where encoding mechanisms are used to reduce the data set size.
4) Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, use of histograms.
5) Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels.

Cube aggregation ⇒

- Data cube of the AllElectronics sales per quarter, for the years 2002 to 2004.
- However interested in the annual sales (total per year), rather than the total per quarter.
- Thus the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter.

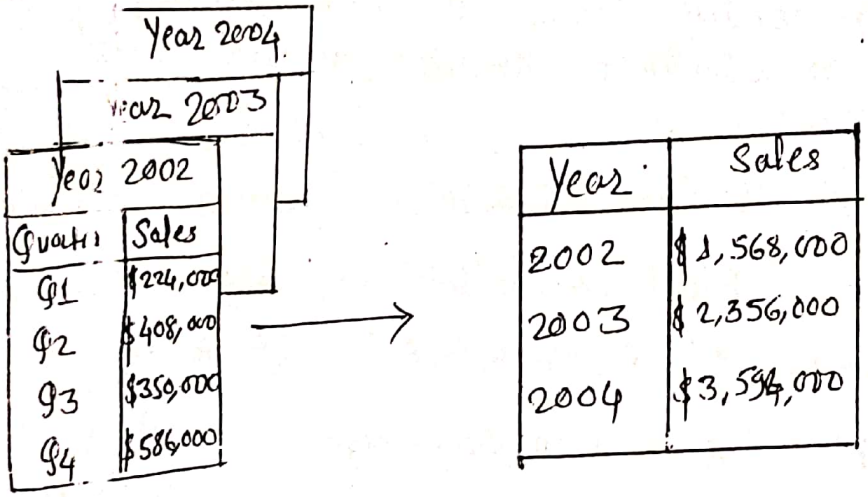


Fig. Sales data for a given branch AllElectronics for the years 2002 to 2004.

- In the left, the sales are shown per quarter.
- In the right, the data are aggregated to provide the annual sales.

Attribute Subset Selection ⇒

- Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or redundant.
- For example, if the task is to classify customers as to whether or not they are likely to purchase a popular new CD at AllElectronics when notified of a sale, attributes such as the customer's telephone number are likely to be irrelevant, unlike attributes such as age or music_taste.
- Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions).

Forward selection

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

Initial reduced set:
{ }

⇒ {A1}

⇒ {A1, A4}

⇒ Reduced attribute set:
{A1, A4, A6}

⇒ Stepwise forward selection:
- The procedure starts with an empty set of attributes as the reduced set.

Backward elimination

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

⇒ {A1, A2, A3, A4, A5, A6}

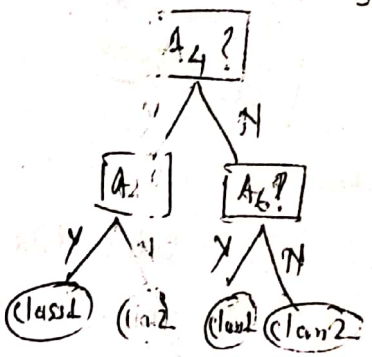
⇒ {A1, A4, A5, A6}

⇒ Reduced attribute set:
{A1, A4, A6}

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

Decision tree induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}



⇒ Reduced attribute set:
{A1, A4, A6}

⇒ Decision tree induction ⇒
- Constructs a flowchart like structure where each nonleaf node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each leaf node denotes a class prediction.

Discretization and concept Hierarchy Generation ⇒

- Discretization techniques can be used to reduce the number of values for given continuous attribute by dividing the range of the attribute into intervals.
- Interval labels can then be used to replace actual data values.
- If the discretization process uses class information then we say it is supervised discretization. Otherwise, it is unsupervised.
- Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy.
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts (such as numerical values for the attribute age) with higher-level concepts (such as youth, middle-aged, or senior).

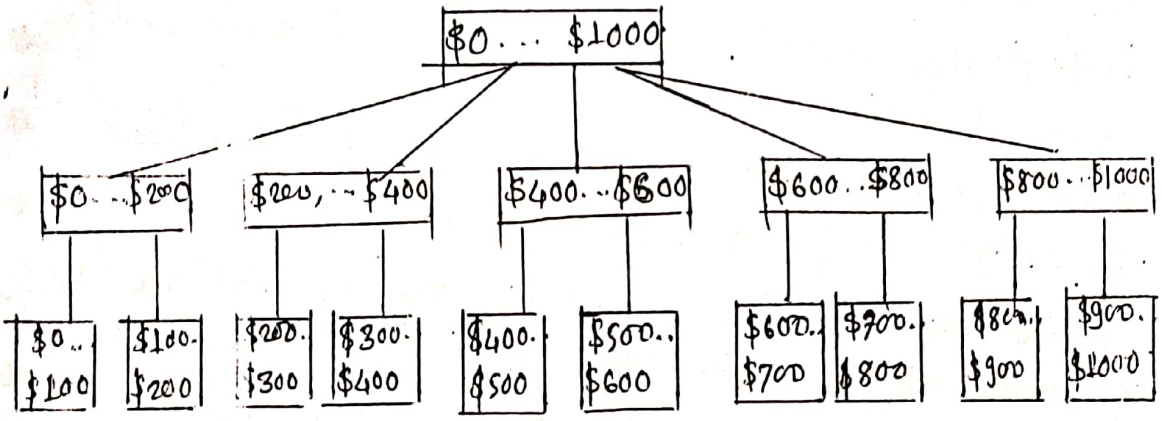


Fig. A concept hierarchy for the attribute price, where an interval ($x ... $y) denotes the range from $x (inclusive) to $y (inclusive).

Data Mining Application Areas ⇒ Sessional (6 Mark 6 points)

- In this section we examine a few application domains.

1) Data mining for financial Data Analysis ⇒

- most banks and financial institutions offer a wide variety of banking services (such as checking and savings accounts for business as individual customers).
- Credit card
- investment services (such as mutual funds)
- some also offer insurance services and stock investment services.
- financial data collected in the financial industry are often relatively complete, reliable, and high quality which facilitates systematic data analysis and data mining.
- few typical cases:

i) Design and construction of data warehouses for multidimensional data analysis and data mining.

- data warehouses need to be constructed for banking and financial data.
- multidimensional data analysis methods should be used to analyze the general properties of such data.

For example, one may like to view the debt and revenue changes by month, by region, by sectors along with maximum, minimum, total, average, trend, and other statistical information.

- Data warehouses, data cubes, multifeature and discovery-driven data cubes, characterization and class comparison and outlier analysis all play important roles in financial data analysis and mining.

payment prediction and customer credit policy analysis ⇒

- example factors related to risk of loan payments include loan-to-value ratio, term of the loan, debt ratio (total amount of monthly debt versus the total monthly income),
- payment-to-income ratio
- customer income level,
- education level
- residence region
- and credit history.
- In these factors data mining methods, such as attribute selection and attribute relevance ranking may help identify important factors and eliminate irrelevant ones.

iii) Classification and clustering of customers for targeted marketing ⇒

- classification and clustering methods used for customer group identification and targeted marketing.

iv) Detection of money laundering and other financial crimes

-

ii) ⇒ Data mining for the Retail Industry ⇒

- major application area for data mining.
- it collect huge amounts of data on sales, customer shopping history, good transportation, consumption and service.
- Today, many stores also have websites where customers can make purchases on-line.
- Retail data mining can help identify customer buying behaviours discover customer shopping patterns and trends, improve quality of customer service, enhance goods consumption ratios.

3) Data mining for the Telecommunication Industry

- In this area, data mining help to understand the business involved, identify telecommunication patterns catch fraudulent activities, make better use of resources and improve quality of service.
- few scenarios :-

i) Multidimensional analysis of telecommunication data ⇒
- multidimensional data available in Tele-comm. data with dimensions such as calling-time, duration, location of caller, location of callee and type of call.
- Analyst in the industry may wish to regularly view charts and graphs regarding calling source, destination, volume and time of day usage patterns.

ii) Fraudulent Pattern analysis and the identification of unusual patterns ⇒
- identify potentially fraudulent users and their typical usage patterns.

4) Data Mining for Biological Data Analysis ⇒
- identification of DNA or amino acid sequence patterns that play roles in various biological functions, genetic diseases and evolution is challenging.
- This requires a great research in computational algorithms, statistics, mathematical programming, data mining, machine learning, information retrieval, and other disciplines to develop effective genomic and proteomic data analysis tools.

Questions (UNIT-II)

1) What are major components of a typical data mining? Draw & explain architecture of data mining system. [7]

Ans ⇒ refer page No. 3

2) What is the need of Data Preprocessing? Explain Data cleaning and Data Transformation in short. [7]

Ans ⇒ refer page No. 11

3) Address all the issues in data mining regarding data mining methodology and user interaction. [7]

Ans ⇒ refer page No. 9 (b)

4) Describe the steps involved in data mining when viewed as a process of knowledge discovery. [7].

Ans ⇒ refer page No. 2