Unit 1 long questions.

Q: What is Data Mining?

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data.

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouse.

Q: What motivated data mining?

We have huge amount of data and now need to find some valuable information of or knowledge from this huge amount of data. We can use data mining technology in fraud detection, customer retention, loan recovery etc

Q: Why data mining is important?

Data mining is highly useful in the following domains –

1. Market Analysis and Management
2. Corporate Analysis &Risk Management
3. Fraud Detection

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid

1. Market Analysis and Management

Listed below are the various fields of market where data mining is used –

- **Customer Profiling** – Data mining helps determine what kind of people buy what kind of products.
- **Identifying Customer Requirements** – Data mining helps in identifying the best products for different customers. It uses prediction to find the factors that may attract new customers.
- **Cross Market Analysis** – Data mining performs Association/correlations between product sales.
- **Target Marketing** - Data mining helps to find clusters of model customers who share the same characteristics such as interests, spending habits, income, etc.
- **Determining Customer purchasing pattern** – Data mining helps in determining customer purchasing pattern.
- **Providing Summary Information** – Data mining provides us various multidimensional summary reports.

2. Corporate Analysis and Risk Management

Data mining is used in the following fields of the Corporate Sector –

- **Finance Planning and Asset Evaluation** – It involves cash flow analysis and prediction, contingent claim analysis to evaluate assets.
- **Resource Planning** – It involves summarizing and comparing the resources and spending.
- **Competition** – It involves monitoring competitors and market directions.

3. Fraud Detection

Data mining is also used in the fields of credit card services and telecommunication to detect frauds. In fraud telephone calls, it helps to find the destination of the call, duration of the call, time of the day or week, etc. It also analyzes the patterns that deviate from expected norms.

Q: Data mining functionalities.

There are two categories of functions involved in Data Mining –

- Descriptive
- Classification and Prediction

Descriptive Function

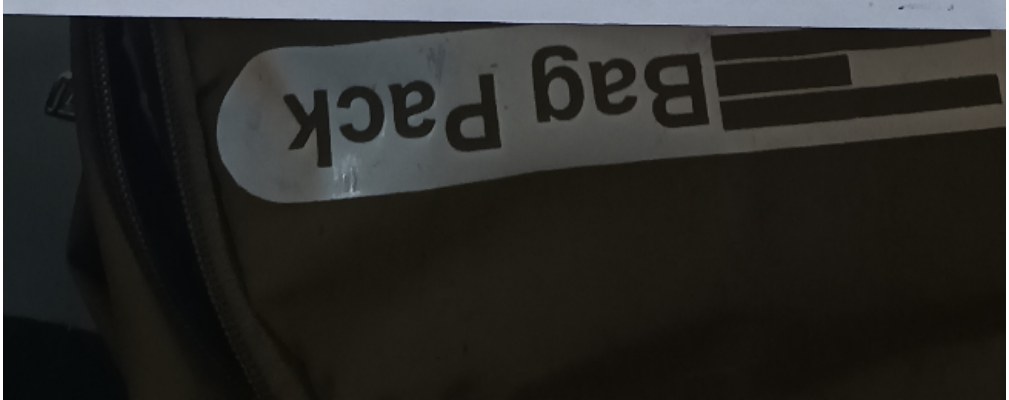
The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions –

1. Class/Concept Description
2. Mining of Frequent Patterns
3. Mining of Associations
4. Mining of Correlations
5. Mining of Clusters

1. Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways.

- **Data Characterization** – This refers to summarizing data of class under study. This class under study is called as Target Class.



- **Data Discrimination** – It refers to the mapping or classification of a class with some predefined group or class.

2. Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns –

- **Frequent Item Set** – It refers to a set of items that frequently appear together, for example, milk and bread.
- **Frequent Subsequence** – A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
- **Frequent Sub Structure** – Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

3. Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

4. Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute–value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

5. Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms.

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

The lists of functions involved in these processes are as follows –

- **Classification** – It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
- **Prediction** – It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
- **Outlier Analysis** – Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
- **Evolution Analysis** – Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

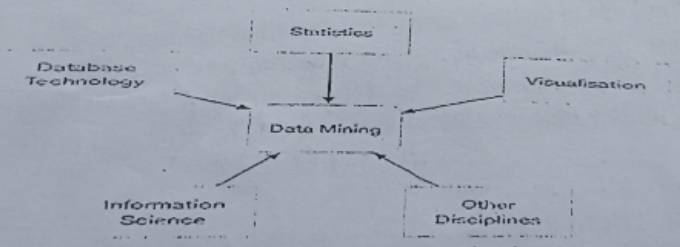
Q: Are all patterns interesting?

- The pattern is interesting if it is easily understood by users, it is valid on test data and new data with some certainty, it is useful and if it validates hypothesis that user sought to confirm.
- Finally an interesting pattern represents a knowledge.

Q: Classification of data mining.

A data mining system can be classified according to the following criteria –

- Database Technology
- Statistics
- Machine Learning
- Information Science
- Visualization
- Other Disciplines



Apart from these, a data mining system can also be classified based on the kind of (a) databases mined, (b) knowledge mined, (c) techniques utilized, and (d) applications adapted.

A. **Classification Based on the Databases Mined**

- We can classify a data mining system according to the kind of databases mined. Database system can be classified according to different criteria such as data models, types of data, etc. And the data mining system can be classified accordingly.
- For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

B. **Classification Based on the kind of Knowledge Mined**

We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified on the basis of functionalities such as–

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Prediction
- Outlier Analysis
- Evolution Analysis

C. **Classification Based on the Techniques Utilized**

- We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

D. **Classification Based on the Applications Adapted**

- We can classify a data mining system according to the applications adapted. These applications are as follows.

- Finance
- Telecommunications
- DNA
- Stock Markets
- E-mail

Q: Data Mining Task primitives.

- We can specify a data mining task in the form of a data mining query.
- This query is input to the system.
- A data mining query is defined in terms of data mining task primitives.

These primitives allow us to communicate in an interactive manner with the data mining system. Here is the list of Data Mining Task Primitives –

A. Set of task relevant data to be mined.
B. Kind of knowledge to be mined.
C. Background knowledge to be used in discovery process.
D. Interestingness measures and thresholds for pattern evaluation.
E. Representation for visualizing the discovered patterns.

A. Set of task relevant data to be mined

This is the portion of database in which the user is interested. This portion includes the following

- Database Attributes
- Data Warehouse dimensions of interest

B. Kind of knowledge to be mined

It refers to the kind of functions to be performed. These functions are

- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

C. Background knowledge

The background knowledge allows data to be mined at multiple levels of abstraction. For example, the Concept hierarchies are one of the background knowledge that allows data to be mined at multiple levels of abstraction.

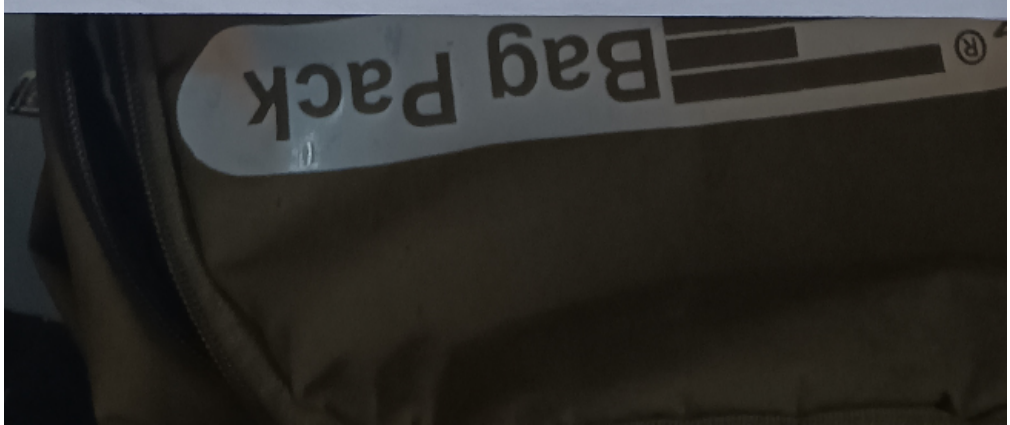
D. Interestingness measures and thresholds for pattern evaluation

This is used to evaluate the patterns that are discovered by the process of knowledge discovery. There are different interesting measures for different kind of knowledge.

E. Representation for visualizing the discovered patterns

This refers to the form in which discovered patterns are to be displayed. These representations may include the following

- Rules
- Tables
- Charts
- Graphs
- Decision Trees
- Cubes

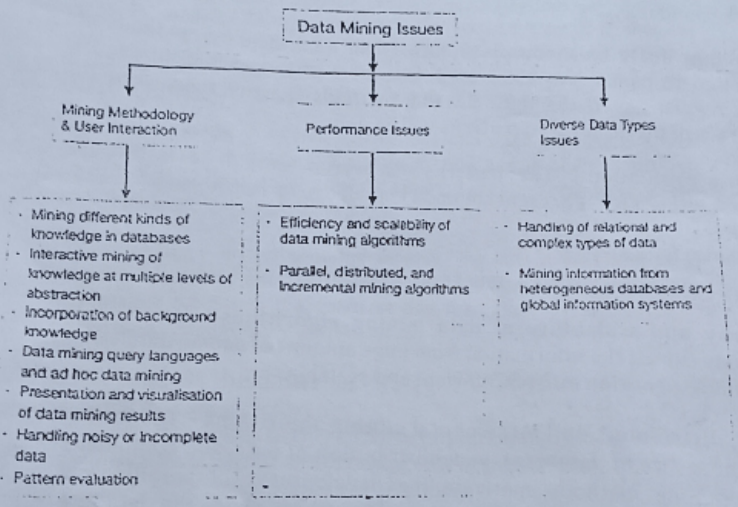


Q: Data Mining – Issues or Q: Challenges in data mining

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding

A. Mining Methodology and User Interaction
B. Performance Issues
C. Diverse Data Types Issues

The following diagram describes the major issues.



A. Mining Methodology and User Interaction Issues

It refers to the following kinds of issues –

- **Mining different kinds of knowledge in databases** – Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

- **Interactive mining of knowledge at multiple levels of abstraction** – The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** – To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

- **Data mining query languages and ad hoc data mining** – Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results** – Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

- **Handling noisy or incomplete data** – The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.

- **Pattern evaluation** – The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

B. Performance Issues

There can be performance–related issues such as follows –

- **Efficiency and scalability of data mining algorithms** – In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

- **Parallel, distributed, and incremental mining algorithms** – The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the result from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

C. Diverse Data Types Issues

- **Handling of relational and complex types of data** – The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.

- **Mining information from heterogeneous databases and global information systems** – The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

Q: Integrating a Data Mining System with a DB/DW System

If a data mining system is not integrated with a database or a data warehouse system, then there will be no system to communicate with. This scheme is known as the non-coupling scheme. In this scheme, the main focus is on data mining design and on developing efficient and effective algorithms for mining the available data sets.

The list of Integration Schemes is as follows

- **No Coupling** – In this scheme, the data mining system does not utilize any of the database or data warehouse functions. It fetches the data from a particular source and processes that data using some data mining algorithms. The data mining result is stored in another file.

- **Loose Coupling** – In this scheme, the data mining system may use some of the functions of database and data warehouse system. It fetches the data from the data respiratory managed by these systems and performs data mining on that data. It then stores the mining result either in a file or in a designated place in a database or in a data warehouse.

- **Semi–tight Coupling** - In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, efficient implementations of a few data mining primitives can be provided in the database.

- **Tight coupling** – in this coupling scheme, the data mining system is smoothly integrated into the database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

Q: Why preprocess the data? – The process of transforming raw da[ta] into an understandable format

Real world data are generally

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- Noisy: containing errors or outliers
- Inconsistent: containing discrepancies in codes or names

Tasks in data preprocessing

A. **Data cleaning**: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

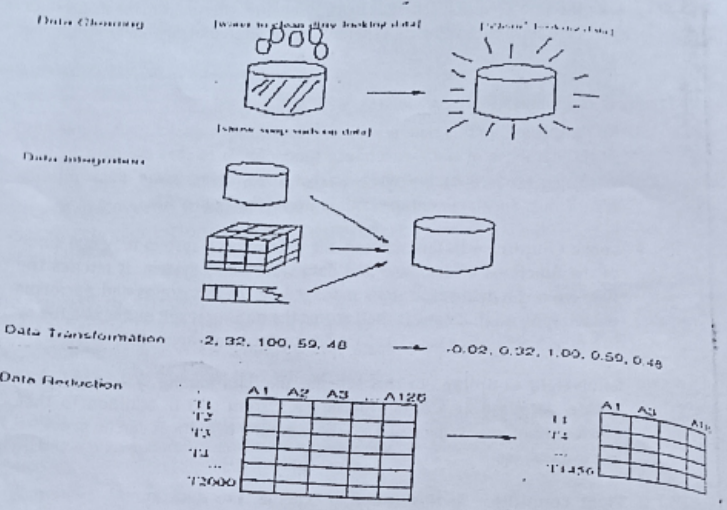
B. **Data integration**: using multiple databases, data cubes, or files.

C. **Data transformation**: normalization and aggregation.

D. **Data reduction**: reducing the volume but producing the same or similar analytical results.

E. **Data discretization**: part of data reduction, replacing numerical attributes with nominal ones.

Forms of data preprocessing



A. Data Cleaning

Data cleaning tasks

- Fill in missing values,
- Identify outliers and smooth out noisy data,
- Correct inconsistent data

- **Missing Data**

✓ Data is not always available
✓ E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
✓ Missing data may be due to equipment malfunction (faulty functioning)
✓ inconsistent with other recorded data and thus deleted
✓ Data not entered due to misunderstanding
✓ certain data may not be considered important at the time of entry
✓ not register history or changes of the data
✓ Missing data may need to be inferred

- **How to Handle Missing Data?**

✓ Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably)
✓ Fill in the missing value manually: tedious + infeasible?
✓ Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
✓ Use the attribute mean to fill in the missing value

- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree
- **Noisy Data**
  - Noise: random error or variance in a measured variable
  - Incorrect attribute values may due to
    - faulty data collection instruments
    - data entry problems
    - data transmission problems
    - technology limitation
    - inconsistency in naming convention
  - Other data problems which requires data cleaning
    - duplicate records
    - incomplete data
    - inconsistent data

- **How to Handle Noisy Data?**
  - Binning method:
    - first sort data and partition into (equi-depth) bins
    - then smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - Clustering
    - detect and remove outliers
  - Combined computer and human inspection
    - detect suspicious values and check by human
  - Regression
    - smooth by fitting the data into regression functions

Data Integration

- Data integration combines data from multiple sources into a coherent store
- Schema integration
- integrate metadata from different sources
- Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ,B.custid
- Detecting and resolving data value conflicts

for the same real world entity, attribute values from different sources are different

possible reasons: different representations, different scales, e.g., metric vs. British units

- **Handling Redundant Data**
- Redundant data occur often when integration of multiple databases
- The same attribute may have different names in different databases

Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

C. **Data Transformation**
- ✓ Smoothing: remove noise from data
- ✓ Aggregation: summarization, data cube construction
- ✓ Generalization: concept hierarchy climbing
- ✓ Normalization: scaled to fall within a small, specified range
- ✓ min-max normalization
- ✓ z-score normalization
- ✓ normalization by decimal scaling

❖ **Data Reduction Strategies**
- ✓ Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- ✓ Data reduction obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

1. Data cube aggregation
2. Dimensionality reduction
3. Numerosity reduction
4. Discretization and concept hierarchy generation

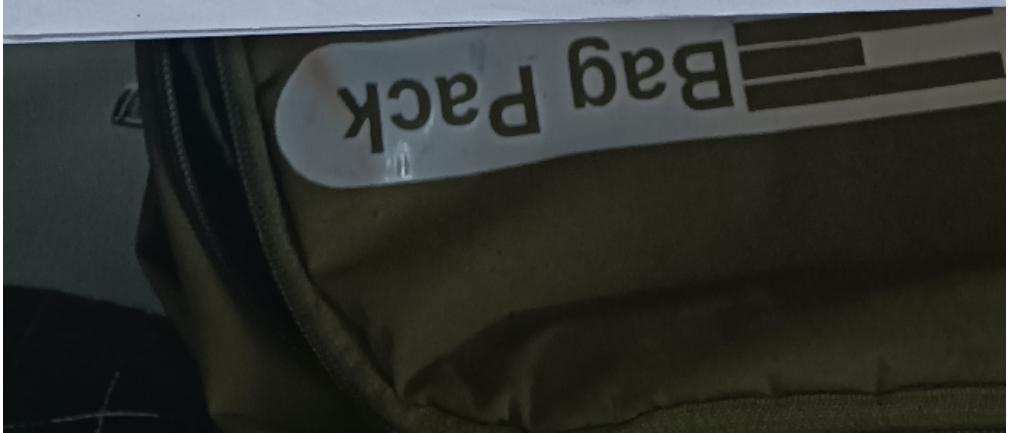
Data Cube Aggregation
- The lowest level of a data cube
  - the aggregated data for an individual entity of interest
  - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task

Dimensionality Reduction
- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand

Discretization and Concept Hierarchy Generation

✓ Discretization techniques can be used to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value labels can be used to actual data values. These methods are typically recursive, where a large amount of time is spent on sorting the data at each step. The smaller the number of distinct values to sort, the faster these methods should be.

Normalization :- Database normalization, or simply normalization, is the process of organizing the columns (attributes) and tables (relations) of a relational database...



- ✓ Many discretization techniques can be applied recursively in order to provide a hierarchical or
- ✓ Multiresolution partitioning of the attribute values known as concept hierarchy.
- ✓ A concept hierarchy for a given numeric attribute attribute defines a discretization of the attribute.
- ✓ Concept hierarchies can be used to reduce the data y collecting and replacing low-level concepts (such as numeric value for the attribute age) by higher level concepts (such as young, middle-aged, or senior). Although detail is lost by such generalization, it becomes meaningful and it is easier to interpret.
- ✓ Manual definition of concept hierarchies can be tedious and time-consuming task for the user or domain expert. Fortunately, many hierarchies are implicit within the database schema and can be defined at schema definition level. Concept hierarchies often can be generated automatically or dynamically refined based on statistical analysis of the data distribution.

Discretization and Concept Hierarchy Generation for Numeric Data:

- ✓ It is difficult and laborious for to specify concept hierarchies for numeric attributes due to the wide diversity of possible data ranges and the frequent updates if data values. Manual specification also could be arbitrary.

Binning:

- ✓ Attribute values can be discredited by distributing the values into bin and replacing each bin by the mean bin value or bin median value.
- ✓ This technique can be applied recursively to the resulting partitions in order to generate concept hierarchies.

Histogram Analysis:

- ✓ Histograms can also be used for discretization. Partitioning rules can be applied to define range of values.
- ✓ The histogram analyses algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a prespecified number of concept levels have been reached.
- ✓ A minimum interval size can be used per level to control the recursive procedure. This specifies the minimum width of the partition, or the minimum member of partitions at each level.

Cluster Analysis:

- ✓ A clustering algorithm can be applied to partition data into clusters or groups. Each cluster forms a node of a concept hierarchy, where all noses are at the same conceptual level.
- ✓ Each cluster may be further decomposed into sub-clusters, forming a lower level in the hierarchy.
- ✓ Clusters may also be grouped together to form a higher-level concept hierarchy.