Unit II

CHAPTER 2

Mathematical Models for Decision Making, Data Mining and Data Preparation

2.1 Modeling

Modeling is building models for the representation of modules which is also called as the entities of a System.

☞ **The needs of modeling are as follows**

- To decompose the system into its basic entities.
- To identify the essential entities and linkages.
- To recompose a selected version of the system with its essential/relevant entities and linkages (i.e. the model).

2.2 Models

A Model is a simplified representation of the essential entities of some specific reality and their characteristics.

☞ **The Models are used for following things :**

- Exploration
- Explanation
- Extrapolation

2.2.1 Mathematical Models

Q. 2.2.1 What are the different types of model? (Ref. Sec. 2.2.1) **(5 Marks)**

Mathematical Models can be classified as follows :

Types of mathematical models

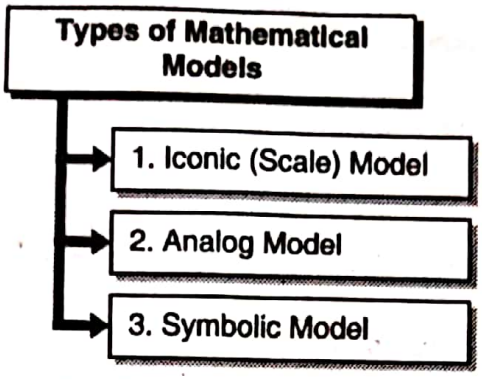

Fig. 2.2.1 : Types of mathematical models

1. Iconic (Scale) Model

- An iconic model is a physical copy of a system usually based on a different scale than the original. These may appear in three dimensions like airplane, car or bridge model to scale.
- Photographs are another type of iconic model but it is only two dimensions. An Iconic Model is a look-alike representation of some specific entity for example house.

Iconic Models can be represented in :

- **Two Dimensions:** e.g. photos, drawings, etc.
- **Three Dimensions :** e.g. scale model.

A scale model can be a

- reduction (scaled down, e.g. the model of a building).
- reproduction (same scale, e.g. copy model, prototype or working model).
- enlargement (scaled up, e.g. the model of an atom).

2. Analog Model

- An analog model does not look like the real system but behaves like it. These are usually two dimensional charts or diagrams for e.g., organization charts, showing structure, authority, and responsibility relationships.
- Analog models are more abstract than iconic ones. An Analogue Model is the representation of entities of a system by analogue entities pertaining to the model (e.g. through diagrams).

- An Analogue Model can be built through :

(a) Two Dimensional Visualization

(b) Three Dimensional Visualization

➜ **(a) Two Dimensional Visualization**

Charts, Graphs, Diagrams

(e.g. the colour coding of a geographical chart for representing different altitudes)

➜ **(b) Three Dimensional Visualization**

Analogue Devices

(e.g. the flow of water in pipes to represent the flow of electricity in wires or the flow of resources in an economic system)

➜ **3. Symbolic Model**

- The complexity of relationships in some systems cannot be represented physically or the physical representation may be cumbersome and take time to construct. Therefore a more abstract model is used with the aid of symbols.
- Most management science analysis is executed with the aid of mathematical models which utilize mathematical symbols. These are general rather than specific and can describe diverse situations.
- Furthermore they can be manipulated easily for purposes of experimentation and prediction.
- When the concept of a model is extended to the area of mathematics, it is useful to know in a quantitative sense how important or how pertinent the variables are in the model with regard to their impact on the solution.
- The mathematical models depict explicit relationships and interrelationships among the variables and other factors deemed important in solving problems.
- A Symbolic Model is the representation of entities of a system through symbols.
- Symbols can be :
  - Mathematical.
  - Logical.
  - ad-hoc.

- A Symbolic Model is used whenever the reality is :
  - too complex or too abstract to be portrayed through an iconic or analogue model
  - the factors of the system (variables) can be represented by symbols that can be manipulated in a meaningful and fruitful way.

Syllabus Topic : Structure of Mathematical Model

2.3 The Structure of Mathematical Models

Q. 2.3.1 Write short note on structure of mathematical model. (Ref. Sec. 2.3) **(5 Marks)**

Mathematical models are typically in the form of equations or other mathematical statements.

For example, the relationship between cost, revenue and profit can be expressed as:

$$P = R - C \quad \text{.... (2.3.1)}$$

Where, P is profit,

R is revenues, and C is cost.

2.3.1 Classification of Mathematical Models

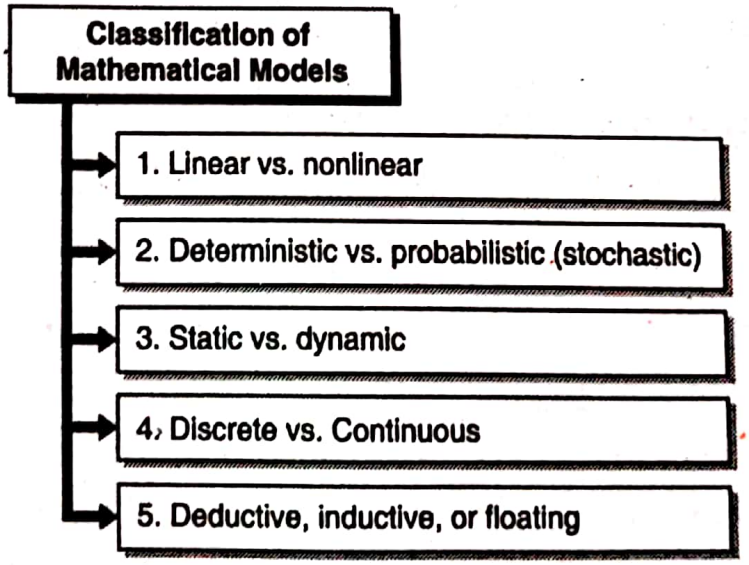

Fig. 2.3.1 : Classification of mathematical models

➜ **1. Linear vs. nonlinear**

- Mathematical models are usually composed by variables, which are abstractions of quantities of interest in the described systems, and operators that act on these variables, which can be algebraic operators, functions, differential operators, etc.

- If all the operators in a mathematical model exhibit linearity, the resulting mathematical model is defined as linear.
- A model is considered to be nonlinear otherwise. The question of linearity and nonlinearity is dependent on context, and linear models may have nonlinear expressions in them.
- For example, in a statistical linear model, it is assumed that a relationship is linear in the parameters, but it may be nonlinear in the predictor variables.
- Similarly, a differential equation is said to be linear if it can be written with linear differential operators, but it can still have nonlinear expressions in it.
- In a mathematical programming model, if the objective functions and constraints are represented entirely by linear equations, then the model is regarded as a linear model.
- If one or more of the objective functions or constraints are represented with a nonlinear equation, then the model is known as a nonlinear model.
- Nonlinearity, even in fairly simple systems, is often associated with phenomena such as chaos and irreversibility. Although there are exceptions, nonlinear systems and models tend to be more difficult to study than linear ones.
- A common approach to nonlinear problems is linearization, but this can be problematic if one is trying to study aspects such as irreversibility, which are strongly tied to nonlinearity.

➜ **2. Deterministic vs. probabilistic (stochastic)**

- A deterministic model is one in which every set of variable states is uniquely determined by parameters in the model and by sets of previous states of these variables.
- Therefore, deterministic models perform the same way for a given set of initial conditions.
- Conversely, in a stochastic model, randomness is present, and variable states are not described by unique values, but rather by probability distributions.

➜ **3. Static vs. dynamic**

- A static model does not account for the element of time, while a dynamic model does.
- Dynamic models typically are represented with difference equations or differential equations.

4. Discrete vs. Continuous

- A discrete model does not take into account the function of time and usually uses time-advance methods, while a Continuous model does.
- Continuous models typically are represented with f (t) and the changes are reflected over continuous time intervals.

5. Deductive, inductive, or floating

- A deductive model is a logical structure based on a theory. An inductive model arises from empirical findings and generalization from them. The floating model rests on neither theory nor observation, but is merely the invocation of expected structure.
- Application of mathematics in social sciences outside of economics has been criticized for unfounded models. Application of catastrophe theory in science has been characterized as a floating model.

Seven Steps of Mathematical Modeling

1. Formulate the Problem.
2. Observe the System.
3. Formulate a Mathematical Model of the Problem.
4. Verify the Model and Use the Model for Prediction.
5. Select a Simulation Alternative.
6. Present the Results and Conclusion of the Study to the Organization.
7. Implement and Evaluate Recommendations.

Characteristics of mathematical models

To be used successfully in a typical Management Science (MS) project, a mathematical model must meet the following criteria:

(i) The model should be as simple and understandable as possible.

(ii) The Model should be reasonable.

(iii) The Model should be easy to maintain and control.

(iv) The model should be adaptive. The parameters and structure of the model should be easy to change as new insights and information evolve.

(v) The model should be complete on important issues, i.e., all important variables and factors should have been taken into consideration.

Advantages of mathematical models

1. Use of models avoids constructing costly plants and warehouses in locations that do not best meet the present and future needs of the customers.
2. A model indicates gaps that are not immediately apparent, and after testing, the character of the failure might give a clue to the model's deficiencies.
3. Models have the advantage of time, since results can be obtained within a relatively-short time.
4. Because of the constant squeeze on profits, the cost and time saving that MS models allow make them decision-making tools of great value to the manager.

Disadvantages of mathematical models

1. A model that oversimplifies may inaccurately reflect the real world situation.
2. If the person who builds a model does not know what he is doing, output from the model will be incorrect.
3. Models can sometimes prove too expensive to originate when their cost is compared to the expected return from their use.

Syllabus Topic : Classes of Models

2.4 Classes of Models

Q. 2.4.1 Explain classes of model. (Ref. Sec. 2.4) (5 Marks)

There are various models which are used for making decisions. The various mathematical models are as follows :

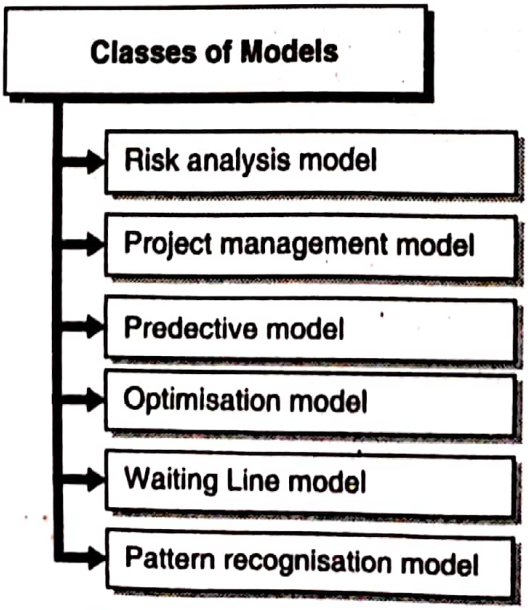

Fig. 2.4.1 : Classes of Models

1. Risk analysis model

- Risk analysis is the process of assessing the likelihood of an adverse event occurring within the corporate, government, or environmental sector.
- Risk analysis is the study of the underlying uncertainty of a given course of action and refers to the uncertainty of forecasted cash flow streams, variance of portfolio/stock returns, the probability of a project's success or failure, and possible future economic states.
- Risk analysts often work in tandem with forecasting professionals to minimize future negative unforeseen effects.

2. Project management model

- Every project is extremely unique which means we cannot have a standard structure to execute our projects and achieve success in our endeavor.
- However, to have a good plan we need some kind of framework or structure to follow depending on the nature of the project.
- Project management models or methodologies provide the framework to execute projects. A framework is something that tells you how often you will meet and discuss the progress, how you will document results, how you will communicate and so on.

3. Predective model

- Predictive modeling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results.
- Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation, or it may be a complex neural network, mapped out by sophisticated software.
- As additional data becomes available, the statistical analysis model is validated or revised.

4. Optimisation model

- The Optimization Model class provides a common API for defining and accessing variables and constraints, as well as other properties of each model.
- We will now discuss each of these components in more detail.

☞ Types of Optimization Models

- Optimization problems can be classified in terms of the nature of the objective function and the nature of the constraints. Special forms of the objective function and the constraints give rise to specialized algorithms that are more efficient.
- From this point of view, there are four types of optimization problems, of increasing complexity.
- An Unconstrained optimization problem is an optimization problem where the objective function can be of any kind (linear or nonlinear) and there are no constraints. These types of problems are handled by the classes discussed in the earlier sections.
- A linear program is an optimization problem with an objective function that is linear in the variables, and all constraints are also linear. Linear programs are implemented by the **Linear Program** class.
- A quadratic program is an optimization problem with an objective function that is quadratic in the variables (i.e. it may contain squares and cross products of the decision variables), and all constraints are linear. A quadratic program with no squares or cross products in the objective function is a linear program. Quadratic programs are implemented by the **Quadratic Program** class.
- A nonlinear program is an optimization problem with an objective function that is an arbitrary nonlinear function of the decision variables, and the constraints can be linear or nonlinear. Nonlinear programs are implemented by the **Nonlinear Program** class.

➜ 5. Waiting Line model

- There are basically two costs that must be balanced in waiting line system - the cost of service and the cost of waiting. Note that I am not considering another possible cost component - the cost of a scheduling system.
- Theoretically, a scheduling system is a management strategy designed to avoid waiting lines (meaning you should never wait in the doctor's office - yeah, right!) and is not covered in this module.
- Scheduling systems are useful when the customer is known to the system and the short and long run costs of waiting are relatively high. We will study scheduling system applications in linear programming later on in the course.
- Operational characteristics of waiting lines include:
  1. The probability that no customers (or units) are in the system.
  2. The average number of customers in the lines.

3. The average number of customers in the system (customers in line plus those being served.
4. The average time a customer spends in the waiting line.
5. The average time a customer spends in the system (waiting time plus time in the service facility.
6. The probability that an arriving customer has to wait for service.

6. Pattern recognisation model

- Pattern recognition deals with identifying a pattern and confirming it again. In general, a pattern can be a fingerprint image, a handwritten cursive word, a human face, a speech signal, a bar code, or a web page on the Internet.
- The individual patterns are often grouped into various categories based on their properties. When the patterns of same properties are grouped together, the resultant group is also a pattern, which is often called a pattern **class**.
- Pattern recognition is the science for observing, distinguishing the patterns of interest, and making correct decisions about the patterns or pattern classes. Thus, a biometric system applies pattern recognition to identify and classify the individuals, by comparing it with the stored templates.

Syllabus Topic : Definition of Data Mining

2.5 Data Mining

Q. 2.5.1 Define Data Mining. (Ref. Sec. 2.5) (2 Marks)

- Data mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers to develop more effective marketing strategies, increase sales and decrease costs.
- Data mining depends on effective data collection, warehousing and computer processing. Data mining is also known as data discovery and knowledge discovery.

Syllabus Topic : Representation of Input Data

2.6 Data Mining Parameters

Q. 2.6.1 Write short note on Data Mining parameters. (Ref. Sec. 2.6) **(5 Marks)**

- In data mining, association rules are created by analysing data for frequent if/then patterns, then using the support and confidence criteria to locate the most important relationships within the data.
- Support is how frequently the items appear in the **database**, while confidence is the number of times if-then statements are accurate.
- Other data mining parameters include Sequence or Path Analysis, Classification, Clustering and Forecasting. Sequence or Path Analysis parameters look for patterns where one event leads to another later event.
- A Sequence is an ordered list of sets of items, and it is a common type of data structure found in many databases. A Classification parameter looks for new patterns, and might result in a change in the way the data is organized. Classification algorithms predict variables based on other factors within the database.
- Clustering parameters find and visually document groups of facts that were previously unknown. Clustering groups a set of objects and aggregates them based on how similar they are to each other.
- There are different ways a user can implement the cluster, which differentiate between each clustering model. Fostering parameters within data mining can discover patterns in data that can lead to reasonable predictions about the future, also known as predictive analysis.

2.6.1 Data Mining Tools and Techniques

- Data mining techniques are used in many research areas, including mathematics, cybernetics, genetics and marketing. While data mining techniques are a means to drive efficiencies and predict customer behavior, if used correctly, a business can set itself apart from its competition through the use of predictive analysis.
- Web mining, a type of data mining used in customer relationship management, integrates information gathered by traditional data mining methods and techniques over the web.
- Other data mining techniques include network approaches based on multitask learning for classifying patterns, ensuring parallel and scalable execution of data mining algorithms,

the mining of large databases, the handling of relational and complex data types, and machine learning. Machine learning is a type of data mining tool that designs specific algorithms from which to learn and predict.

Syllabus Topic : Data Mining Process

2.7 Data Mining Architecture

Q. 2.7.1 Draw and explain architecture of data mining. (Ref. Sec. 2.7) (5 Marks)

- The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

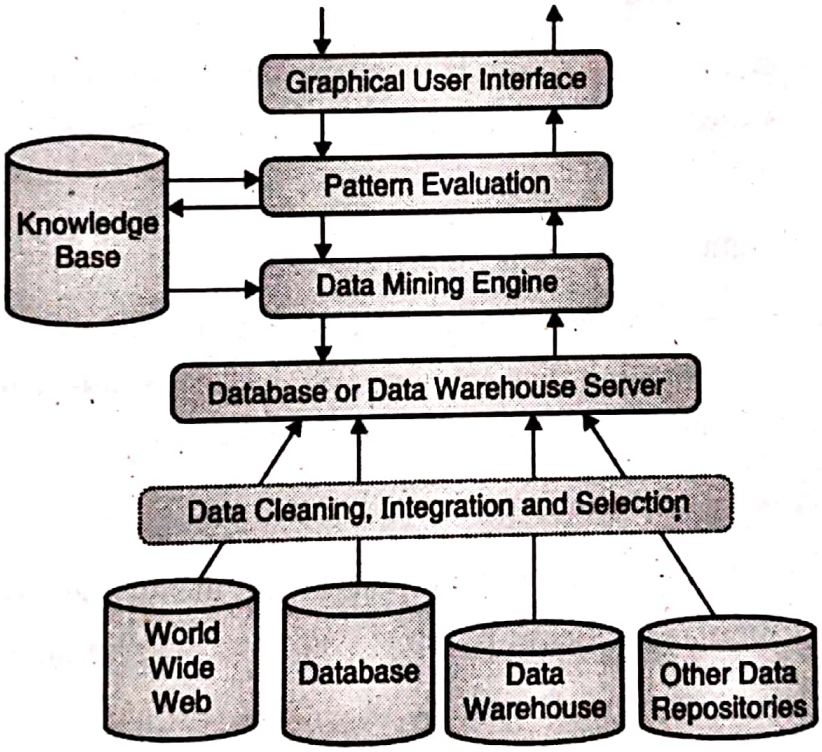

Fig. 2.7.1 : Data Mining System

(a) Data sources

- Database, data warehouse, World Wide Web (WWW), text files and other documents are the actual sources of data. You need large volumes of historical data for data mining to be successful.
- Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

☞ **Different processes**

- The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable.
- So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server.
- These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

➜ **(b) Database or Data warehouse server**

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.

➜ **(c) Data mining engine**

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

➜ **(d) Pattern evaluation modules**

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

➜ **(e) Graphical user interface**

- The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process.
- When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

➜ **(f) Knowledge base**

- The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns.

- The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable.
- The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.

2.7.1 Four Types of Data Mining Architecture

Types of Data Mining Architecture

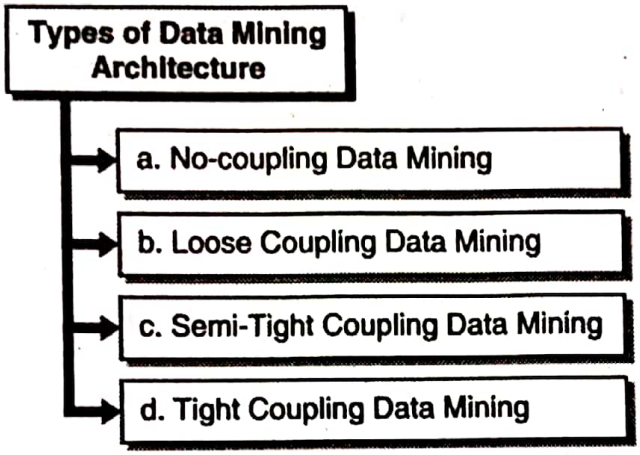

Fig. 2.7.2 : Types of Data Mining Architecture

(a) No-coupling data mining

- In this architecture, data mining system does not use any functionality of a database. A no-coupling data mining system retrieves data from a particular data sources.
- The no-coupling data mining architecture does not take any advantages of a database. That is already very efficient in organizing, storing, accessing and retrieving data.
- The no-coupling architecture is considered a poor architecture for data mining system. But it is used for simple data mining processes.

(b) Loose coupling data mining

- In this architecture, data mining system uses a database for data retrieval. In loose coupling, data mining architecture, data mining system retrieves data from a database. And it stores the result in those systems.
- Data mining architecture is for memory-based data mining system. That does not must high scalability and high performance.

➜ (c) Semi-Tight coupling data mining

- In semi-tight coupling, data mining system uses several features of data warehouse systems. That is to perform some data mining tasks. That includes sorting, indexing, aggregation.
- In this, some intermediate result can be stored in a database for better performance.

➜ (d) Tight coupling data mining

- In tight coupling, a data warehouse is treated as an information retrieval component. All the features of database or data warehouse are used to perform data mining tasks.
- This architecture provides system scalability, high performance, and integrated information.

There are three tiers in the tight-coupling data mining architecture

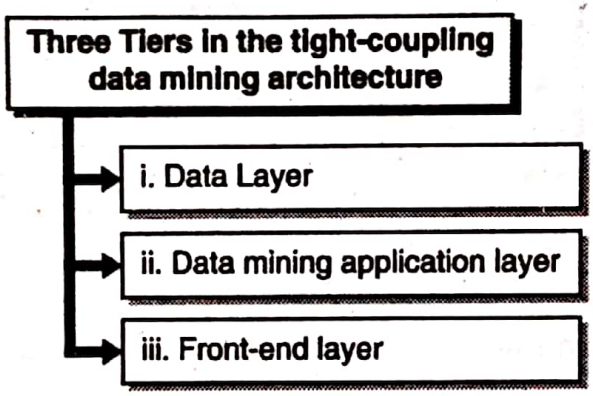

Fig. 2.7.3 : Three Tiers in the tight-coupling data mining architecture

➜ (i) Data layer

- We can define data layer as a database or data warehouse systems. This layer is an interface for all data sources.
- Data mining results are stored in the data layer. Thus, we can present to end-user in form of reports or another kind of visualization.

➜ (ii) Data mining application layer

- It is to retrieve data from a database. Some transformation routine has toper form here. That is to transform data into the desired format.
- Then we have to process data using various data mining algorithms.

➜ (iii) Front-end layer

- It provides the intuitive and friendly user interface for end-user. That is to interact with data mining system.

- Data mining result presented in visualization form to the user in the front-end layer.

2.7.2 Types of Data Mining Processes

- Different data mining processes can be classified into two types: data preparation or data preprocessing and data mining. In fact, the first four processes, that are data cleaning, data integration, data selection and data transformation, are considered as data preparation processes.
- The last three processes including data mining, pattern evaluation and knowledge representation are integrated into one process called data mining.

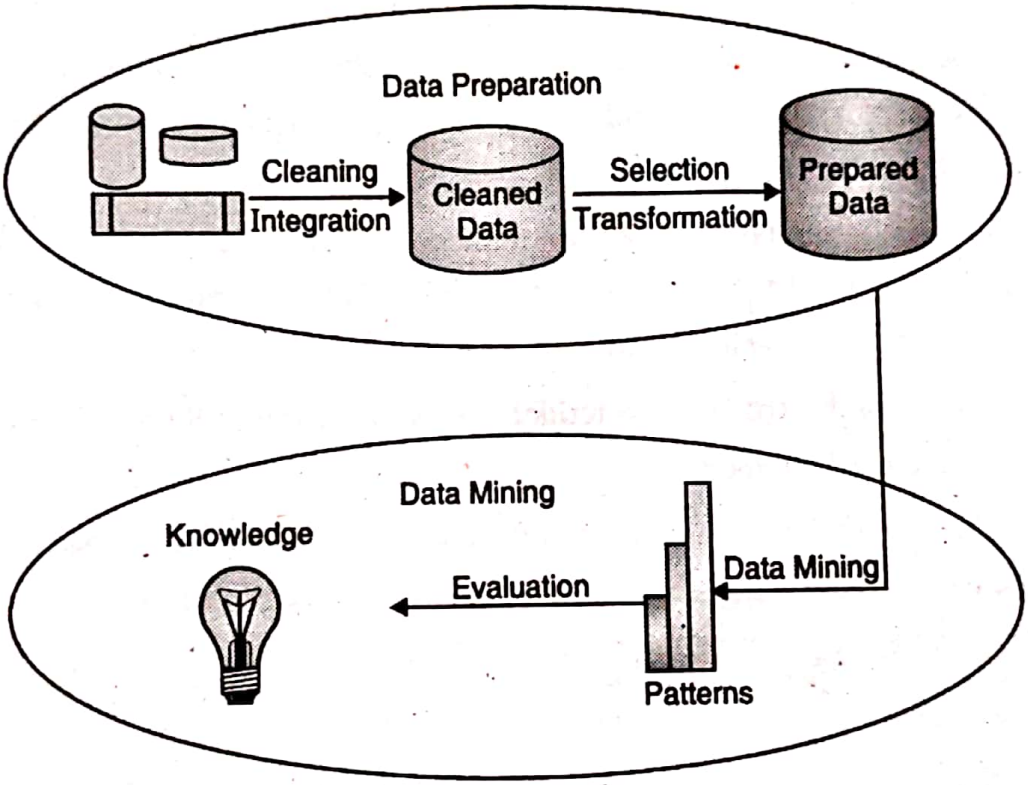

Fig. 2.7.4

(a) Data cleaning

- Data cleaning is the process where the data gets cleaned. Data in the real world is normally incomplete, noisy and inconsistent.
- The data available in data sources might be lacking attribute values, data of interest etc. For example, you want the demographic data of customers and what if the available data does not include attributes for the gender or age of the customers? Then the data is of course incomplete. Sometimes the data might contain errors or outliers.
- An example is an age attribute with value 200. It is obvious that the age value is wrong in this case. The data could also be inconsistent.

- For example, the name of an employee might be stored differently in different data tables or documents. Here, the data is inconsistent. If the data is not clean, the data mining results would be neither reliable nor accurate.
- Data cleaning involves a number of techniques including filling in the missing values manually, combined computer and human inspection, etc. The output of data cleaning process is adequately cleaned data.

➜ **(b) Data integration**

- Data integration is the process where data from different data sources are integrated into one. Data lies in different formats in different locations.
- Data could be stored in databases, text files, spreadsheets, documents, data cubes, Internet and so on. Data integration is a really complex and tricky task because data from different sources does not match normally.
- Suppose a table A contains an entity named customer_id where as another table B contains an entity named number. It is really difficult to ensure that whether both these entities refer to the same value or not.
- Metadata can be used effectively to reduce errors in the data integration process. Another issue faced is data redundancy.
- The same data might be available in different tables in the same database or even in different data sources. Data integration tries to reduce redundancy to the maximum possible level without affecting the reliability of data.

➜ **(c) Data selection**

- Data mining process requires large volumes of historical data for analysis. So, usually the data repository with integrated data contains much more data than actually required.
- From the available data, data of interest needs to be selected and stored. Data selection is the process where the data relevant to the analysis is retrieved from the database.

➜ **(d) Data transformation**

- Data transformation is the process of transforming and consolidating the data into different forms that are suitable for mining. Data transformation normally involves normalization, aggregation, generalization etc.
- For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. After data integration, the available data is ready for data mining.

➜ **(e) Data mining**

- Data mining is the core process where a number of complex and intelligent methods are applied to extract patterns from data.
- Data mining process includes a number of tasks such as association, classification, prediction, clustering, time series analysis and so on.

➜ **(f) Pattern evaluation**

- The pattern evaluation identifies the truly interesting patterns representing knowledge based on different types of interestingness measures.
- A pattern is considered to be interesting if it is potentially useful, easily understandable by humans, validates some hypothesis that someone wants to confirm or valid on new data with some degree of certainty.

➜ **(g) Knowledge representation**

- The information mined from the data needs to be presented to the user in an appealing way.
- Different knowledge representation and visualization techniques are applied to provide the output of data mining to the users.

☞ **Benefits of data mining**

1. Data mining technique helps companies to get knowledge-based information.
2. Data mining helps organizations to make the profitable adjustments in operation and production.
3. The data mining is a cost-effective and efficient solution compared to other statistical data applications.
4. Data mining helps with the decision-making process.
5. Facilitates automated prediction of trends and behaviors as well as automated discovery of hidden patterns.
6. It can be implemented in new systems as well as existing platforms.
7. It is the speedy process which makes it easy for the users to analyze huge amount of data in less time.

☞ **Disadvantages of data mining**

1. There are chances of companies may sell useful information of their customers to other companies for money. For example, American Express has sold credit card purchases of their customers to the other companies.

2. Many data mining analytics software is difficult to operate and requires advance training to work on.
3. Different data mining tools work in different manners due to different algorithms employed in their design. Therefore, the selection of correct data mining tool is a very difficult task.
4. The data mining techniques are not accurate, and so it can cause serious consequences in certain conditions.

Syllabus Topic : Analysis Methodologies

2.8 Analysis Methodologies

Q. 2.8.1 Write various application of data mining. (Ref. Sec. 2.8) **(5 Marks)**

Data Mining Applications

Data mining is highly useful in the following domains :

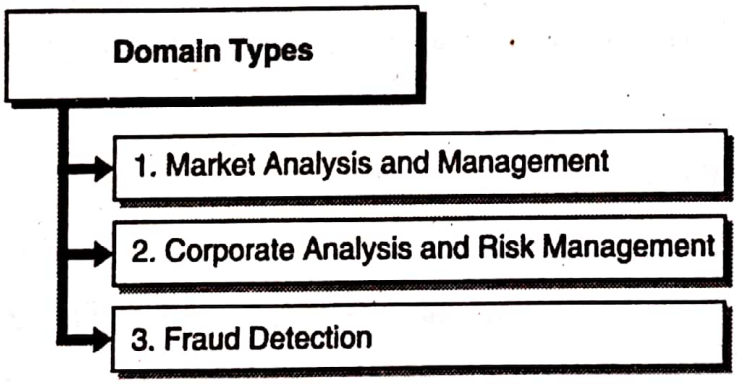

Fig. 2.8.1 : Domain Types

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

2.8.1 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.

2.8.2 Corporate Analysis and Risk Management

Q. 2.8.2 Write short note on Corporate Analysis and Risk Management. (Ref. Sec. 2.8.2) **(5 Marks)**

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.

2.8.3 Fraud Detection

Q. 2.8.3 Write short note on fraud detection. (Ref. Sec. 2.8.3) **(5 Marks)**

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.

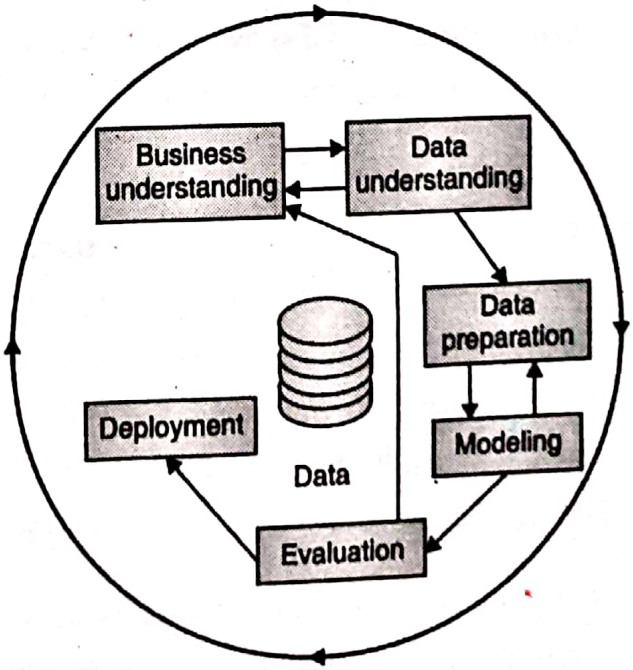

Fig. 2.8.2

➜ 1. Business understanding

In the business understanding phase :

- First, it is required to understand business objectives clearly and find out what are the business's needs.
- Next, we have to assess the current situation by finding the resources, assumptions, constraints and other important factors which should be considered.
- Then, from the business objectives and current situations, we need to create data mining goals to achieve the business objectives within the current situation.
- Finally, a good data mining plan has to be established to achieve both business and data mining goals. The plan should be as detailed as possible.

➜ 2. Data understanding

- First, the data understanding phase starts with initial data collection, which we collect from available data sources, to help us get familiar with the data.
- Some important activities must be performed including data load and data integration in order to make the data collection successfully.
- Next, the "gross" or "surface" properties of acquired data need to be examined carefully and reported.
- Then, the data needs to be explored by tackling the data mining questions, which can be addressed using querying, reporting, and visualization.
- Finally, the data quality must be examined by answering some important questions such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

➜ 3. Data preparation

- The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set.
- Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding.

➜ 4. Modeling

- First, modeling techniques have to be selected to be used for the prepared dataset.
- Next, the test scenario must be generated to validate the quality and validity of the model.

- Then, one or more models are created by running the modeling tool on the prepared dataset.
- Finally, models need to be assessed carefully involving stakeholders to make sure that created models are met business initiatives.

➜ 5. Evaluation

- In the evaluation phase, the model results must be evaluated in the context of business objectives in the first phase. In this phase, new business requirements may be raised due to the new patterns that have been discovered in the model results or from other factors.
- Gaining business understanding is an iterative process in data mining. The go or no-go decision must be made in this step to move to the deployment phase.

➜ 6. Deployment

- The knowledge or information, which we gain through data mining process, needs to be presented in such a way that stakeholders can use it when they want it.
- Based on the business requirements, the deployment phase could be as simple as creating a report or as complex as a repeatable data mining process across the organization.
- In the deployment phase, the plans for deployment, maintenance, and monitoring have to be created for implementation and also future supports.
- From the project point of view, the final report of the project needs to summary the project experiences and reviews the project to see what need to improved created learned lessons.
- The CRISP-DM offers a uniform framework for experience documentation and guidelines. In addition, the CRISP-DM can apply in various industries with different types of data.

Syllabus Topic : Data Preparation

2.9 What is Data Preparation ?

Q. 2.9.1 Draw diagram and explain data preparation. (Ref. Sec. 2.9) (5 Marks)

- Data preparation (or data pre-processing) in this context means manipulation of data into a form suitable for further analysis and processing. It is a process that involves many different tasks and which cannot be fully automated.

- Many of the data preparation activities are routine, tedious, and time consuming. It has been estimated that data preparation accounts for 60%-80% of the time spent on a data mining project.
- Data preparation is essential for successful data mining. Poor quality data typically result in incorrect and unreliable data mining results.
- Data preparation improves the quality of data and consequently helps improve the quality of data mining results. The well-known saying "garbage-in garbage-out" is very relevant to this domain.

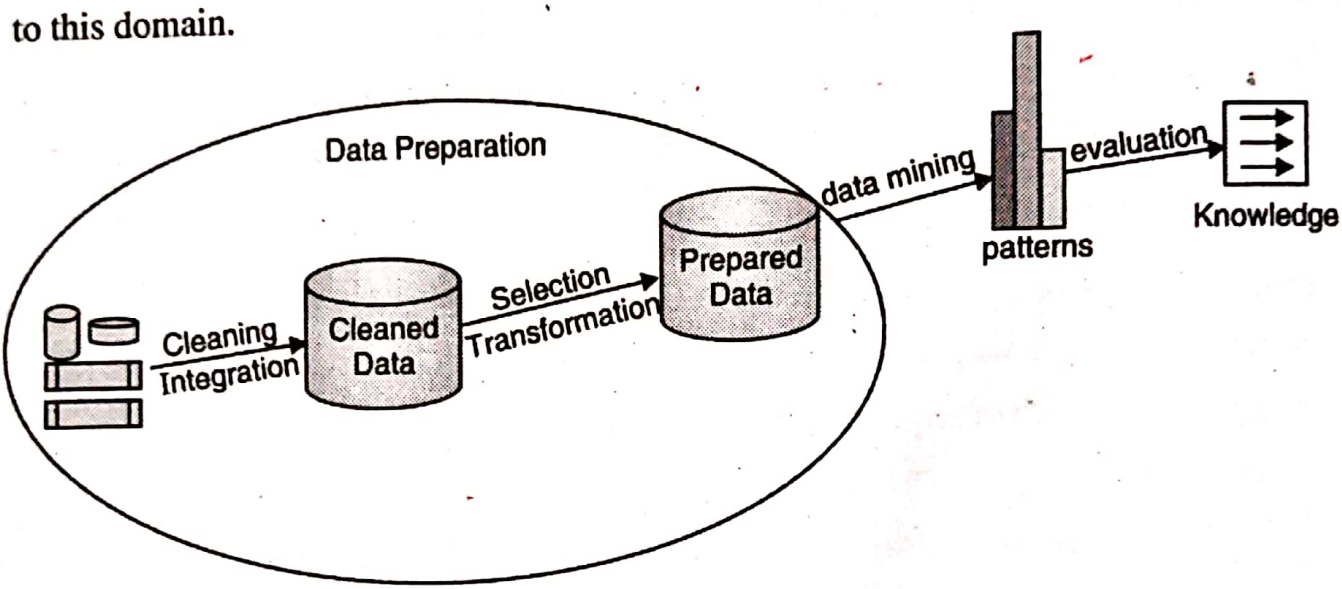

Fig. 2.9.1

Syllabus Topic : Data Validation

2.10 Data Validation

Q. 2.10.1 Write note on Data validation. (Ref. Sec. 2.10) (5 Marks)

- Data validation is about checking the information and to ensure that it complements the data needs of the system. This removes the chances of errors. One of the many examples of data validation is range check.
- Data validation has nothing to do with what the user wants to input. Validation is about checking the input data to ensure it conforms to the data requirements of the system to avoid data errors.
- An example of this is a **range check** to avoid an input number that is greater or smaller than the specified range.

Syllabus Topic : Data Transformation

2.11 Data Transformation

Q. 2.11.1 Explain data transformation with suitable diagram. (Ref. Sec. 2.11) (5 Marks)

In data transformation process data are transformed from one format to another format, that is more appropriate for data mining.

☞ **Some data transformation strategies**

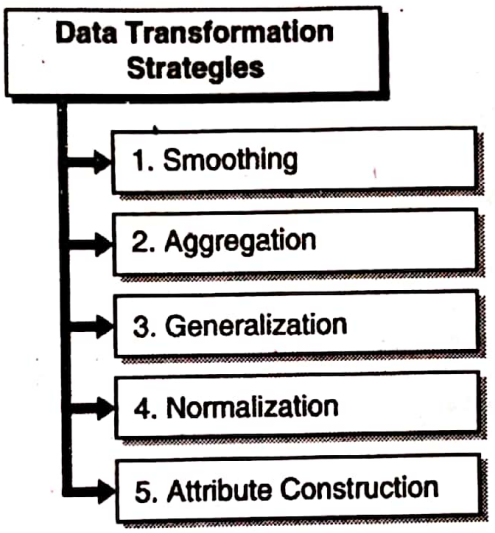

Fig. 2.11.1 : Data Transformation Strategies

➜ **1. Smoothing**

Smoothing is a process of removing noise from the data.

➜ **2. Aggregation**

Aggregation is a process where summary or aggregation operations are applied to the data.

➜ **3. Generalization**

In generalization low-level data are replaced with high-level data by using concept hierarchies climbing.

➜ **4. Normalization**

Normalization scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0.

➜ **5. Attribute Construction**

In Attribute construction, new attributes are constructed from the given set of attributes, database or date warehouse may store terabytes of data. So it may take very long to perform data analysis and mining on such huge amounts of data.

Syllabus Topic : Data Reduction

2.12 Data Reduction

Q. 2.12.1 Write short note on data Reduction. (Ref. Sec. 2.12) **(5 Marks)**

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but still contain critical information.

☞ **Data reduction strategies**

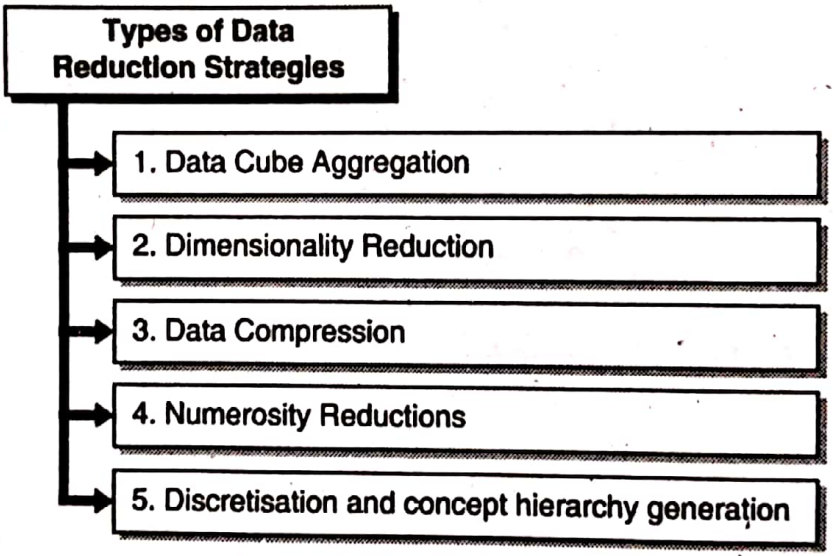

Fig. 2.12.1 : Types of data reduction strategies

➜ **1. Data cube aggregation**

Aggregation operations are applied to the data in the construction of a data cube.

➜ **2. Dimensionality reduction**

In dimensionality reduction redundant attributes are detected and removed which reduce the data set size.

➜ **3. Data compression**

Encoding mechanisms are used to reduce the data set size.

➜ **4. Numerosity reductions**

In numerosity reduction where the data are replaced or estimated by alternative.

➜ **5. Discretisation and concept hierarchy generation**

Where raw data values for attributes are replaced by ranges or higher conceptual levels.

2.13 Exam Pack (Review Questions)

Q. 1 What are the different types of model? (Refer Section 2.2.1) **(5 Marks)**

☞ **Syllabus Topic : Structure of Mathematical Model**

Q. 2 Write short note on structure of mathematical model. (Refer Section 2.3) **(5 Marks)**

☞ **Syllabus Topic : Classes of Models**

Q. 3 Explain classes of model. (Refer Section 2.4) **(5 Marks)**

☞ **Syllabus Topic : Definition of Data Mining**

Q. 4 Define Data Mining. (Refer Section 2.5) **(2 Marks)**

☞ **Syllabus Topic : Representation of Input Data**

Q. 5 Write short note on Data Mining parameters. (Refer Section 2.6) **(5 Marks)**

☞ **Syllabus Topic : Data Mining Process**

Q. 6 Draw and explain architecture of data mining. (Refer Section 2.7) **(5 Marks)**

☞ **Syllabus Topic : Analysis Methodologies**

Q. 7 Write various application of data mining. (Refer Section 2.8) **(5 Marks)**

Q. 8 Write short note on Corporate Analysis and Risk Management. (Refer Section 2.8.2) **(5 Marks)**

Q. 9 Write short note on fraud detection. (Refer Section 2.8.3) **(5 Marks)**

☞ **Syllabus Topic : Data Preparation**

Q. 10 Draw and explain data preparation. (Refer Section 2.9) **(5 Marks)**

☞ **Syllabus Topic : Data Validation**

Q. 11 Write note on Data validation. (Refer Section 2.10) **(5 Marks)**

☞ **Syllabus Topic : Data Transformation**

Q. 12 Explain data transformation with suitable diagram. (Refer Section 2.11) **(5 Marks)**

☞ **Syllabus Topic : Data Reduction**

Q. 13 Write short note on data Reduction. (Refer Section 2.12) **(5 Marks)**

❑❑❑

Chapter Ends...