**UNIT-I**

Introduction: Fundamentals of data mining, Data Mining Functionalities, Classification of Data Mining systems, Data Mining Task Primitives, Integration of a Data Mining System with a Database or Data Warehouse System, Major issues in Data Mining.

**Why Data Mining. 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. The information or knowledge extracted so can be used for any of the following applications −

* Market Analysis
* Fraud Detection
* Customer Retention
* Production Control
* Science Exploration

It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

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.

The key properties of data mining are:

* Automatic discovery of patterns
* Prediction of likely outcomes
* Creation of actionable information
* Focus on large datasets and databases

# The Scope of Data Mining

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities.

**Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands- on analysis can now be answered directly from the data — quickly.

A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

**Automated discovery of previously unknown patterns.** Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

# Kinds of Data Can Be Mined

Types of data mining

* Flat Files.
* Relational Databases.
* Data Warehouse.
* Transactional Databases.
* Multimedia Databases.
* Spatial Databases. (Maps)
* Time Series Databases. (Temporal data)
* World Wide Web(WWW)

**Data Mining Functionalities.**

Data mining involves six common classes of tasks:

* Anomaly detection (Outlier/change/deviation detection)
* Association rule learning (Dependency modelling)
* Clustering
* Classification
* Regression
* Summarization

Anomaly detection (Outlier/change/deviation detection) - The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (Dependency modelling) - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification - is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression - attempts to find a function which models the data with the least error.

Summarization - providing a more compact representation of the data set, including visualization and report generation.

**Kinds of Patterns that can Be Mined**

Data mining functionality can be broken down into 4 main "problems," namely:

* + - Classification, Prediction and Regression (together: predictive analysis);
    - Cluster Analysis;
    - Frequent Pattern Mining; and
    - Outlier Analysis.

# Classification

Classification is one of the main methods of supervised learning, and the manner in which prediction is carried out as relates to data with class labels. Classification involves finding a model which describes data classes, which can then be used to classify instances of unknown data. The concept of training data versus testing data is of integral importance to classification.

Popular classification algorithms for model building, and manners of presenting classifier models, include (but are not limited to):

* + - * Decision Trees
      * Support Vector Machines
      * Neural Networks
      * Nearest Neighbors

Examples of classification abound. A sample of such opportunities include:

* + - Identifying credit risks at multiple levels (low, medium, high)
    - Loan approvals (binary classification: loan versus no loan)
    - Classifying news stories based on multiple topics (politics, sports, business, , etc.)

# Regression

Regression is similar to classification, in that it is another dominant form of supervised learning and is useful for predictive analysis. They differ in that classification is used for predictions of data with distinct finite classes, while regression is used for predicting continuous numeric data. As a form of supervised learning, training/testing data is an important concept in regression as well. Linear regression is a common form of regression "mining."

A few particular examples include:

* + - Predicting home prices, as houses tend to be priced on the financial continuum, as opposed to being categorical
    - Trend estimation, in the fitting of trend lines to time series data
    - Multivariate estimation of health related indicators, such as life expectancy

# Cluster Analysis

Clustering is used for analyzing data which does not include pre-labeled classes. Data instances are grouped together using the concept of maximizing intraclass similarity and minimizing the similarity between differing classes. This translates to the clustering algorithm identifying and grouping instances which are very similar, as opposed to ungrouped instances which are much less-similar to one another. As clustering does not require the pre-labeling of classes, it is a form of unsupervised learning.

*k*-means Clustering is perhaps the most well-known example of a clustering algorithm. Different clustering schemes exist, including hierarchical clustering, fuzzy clustering, and density clustering,

In marketing, clustering can be of particular use in identifying distinct groups of customer bases, allowing for targeting based on what techniques may be known to have worked with other similar customers in said groups.

# Frequent Pattern Mining

Frequent pattern mining is a concept that has been used for a very long time to describe an aspect of data mining that many would argue is the very essence of the term data mining: taking a set of data and applying statistical methods to find interesting and previously- unknown patterns within said set of data. We aren't looking to classify instances or perform instance clustering; we simply want to learn patterns of subsets which emerge within a dataset and across instances, which ones emerge frequently, which items are **associated**, and which items **correlate** with others. It's easy to see why the above terms become conflated.

Frequent pattern mining is most closely identified with market basket analysis, which is the identification of subsets of finite superset of products that are purchased together with some level of both absolute and correlative frequency. This concept can be generalized beyond the purchase of items; however, the underlying principle of item subsets remains unchanged.

# Outlier Analysis

Outlier analysis, also called anomaly detection, is a bit different than the other data mining "problems," and is often not considered on its own, for a few specific reasons.

First, and most importantly to this discussion, outlier analysis is not its own method of mining as are the other problems above, but instead can actually use the above methods for its own goals (it's an end, as opposed to a means). Second, outlier analysis can also be approached as an exercise in descriptive statistics, which some would argue is not data mining at all (holding that data mining consists of, by definition, predictive statistical methods). However, in the interests of being exhaustive, it has been included here.

Outliers are data instances which do not seem to readily fit the behavior of the remaining data or a resulting model. Though many data mining algorithms intentionally do not take outliers into account, or can be modified to explicitly discard them, there are times when outliers themselves are where the money is.

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

Characterization Discrimination





Association and Correlation Analysis Classification





Prediction Clustering Outlier Analysis







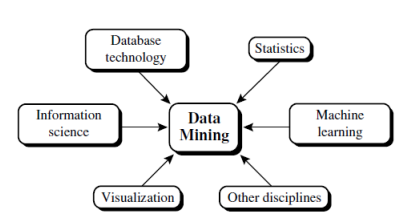
Evolution Analysis



**Classification of Data Mining Systems.**

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science .Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing. 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 datamining, datamining research is expected to generate a large variety of data mining systems. Therefore, it is necessary to provide a clear classification of data mining systems, which may help potential users distinguish between such systems and identify those that best match their needs.



Data mining systems can be categorized according to various criteria, as follows:

Classification according to the kinds of databases mined: A data mining system can be classified according to the kinds of databases mined. Database systems 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. Data mining systems can therefore be classified accordingly.

For instance, if classifying according to data models, we may have a relational, transactional, object-relational, or data warehouse mining system. If classifying according to the special types of data handled, we may have a spatial, time-series, text, stream data, multimedia data mining system, or a WorldWideWeb mining system.

Classification according to the kinds of knowledge mined: Data mining systems can be categorized according to the kinds of knowledge they mine, that is, based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis. A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.

Moreover, data mining systems can be distinguished based on the granularity or levels of abstraction of the knowledge mined, including generalized knowledge (at a highlevel of abstraction),primitive-level knowledge (at a rawdata level), or knowledge atmultiple levels (considering several levels of abstraction). An advanced data mining system should facilitate the discovery of knowledge at multiple levels of abstraction.

Data mining systems can also be categorized as those that mine data regularities (commonly occurring patterns) versus those that mine data irregularities (such as exceptions, or outliers). In general, concept description, association and correlation analysis, classification, prediction, and clustering mine data regularities, rejecting outliers as noise. These methods may also help detect outliers.

Classification according to the kinds of techniques utilized: Data mining systems can be categorized according to the underlying data mining techniques employed. 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). A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique that combines the merits of a few individual approaches.

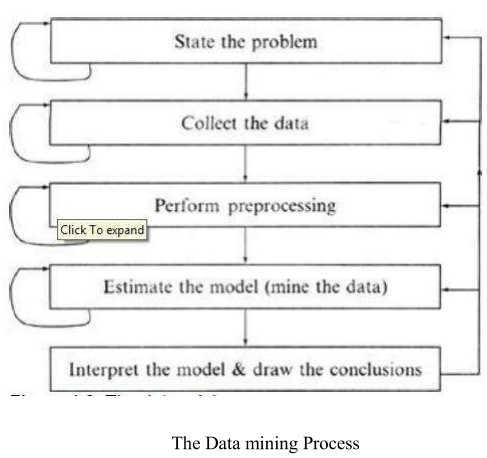
Classification according to the applications adapted: Data mining systems can also be categorized according to the applications they adapt. For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and so on. Different applications often require the integration of application-specific methods. Therefore, a generic, all-purpose data mining system may not fit domain-specific mining tasks.

**Data Mining Process**

Data Mining is a process of discovering various models, summaries, and derived values from a given collection of data.

The general experimental procedure adapted to data-mining problems involves the following steps:

* State the problem and formulate the hypothesis
* Collect the data
* Preprocessing the data
* Estimate the model
* Interpret the model and draw conclusions



**1. State the problem and formulate the hypothesis**

Most data-based modeling studies are performed in a particular application domain. Hence, domain-specific knowledge and experience are usually necessary in order to come up with a meaningful problem statement. Unfortunately, many application studies tend to focus on the data-mining technique at the expense of a clear problem statement. In this step, a modeler usually specifies a set of variables for the unknown dependency and, if possible, a general form of this dependency as an initial hypothesis. There may be several hypotheses formulated for a single problem at this stage. The first step requires the combined expertise of an application domain and a data-mining model. In practice, it usually means a close interaction between the data-mining expert and the application expert. In successful data-mining applications, this cooperation does not stop in the initial phase; it continues during the entire data-mining process.

**2. Collect the data**

This step is concerned with how the data are generated and collected. In general, there are two distinct possibilities. The first is when the data-generation process is under the control of an expert (modeler): this approach is known as a designed experiment. The second possibility is when the expert cannot influence the data- generation process: this is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Typically, the sampling distribution is completely unknown after data are collected, or it is partially and implicitly given in the data-collection procedure. It is very important, however, to understand how data collection affects its theoretical distribution, since such a priori knowledge can be very useful for modeling and, later, for the final interpretation of results. Also, it is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. If this is not the case, the estimated model cannot be successfully used in a fmal application of the results.

**3. Preprocessing the data**

In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts. Data preprocessing usually includes at least two common tasks:

1. Outlier detection ( and removal) - Outliers are unusual data values that are not consistent with most observations. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such non representative samples can seriously affect the model produced later. There are two strategies for dealing with outliers:

a.Detect and eventually remove outliers as a part of the preprocessing phase, or

b.Develop robust modeling methods that are insensitive to outliers.

2 Scaling, encoding, and selecting features - Data preprocessing includes several steps such as variable scaling and different types of encoding. For example, one feature with the range [0, 1] and the other with the range [-100, 1000] will not have the same weights in the applied technique; they will also influence the final data-mining results differently. Therefore, it is recommended to scale them and bring both features to the same weight for further analysis. Also, application-specific encoding methods usually achieve dimensionality reduction by providing a smaller number of informative features for subsequent data modeling.

These two classes of preprocessing tasks are only illustrative examples of a large spectrum of preprocessing activities in a data-mining process.

Data-preprocessing steps should not be considered completely independent from other data-mining phases. In every iteration of the data-mining process, all activities, together, could define new and improved data sets for subsequent iterations. Generally, a good preprocessing method provides an optimal representation for a data-mining technique by incorporating a priori knowledge in the form of application-specific scaling and encoding.

**4. Estimate the model**

The selection and implementation of the appropriate data-mining technique is the main task in this phase. This process is not straightforward; usually, in practice, the implementation is based on several models, and selecting the best one is an additional task.

**5. Interpret the model and draw conclusions**

In most cases, data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory. Usually, simple models are more interpretable, but they are also less accurate. Modem data-mining methods are expected to yield highly accurate results using high dimensional models. The problem of interpreting these models, also very important, is considered a separate task, with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He does not understand them; he cannot summarize, interpret, and use them for successful decision making.

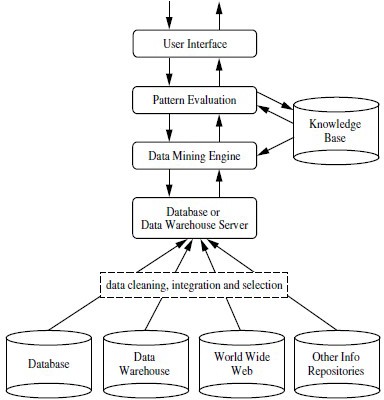
**Architecture of Data Mining**

A typical data mining system may have the following major components.

# Knowledge Base

# Data Mining Engine

* Pattern Evaluation Module
* User interface



A typical data mining system Architectures

# Knowledge Base:

This is the domain knowledge that is used to guide the search ore valuate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction. Knowledge such as user beliefs, which can be used to assess a pattern’s interestingness based on its unexpectedness, may also be included. Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

# Data Mining Engine:

This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

# Pattern Evaluation Module:

This component typically employs interestingness measures interacts with the data mining modules so as to focus the search toward interesting patterns. It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the datamining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process so as to confine the search to only the interesting patterns.

# User interface:

This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory datamining based on the intermediate data mining results. In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

**data integration**

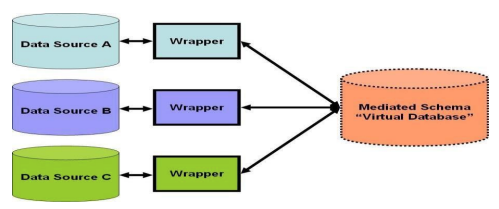
Data Integration: It combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files.

The data integration systems are formally defined as triple <G,S,M>

Where G: The global schema

S: Heterogeneous source of schemas

M: Mapping between the queries of source and global schema



**Issues in Data integration:**

1. Schema integration and object matching: How can the data analyst or the computer be sure that customer id in one database and customer number in another reference to the same attribute.

2. Redundancy: An attribute (such as annual revenue, for instance) may be redundant if it can be derived from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

3. Detection and resolution of data value conflicts: For the same real-world entity, attribute values from different sources may differ.

**Data Transformation:** In data transformation, the data are transformed or consolidated into forms appropriate for mining.

Data transformation can involve the following:

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. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities. Generalization of the data, where low-level or ―primitive‖ (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higherlevel 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.

# Major Issues 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 –

* + - Mining Methodology and User Interaction
    - Performance Issues
    - Diverse Data Types Issues

# Mining Methodology and User Interaction

* **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.

**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 results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

# 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.

# Applications that are targeted in Data mining

Here is the list of areas where data mining is widely used −

* + - Financial Data Analysis
    - Retail Industry
    - Telecommunication Industry
    - Biological Data Analysis
    - Other Scientific Applications
    - Stock market
    - Intrusion Detection

# Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

* Design and construction of data warehouses for multidimensional data analysis and data mining.
* Loan payment prediction and customer credit policy analysis.
* Classification and clustering of customers for targeted marketing.
* Detection of money laundering and other financial crimes.

# Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

* Design and Construction of data warehouses based on the benefits of data mining.
* Multidimensional analysis of sales, customers, products, time and region.
* Analysis of effectiveness of sales campaigns.
* Customer Retention.
* Product recommendation and cross-referencing of items.

# Telecommunication Industry

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

* Multidimensional Analysis of Telecommunication data.
* Fraudulent pattern analysis.
* Identification of unusual patterns.
* Multidimensional association and sequential patterns analysis.
* Mobile Telecommunication services.
* Use of visualization tools in telecommunication data analysis.

# Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

* Semantic integration of heterogeneous, distributed genomic and proteomic databases.
* Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
* Discovery of structural patterns and analysis of genetic networks and protein pathways.
* Association and path analysis.
* Visualization tools in genetic data analysis.

# Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modelling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

* Data Warehouses and data preprocessing.
* Graph-based mining.
* Visualization and domain specific knowledge.