**UNIT-02**

**INTRODUCTION TO DATAMINING**

**DATAMINING:**

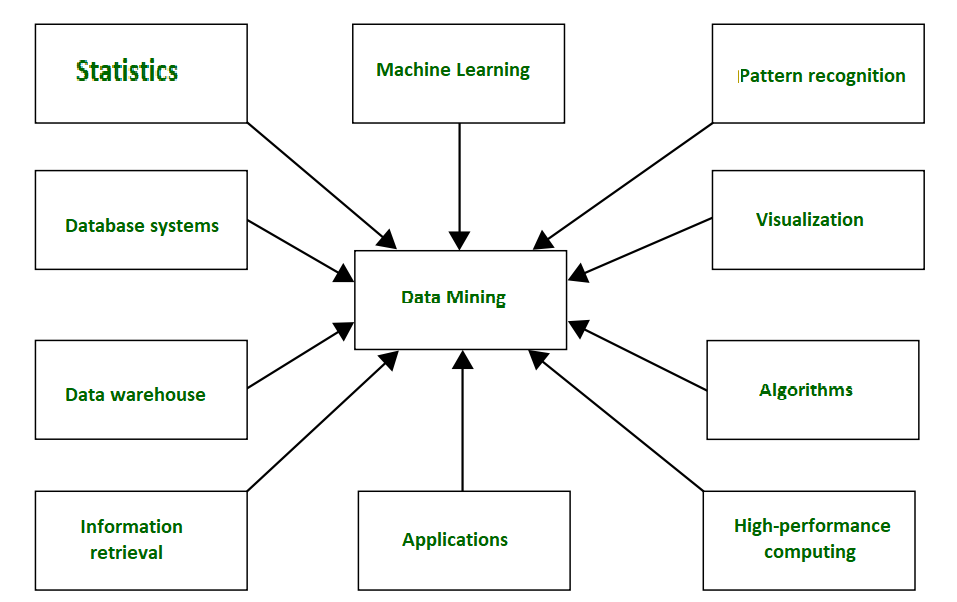
**Data mining** is the process of extracting useful information from large sets of data. It involves using various techniques from statistics, machine learning, and database systems to identify patterns, relationships, and trends in the data. This information can then be used to make data-driven decisions, solve business problems, and uncover hidden insights. Applications of data mining include customer profiling and segmentation, market basket analysis, anomaly detection, and predictive modeling. Data mining tools and technologies are widely used in various industries, including finance, healthcare, retail, and telecommunications.

In general terms, “**Mining**” is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining, etc. In the context of computer science, “**Data Mining”** can be referred to as **knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging**.  It is basically the process carried out for the extraction of useful information from a bulk of data or[data warehouses.](https://www.geeksforgeeks.org/data-warehousing/) One can see that the term itself is a little confusing. In the case of coal or diamond mining, the result of the extraction process is coal or diamond. But in the case of Data Mining, the result of the extraction process is not data!! Instead, data mining results are the patterns and knowledge that we gain at the end of the extraction process. In that sense, we can think of Data Mining as a step in the process of Knowledge Discovery or Knowledge Extraction.

**Gregory Piatetsky-Shapiro** coined the term**“Knowledge Discovery in Databases”** in 1989. However, the term**‘data mining’** became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

Nowadays, data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use ‘data mining’ to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

**Main Purpose of Data Mining**



*Data Mining*

Basically, Data mining has been integrated with many other techniques from other domains such as **statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization,** etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions.

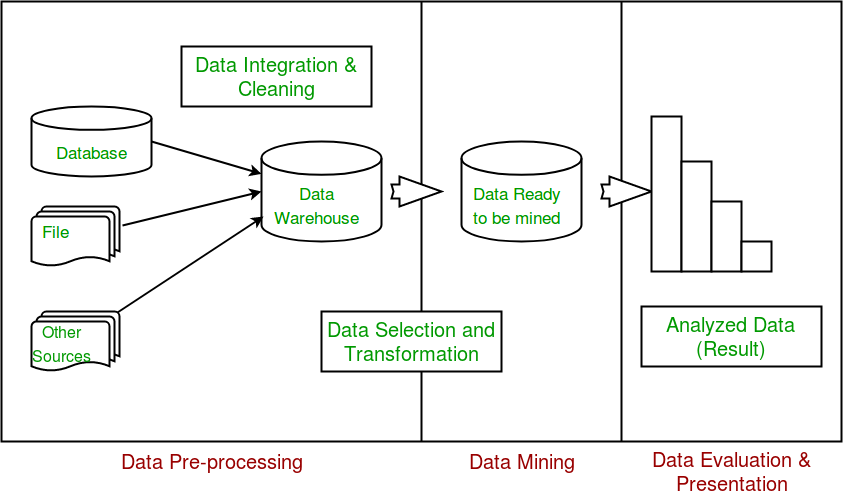
Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

Data Mining can be applied to any type of data e.g., **Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.**

**Data Mining as a whole process**

The whole process of Data Mining consists of three main phases:

1. Data Pre-processing – Data cleaning, integration, selection, and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results



**Applications of Data Mining**

1. Financial Analysis
2. Biological Analysis
3. Scientific Analysis
4. Intrusion Detection
5. Fraud Detection
6. Research Analysis

**There are several benefits of data mining, including:**

1. **Improved decision making**: Data mining can provide valuable insights that can help organizations make better decisions by identifying patterns and trends in large data sets.
2. **Increased efficiency:**Data mining can automate repetitive and time-consuming tasks, such as data cleaning and preparation, which can help organizations save time and resources.
3. **Enhanced competitiveness:** Data mining can help organizations gain a competitive edge by uncovering new business opportunities and identifying areas for improvement.
4. **Improved customer service:** Data mining can help organizations better understand their customers and tailor their products and services to meet their needs.
5. **Fraud detection:** Data mining can be used to identify fraudulent activities by detecting unusual patterns and anomalies in data.
6. **Predictive modeling:**Data mining can be used to build models that can predict future events and trends, which can be used to make proactive decisions.
7. **New product development:** Data mining can be used to identify new product opportunities by analyzing customer purchase patterns and preferences.
8. **Risk management:** Data mining can be used to identify potential risks by analyzing data on customer behavior, market conditions, and other factors.

**Real-life examples of Data Mining**

**Market Basket Analysis**: It is a technique that gives the careful study of purchases done by a customer in a supermarket. The concept is basically applied to identify the items that are bought together by a customer. Say, if a person buys bread, what are the chances that he/she will also purchase butter. This analysis helps in promoting offers and deals by the companies. The same is done with the help of data mining.

**Protein Folding:**It is a technique that carefully studies the biological cells and predicts the protein interactions and functionality within biological cells. Applications of this research include determining **causes and possible cures for Alzheimer’s, Parkinson’s,** and cancer caused by Protein misfolding.

**Fraud Detection:**Nowadays, in this land of cell phones, we can use data mining to analyze cell phone activities for comparing suspicious phone activity. This can help us to detects calls made on cloned phones. Similarly, with credit cards, comparing purchases with historical purchases can detect activity with stolen cards.

Data mining also has many successful applications, such as business intelligence, Web search, bioinformatics, health informatics, finance, digital libraries, and digital governments.

**Knowledge Discovery in Databases** (KDD).

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process in data mining typically involves the following steps:

1. **Selection**: Select a relevant subset of the data for analysis.
2. **Pre-processing:** Clean and transform the data to make it ready for analysis. This may include tasks such as data normalization, missing value handling, and data integration.
3. **Transformation:**Transform the data into a format suitable for data mining, such as a matrix or a graph.
4. **Data Mining:** Apply data mining techniques and algorithms to the data to extract useful information and insights. This may include tasks such as clustering, classification, association rule mining, and anomaly detection.
5. **Interpretation**: Interpret the results and extract knowledge from the data. This may include tasks such as visualizing the results, evaluating the quality of the discovered patterns and identifying relationships and associations among the data.
6. **Evaluation**: Evaluate the results to ensure that the extracted knowledge is useful, accurate, and meaningful.
7. **Deployment**: Use the discovered knowledge to solve the business problem and make decisions.

The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.

**Why do we need Data Mining?**

Volume of information is increasing everyday than we can handle from business transactions, scientific data, sensor data, Pictures, videos, etc. So, we need a system that will be capable of extracting essence of information available and that can automatically generate report,   
views or summary of data for better decision-making.

**Why Data Mining is used in Business?**   
Data mining is used in business to make better managerial decisions by: 

* **Automatic summarization of data**
* **Extracting essence of information stored**.
* **Discovering patterns in raw data.**

**Data Mining** also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

**Steps Involved in KDD Process:**



 KDD process

1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
   * Cleaning in case of ***Missing values***.
   * Cleaning ***noisy*** data, where noise is a random or variance error.
   * Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.
2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(Datawarehouse).
   * Data integration using ***Data Migration tools***.
   * Data integration using ***Data Synchronization tools***.
   * Data integration using ***ETL***(Extract-Load-Transformation) process.
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   * Data selection using ***Neural network***.
   * Data selection using ***Decision Trees***.
   * Data selection using ***Naive bayes***.
   * Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

Data Transformation is a two step process:

* + ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
  + ***Code generation***: Creation of the actual transformation program.

1. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   * Transforms task relevant data into ***patterns***.
   * Decides purpose of model using ***classification*** or ***characterization***.
2. ***Pattern Evaluation***: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
   * Find ***interestingness score*** of each pattern.
   * Uses ***summarization*** and ***Visualization*** to make data understandable by user.
3. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   * Generate ***reports***.
   * Generate ***tables***.
   * Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

**Note**: 

* KDD is an ***iterative* process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
* ***Preprocessing* of databases** consists of **Data cleaning** and **Data Integration**.

**ADVANTAGES OR DISADVANTAGES:  
Advantages of KDD:**

1. **Improves decision-making:** KDD provides valuable insights and knowledge that can help organizations make better decisions.
2. Increased efficiency: KDD automates repetitive and time-consuming tasks and makes the data ready for analysis, which saves time and money.
3. **Better customer service:** KDD helps organizations gain a better understanding of their customers’ needs and preferences, which can help them provide better customer service.
4. **Fraud detection:** KDD can be used to detect fraudulent activities by identifying patterns and anomalies in the data that may indicate fraud.
5. **Predictive modeling:** KDD can be used to build predictive models that can forecast future trends and patterns.

**Disadvantages of KDD:**

1. **Privacy concerns:** KDD can raise privacy concerns as it involves collecting and analyzing large amounts of data, which can include sensitive information about individuals.
2. **Complexity:** KDD can be a complex process that requires specialized skills and knowledge to implement and interpret the results.
3. **Unintended consequences:** KDD can lead to unintended consequences, such as bias or discrimination, if the data or models are not properly understood or used.
4. **Data Quality:** KDD process heavily depends on the quality of data, if data is not accurate or consistent, the results can be misleading
5. **High cost:** KDD can be an expensive process, requiring significant investments in hardware, software, and personnel.
6. **Overfitting:** KDD process can lead to overfitting, which is a common problem in machine learning where a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new unseen data.

**Difference Between KDD and Data Mining**

|  |  |  |
| --- | --- | --- |
| **Parameter** | **KDD** | **Data Mining** |
| Definition | KDD refers to a  process of identifying valid, novel, potentially useful, and ultimately understandable patterns and relationships in data. | Data Mining refers to a  process of extracting useful and valuable information or patterns from large data sets. |
| Objective | To find useful knowledge from data. | To extract useful information from data. |
| Techniques Used | Data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge representation and visualization. | Association rules, classification, clustering, regression, decision trees, neural networks, and dimensionality reduction. |
| Output | Structured information, such as rules and models, that can be used to make decisions or predictions. | Patterns, associations, or insights that can be used to improve decision-making or understanding. |
| Focus | Focus is on the discovery of useful knowledge, rather than simply finding patterns in data. | Focus is on the discovery of patterns or relationships in data. |
| Role of domain expertise | Domain expertise is important in KDD, as it helps in defining the goals of the process, choosing appropriate data, and interpreting the results. | Domain expertise is less critical in data mining, as the algorithms are designed to identify patterns without relying on prior knowledge. |

# Challenges of Data Mining

Nowadays [Data Mining](https://www.geeksforgeeks.org/data-mining/) and knowledge discovery are evolving a crucial technology for business and researchers in many domains.Data Mining is developing into established and trusted discipline, many still pending challenges have to be solved.

Some of these challenges are given below.

1. **Security and Social Challenges:**  
   Decision-Making strategies are done through data collection-sharing, so it requires considerable security. Private information about individuals and sensitive information are collected for customers profiles, user behaviour pattern understanding. Illegal access to information and the confidential nature of information becoming an important issue.
2. **User Interface:**  
   The knowledge discovered is discovered using data mining tools is useful only if it is interesting and above all understandable by the user. From good visualization interpretation of data, mining results can be eased and helps better understand their requirements. To obtain good visualization many research is carried out for big data sets that display and manipulate mined knowledge.  
   **(i) Mining based on Level of Abstraction:**Data Mining process needs to be collaborative because it allows users to concentrate on pattern finding, presenting and optimizing requests for data mining based on returned results.  
   **(ii) Integration of Background Knowledge:** Previous information may be used to express discovered patterns to direct the exploration processes and to express discovered patterns.
3. **Mining Methodology Challenges:**  
   These challenges are related to data mining approaches and their limitations. Mining approaches that cause the problem are:
4. **(i)** Versatility of the mining approaches,
5. **(ii)** Diversity of data available,
6. **(iii)** Dimensionality of the domain,

**(iv)** Control and handling of noise in data, etc.

Different approaches may implement differently based upon data consideration. Some algorithms require noise-free data. Most data sets contain exceptions, invalid or incomplete information lead to complication in the analysis process and some cases compromise the precision of the results.

1. **Complex Data:**  
   Real-world data is heterogeneous and it could be multimedia data containing images, audio and video, complex data, temporal data, spatial data, time series, natural language text etc. It is difficult to handle these various kinds of data and extract the required information. New tools and methodologies are developing to extract relevant information.  
   **(i) Complex data types:** The database can include complex data elements, objects with graphical data, spatial data, and temporal data. Mining all these kinds of data is not practical to be done one device.  
   **(ii) Mining from Varied Sources:**The data is gathered from different sources on Network. The data source may be of different kinds depending on how they are stored such as structured, semi-structured or unstructured.
2. **Performance:**  
   The performance of the data mining system depends on the efficiency of algorithms and techniques are using. The algorithms and techniques designed are not up to the mark lead to affect the performance of the data mining process.  
   **(i) Efficiency and Scalability of the Algorithms:** The data mining algorithm must be efficient and scalable to extract information from huge amounts of data in the database.  
   **(ii) Improvement of Mining Algorithms:** Factors such as the enormous size of the database, the entire data flow and the difficulty of data mining approaches inspire the creation of parallel & distributed data mining algorithms.

# **Data Preprocessing in Data Mining**

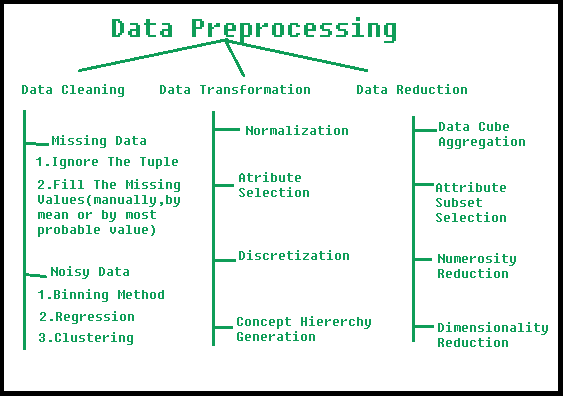
Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

**Some common steps in data preprocessing include:**

* **Data cleaning:** this step involves identifying and removing missing, inconsistent, or irrelevant data. This can include removing duplicate records, filling in missing values, and handling outliers.
* **Data integration:** this step involves combining data from multiple sources, such as databases, spreadsheets, and text files. The goal of integration is to create a single, consistent view of the data.
* **Data transformation:** this step involves converting the data into a format that is more suitable for the data mining task. This can include normalizing numerical data, creating dummy variables, and encoding categorical data.
* **Data reduction:** this step is used to select a subset of the data that is relevant to the data mining task. This can include feature selection (selecting a subset of the variables) or feature extraction (extracting new variables from the data).
* **Data discretization:** this step is used to convert continuous numerical data into categorical data, which can be used for decision tree and other categorical data mining techniques.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

**Pre-processing in Data Mining:**   
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



**Steps Involved in Data Preprocessing:**

**1. Data Cleaning:**   
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc. 

* **(a). Missing Data:**   
  This situation arises when some data is missing in the data. It can be handled in various ways.   
  Some of them are:
  1. **Ignore the tuples:**   
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**   
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
* **(b). Noisy Data:**   
  Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**   
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**   
     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).
  3. **Clustering:**   
     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

**2. Data Transformation:**   
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**   
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
2. **Attribute Selection:**   
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**   
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**   
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

**3. Data Reduction:**   
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**   
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**   
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**   
   This enable to store the model of data instead of whole data, for example: Regression Models.
4. **Dimensionality Reduction:**   
   This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

# Feature Subset Selection Process

**Feature Selection** is the most critical pre-processing activity in any machine learning process. It intends to select a subset of attributes or features that makes the most meaningful contribution to a machine learning activity. In order to understand it, let us consider a small example i.e. **Predict the weight of students based on the past information about similar students**, which is captured inside a ‘Student Weight’ data set. The data set has 04 features like **Roll Number, Age, Height & Weight.**Roll Number has no effect on the weight of the students, so we eliminate this feature. So now the new data set will be having only 03 features. This subset of the data set is expected to give better results than the full set.

|  |  |  |
| --- | --- | --- |
| **AGE** | **Height** | **Weight** |
| 12 | 1.1 | 23 |
| 11 | 1.05 | 21.6 |
| 13 | 1.2 | 24.7 |
| 11 | 1.07 | 21.3 |
| 14 | 1.24 | 25.2 |
| 12 | 1.12 | 23.4 |

**Data discretization**

It refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Now, we can understand this concept with the help of an example

Suppose we have an attribute of Age with the given values

|  |  |
| --- | --- |
| Age | 1,5,9,4,7,11,14,17,13,18, 19,31,33,36,42,44,46,70,74,78,77 |

Table before Discretization

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Attribute | Age | Age | Age | Age |
|  | 1,5,4,9,7 | 11,14,17,13,18,19 | 31,33,36,42,44,46 | 70,74,77,78 |
| After Discretization | Child | Young | Mature | Old |

Another example is analytics, where we gather the static data of website visitors. For example, all visitors who visit the site with the IP address of India are shown under country level.

## **Some Famous techniques of data discretization**

**Histogram analysis**

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

**Binning**

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

**Cluster Analysis**

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

**Data discretization using decision tree analysis**

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

**Data discretization using correlation analysis**

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

## **Data discretization and concept hierarchy generation**

The term hierarchy represents an organizational structure or mapping in which items are ranked according to their levels of importance. In other words, we can say that a hierarchy concept refers to a sequence of mappings with a set of more general concepts to complex concepts. It means mapping is done from low-level concepts to high-level concepts. For example, in computer science, there are different types of hierarchical systems. A document is placed in a folder in windows at a specific place in the tree structure is the best example of a computer hierarchical tree model. There are two types of hierarchy: top-down mapping and the second one is bottom-up mapping.

Let's understand this concept hierarchy for the dimension location with the help of an example.

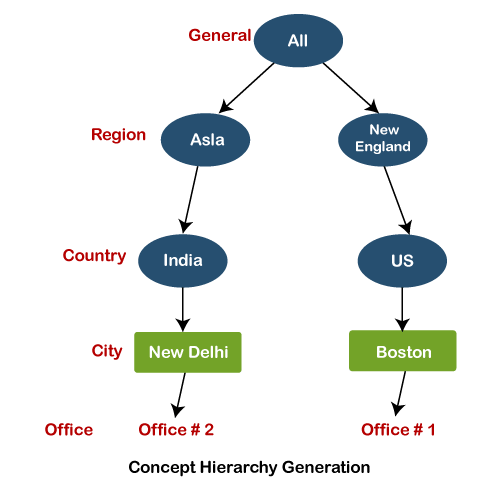
A particular city can map with the belonging country. For example, New Delhi can be mapped to India, and India can be mapped to Asia.

**Top-down mapping**

Top-down mapping generally starts with the top with some general information and ends with the bottom to the specialized information.

**Bottom-up mapping**

Bottom-up mapping generally starts with the bottom with some specialized information and ends with the top to the generalized information.



Data discretization and binarization in data mining

Data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. In contrast, data binarization is used to transform the continuous and discrete attributes into binary attributes.

Why is Discretization important?

As we know, an infinite of degrees of freedom mathematical problem poses with the continuous data. For many purposes, data scientists need the implementation of discretization. It is also used to improve signal noise ratio.

**Data transformation**

Data transformation in data mining refers to the process of converting raw data into a format that is suitable for analysis and modeling. The goal of data transformation is to prepare the data for data mining so that it can be used to extract useful insights and knowledge. Data transformation typically involves several steps, including:

1. **Data cleaning:**Removing or correcting errors, inconsistencies, and missing values in the data.
2. **Data integration:**Combining data from multiple sources, such as databases and spreadsheets, into a single format.
3. **Data normalization:**Scaling the data to a common range of values, such as between 0 and 1, to facilitate comparison and analysis.
4. **Data reduction:** Reducing the dimensionality of the data by selecting a subset of relevant features or attributes.
5. **Data discretization**: Converting continuous data into discrete categories or bins.
6. **Data aggregation:** Combining data at different levels of granularity, such as by summing or averaging, to create new features or attributes.
7. Data transformation is an important step in the data mining process as it helps to ensure that the data is in a format that is suitable for analysis and modeling, and that it is free of errors and inconsistencies. Data transformation can also help to improve the performance of data mining algorithms, by reducing the dimensionality of the data, and by scaling the data to a common range of values.

The data are transformed in ways that are ideal for mining the data. The data transformation involves steps that are:

**1. Smoothing:** It is a process that is used to remove noise from the dataset using some algorithms It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form. The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn’t see otherwise.

**2. Aggregation:** Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results. The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies. For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

**3. Discretization:** It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes. Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values. For **example**, (1-10, 11-20) (age:- young, middle age, senior).

**4. Attribute Construction:** Where new attributes are created & applied to assist the mining process from the given set of attributes. This simplifies the original data & makes the mining more efficient.

**5. Generalization:** It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example Age initially in Numerical form (22, 25) is converted into categorical value (young, old). For **example**, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

**6. Normalization:** Data normalization involves converting all data variables into a given range. Techniques that are used for normalization are:

* **Min-Max Normalization:**
  + This transforms the original data linearly.
  + Suppose that: min\_A is the minima and max\_A is the maxima of an attribute, P
  + Where v is the value you want to plot in the new range.
  + v’ is the new value you get after normalizing the old value.
* **Z-Score Normalization:**
  + In z-score normalization (or zero-mean normalization) the values of an attribute (A), are normalized based on the mean of A and its standard deviation
  + A value, v, of attribute A is normalized to v’ by computing
* **Decimal Scaling:**
  + It normalizes the values of an attribute by changing the position of their decimal points
  + The number of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.
  + A value, v, of attribute A is normalized to v’ by computing
  + where j is the smallest integer such that Max(|v’|) < 1.
  + Suppose: Values of an attribute P varies from -99 to 99.
  + The maximum absolute value of P is 99.
  + For normalizing the values we divide the numbers by 100 (i.e., j = 2) or (number of integers in the largest number) so that values come out to be as 0.98, 0.97 and so on.

**ADVANTAGES OR DISADVANTAGES:**

**Advantages of Data Transformation in Data Mining:**

1. Improves Data Quality: Data transformation helps to improve the quality of data by removing errors, inconsistencies, and missing values.
2. Facilitates Data Integration: Data transformation enables the integration of data from multiple sources, which can improve the accuracy and completeness of the data.
3. Improves Data Analysis: Data transformation helps to prepare the data for analysis and modeling by normalizing, reducing dimensionality, and discretizing the data.
4. Increases Data Security: Data transformation can be used to mask sensitive data, or to remove sensitive information from the data, which can help to increase data security.
5. Enhances Data Mining Algorithm Performance: Data transformation can improve the performance of data mining algorithms by reducing the dimensionality of the data and scaling the data to a common range of values.

**Disadvantages of Data Transformation in Data Mining:**

1. Time-consuming: Data transformation can be a time-consuming process, especially when dealing with large datasets.
2. Complexity: Data transformation can be a complex process, requiring specialized skills and knowledge to implement and interpret the results.
3. Data Loss: Data transformation can result in data loss, such as when discretizing continuous data, or when removing attributes or features from the data.
4. Biased transformation: Data transformation can result in bias, if the data is not properly understood or used.
5. High cost: Data transformation can be an expensive process, requiring significant investments in hardware, software, and personnel.

Overfitting: Data transformation can lead to overfitting, which is a common problem in machine learning where a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new unseen data.

**Similarity** measure

* is a numerical measure of how alike two data objects are.
* higher when objects are more alike.
* often falls in the range [0,1]

Similarity might be used to identify

* duplicate data that may have differences due to typos.
* equivalent instances from different data sets. E.g. names and/or addresses that are the same but have misspellings.
* groups of data that are very close (clusters)

**Dissimilarity** measure

* is a numerical measure of how different two data objects are
* lower when objects are more alike
* minimum dissimilarity is often 0 while the upper limit varies depending on how much variation can be

Dissimilarity might be used to identify

* outliers
* interesting exceptions, e.g. credit card fraud
* boundaries to clusters

**Proximity** refers to either a similarity or dissimilarity