UNIT-IV

DATA PRE-PROCESSING

Why to pre-process data? - Data cleaning: Missing Values, Noisy Data - Data Integration and transformation - Data Reduction: Data cube aggregation, Dimensionality reduction - Data Compression - Numerosity Reduction - Data Mining Primitives - Languages and System Architectures: Task relevant data - Kind of Knowledge to be mined - Discretization and Concept Hierarchy.

What is Data Preprocessing?

Data preprocessing is the process of preparing raw data for analysis by cleaning and transforming it into a usable format. In data mining it refers to preparing raw data for mining by performing tasks like cleaning, transforming, and organizing it into a format suitable for mining algorithms.

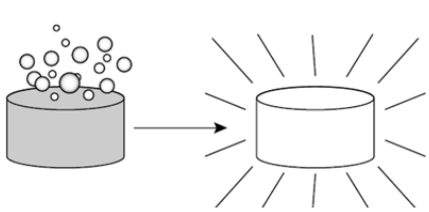
- Goal is to improve the quality of the data.
- Helps in handling missing values, removing duplicates, and normalizing data.
- Ensures the accuracy and consistency of the dataset.

Steps in Data Preprocessing

Some key steps in data preprocessing are:

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation

1. Data Cleaning: It is the process of identifying and correcting errors or inconsistencies in the dataset. It involves handling missing values, removing duplicates, and correcting incorrect or outlier data to ensure the dataset is accurate and reliable. Clean data is essential for effective analysis, as it improves the quality of results and enhances the performance of data models.



Missing Values: This occur when data is absent from a dataset. You can either ignore the rows with missing data or fill the gaps manually, with the attribute mean, or by using the most probable value. This ensures the dataset remains accurate and complete for analysis.

Noisy Data: It refers to irrelevant or incorrect data that is difficult for machines to interpret, often caused by errors in data collection or entry. It can be handled in several ways:

a) Binning: Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it. The sorted values are distributed into several "buckets," or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing.

There are three kinds of binning. They are:

- **Smoothing by Bin Means**: In this method, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9.
- **Smoothing by Bin Medians:** In this method, each value in a bin is replaced by the median value of the bin. For example, the median of the values 4, 8, and 15 in Bin 1 is 8. Therefore, each original value in this bin is replaced by the value 8.
- **Smoothing by Bin Boundaries:** In this method, the minimum and maximum values in each bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

  For example, the middle value of the values 4, 8, and 15 in Bin 1is replaced with nearest boundary i.e., 4.

Example:

Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

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

Smoothing by bin medians:

Bin 1: 8, 8, 8
Bin 2: 21, 21, 21
Bin 3: 28, 28, 28

Smoothing by bin boundaries:

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

b) Regression: Data smoothing can also be done by regression, a technique that used to predict the numeric values in a given data set. It analyses the relationship between a target variable (dependent) and its predicate variable (independent).

- Regression is a form of a supervised machine learning technique that tries to predict any continuous valued attribute.
- Regression done in two ways;

    a. **Linear regression** involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.

    b. **Multiple linear regression** is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.

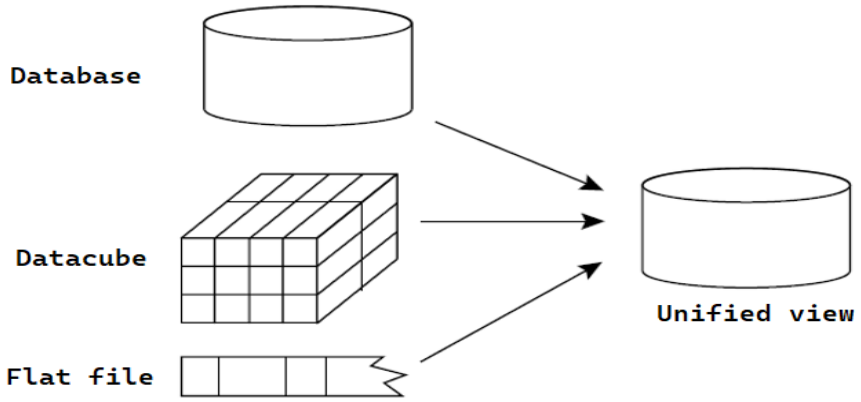
c) Clustering: It supports in identifying the outliers. The similar values are organized into clusters and those values which fall outside the cluster are known as outliers.

2. Data Integration

Data integration is the process of combining data from multiple sources into a single, unified view. This process involves identifying and accessing the different data sources, mapping the data to a common format. Different data sources may include multiple data cubes, databases, or flat files.

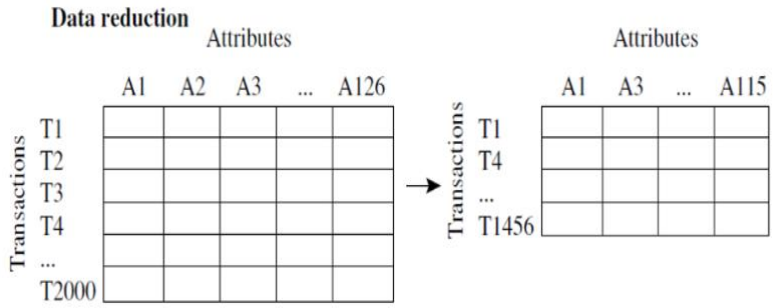
The goal of data integration is to make it easier to access and analyze data that is spread across multiple systems or platforms, in order to gain a more complete and accurate understanding of the data.

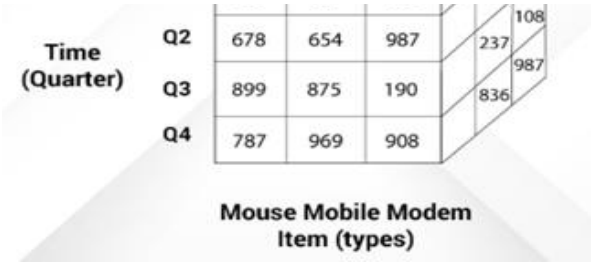
Data integration strategy is typically described using a triple (G, S, M) approach, where G denotes the global schema, S denotes the schema of the heterogeneous data sources, and M represents the mapping between the queries of the source and global schema.



3. Data Reduction: It reduces the dataset's size while maintaining key information. This can be done through feature selection, which chooses the most relevant features, and feature extraction, which transforms the data into a lower-dimensional space while preserving important details.

It uses various reduction techniques such as,

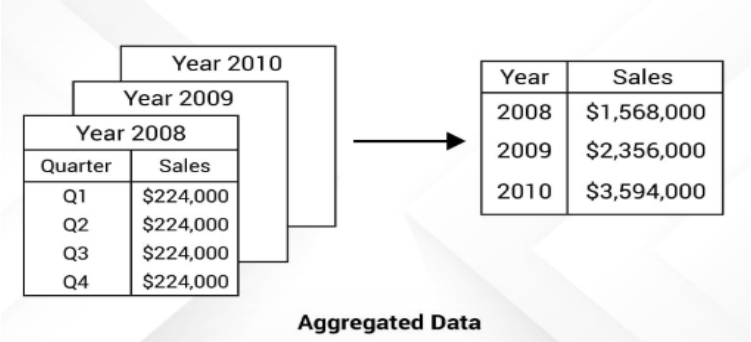




A) Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.3-dimensional data in a tabular structure, representing this data in a cube format increases the readability of the data:

Each side of the cube represents one dimension- Time, Location, and Item Type (Mouse mobile modem).

Another usability of the data cube aggregation in data mining is when we want to aggregate the data values. In the example below, we have quarterly sales data for different years from 2008 to 2010



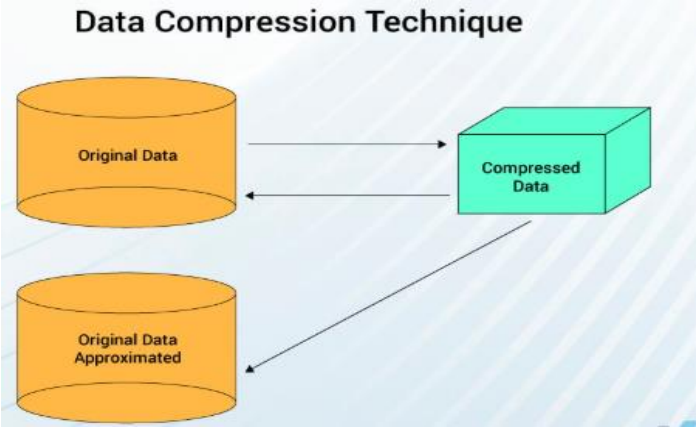
B) Data Compression

Data compression in data mining as the name suggests simply compresses the data. This technique encapsulates the data or information into a condensed form by eliminating duplicate, not needed information. It changes the structure of the data without taking much space and is represented in a binary form.

There are two types of data compression:

1. Lossless Compression: When the compressed data can be restored or reconstructed back to its original form without the loss of any information then it is referred to as lossless compression.

2. Lossy Compression: When the compressed data cannot be restored or reconstructed back into its original form then referred to as Lossy compression.



Data compression technique varies based on the type of data.

String data: In string compression, the data is modified in a limited manner without complete expansion; hence the string is mostly lossless as the data can be retrieved back to its original form. Therefore it is lossless data compression. There are extensive theories and well-tuned algorithms that are used for data compression.

Audio or video data: Unlike string data, audio or video data cannot be recreated to its original shape, hence is lossy data compression. At times, it may be possible to reconstruct small bits or pieces of the signal data but you cannot restore it to its whole form.

Time Sequential data: The time-sequential data is not audio data. It is by large, usually short data fragments and it varies slowly with time as is used for data compression.

Data Compression Methodologies

There are two methodologies:

1. Dimensionality
2. Numerosity reduction

1.Dimensionality Reduction

Dimensions are also known as features or attributes . These dimensions are nothing else but properties of the data, i.e., describing what the data is about.

For instance, we have the data of employees of a company. We have their name, age, gender, location, education, and income. All these variables do nothing but help us to *know, understand and describe the data point*.

As the features increase, the sparsity of the dataset also increases. The sparsity indicates that there is a relatively higher percentage of the variables that do not contain actual data. These "empty" cells or NA values take up unnecessary storage.

Dimensionality Reduction in data mining is the process of reducing the data by removing these features from the data. There are three techniques for this:

- **Wavelet Transformation**

  Wavelet Transform in Data Mining is a form of lossy data compression.

  Let's say we have a data vector Y, by applying the wavelet transform on this vector Y, we would receive a different numerical data vector Y', where the length of both the vectors Y and Y' are the same. Now, you may be wondering how transforming Y into Y' helps us to reduce the data. This Y' data can be trimmed or truncated whereas the actual vector Y cannot be compressed.

  Let's say we have a data vector Y. When we apply wavelet transformation on this vector Y, we get a different numerical data vector Y' where the length of both the vectors Y and Y' are the same now.

  The reason it is called 'wavelet transform' is that the information here is present in the form of waves, like how a frequency is depicted graphically as signals. The wavelet transform

also has efficiency for data cubes, sparse or skewed data. It is mostly applicable for image compression and for signal processing.

- **Principal Component Analysis**

  (e.g., Principal Component Analysis): A technique that reduces the number of variables in a dataset while retaining its essential information.

  Principal component analysis – PCA in data mining, a technique for data reduction in data mining, groups the important variables into a component taking the maximum information present within the data and discards the other, not important variables.

  Now, let's say out of total n variables, k are such variables that are identified and are part of this new component. This component is now what is representative of the data and used for further analysis.

  In short, PCA in data mining is applied to reducing multi-dimensional data into lower-dimensional data. This is done by eliminating variables containing the same information as provided by other variables and combining the relevant variables into components. The principal component analysis is also useful for sparse, and skewed data.

- **Feature Selection or Attribute Subset Selection**

  The attribute subset selection or feature selection method decreases the data volume by removing unnecessary variables. Hence, the name feature selection. This is done in such a way that the probability distribution of the reduced data is similar to that of the actual data, given the original variables.

2. Numerosity Reduction: Reducing the number of data points by methods like sampling to simplify the dataset without losing critical patterns.

Another methodology in data reduction in data mining is numerosity reduction in which the volume of the data is reduced by representing it in a lower format. There are two types of this technique: parametric and non-parametric numerosity reduction.

1. Parametric Reduction

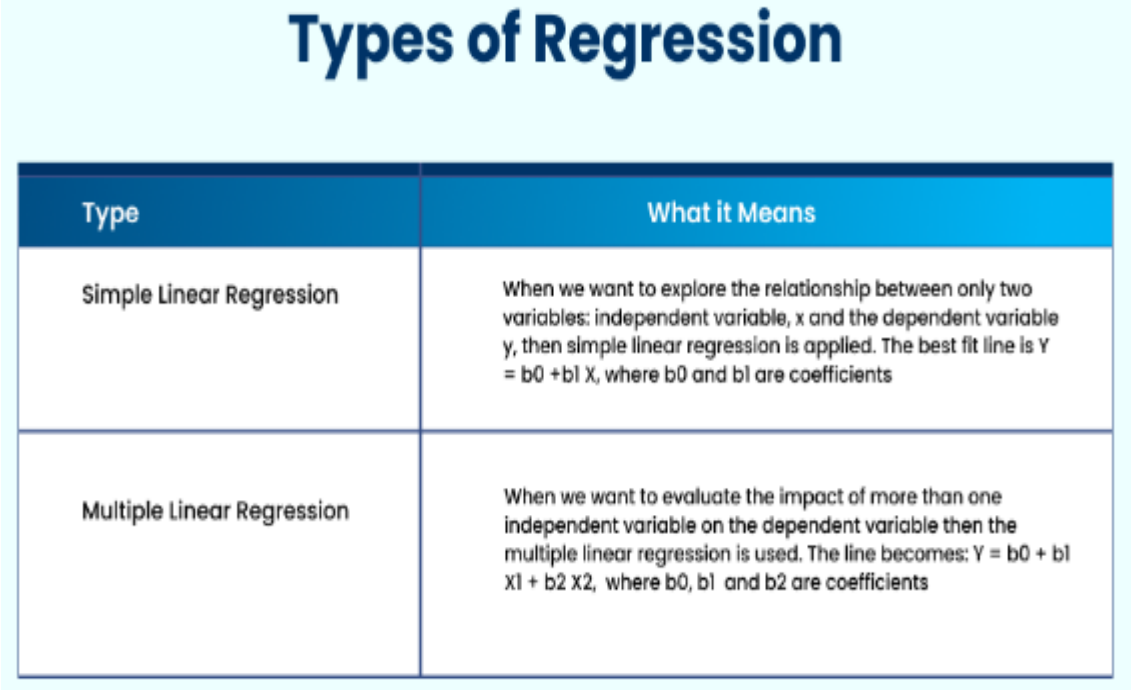
The parametric numerosity reduction technique holds an assumption that the data fits into the model. Hence, it estimates the model parameters, and stores only these estimated parameters,

and not the original or the actual data. The other data is discarded, leaving out the potential outliers.

The ways to perform parametric numerosity reduction are: Regression and Log-Linear. Both the parametric methods of regression and log-linear methods are applicable for sparse and skewed data.

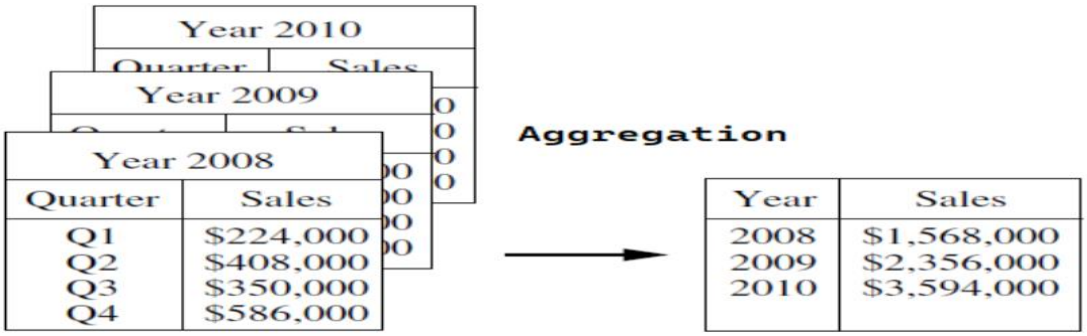
Regression: Linear Regression analysis is used for studying or summarizing the relationship between variables that are linearly related. The regression is also of two kinds:

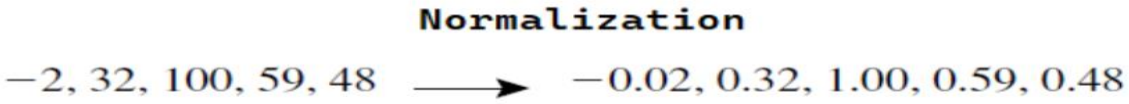
Simple Linear regression and Multiple Linear regression.



4. Data Transformation:

Data transformation in data mining refers to the process of converting raw data into a format that is suitable for **analysis and modelling**. 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. **Smoothing:** It is a process that is used to remove noise from the dataset using techniques include binning, regression, and clustering.

2. **Attribute construction (or feature construction):** In this, new attributes are constructed and added from the given set of attributes to help the mining process.

3. **Aggregation:** In this, summary or aggregation operations are applied to the data. **For example**, the daily sales data may be aggregated to compute **monthly and annual total amounts.**

4. **Data normalization**: This process involves converting all data variables into a small range. such as -1.0 to 1.0, or 0.0 to 1.0.

5. **Generalization:** It converts low-level data attributes to high-level data attributes using concept hierarchy. For Example, Age initially in Numerical form (22,) is converted into categorical value (young, old).

Method Name	**Irregularity**	**Output**
Data Cleaning	Missing, Nosie, and Inconsistent data	Quality Data before Integration
Data Integration	Different data sources (data cubes, databases, or flat files)	Unified view

Method Name	Irregularity	Output
Data Reduction	Huge amounts of data can take a long time, making such analysis impractical or infeasible.	Reduce the size of a dataset and maintains the integrity.
Data Transformation	Raw data	Prepare the data for data mining

4.1 Data Mining Primitives

A data mining query is defined in terms of the following primitives

Task-relevant data: "This is the database portion to be investigated. For example, suppose that" you are a manager of All Electronics in charge of sales in the United States and Canada. In particular, you would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes

The kinds of knowledge to be mined: "This specifies the data mining functions to be performed," such as characterization, discrimination, association, classification, clustering, or evolution analysis. For instance, if studying the buying habits of customers in Canada, you may choose to mine associations between customer profiles and the items that these customers like to buy.

Background knowledge:"Users can specify background knowledge, or knowledge about the"domain to be mined. This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found. There are several kinds of background knowledge.

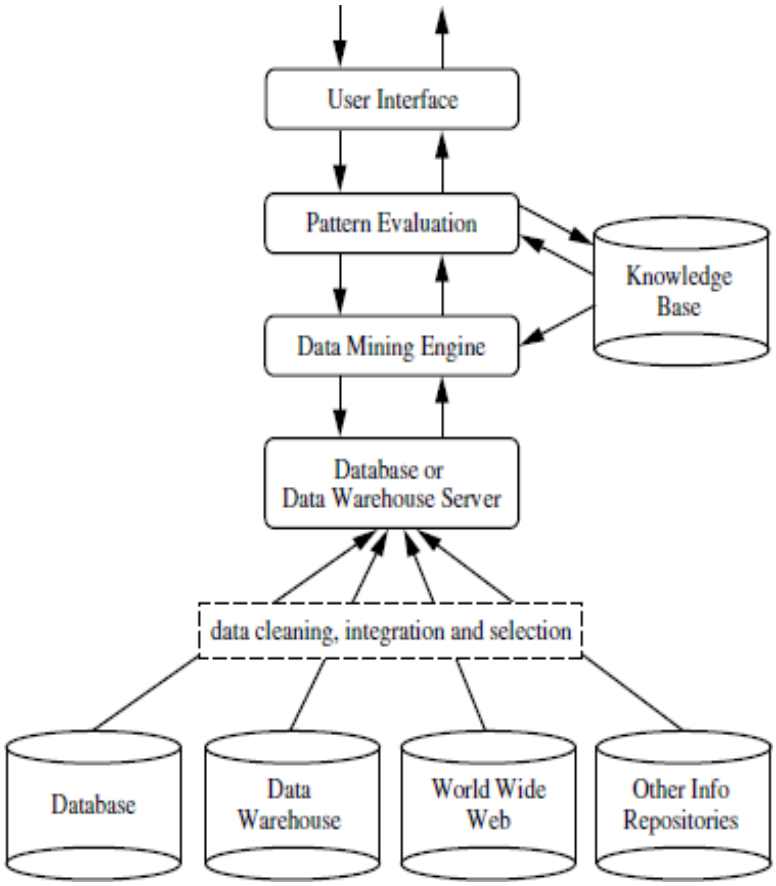
Interestingness measures:"These functions are used to separate uninteresting patterns from"knowledge. They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures.

Presentation and visualization of discovered patterns: "This refers to the form in which" discovered patterns are to be displayed. Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

Task-Relevant Data

Architecture of DataMining

A typical datamining system may have the following major components.



KnowledgeBase:

This is the domain knowledge that is used to guide the search or evaluate 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).

1. *DataMining Engine:*
   This is essential to the data mining systemand ideally consists ofa set of functional modules for tasks such as characterization, association and correlationanalysis, classification,

prediction, cluster analysis, outlier analysis, and evolutionanalysis.

2. *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.
3. *Userinterface:*
   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 Discretization

- Dividing the range of a continuous attribute into intervals.
- Interval labels can then be used to replace actual data values.
- Reduce the number of values for a given continuous attribute.
- Some classification algorithms only accept categorically attributes.
- This leads to a concise, easy-to-use, knowledge-level representation of mining results.
- Discretization techniques can be categorized based on whether it uses class information or not such as follows:
  - ***Supervised Discretization*** **-** This discretization process uses **class** information.
  - ***Unsupervised Discretization*** **-** This discretization process does **not use class** information.

- Discretization techniques can be categorized based on which direction it proceeds as follows:

Top-down Discretization -

- If the process starts by first finding one or a few points called split points or cut points to split the entire attribute range and then repeat this recursively on the resulting intervals.

Bottom-up Discretization -

- Starts by considering all of the continuous values as potential split-points.
- Removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals.

Concept Hierarchies

- Discretization can be performed rapidly on an attribute to provide a hierarchical partitioning of the attribute values, known as a ***Concept Hierarchy.***
- Concept hierarchies can be used to reduce the data by collecting and replacing low-level concepts with higher-level concepts.
- In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies.
- This organization provides users with the flexibility to view data from different perspectives.
- Data mining on a reduced data set means fewer input and output operations and is more efficient than mining on a larger data set.
- Because of these benefits, discretization techniques and concept hierarchies are typically applied before data mining, rather than during mining.

Typical Methods of Discretization and Concept Hierarchy Generation for Numerical Data

1] Binning

- Binning is a top-down splitting technique based on a specified number of bins.

- Binning is an unsupervised discretization technique because it does not use class information.
- In this, The sorted values are distributed into several buckets or bins and then replaced with each bin value by the bin mean or median.
- It is further classified into
    - ***Equal-width (distance) partitioning***
    - ***Equal-depth (frequency) partitioning***

2] Histogram Analysis

- It is an unsupervised discretization technique because histogram analysis does not use class information.
- Histograms partition the values for an attribute into disjoint ranges called buckets.
- It is also further classified into
    - ***Equal-width histogram***
    - ***Equal frequency histogram***
- The histogram analysis algorithm can be applied recursively to each partition to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre-specified number of concept levels has been reached.

3] Cluster Analysis

- Cluster analysis is a popular data discretization method.
- A clustering algorithm can be applied to discretize a numerical attribute of A by partitioning the values of A into clusters or groups.
- Clustering considers the distribution of A, as well as the closeness of data points, and therefore can produce high-quality discretization results.
- Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy.

4] Entropy-Based Discretization

- Entropy-based discretization is a supervised, top-down splitting technique.

- It explores class distribution information in its calculation and determination of split points.
- Let D consist of data instances defined by a set of attributes and a class-label attribute.
- The class-label attribute provides the class information per instance.
- In this, the interval boundaries or split-points defined may help to improve classification accuracy.
- The entropy and information gain measures are used for decision tree induction.

5] Interval Merge by χ2 Analysis

- It is a bottom-up method.
- Find the best neighboring intervals and merge them to form larger intervals recursively.
- The method is supervised in that it uses class information.
- Chi Merge treats intervals as discrete categories.
- The basic notion is that for accurate discretization, the relative class frequencies should be fairly consistent within an interval.
- Therefore, if two adjacent intervals have a very similar distribution of classes, then the intervals can be merged.
- Otherwise, they should remain separate.

(Refer the Assignment 3 :Data Mining Primitives - Languages and System Architectures: Task relevant data - Kind of Knowledge to be mined - Discretization and Concept Hierarchy.)