**UNIT - II**

**Data Preprocessing:** Data cleaning, Data Integration & Transformation, Data Reduction, Discritization & Concept Hierarchy Generation, Data Mining Primitives.

**Objectives:**

To conceptualize the need of pre-processing.

To identify the interesting patterns and association rules in large databases

**UNIT 2**

**Data Preprocessing**

Real world data bases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size. Data cleaning can be applied to the remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store such as data warehouse or data cube. Data transformation such as normalization applied to improve the accuracy and efficiency of mining algorithms. Data reduction can reduce the data size by aggregating, eliminating redundant features.

Noise in the data is possible with following reasons:

1. Data collection instruments used may be faulty.
2. Human or computer errors occurring at data mining.
3. Errors in data transmission.
4. Technology limitations such as limited the buffer size.

**Data cleaning:**

Real world data tend to be incomplete, noisy and inconsistent data cleaning routines attempt to fill in missing values. Smooth out noise while identifying outliers, and correct inconsistencies in the data

Filling in the missing values of an attribute:

1. Ignore the tuples: This is done when the class label is missing it is not very efficient.
2. Fill in the missing value manually: It is a time consuming and may not be feasible with many missing values in large data set.
3. Use a global constant to fill the missing value: Replace all missing attribute values by the same constant. It is a simple method but not recommendable.
4. Use the attribute mean to fill in the missing value: Use the average value to fill the missing value of an attitude
5. Use the attribute mean for all samples belonging to the same class as a given tuple: Use the average value of all same class belonging to the same class of the missing value category
6. Use the most probable value to the messenger decoration inference based tools on decision tree induction are used to predict the missing values

**Noisy Data:** Noise is random error or variance in a measured variable.

*Binning:* Binning methods smooth sorted data value by consulting its neighborhood. i.e., the values around it. Sorted values are distributed into the number of buckets or the bins.

*Smoothing:* In smoothing by bin means, each value in bin is replaced by the mean value of the bin. In smoothing by bin medians, each bin value is replaced by median value. In smoothing by bin boundaries, each bin value is replaced by the closest boundary value.

Sorted data: 4,8,15,21,21,24,25,28,34

|  |  |  |
| --- | --- | --- |
| Partition into bins | Smooth by bin means | Smooth by bin boundaries |
| Bin 1: 4,8,15 | Bin1: 9,9,9 | Bin1: 4,4,15 |
| Bin 2 : 21,21,24 | Bin 2 : 22,22,22 | Bin 2 : 21,21 24 |
| Bin 3: 25,28,34 | Bin 3 : 29,29,29 | Bin 3 : 25,25,34 |

**Clustering:** Outliers may be detected by clustering, where similar values are organized into groups, values that fall outside of the set of clusters may be considered as outliers.

Combine computer and human inspection: outliers may be identified through combination of computer and human inspection. Outlier patterns may be informative or garbage. Human can sort the patterns in the list to identify the actual garbage ones.

**Regression:** Data can be smoothed by fitting the data to a function called regression. Linear regression involves finding the best line to fit two variables, so that one variable used to predict the other. Multiple linear regressions is an extension of linear regressions where more than two variables are involved and the data are fit to a multidimensional.

**Inconsistent data**: Inconsistency may be corrected manually using external reference. Knowledge engineering tools may also be used to detect the violation of known data constraints.

**2.2 Data Integration and transformation**: Data integration combines data from multiple sources into coherent data store, as in data warehousing. Schema integration define how the real world entities from multiple data sources be matched up. This is referred as Entity Identification Problem.

Redundancy can be detected using correlation analysis. The correlation between attributes A and B can be measured by

n is the number of tuples, and are respective mean values of A and B. are respective standard deviations of A and B. If the resulting value is greater than 0 then A and B are positively correlated, means that the value of A increases as the values of the B increases

If the resulting value is equal to 0 then A and B are independent and there is no correlation between them. If the resulting value is less than 0, then A and B are negatively correlated with the value of one attribute increases Where the value of other attributes decreases.

**2.3 Data transformation:**

The data are transformed or consolidated into forms which are appropriate for mining.

*Smoothing:* Used to remove noise from data. Techniques like the binning, clustering and regression are used.

*Aggregation:* Aggregation or summarization operations are applied to the data.

*Generalization:* Where low level or primitive data is replaced by higher level concepts through use of concept hierarchies.

*Normalization:* Where the attribute data are scaled so as to fall within a small specified range such as 0 to 1

*Min Max Normalization:* Performs linear transform on the originals data. MinA and MaxA are the minimum and maximum values of an attribute A. The normalization maps a value V of A to of new range [new\_minA, new\_maxA].

It encounters out of bounds error if a future input falls outside original data range.

**Problem**: Min and Max values of attribute income are Rs 12000/- and Rs 98000/- respectively. Map income to the range [0,1]. Find the transformed value of Rs 73600/-

*Z- Score Normalization (Zero – Mean):* The values of an attribute are normalized based on mean and standard deviation.

This method is useful when actual minimum and maximum are not known. This technique is used to handle outliers.

**Problem**: Mean and standard deviation of the values of income are Rs54000/- and Rs 16000/- respectively. With Z-score normalization a value of Rs 73,600/- is transformed to

Decimal scaling normalizes the value by moving the decimal point of attribute A. The number of decimal points moved depends on the maximum absolute value of A.

Where j is the smallest integer for which max (|) <1

**2.4 Data Reduction**

Data reduction techniques are applied to obtain a reduced representation of data which is smaller in volume yet closely maintains the integrity of the original data.

*Data Cube Aggregation:* Data Cube store multidimensional aggregated information. Data cubes fast access to pre computing summarized data there by benefiting online analytical processing. The cube created the lowest level of abstraction is referred to as the base cuboids. A cube for the highest level of abstraction is called apex cuboids.

*Dimensionality Reduction:* Dimensionality reduction reduces the data set by removing such attribute from it. The goal of attribute subset selection is to find

a minimum set of attributes such that the resulting probability distribution of data classes is close to original distribution of all attributes. Mining on reduced set of attributes reduces the number of attributes appearing in discovered patterns.

*Stepwise forward selection:* Procedure starts with empty set. The best of the original attributes is determined and added to the set. At each iteration the best of the reminder original attributes is added to the set.

Stepwise backward elimination: Procedure starts with full set of attributes. At each iteration, it removes worst attribute remain in the set.

Combination of forward selection and backward elimination: Forward selection and backward elimination methods can be combined so that, at each step the procedure selects best attribute and worst attributes are removed.

Data compression: Data encoding or transformations are applied so as to obtain a reduced or compress representation of original data. If original data can be reconstructed from compressed data without loss of information called lossless compression. Approximation of original data is reconstructed then it is called lossy.

Wavelet transforms: Discrete wavelet transform is a linear signal processing technique, applied to a date vector of D, transforms it to a numerically different vector of wavelet coefficient. L is the length of the data vector and it must be an integer’s power of 2, if necessary padding the data vector with zeroes.

Data smoothing such as sum or weighted average and then weighted difference is applied to bring out detailed features of the data. Two functions are applied to pairs of input data, results two sets of data of length L/2. Represents low frequency version of data and high frequency version of data. Two functions are recursively applied until dataset obtained of length 2. The selection of values is designated as wavelet coefficients.

**Principal Component Analysis:** PCA Searches for c k-dimensional orthogonal vectors that can best be used to represent the data where is c<= k.

*Basic procedure of PCA is as follows:*

1. The input data are normalized, so each attribute falls within the same range.

2. PCA computes c ortho normal vectors that provide basis for normalized input data referred as principal components.

3. Principal Components are stored in decreasing order of significance; the size of data can be reduced by eliminating weaker component.

*Numerosity Reduction:* Parametric methods, a model are used to estimate the data, so that typically data parameters need to be stored, instead of actual data. Non parametric methods used for storing reduced representation of data include histograms, clustering.

*Regression and log linear models:* Regression and log linear models used to approximate the given data. In linear regression data modeled to fit a straight line. Response variable (Y) modeled as a linear function of another random variable X (Predictor) with

Y=α+βX

α and β can be called as regression coefficients.

Method of least squares minimizes the error between actual line separating the data and estimate of the line. Log linear models approximate discrete multidimensional probability distributions. This method estimate the probability of each cell in base cuboid for a set of discretized attributes based on smaller cuboids making up data cube little.

Histograms: Histograms use binning to approximate data distributions and histogram partitions the data distribution into disjoint subsets called buckets. Buckets displayed on horizontal axis, height of a bucket reflects the average frequency of the values represented by the bucket.

Histograms are defined according to partitioning rules:

Equi width: The width of each bucket range is uniform.

Equi Depth: The frequency of each bucket is constant.

V optimal: It is a histogram with least variance.

Max Diff: Bucket boundary is established between pair of pairs having β-1 largest differences.

**Clustering**: Clustering partitions the data objects into groups or clusters objects in the cluster are similar to one another and dissimilar to objects in other clusters. The quality of a cluster represented by its diameter, the maximum distance between any two objects in a cluster.

Sampling: Sampling can be used as a data reduction technique it allows a large data set to be represented by a smaller random sample of the data.

Simple random sample without replacement of size n: This is created by drawing n of the H tuples from D, where the probability of drawing any tuple is i/N, i.e., all tuples are equally likely.

Simple random sample with replacement: A tuple is drawn from D then it is recorded and then replaced, i.e., after a tuple is drawn, it is placed back in D and it may be drawn again.

Cluster sample: If the tuples in D are grouped into mutually disjoint clusters then SRS of m clusters can be obtained.

Stratified sample: D is divided into mutually disjoint parts called strata, a stratified sample is generated by obtaining SRS at each stratum.

**Discretization and concept hierarchy generation**

Discretization techniques used to reduce the number of values for given continuous attribute by dividing the range of the attribute into intervals. Interval labels are used to replace actual data values.

A concept hierarchy describes discretization of attribute. Concept hierarchies used to reduce the data by collecting and replacing low level concepts by higher level concepts.

Discretization and concept hierarchy generation for numeric data

Binning: Attribute values can be discretized by distributing the values into bins, and replacing each bin value by bin mean, median. These techniques are recursively applied to generate concept hierarchy.

Histogram analysis: Histogram analysis applied recursively to each partition in order to generate multi level concept hierarchy procedure terminating once a prespecified number of concept levels has been reached.

Cluster analysis: A clustering algorithm partition the data into clusters. Each cluster forms a node of a concept hierarchy, where all nodes are at the same conceptual level. Each cluster may be decomposed into sub clusters to form a lower level of hierarchy.

Entropy based Discretization: Entropy recursively partition the values of a numeric attribute to result a hierarchical discretization.

1. A value of an attribute partition the samples into two subsets A< V & A>=V, called binary discretization.

2. The threshold value maximizes the information gain result from subsequent partitioning.

Where S1 and S2 correspond to samples in s satisfying the conditions A<T and A>=T respectively.

Ent(S1) = -

Pi is the probability of class i in S1.

The process of determining is threshold value is recursively applied to each partition until stopping criterion is met.

Ent(S)-I(S,T)>δ

Segmentation by Natural Partitioning:

3-4-5 rule used to segment numeric data into relatively uniform natural intervals. If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit then partition the range into 4 equal width intervals. If it covers 2, 4 or 8 distinct values at the most significant digit then petition the range into 5 equi width intervals.

**Concept hierarchy generation for categorical data**: Categorical attributes have a finite number of distinct values with no ordering among the values. Specification of a partial ordering of an attributes explicitly at the schema level by users or experts. A User or expert can easily define a concept hierarchy by specifying partial or total ordering of attributes at schema level.

Street < city < state < country

**Specification of a petition a hierarchy by explicit data grouping:** it is needed to specify explicit groups for a small partition of intermediate level data

{Hyderabad, Vizag} <Andhra Pradesh.

**Specification of a set of attributes but not of their partial ordering**: A user may specify a set of attributes forming a concept hierarchy but omit to explicitly state their partial ordering.

**Specification of only a partial set of attributes**: Users sometimes have a vague idea about the attitudes and include a small set of attributes to specify a concept hierarchy. The specification of one attribute may trigger a whole group of semantically tightly linked attributes to be dragged in to form a complete hierarchy.

**SUMMARY**

Data preparation is an important issue for both data warehousing and data mining, as real-world data tends to be incomplete, noisy, and inconsistent. Data preparation includes data cleaning, data integration, data transformation, and data reduction.

Data cleaning routines can be used toll in missing values, smooth noisy data, identify outliers, and correct data inconsistencies.

Data integration combines data from multiples sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration.

Data transformation routines confirm the data into appropriate forms for mining. For example, attribute data may be normalized so as to fall between a small range, such as 0 to 1.0.

Data reduction techniques such as data cube aggregation, dimension reduction, data compression, Numerosity reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of information content.

Concept hierarchies organize the values of attributes or dimensions into gradual levels of abstraction. They are a form a discretization that is particularly useful in multilevel mining.

Automatic generation of concept hierarchies for categoric data may be based on the number of distinct values of the attributes defining the hierarchy. For numeric data, techniques such as data segmentation by partition rules, histogram analysis, and clustering analysis can be used.