UNIT-II

INTRODUCTION to Data Mining

Introduction:-

Data mining is one of the most useful techniques that helps entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called as Knowledge Discovery in Database (KDD). The knowledge discovery process includes Data cleaning, Data integration, Data Selection, Data transformation, Data mining, Pattern evaluation and knowledge Presentation.

What is data mining?

The process of extracting information to identify patterns, trends and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data mining.

Data mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data which is collected and assembled in particular areas such as data warehouses, efficient data mining algorithm, helping decision making and often data requirement to eventually cost-cutting and generating revenue.

Data mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information

KDD Process:-

1. Data cleaning - to remove noise and inconsistent data
2. Data integration - where multiple data sources may be combined.
3. Data selection - where data relavent to the analysis task are retrieved from the database
4. Data transformation - where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining:- an essential process where intelligent methods are applied to extract data patterns
6. Pattern evaluation:- to identify the truly interesting patterns representing knowledge based on interestingness measures.
7. Knowledge presentation:- where visualization and knowledge representation techniques are used to present mined knowledge to users

Advantages of Datamining :-

→ The Datamining technique enables organizations to obtain knowledge based Data

→ Datamining enables organizations to make lucrative modification in operation and Production.

→ Compared with other statistical data applications data mining is a cost efficient.

→ Data mining helps the decision making process of an organization

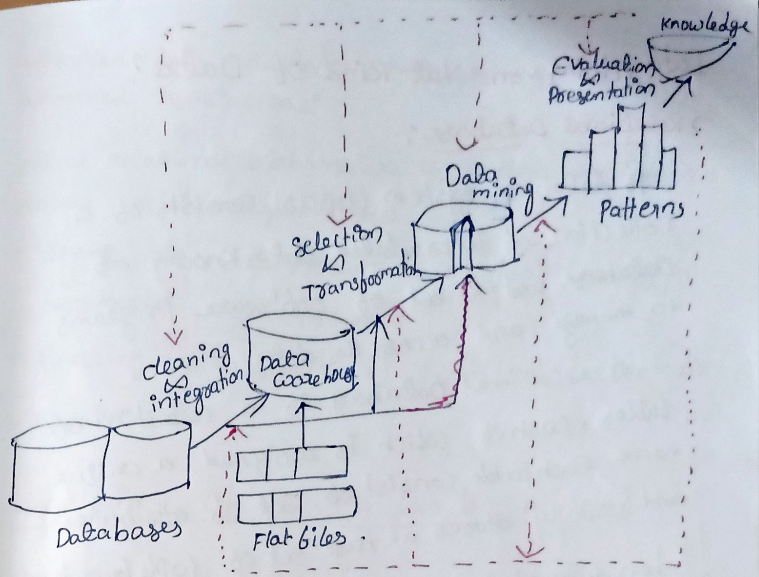
→ It is a quick process that makes it easy for new users to analyze enormous amount of data in short time.

Disadvantages :-

→ There is a probability that the organization may sell useful data of customers to other organization for money

→ Many data mining analytics software is difficult to operate and needs advance training to work on.

→ The Data mining techniques are not precise so that it may lead to severe consequences in certain conditions.



Datamining as a step in the process of knowledge Discovery.

KDD :- The term KDD stands for knowledge discovery in DataBases. It refers to the broad procedure of discovering knowledge in data and emphasize the high-level applications of specific Data mining techniques.

The main objective of KDD process is to extract information from data in the context of large databases. It does this by using data mining algorithms to identify what is deemed knowledge

Data Mining - on what kind of Data?

→ Relational Database:-

A database system (DBMS) consists of a collection of interrelated data known as Database and a set of software programs to manage and access the data.

A relational Database is a collection of tables, each of which is assigned a unique name. Each table consist of set of attributes and usually stores a large set of tuple (records or rows). When data mining is applied to relational database, we can go further by searching for trends or data patterns.

Datawarehouse:-

A datawarehouse usually modeled by a multidimensional data base structure, where each dimension corresponds to an attribute or a set of attributes in the schema.

Transactional data:-

Transactional Database consists of a file where each record represents a transaction. A transaction typically include a unique transaction identity number and a list of the items making up the transaction.

Advanced data and Information systems and Advanced Applications:-

object relational databases are constructed based on an object-relational data model.

Sequence data, data streams, spatial data (maps),
↓ stock market ↓ continuous transmit

engineering design data (IC's), hypertext, multimedia, webdata etc.

→ what kind of Patterns can be mined?

Descriptive: It describe the general properties of data

Predictive:- it makes predictions based on current data.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

→ Concept/class Description: Characterization and discrimination eg: Electronic store

→ Mining frequent patterns, associations and correlation

→ classification and prediction.

→ Cluster analysis

→ outlier Analysis - out of group.

→ Evolution Analysis. - changes over time.

Example for characterization :-

A data mining system should be able to produce a description summarizing the characteristics of customers who spend more than $1000 a year at AllElectronics store

The result should be General profile of the customers

→ 40-50 years old
→ Employed
→ Excellent credit rating

Example for Data discrimination :-

A data mining system should be able to Compare two groups of AllElectronics store customers

(i) Regular customers
(ii) Irregular customers

→ Comparative Profile of the customers
80% of the customers who frequently Purchase computer products are between 20 and 40 only.
→ Have a university education.
→ whereas 60% of the customers who infrequently buy such products are either seniors or youth and have no university degree.

Mining frequent Patterns, Associations and Correlation

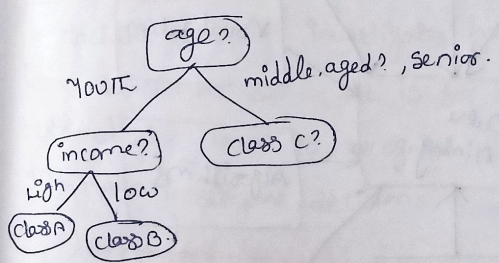
→ Frequent Patterns as the name suggest are patterns that occur frequently data

buys (x ; " computer ")) buys (" software ")

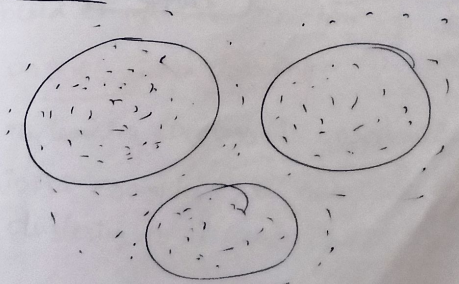
Classification & Prediction :-

Classification is the Process of finding a model that describes and distinguishes data classes or Concepts for the Purpose of being able to use the model to Predict

the class of objects whose class label is unknown

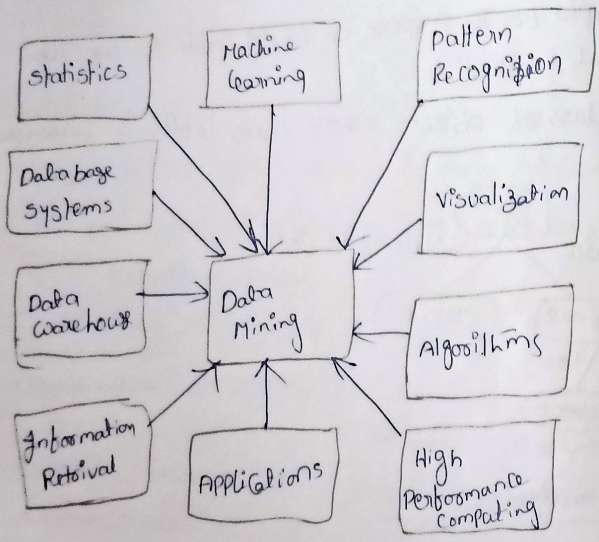


Cluster analysis :-



Which technologies Are used?:-

Data mining has incorporated many techniques from other domains such as statistics, machine learning, pattern recognition, database & Datawarehouse systems, information retrival, visualization, algorithms, high performance computing & many application Domains



statistics:-

statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics.

A statistical model is a set of mathematical functions that describe the behaviour of the objects in a target class in terms of random variable and their associated probability distributions. statistical models are widely used to model data and data classes.

Machine learning:-

Machine learning investigates how Computers Can learn based on data. ML is for Computer Programs to automatically learn to recognize complex patterns and make intelligent decisions based on data.

a) supervised learning:

The supervision in the learning comes from the labeled examples in the training dataset.

b) unsupervised learning:

The learning process is unsupervised since the input examples are not class labeled. We may have clustering to discover classes within the data.

c) Semi-supervised learning :-

using both labeled & unlabeled (half, half)

d) Active learning :

The algorithm is designed in such a way that the desired output should be decided by the algorithm itself

Algorithm :-

An algorithm can be written for any application in data mining

Application :-

Applications can be built & used as a Technology

Pattern Recognation :-

By observing various patterns we can perform datamining.

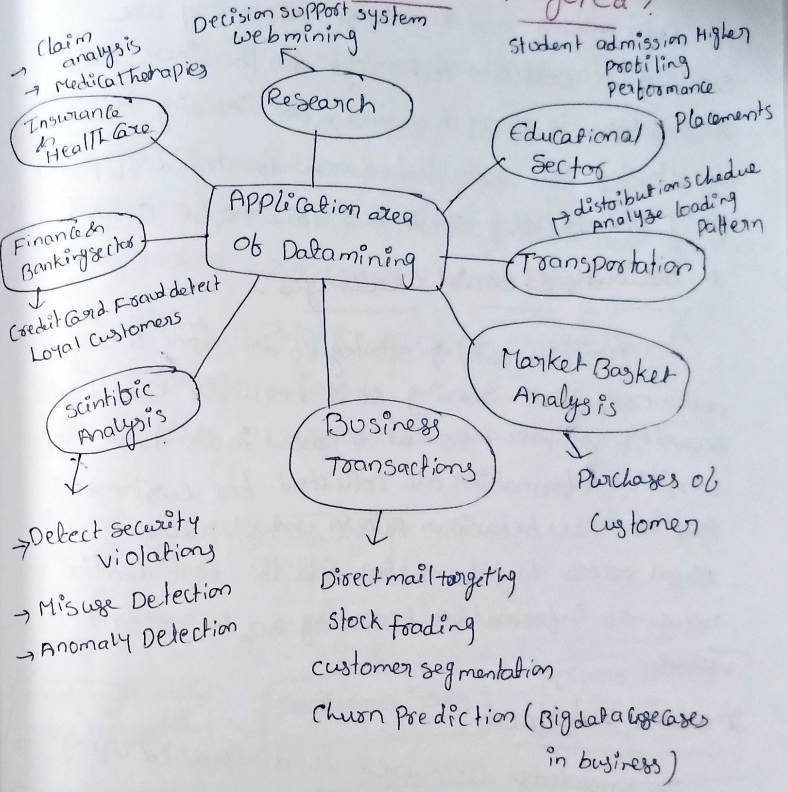
Visualization :-

To depending analysis of data by Putting into graph we do visualization.

Database technology :

using database SQL we extract data from database

which kind of applications Targeted ?



Challenges of Datamining :-

Data mining and knowledge discovery are evolving a crucial technology for business and researchers in many domains. Datamining is developing into established and trusted disciplaine. many still pending challenges have to be solved.

1. Security & social Challenges :-

Decision making strategies are done through data collection - sharing, so it requires considerable security. Private information about individual 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 :-

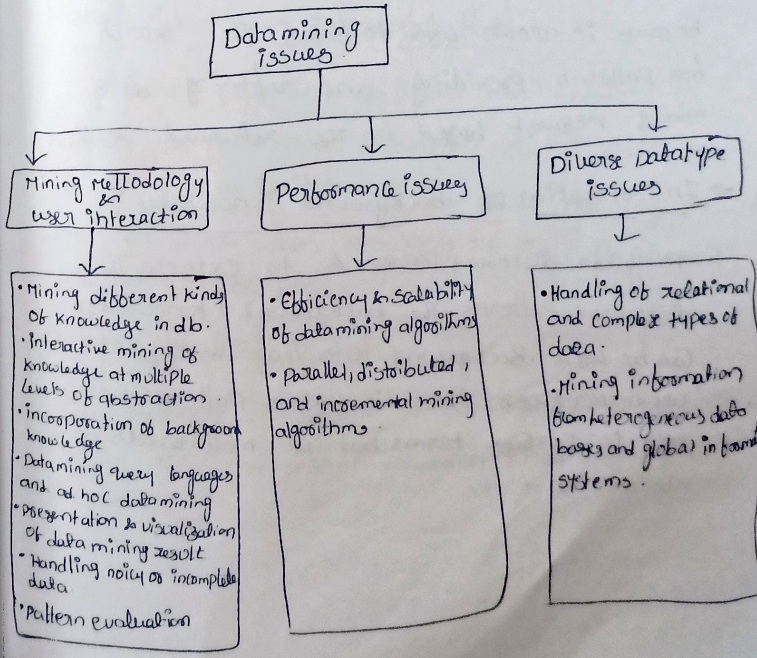
The knowledge discovered is discovered using datamining tools is useful only if it is interesting and above all understandable by the user.

From good visualization interpretation of data, mining results can be eased and helps better understand their requirements

Challenges & issues in Datamining :-

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.

- → Mining Methodology & User interaction
- → Performance issues
- → Diverse Datatypes issues.



Mining Methodology & User Interaction issues:

It refers following kinds of issues.

→ 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 Datamining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining request based on the returned result

→ Incorporation of background knowledge:-

To guide discovery process & 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 multilevels of abstraction.

Datamining query languages and adhoc datamining:

Datamining query language that allows the user to describe adhoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient & flexible Data mining.

Presentation & visualization of datamining results:

once the patterns are discovered it needs to be expressed in high level languages, and visual representation. 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 accuracy of the discovered pattern will be poor.

Pattern evaluation:-

The pattern discovered should be interesting because either they represent common knowledge or lack novelty.

Performance issues :-

There can be performance-related issues :

Efficiency & 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 result from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Datatype issues :-

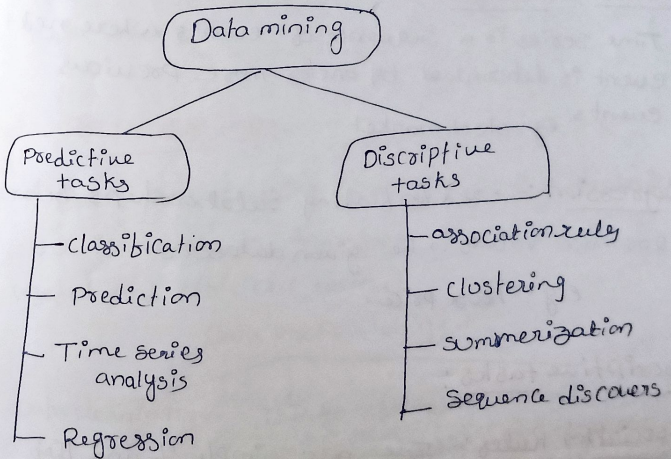
Handling of relational and complex types of data :-

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 heterogenious databases and global information systems :-

The data is available at different data sources on LAN or WAN. These data sources may be structured semi structured or unstructured, therefore mining the knowledge from them adds challenges to data mining.

Data Mining Tasks :-



Prediction :- Prediction uses same variables or field in the database to predict unknown or future values of other variables.

Discription :- Characterize the general properties of data and used for finding patterns that describes the data

Classification:-

used to classify data into one of several predefined classes. eg: Loan < accept / Reject > Based on data

Prediction:- Predictive task come up with a model from the available data set that is helpful in predicting unknown or future values of another dataset

eg: Based on medical reports Doctor predict the diseases

Time series analysis:-

Time series is a sequence of events where next event is determined by one or more previous events eg: stock market

Regression:- used for finding relationships between 2 or more values in the given datasets

e.g:- house price.

Discriptive tasks:

Association Rules:- These are simple if and then statements that help to discover relationships between datasets

e.g: if person buy milk then he buy buy bread

Clustering:- It is a divison of information into groups of connected objects. Cluster is used to identify similar data.

e.g: Library books

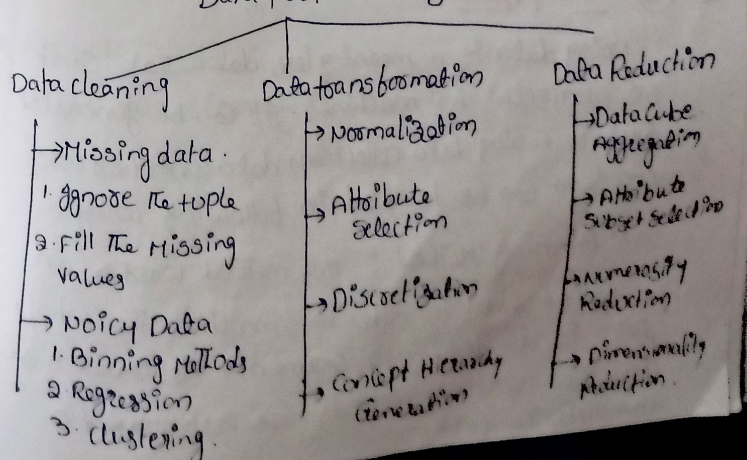
summerization:- Used for finding compact description for subset of data.

sequence discovery:

Helps to discovery similar patterns in transactional data over time.

Data Preprocessing:-

Data preprocessing is a datamining technique which is used to transform the raw data in a useful and efficient format.



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.

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.

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.

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 "city" can be converted to "country"

(3) Data Reduction:

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 use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps of 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 relevent 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 eg: 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.

Two effective methods of dimensionality reduction are:

→ wavelet transforms &

→ PCA (Principal Component Analysis)

Data Cleaning:

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

→ Missing values:

i) Ignore tuple

which row having more missing values we can ignore that.

2) Fill in the missing values manually

3) Use a global constant to fill in the missing values (e.g. NA)

4) Use a measure of central tendency for the attribute to fill the missing value

5) Use the most probable value to fill in the missing value (e.g. decision tree)

S.No	Name	occupation	Branch	Date	Price	Address	Pin code
1	Ramesh	govt	TPG	10 May		TPG	524121
2	Suresh	Self	JNK	11 May	2500	JNK	
3	Prasad	Private	TPG	11-May	500	TPG	534101
4	Govind	Private	VJWD	12 May	600	VJWD	520010
5	Ravi						
6	Naresh	busines	VJWD	14 May	1400	VJWD	
7	Bhuvan	busines	TPG	14-May	2300	TPG	
8	Kiran	govt	JNK			JNK	534221
9		Self	AJY	15 May	1100	AJY	533101
10	Yogi	business	JNK	16-May	1800	JNK	534221

→ Noisy Data:

Noise is a random error or variance in a measured variable

Approaches in Noisy data:

(i) Binning

(ii) Regression

(iii) outlier analysis

→ Binning:

a) Partition into equal frequency bins

b) smoothing by bin means

c) smoothing by bin boundries.

eg: 6, 10, 17, 22, 22, 25, 27, 30, 36

Partition into equal frequency bins

Bin 1: 6, 10, 17

Bin 2: 22, 22, 25

Bin 3: 27, 30, 36

$$m = \frac{\text{Sum of the terms}}{\text{no: of terms}}$$

Smoothing by bin means finding mean value

Bin 1: 11, 11, 11

Bin 2: 23, 23, 23

Bin 3: 31, 31, 31

Smoothing by bin boundaries

~~smoothing by~~

Bin1 : 6, 6, 17

Bin2 : 22, 22, 25

Bin3 : 27, 27, 36

→ Regression :-

Linear regression involves finding the "best" line to fit two attributes so that one attribute can be used to predict the other.

Multiple Linear regression is an extension of Linear regression where more than two attributes are involved and the data are fit to multidimensional surface.

→ outlier analysis :-

outlier may be detected by clustering for example, where similar values are ~~g~~ organized into groups, or 'clusters'. Intuitively, values that fall outside of the set of clusters may be considered outliers.

Data Transformation :-

It is a data preprocessing technique that transforms (or) consolidate the data into alternate forms appropriate for mining.

(i) smoothing :- Remove the noise from data [binning, regression, clustering]

(ii) Aggregation :- summary or aggregate function. It is used to constructing a data cube

(iii) Generalization :- low level concepts are replaced with high level. eg: city to country

(iv) normalization :-

attribute values are normalized by scaling their values so that they fall in specified range

eg: {2, 40, 500, 1, 3, 900}

we should change these values into 0-1 range

That means we can normalize the attribute values [0.0 : 1.0]

In this we have two types

(i) Min-Max normalization :

$$v' = \frac{v - min_A}{max_A - min_A} \times (new\ max - new\ min) + new\ min$$

v' - new value

min_A - minimum value of attribute

max_A - maximum " " "

v - original attribute value.

(ii) z-score normalization.

$$v' = \frac{v - \bar{x}}{\sigma_x}$$

$\bar{x}$ - mean of an attribute
σ_x - standard deviation
v - original value
v' - new value

e.g: min-max normalization.

8
10
15
20 } Marks. It will scale the data into 0 & 1

$$v' = \frac{v - min_n}{max_n - min_n}$$

In the given attribute $min_n = 8$
$max_n = 20$

(i) original value v = 8

$$v' = \frac{8-8}{20-8} = \frac{0}{12} = 0$$

v' for 8 = 0

(ii) 10

$$v' = \frac{10-8}{20-8} = \frac{2}{12} = \frac{1}{6} = 0.16$$

(iii) 15

$$v' = \frac{15-8}{20-8} = \frac{7}{12} = 0.58$$

(iii) 20

$$v' = \frac{20-8}{20-8} = \frac{12}{12} = 1$$

new value = 1

new values are

8	0	new min = 0
10	0.16	new max = 1
15	0.58	
20	1	

eg: z score normalization.

$$v' = \frac{v - \bar{x}}{\sigma_x}$$

mean of marks = total of all attributes / no. of attributes

standard deviation $= \sqrt{\frac{\Sigma (x - \bar{x})^2}{n}}$

$= \sqrt{\frac{\Sigma (\text{every individual marks} - \text{mean of marks})^2}{n}}$

=

8
10
15
20
} marks

mean of marks = $8+10+15+20/4 = 13.25$

standard deviation = $\sqrt{\frac{(8-13.25)^2+(10-13.25)^2+(15-13.5)^2+(20-13.25)^2}{4}}$

$= \sqrt{\frac{(-5.25)^2+(-3.25)^2+(1.75)^2+(6.75)^2}{4}}$

$= \sqrt{\frac{27.56+10.56+3.06+45.56}{4}}$

$= \sqrt{\frac{86.74}{4}} = \sqrt{21.6} = 4.6$

standard deviation = 4.6

(i) $8 = \frac{8-13.25}{4.6} = -1.14$

(ii) $10 = \frac{10-13.25}{4.6} = -0.70$

(iii) $15 = \frac{15-13.25}{4.6} = 0.38$

(iv) $20 = \frac{20-13.25}{4.6} = 1.46$

Data Reduction:-

It is a PreProcessing technique that helps in obtaining reduced representation of data set from the available dataset

→ Integrity of the original data should even after reduction in data volume

→ It should produce same analytics result as an original data.

Data cube aggregation:- It is a process in which information is gathered and expressed in a summary form.

e.g.:- Year 2017

H.Y	Sales
H1	500
H2	300

Year 2018

H.Y	Sales
H1	600
H2	100

after aggregation

Year	Sales
2017	800
2018	700

Dimensionality reduction:-

It eliminates the redundant attributes which are weakly important across the Data

DOB → age

(i) stepwise Forward Selection

(ii) stepwise Backward Selection

(iii) Decision tree induction.

(i) ~~stepwise forward~~ Selection

It represents the original data in the compress or reduced form by applying data encoding (or) transformation. In this we have Lossless & Lossy.

Lossless :- if original data can be reconstructed from compressed data without loosing any information

Lossy :- if Reconstructed data is the approximation of compressed Data.

Two effective methods of dimensionality Reduction are:

→ wavelet transforms :- (DWT) - discrete wavelet transform

It is a linear signal processing technique that when applied to data vectors x transforms it numerically different vector 'x' of wavelet coefficient

→ when applying this technique we consider each tuple as n-dimensional data vector depicting n measurements made on the tuple from n different database attributes.

→ It removes noise without smoothing the data.

There are several families of DWTs, some Popular wavelet transforms include the Haar-2, Daubechies-4 and Daubechies-6. The General Procedure for applying a discrete wavelet transform uses a hierachical Pyramid algorithm that halves the data at each iteration, result is fast computational speed.

The method follows

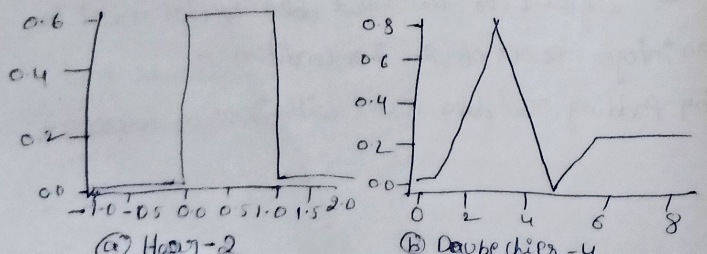
1. The length L, of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary.

2. Each transform involves applying two functions. The first applies some data smoothing, such as a sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.

3. The two functions are applied to pairs of data points in X, that is to all pairs of measurements (x_{2i}, x_{2i+1}). This results in two data sets of length $L/2$. In general these represents a smoothed or low-frequency version of the input data and the high frequency content of it.

4. The two functions are recursively applied to the data sets obtained in the previous loop, until the resulting data sets obtained are of length 2.

5. Selected values from data sets obtained in the previous iteration are designated the wavelet coefficients of the transformed data.



(a) Haar-2 (b) Daubechies-4

a matrix multiplication can be applied to the input data in order to obtain the wavelet coefficients where the matrix used depends on the given DWT.

Principal Components Analysis:

Principle components analysis as a method of dimensionality reduction.

The data to be reduced consists of tuples or data vectors described by n attributes or dimensions. PCA is also called Karhunen-Loeve, or K-L method) searches for k n-dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$. The original data are thus projected on to a much smaller space, resulting in dimensionality reduction.

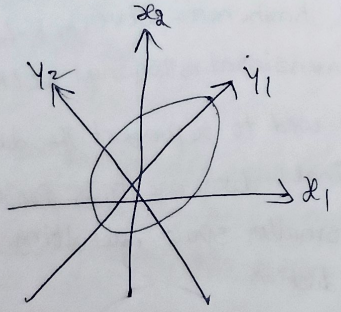
PCA combines the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set.

The basic procedure is

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

2. PCA computes K orthogonal vectors that provides a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others.
These vectors are referred to as the Principal Components.

3. The Principal components are sorted in order of decreasing "significance" (or) strength.
The Principal components essentially serves as a new set of axes for the data provide important information about variance



The sorted axes are such that the axis shows the most variance among the data, the second axis shows the next highest variance.
The First two principal components Y_1 and Y_2 for the given set of data originally mapped to the axes x_1 and x_2. This information helps identify groups or patterns with in data

4. Because the components are sorted in decreasing order of "significance," the data size can be reduced by eliminating weaker components that is those with low variance.

Feature subset selection:-

→ Another way to reduce the dimensionality is to use only a subset ~~selection~~ of the features.

→ while it might seem that such an approach would lose information, this is not the case if redundent & irrelavant features are present

→ Redundent features duplicate much or all of the information contained in one or more other ~~the~~ attributes. For example the purchase price of a product and the amount of sales tax paid contain much of the same information

→ Irrelavant features contain almost no useful information for the data mining task at hand. For instance, students' ID numbers are irrelavant to the task of predicting student's gradepoint averages.

→ Redundant and irrelavant features can reduce classification accuracy and the quality of the clusters that are found.

→ While some irrelavant and redundant attributes can be eliminated immediatly by using common sense or domain knowledge, selecting the best subset of features frequently requires a systamatic approch.

→ The ideal approch to feature selection is to try all possible subsets of features as input to the datamining algorithm of interest, and then take the subset that produces the best results. This method has the advantage of reflecting the objective and bias of the datamining algorithm that will eventually be used. Unfortunatly since the number of subsets involving n attributes is 2^n such an approach is impractical in most situations.

There are three standard approaches to feature selection: embedded, filter & wrapped.

1. Embedded approches:-

→ Feature selection occur naturally as part of the data mining algorithms, the algorithm specifically, during the operation of the data mining algorithm, the algorithm itself decides which attribute to use and which to ignore.

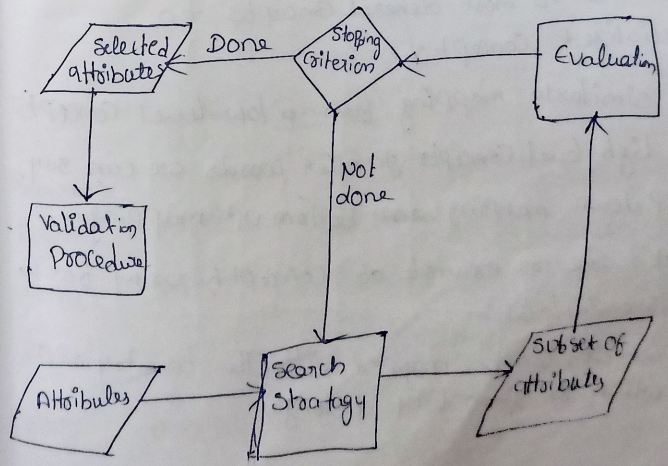
2. Filter approches:-

Features are selected before the data mining algorithm is run, using some approach that is independent of the data mining task

3. Wrapped approches:-

These methods use the target datamining algorithm as a black box to find the best subset of attributes. in a way similar to that of the ideal algorithm described above but typically with out enumarating all possible subsets

An Architecture for Feature Subset Selection:-



Discritization & Binarization:-

Discritization is used to transform the attributes that are in continuous format

Data Discritization converts a large number of data values into smaller ones, so that data evaluation and data management becomes very easy

eg: we have an attribute of age with the following values

Age: 10, 11, 13, 14, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75

Attribute	age 1	age 2	age 3
	10, 11, 13, 14, 17, 19	30, 31, 32, 38, 40, 42	70, 72, 73, 75
After Discritization	Young	Mature	old

Hierarchy Generation data discritization and Concept:-

A Concept Hierarchy represents a sequence of mapping with a set of more General Concepts to specialized Concepts:

Similarly mapping from a low-level Concepts to high-level Concepts. In other words we can say top down mapping and bottom up mapping.

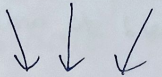
Let's see an example of Concept hierarchy for the dimension location.

Each city can be mapped with the country with which the given city belongs

For example Delhi can be mapped to INDIA and INDIA can be mapped to Asia

Top-Down Mapping:-

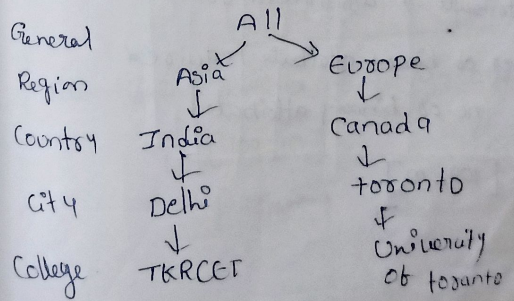
Top down mapping starts from top with General Concepts and make to the bottom to the specialized Concepts



Bottom up mapping:-

Bottom up mapping starts from Bottom with specialized Concept and move to the top to the Generalized Concepts

Concept Hierarchy Generation:-



Binarization :-

Binarization is used to transform both the discrete attributes and the continuous attributes into binary attributes in data mining with respect to feature selection.

Best Binarization approach is the one that "produces the best result for the data mining algorithm that will be used to analyze the data.

Simple techniques for binarization:

→ Assigning numerical value
→ Finding number of binary attribute require
→ Conversion into binary

eg: say there is an categorical attribute with 'm' number of values.

→ Assigning numerical value:

number assigned will be between [0, m-1]

For ordinal attribute → assignment follows order

→ Finding number of binary attribute required.

say n be the no. of binary attributes

$$n = [\log_2 m]$$

Converting the number assigned into binary value.

eg: if number of binary attribute is $n = 3$ in numbers then we can write three bit binary number,

2 - 010

if the number of binary attribute is n = 4 in numbers then

2 = 0010

eg: Let us consider an example to learn

{awful, poor, ok, good, great} - ordinal form

Attribute values	Integer value
awful	0
poor	1
ok	2
good	3
great	4

Identifying number of binary attributes;

$$n = [\log_2 m]$$

$$n = [\log_2 5] = 3$$

Binary conversion

Attribute value	Integer values	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
ok	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Binarization

If we have mutual relation overcoming the issues

No. of binary attribute = No. of values

attribute values	Integer value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
ok	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data transformation:

→ Decimal scaling:

→ It normalizes the values of an attribute by changing the position of their decimal points.

→ The no. 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

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max(|v'|) < 1$

eg: suppose values of an attribute P varies from −99 to 99

The maximum absolute value $p = 99$

For normalization the values we divide the numbers by 100 (i.e $j = 2$) (or) (no of integers in largest number so that values come out to be as 0.98, 0.97 and so on.

Binary conversion

Attribute Value	Integer values	x_1	x_2	x_3
awful	0	0	0	0
poor	1	0	0	1
ok	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Binarization

If we have mutual relation overcoming the issues

No. of binary attribute = No. of values

attribute values	Integer value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
ok	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data transformation:-

→ Decimal scaling :-

→ It normalizes the values of an attribute by changing the position of their decimal points.

→ The no. 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

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max(|v'|) < 1$

eg:- suppose values of an attribute P varies from -99 to 99

The maximum absolute value $p = 99$

For normalization the values we divide the numbers by 100 (i.e $j = 2$) (or) [no of integers in largest number so that values come out to be as 0.98, 0.97 and so on.

Measures of Similarity & Dissimilarity - Basics :-

Similarity:-

Numerical measures of how alike two data objects are is higher when objects are more alike often falls in the Range $[0, 1]$

Dissimilarity:-

Numerical measures of how different are two data objects lower when objects are more alike.
Maximum dissimilarity is often 0
Upper limit varies.

Data matrix vs Dissimilarity matrix:-

Suppose that we have n objects (eg: persons, items or courses) described by p attributes (eg: age, height, weight, or gender)

The objects are $x_1 = x_{11}, x_{12}, x_{13} \ldots x_{1p}$
$x_2 = x_{21}, x_{22}, x_{23} \ldots x_{2p}$ and soon

When x_{ij} is the value for object x_i of the j^{th} attribute.

Then Data matrix is $\begin{bmatrix} x_{11} \cdots x_{1f} \cdots x_{1p} \\ \cdots \cdots \cdots \cdots \\ x_{i1} \cdots x_{if} \cdots x_{ip} \end{bmatrix}$

Similarity / Dissimilarity for objects with single attribute

p and q are the attribute values for two data objects

Attribute type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to n-1 where n is the no. of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval (or) Ratio	$d = \lvert p-q \rvert$	$s = -d, s = \frac{1}{1+d}$ (or) $s = 1 - \frac{d - min\,d}{max\,d - min\,d}$

Dissimilarities between Data objects with multiple numeric attributes.

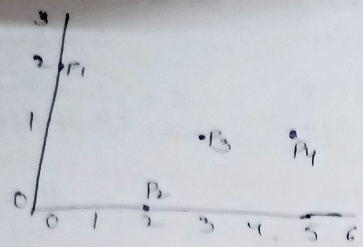
Euclidean Distance.

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where n is the no. of dimensions & p_k and q_k are respectively k^{th} attributes of data objects p and q.

→ Standardization is necessary, if scales differ.

Euclidean Distance:



point	x	y
p_1	0	2
p_2	2	0
p_3	3	1
p_4	5	1

	p_1	p_2	p_3	p_4
p_1	0	2.828	3.162	5.099
p_2	2.828	0	1.414	3.162
p_3	3.162	1.414	0	2
p_4	5.099	3.162	2	0

Distance matrix

Minkowski distance:-

Minkowski distance is a Generalization of Euclidean Distance. Given two objects p and q

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{1/r}$$

where r is a parameter, n is the number of dimensions and p_k and q_k are respectively, the k^{th} attributes of data objects p and q.

eg:

r = 1 city block (manhattan, taxicab, L_1 norm) distance

A common example of this the Hamming distance which is just the number of bits that are different between two binary vectors.

r = 2 euclidean distance

$r = \infty$, "supremum" (L_{max} norm, L_∞ norm) distance

This is the maximum difference between any attribute of the vectors

$$d(x,y) = \lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Do not confuse r with n i.e all these distances are defined for all number of dimensions

Minkowski Distance:

Point	x	y
P_1	0	2
P_2	2	0
P_3	3	1
P_4	5	1

L_1	P_1	P_2	P_3	P_4
P_1	0	4	4	6
P_2	4	0	2	4
P_3	4	2	0	2
P_4	6	4	2	0

L_2	P_1	P_2	P_3	P_4
P_1	0	2.828	3.162	5.099
P_2	2.828	0	1.414	3.162
P_3	3.162	1.414	0	2
P_4	5.099	3.162	2	0

Manhattan Distance formula:-

$$Dm(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

P_1 at (x_1, y_1) and P_2 at (x_2, y_2)

it is $|x_1 - x_2| + |y_1 - y_2|$

$= |0 - 2| + |2 - 0|$

$= 2 + 2$

$= 4$

UNIT - III

Association Rules

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable it tries to find some interesting relations or association among the variable of dataset.

It is based on different rules to discover the interesting relations between variables in the database.

- Market basket analysis is a technique used by the various big retailer to discover the associations between items. we can understand it by taking an example of a supermarket as in a supermarket all products that are purchased together are put together

For example if a customer buys bread, he most likely can also buy butter, eggs, or milk so these products are stored with in a shelf or mostly near by

Customer 1	Cust... 2	Customer 3	Customer 4
milk	mil...	milk	Sugar
bread	br...	bread	eggs
bu...	b...	butter	