mid2

Data Reduction :-

It is a technique to reduce the size of data but provides the same Analytical results if you work on original data is called as data reduction.

Advantages :-

- To improve the performance of data mining technique.
- To reduce the network bandwidth to transfer.
- Reduce the space.

Types of data Reduction :-

* Dimensionality Reduction
* Numerosity Reduction
* Data compression

→ Dimensionality Reduction :-

It is the process of reducing the number of random variables or attributes under consideration.

Methods in dimensional Reduction are :-

↳ wavelet transforms
↳ principal component Analysis (PCA) ← Transfo
original data to sample data
(Smaller space)
↳ Attribute Subset selection → irrelavant, weakly relavant or redundant attribut
(or) dimensions are deleted (or) remov

→ Numerosity Reduction :-

It is the technique to replace the original data size by alternative smaller forms of data representation.

There are two types in numerosity reduction

- parametric Numerosity Reduction
- Non parametric Numerosity Reduction

Parametric :- In this model, we used to Estimate the data so that, the required parameters needs to be stored in

memory. rather than actual data.

Techniques:- Regression
- log linear models

Non-parametric :- It is used to storing reduce representation of data

Techniques:- Histogram
- clustering
- sampling
- Data cube aggregation.

→ Data compression :-

In this, transformations are applied to obtain reduced or compressed representation of data.
It is a reduction in the no: of bits need to be represent the data.

There are two types → lossless (If the original data can be reconstructed from the compress data without any information loss)

→ lossy. (Instead of getting original data we can reconstruct only an approximate of the original data)

Techniques in loss less < String compression
All the dimensionality & Numerosity reductions.

Wavelet Transformation :-

used for data Analysis to make more accurate data. The dif b/w Fourier transformation and wavelet transformation is, The fourier transformation will not produce accurate image/data compare to the wavelet transformation

- It is not easy to work on continuous signal but in wavelet transformation signal.
- In wavelet transformation. both areas of +ve & -ve will have the common factors then we can apply the transformation only one side

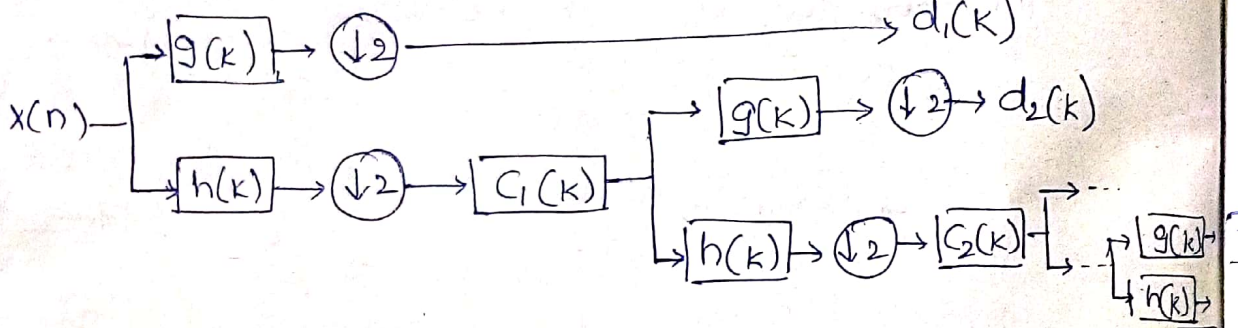
- There are 2 types of wavelet transformation
  - continuous wavelet
  - discrete wavelet

Discrete Wavelet Transformation :-

In DWT, we given a data of size N and divided into $x = (x_1, x_2, x_3 \cdots x_n)$ and each data part contains same length, we call this data parts are wavelet coefficients.

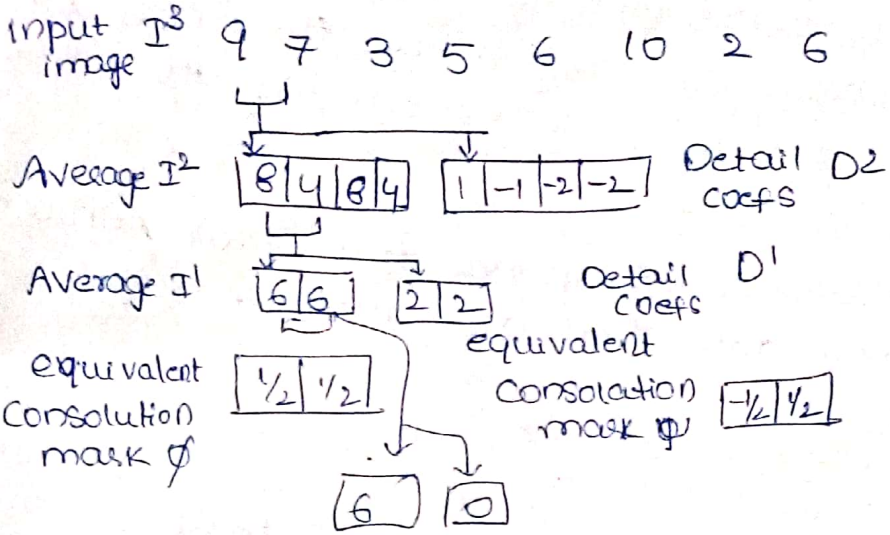
In DWT, There are many Algorithms used to Analyse the data and one among is pyramid algo.

Pyramid Algorithm :-



* The length l of input data must be an integer power (2). This condition can be padding zeros to the data, if it is not sufficient in the power(2)
* Each transformation involves applying 2 functions. The first applies some data smoothing such as sum or aggregation.
  The second performs weited diff which acts to bring out the detail features of data.
* The two functions are applied to pairs of data points in x which is the length $4/2$
* The 2 functions are recursively applied to the data sets obtained in the previous loop until the result data sets which is transformed data of the power (2)

* The selected values from the dataset obtained in previous iterations are decode the wavelength coefficie of the transfered data.



Wavelet transformed image 6 0 2 2 1 -1 -2 -2

Principal component Analysis :-

Ex:-

point	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
A	1	1	-3.17	-2.83	8.97
B	2	1	-2.17	-2.83	6.14
C	4	5	-0.17	1.17	0.19
D	5	5	0.83	1.17	0.97
E	5	6	0.83	2.17	1.80
F	8	5	3.83	1.17	4.48
					$\Sigma = 22.55$

$\bar{x} = 4.16$ $\bar{y} = \frac{23}{6} = 3.83$

5.00 − 3.83 = 1.17

6.00 − 3.83 = 2.17

$$\text{Co-variance} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

$$= \frac{22.55}{6} = 3.75$$

PCA is also called as K-L-method (Karhunen Loeve). It summarizes the data with less characteristics that represents the entire data set.

- PCA reduces the dimensionality of data containing large set of variables by transforming the initial variables into new set of variables without loosing the important information of data.
- These new variables corresponds to a linear combination of the original data and are called principal components.

Usage of PCA:-

- To identify the hidden patterns in a dataset
- To reduce the dimensionality of data by removing noise and redundancy in the data.
- It is used to identify correlative variables

Algorithm for PCA:-

- The i/p data are normalized so that each attribute falls within the same range.
- PCA computes or calculates K orthogonal vectors that provides a basis for normalized i/p data.
- These are unit vectors that each point in a direction similar to other one.
- The principal components are stored in ascending order. It is provided a new set of axis for data providing imp information about variance.
- The components are sorted in descending order of significance the data size can be reduced by eliminating the weaker components.

Ex:-

$$A \begin{pmatrix} x & y \\ 2 & 7 \\ 3 & 2 \\ 4 & 3 \end{pmatrix}$$

$\frac{2+3+4}{3} \rightarrow \frac{3}{x} \quad \frac{}{y}$

C: $A - \underline{M}$ → mean

$$C = \begin{pmatrix} -1 & 3 \\ 0 & -2 \\ 1 & -1 \end{pmatrix}$$ → Normalized

characteristic equation $A\bar{V} = \lambda\bar{V}$

$$A\bar{V} - \lambda\bar{V} = 0$$

$$(A-\lambda)\bar{V} = 0$$

first find covariance.

$$(A-\lambda I)\bar{V} = 0$$

$$A - \lambda I = 0$$

covariance (c) : $A^T \cdot A / \frac{1}{n-1}$

$$= \frac{1}{2}\begin{pmatrix} -1 & 0 & 1 \\ 3 & -2 & -1 \end{pmatrix}\begin{pmatrix} -1 & 3 \\ 0 & -2 \\ 1 & -1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & -2 \\ -2 & 7 \end{pmatrix}$$

characteristic eq $A\bar{V} = \lambda\bar{V}$

$$A\bar{V} - \lambda\bar{V} = 0$$

$$(A-\lambda)\bar{V} = 0$$

$$(A-\lambda I)\bar{V} = 0$$

$$\begin{pmatrix} 1 & -2 \\ -2 & 7 \end{pmatrix} - \lambda\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 0$$

$|ad-bc| = 0$

$$\begin{pmatrix} 1-\lambda & -2-0 \\ -2-0 & 7-\lambda \end{pmatrix} = 0$$

$$(1-\lambda)(7-\lambda) - (-2)(-2) = 0$$

$$\lambda^2 - 8\lambda + 3 = 0$$

$$\frac{-b \pm \sqrt{b^2-4ac}}{2a} = \frac{8 + \sqrt{64-4(1)(3)}}{2\times 1}$$

$$= \frac{8 \pm \sqrt{52}}{2} = \frac{8 \pm 7.2}{2} = \frac{8+7.2}{2}, \frac{8-7.2}{2}$$

$\lambda = 7.6$

$$\lambda = \frac{15.2}{2}, \frac{0.8}{2} \Rightarrow 7.6, 0.4.$$

$$\left[\begin{pmatrix} 1 & -2 \\ -2 & 7 \end{pmatrix} - 7.6\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]\begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$$\begin{pmatrix} 1-7.6 & -2 \\ -2 & 7-7.6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0$$

$$\begin{pmatrix} -6.6 & -2 \\ -2 & -0.6 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 0$$

$-6.6x - 2y = 0$

$\frac{x}{2} = \frac{y}{-6.6} \Rightarrow (2, -6.6)$

$-2x - 0.6y = 0$

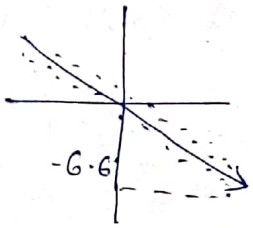
$\frac{x}{0.6} = \frac{y}{-2} \Rightarrow (0.6, -2)$

Eigen vector

Decreasing order :-

$(2, -6.6)$ ✓

$(0.6, -2)$ ×



$\lambda = 0.4$

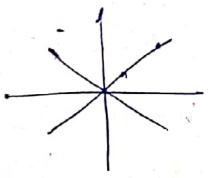
$$\left[\begin{pmatrix} 1 & -2 \\ -2 & 7 \end{pmatrix} - 0.4\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right]\begin{bmatrix} x \\ y \end{bmatrix} = 0$$

$0.6x - 2y = 0 \Rightarrow (2, 0.6)$

$-2x + 6.6y = 0 \Rightarrow (6.6, 2)$

Decreasing order:

$(6.6, 2)\ (2, 0.6)\ (2, -6.6)\ (0.6, -2)$



Attribute subset selection (Feature selection)

The attribute subset selection also called as feature selection. It reduces the data set size by removing irrelevant or redundant attributes or dimensions.

- The Goal of attribute feature selection to find a minimum set of attributes that resulting probability distribution of data & classes is as close as possible to original data set.

Advantages:-

- It enables the machine learning algorithms to train very faster.
- It reduces the complexity of model and makes it easier to interpret.
- It improves accuracy of a model.

Feature selection

Filter method	wrapper method	Embedded method
• Pearson's correlation	• Forward selection	• LASSO Regression
• LDA (Linear discriminant analysis)	• Backward selection	• RIDGE regression
• ANOVA (Analysis of variables)	• Combination of FS & BS	
• Chi-square	• Decision tree induction	
	• Recursive feature combination	

Forward selection :-

This procedure start with an empty set of attributes as a reduced set. The best of the original attributes is determined & added to the reduced set. At each subsequent iteration, the best of the remaining original attributes is added to the set.

Backward (selection): ~~Elimination~~

The procedure start with full set of attributes at each step it removes the worst set of attributes remaining in the set.

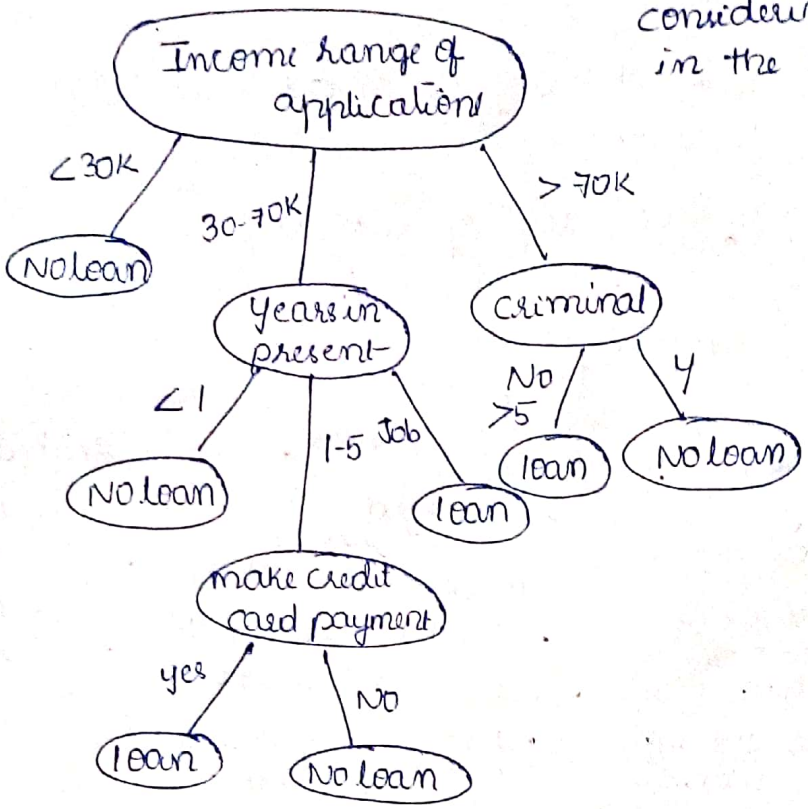
Combination of forward selection & Backward Elimination

The step-wise forward selection & Backward Elimination methods can be combined. So-that Each step the procedure select the best attribute & remove worst attribute in the set.

Decision tree elimination :-

- ID3 C4.5 & CART

Decision tree is used to reduce the dimensionality by removing or not considering the attributes in the original data set



Parametric data reduction Technique :-

↳ Regression

↳ Log linear method

we can also called as numerousity reduction

- These two models used to reduce the size of data.
- In linear regression, the data or model to fit a St. line.

$y = wx + b$ where x is independent variable or predictor variable

y is Response variable

- The data which is not fit in the line is called as outlier and this will be deleted from original dataset.

Loglinear model :-

This is used to approximate discrete, multi dimensional probability distribution.

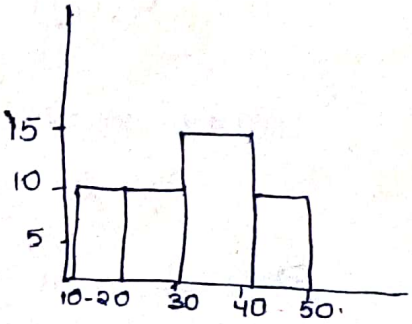
- Estimate the probability of each point in a multi dimensional space for a set of discritize attributes based on a smaller subset of dimensional combination. $F = ae^{bx}$

where a,b - log linear coeff

Non parametric Data Reduction:-

- Histogram - The graphical representation of frequency distribution of a continuous series.

Range	10-20	20-30	30-40	40-50
No. of persons	10	10	15	10



(~~Singleton~~ bucket)

Similarity Measurements :-

The Similarity between the two Obj will indicates the relationship b/w the two objects.

Similarity of (i,j) = 1 − dissimilar of (i,j)

Similarity = 0 means there is no relation b/w obj also indicated by dissimilarity = 1.

Sim = 1 means more similarity b/w data objects also seem as dissim = 0.

* TO represent the similarity b/w Objects, we use 2 data Structures – Data Matrix
– Dissimilarity matrix

↙ Ex

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

nominal – there is no order, the val may contain more than one.

Ex:-

order is Imp →

Object identifier	Test-1 (nominal)	Test-2 (ordinal)	Test-3 (numeric)
1	CodeA	Excellent	45
2	CodeB	fair	22
3	CodeC	good	64
4	CodeA	Excellent	28

Proximity measures b/w nominal attributes :-

$$d(i,j) = \frac{p-m}{p}$$

where m is no of matches
'p' is number of attributes describe the object

From above Ex, p = (1) → which shows the no of col to nominal

Obj	test 1
1	codeA
2	codeB

NO matches

So, m = 0

If codeA ≠ codeA then m = 1.

NOW, $d(2,1) = \frac{p-m}{p}$

$= \frac{1-0}{1} = 1$

For $d(3,1) = \frac{p-m}{p}$

r	codeA
3	codeC

$= \frac{1-0}{1} = 1$

For $d(3,2) = \frac{p-m}{p} = \frac{1-0}{1} = 1$

$d(4,1) = \frac{p-m}{p} = \frac{1-1}{1} = 0$

$d(4,2) = 1$

$d(4,3) = 1$

Now dissimilar matrix,

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Four columns for four attributes

As dissimilarity = 0
So, similarity for d(4,1) is high.

Proximity measures for numeric value :-

To find out the similarity b/w numeric values we use three methods :- Euclidean distance

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2 \ldots (x_{ip}-x_{jp})^2}$$

- Manhattan distance (or) city block distance :-

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots |x_{ip} - x_{jp}|$$

Ex:- $x_1 = (1,2)$, $x_2 = (3,5)$
$x_{i1}\ x_{i2}$ $\quad x_{j1}\ x_{j2}$

Euclidean :-

$$d(i,j) = \sqrt{(1-3)^2 + (2-5)^2}$$

$$= \sqrt{4+9} = \sqrt{13} = 3.61$$

For weight Euclidean distance just multiply with w values

$$\sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots}$$

Manhattan :-

$$d(i,j) = |1-3| + |2-5| = 5$$

- Minkowski distance :-

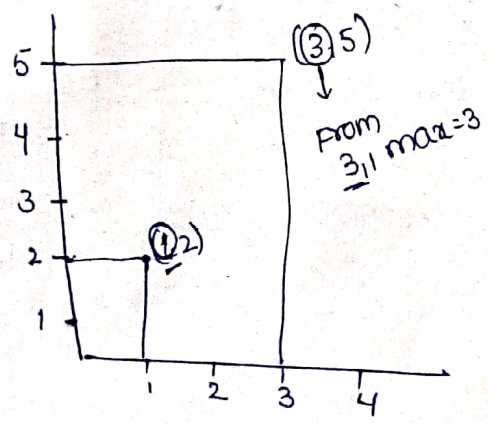
$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots |x_{ip} - x_{jp}|^h}$$

where h is a Real number it represents the distance Of the (~~manhattan~~ Euclidean) when (h = 1). and for (h = 2), it represents (~~Euclidean~~ manhattan) distance.
(h = 1) → manhattan
(h = 2) → Euclidean

- Supreme Distance :- (or) Lmax (or) L_∞ norm (or) Chebyshev distance

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{1/h}$$

$$= \max_{f}^{p} |x_{if} - x_{jf}|$$



Ex:- $x_1 = (1,2)$
$x_2 = (3,5)$

Supreme distance $= \max_{f}^{p} |x_{if} - x_{jf}|$

$= |3-5|_{max} = 2$

To find out dissimilarities for numerical values

$$d_{ij}^{f} = \frac{|x_{if} - x_{jf}|}{\max_h x_h^f - \min_h x_h^f}$$ where f is numeric

From above Ex, table

	type 3
1	45
2	22

$$d(2,1) = \frac{|22-45|}{64-22} = 0.55$$

64 — max value from entire table

22 — min value

$$d(3,1) = \frac{|64-45|}{42} = \frac{19}{42} = 0.452$$

$$d(3,2) = \frac{|64-22|}{42} = \frac{42}{42} = 1$$

$$d(4,1) = \frac{28-45}{42} = \frac{17}{42} = 0.404$$

$$d(4,2) = \frac{28-22}{42} = \frac{6}{42} = 0.14$$

$$d(4,3) = \frac{|28-64|}{42} = \frac{36}{42} = 0.85$$

1.00
0.14
0.86

dissimilarity

$$\begin{bmatrix} 0 & & \\ 0.55 & 0 & \\ 0.45 & 1 & 0 \\ 0.40 & 0.14 & 0.860 \end{bmatrix}$$

For this Sim = 0

So, consider this

Similarity = 1 - dissimi

Similarity = 0.9

1-014

For Proximity measures for Ordinal

From above Ex:- 1 (3) Excellent (1)
2 (1) fair (0)
we provide 3 (2) good (0.5)
ranks 4 (3) Excellent (1)
} values got from formulas

refers to rank → $\frac{r_{if}-1}{m_f-1}$ ← max rank

dissimilarity matrix

$$\begin{bmatrix} 0 & & \\ 1 & 0 & \\ 0.5 & 0.5 & 0 \\ 0 & 1 & 0.5 \quad 0 \end{bmatrix}$$

FOR d(3,1)

check 3 value = 0.5 diff = 0.5
1 value 1

Proximity measures for Binary Attributes :

The Binary attribute values may be 0 or 1. Yes or no
'0' indicates obj is absent '1' refers to obj is present.

$p = q + r + s + t$

$i=1, j=1$ } q $\quad$ $i=1, j=0$ } r

$i=0, j=1$ } s $\quad$ $i=0, j=0$ } t

4, Pi $\begin{matrix} Y=1 \\ N=0 \end{matrix}$

Name	gender	fever	cough	t-1	t-2	t-3	t-4
Jack	m	Y	N	P	N	N	N
Jim	m	Y	Y	N	N	N	N
mary	F	Y	N	P	N	P	N

FOR r $i=1$, $j=0$

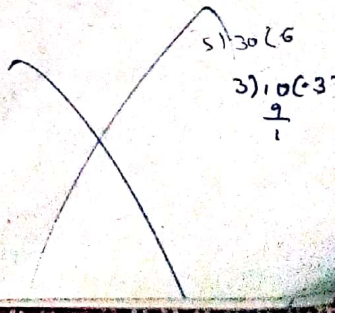
check for Jack, Jim P or Y, N or N

$$d(i,j) = \frac{r+s}{q+r+s}$$

dissimilar mord (jack, jim) $= \frac{1+1}{1+1+1} = \frac{2}{3} = 0.67$

$d(jim, mary) = \frac{1+2}{1+1+2} = \frac{3}{5} = 0.6$

Similarity mord $d(jack, mary) = \frac{0+1}{2+1} = \frac{1}{3} = 0.33$



Proximity measures for mixed attributes

Nominal

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

ordinal

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

(Nominal) Numerical

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1 & 0 & \\ 04 & 0.14 & 0.86 & 0 \end{bmatrix}$$

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

p → No. of attributes

$d_{ij}^{(f)}$ → distance of par position

$\delta_{ij}^{(f)}$ → If all values are present in par attributes then $\delta = 1$ else $\delta = 0$

$(\delta_{ij})(d_{ij})$ $1(d_{2,1})$ From nominal $= \frac{1(1) + 1(1) + 1(0.55)}{1+1+1}$ $(\delta_{ij})(d_{ij})$ $(1)(d_{2,1})$ from numeric

$= \frac{2+55}{3} = 0.85$

3) 255 (8
24
15

For $d(3,1) = \frac{1(1) + 1(0.5) + 1(0.45)}{3} = \frac{1+0.5+0.45}{3} = 0.65$

$d(3,2) = \frac{1(1) + 1(0.5) + 1(1)}{3} = \frac{2.5}{3} = 0.83$

$d(4,1) = \frac{1(0) + 1(0) + 1(0.4)}{3} = \frac{0.4}{3} = 0.133$

1.5
0.86
+
2.36

$d(4,2) = \frac{1(1) + 1(1) + 1(0.14)}{3} = \frac{2.14}{3} = 0.713$

$d(4,3) = \frac{1(1) + 1(0.5) + 1(0.86)}{3} = \frac{1.5+0.86}{3} = 0.78$

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.133 & 0.71 & 0.78 & 0 \end{bmatrix}$$

dissimilar more → 0.85

Similar more → 0.133

To find out the similarity blw 2 obj in asymmetric data we use Jaccard Quotient

$sim(i,j) = \frac{q}{q+r+s}$ $dis(i,j) = \frac{r+s}{q+r+s}$

Cosine similarity :-

It is used to handle the sparse data (the most of the data be zero in matrix) In information retrieval system the most of the data is absent. For those, documents the normal similarity measures will not produce accurate similarity.

The cosine similarity is used to handle this sparse data if the resultant value equal to 1 (angle b/w x, y is 0°). Hence, the documents are more similar. If the cosine value is nearer to zero (90°) then the documents are dissimilar.

$$\text{Similarity}(x,y) = \frac{x \cdot y}{(|x| \cdot |y|)}$$

Ex :-

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document 1	5	0	3	0	2	0	0	2	0	0
2	3	0	2	0	1	1	0	1	0	1
3	0	7	0	2	1	0	0	3	0	0
4	0	1	0	0	1	2	2	0	3	0

Term-frequency table or Document vector

$$\text{Sim}(x,y) = \frac{x \cdot y}{|x| \cdot |y|} = \frac{25}{(6.48)(4.12)} = 0.94$$ No similarity

$$x \cdot y = 5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1$$

$$= 25$$

$$|x| = \sqrt{5^2+0^2+3^2+0^2+2^2+2^2} = 6.48$$

$$|y| = \sqrt{3^2+0^2+2^2+1^2+1^2+1^2+1^2} = 4.12$$

· If the document Attributes are binary attributes then the

$$Sim(x,y) = \frac{x \cdot y}{x \cdot x + y \cdot y - x \cdot y}$$

Data Transformation :-

Types :- · Smoothing
· Attribute construction
· Aggregation
· Generalization
· Normalization

Data are transferred or consolidated into the req format. so that, the resultant mining process may be more Efficient and the patterns are Easier to understand.

* Normalization :-

In this normalization, the attribute data are scaled with in a specified range [-1, 1] or [0, 1]

Types of normalization – min-max normalization
– z-score normalization
– Decimal scalling

→ min-max Normalization :-

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

Ex:- marks 8 10 15 20

new max = 1
new min = 0

$$v' = \frac{8-8}{20-8}(1-0)+0 = 0$$

$$v' = \frac{10-8}{20-8}(1-0)+0 = 1/6 = 0.1$$

$$v' = \frac{15-8}{20-8}(1-0)+0 = \frac{7}{12} = 0.5$$

$$v' = \frac{20-8}{20-8}(1-0)+0 = 1$$

[0, 0.1, 0.5, 1]

→ Z-score normalization :-

$$S.D = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

$$Z\text{-Score} = \frac{x-\mu}{\sigma}$$

Ex:- marks 8 10 15 20

$$\bar{x} = \frac{8+10+15+20}{4} = 13.25$$

For 8

$$S.D = \sqrt{\frac{(8-13.25)^2 + (10-13.25)^2 + (15-13.25)^2 + (20-13.25)^2}{3}}$$

$$= 5.37$$

$$Z\text{-Score} = \frac{8 - 13.25}{5.37} = -0.97$$

For 10

$$Z\text{-score} \quad \frac{10-13.25}{5.37} = -0.60$$

For 15 : $\frac{15-13.25}{5.37} = 0.32$

For 20 : $\frac{20-13.25}{5.37} = \boxed{1.25}$

data must be in $[-1, 1]$ range but it exceeds so we didn't consider.

→ Decimal Scaling :-

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

j represents no. of digits

10 No. of digits = 2

$$v' = \frac{8}{10^1} = \frac{8}{10} = 0.8$$

$$v' = \frac{10}{10^2} = 0.1$$

$$v' = \frac{15}{10^2} = 0.15$$

$$v' = \frac{20}{10^2} = 0.2$$

Smoothing :- which works to remove the noise from the data by using the techniques

1. clustering
2. Binning
3. Regression

Aggregation :- Where the summary and aggregation operation are applied to the given data.

Ex:- Rollup & cube operation in data ware house

Generalization :-

The data where low level are primitive data are placed by the higher level concept by using hierarchy.

Country → higher level
↓
Address → low level
↓
Street → low level

Attribute construction :-

Where a new attributes are constructed and added from the given set of attributes to help the mining aspects Ex:- PCA

UNIT-3

ASSOCIATION RULES

Association rule:- This is used to findout the hidden patterns in the given data.

It indicates the probability of buying one obj if you buy another obj. In association rules we use two terminologies:

- Support
- confidence

* Support means the probability to participate in a transaction or event out of all the transactions.
* confidence indicates the probability of two events occured in a transaction

Ex:- Bread ⇒ Butter (20%, 40%)

Here, 20% indicates out of 100 events the 20 people are going to buy the bread is called support out of 20 members 40% of the people will take butter along with bread we called as confidence.

Basic Algorithm :-

- Apriori Algorithm
- sampling Algorithm
- partitioning Algorithm.

Apriori Algo :-

It is an algorithm for data mining to find out frequent-item sets for bodean association rules. It is follows an bottom up approach where frequent subsets are entried one item at a time which is known as a candidate key generation and grouping of candidates are tested against the data.

Support - A support for an association rule x desives y is the percentage of transaction in the database that contains x Uy:

$$\text{Support}(x, y) = \frac{S(x \cup y)}{n}$$

Confidence: (α) or Strength:

Confidence for an association rule $x \to y$ is the ratio of no. of transactions that contains $x \cup y$ to the no. of transactions that contains 'x'.

$$\alpha = \frac{S(x \cup y)}{S(x)}$$

By using large item set or frequent set we find an association rules for the given data. It indicates the no. of occurences of an item set above the threshold of S. we also called as downward closure

Ex:- For the following given transaction data set generate rules using Apriori Algorithm. consider the value as support = 50% and confidence = 75%.

Transaction id	Items purchased
1	Bread, cheese, Egg, Juice
2	Bread, cheese, Juice
3	Bread, milk, yogurt
4	Bread, Juice, milk
5	Cheese, Juice, milk

↔ Find the frequency of each item

frequent item set :-

items	frequency	Support
Bread	4	4/5 = 80%
cheese	3	3/5 = 60%
Egg	1	1/5 = 20% ✓ least <50% of supp
Juice	4	4/5 = 80%
milk	3	3/5 = 60%
yogurt	1	1/5 = 20% ✓

egg & yogurt is not satisfy the threshold S we can remove.

→ make 2 items as candidate set and write their frequency.

	frequency	Support
{Bread, cheese}	2	2/5 = 40%
{Bread, Juice}	3	3/5 = 60% ✓ > 50% of support
{Bread, milk}	2	2/5 = 40%
{cheese, Juice}	3	3/5 = 60% ✓
{cheese, milk}	1	1/5 = 20%
{Juice, milk}	2	2/5 = 40%

{Bread, juice}
- Bread ⇒ Juice: $\alpha = \frac{S(\text{Bread} \cup \text{Juice})}{S(\text{Bread})} = \frac{3}{4} = 75\%$
- Juice ⇒ Bread: $\alpha = 3/4 = 75\%$ ✓

{cheese, juice}
- cheese ⇒ Juice: $\alpha = 3/3 = 100\%$ ✓
- Juice ⇒ cheese: $\alpha = 3/4 = 75\%$ ✓

all satisfied the confidence.

Ex :-

1	A, B, C
2	A, C
3	A, D
4	B, E, F

Support = 2%.
Confidence = 50%.

A	3	3/6 = 50%
B	2	2/6 = 30%
C	2	2/6 = 30%
~~D~~	1	1/6 = 16%
~~E~~	1	1/6 = 16%
~~F~~	1	1/6 = 16%

As these are < Support

~~{A, B}~~	1	
{A, C}	2	2/6 = 30%
~~{B, C}~~	1	

{A, C}
- A ⇒ C: $\alpha = \frac{S(A \cup C)}{S(A)} = \frac{2}{3} = 60\%$ ✓
- C ⇒ A: $\alpha = \frac{S(C \cup A)}{S(C)} = \frac{2}{2} = 100\%$ ✓

all satisfied

Notations :-

$\alpha \rightarrow$ confidence

$S \rightarrow$ support

$D \rightarrow$ Database of transaction

$t_i \rightarrow$ Transaction in D

$x, y \rightarrow$ Item sets

$x \Rightarrow y \rightarrow$ Association rule

$L \rightarrow$ large item set

$l \rightarrow$ large item set in L

$C \rightarrow$ Set of candidate item sets

$p \rightarrow$ Number of partitions

Ex:- B_1 (M, C, B)

B_2 (M, P, J)

B_3 (M, C, B, J)

B_4 (C, J)

B_5 (M, P, B)

B_6 (M, C, B, J)

B_7 (C, B, J)

B_8 (B, C)

~~Support~~

~~5/8~~

M	5
C	6
B	6
P	~~2~~
J	5

MCB $\Rightarrow$ MC $\rightarrow$ B = 3/3 = 1

$\Rightarrow$ MB $\rightarrow$ C = 3/4 = 0.7

$\Rightarrow$ BC $\rightarrow$ M = 3/5 = 0.6

CBJ $\Rightarrow$ CB $\rightarrow$ J = 3/5

CJ $\rightarrow$ B = 3/4

BJ $\rightarrow$ C = 3/3

2 combni

M, C	3
M, B	4
M, J	3
C, B	5
C, J	4
B, J	3

3-combi

MCB - 3

MCJ - 2 ✗

CBJ - 3

MBJ - 2 ✗

~~MCB - 3~~

CBJ - 3.

~~BCBJ - 2 ✗~~

$S(M \cup C \cup B)$

$S(M)$

FP-Growth Algo :- (Frequency pattern)

- The Drawback of Apriori Algorithm is, we repeatedly scan the database and it will take high Expensive when the datasets are very large.
- FP-growth means Frequency pattern growth Algo used for avoid the drawback for Apriori Algo by considering FP tree where each node represent the support count of itemset and the branch b/w the nodes will indicate the Association of the itemsets.

Ex:- Generate FP-tree for the following transaction dataset (min support = 30%).

Transaction-id	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, E

Items	frequency	priority
A	5	3
B	6	1
C	3	5
D	6	2
E	4	4

All → frequencies are greater than 3 which is Supp
If value is less than suppos then remove it.

Order set of frequency item ⇒ B, D, A, E, C

* Order the items According to priority

By prepering this order

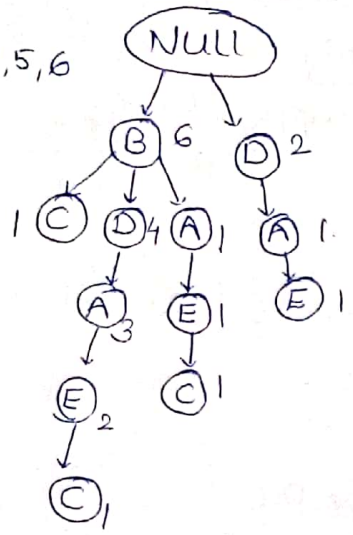
Trans id	items	ordered items
1	EADB	BDAE
2	DACEB	BDAEC
3.	CABE	BAEC
4	BAD	BDA
5	D	D

6	D, B	B, D
7	A, D, E	DAE
8	B, C	B, C

FP-tree :-

B : 1, 2, 3, 4, 5, 6
D : 1, 2, 3, 4
A : 1, 2, 3
E : 1, 2
C : 1

A : 1,
E : 1
C : 1
D : 1, 2
A : 1
E : 1
C : 1



Conditional item sets

B	NULL
D	<B:4>, NULL
A	<D:1> <BD:3>, <B:1>
A ← E	<BA:1> <DA:1> <BDA:1>
B ← C	<B:1>, <BDAE:1> <BAE:1>

Common letters in all

Frequency pattern

E → A
A → E

C → B
B → C

Ex-2 :-

Trans id	itemsets
1	MONKEY
2	DONKEY
3	MAKE
4	MOCKY
5	COOKIE

Support → 3

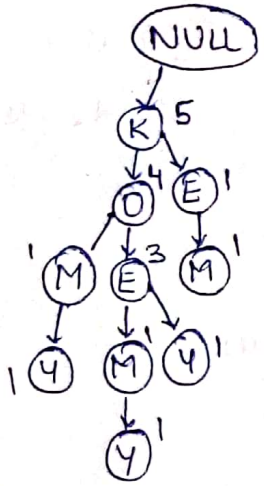
item	frequency	priority
I	1 ✗ support 3	8
M	3	4
O	4	2
N	2 ✗ < support	6
K	5	1
E	4	3
Y	3	5
D	1 ✗	9
A	1 ✗	10
C	2 ✗	7

order of frequency item

~~K~~ ~~D~~ E M Y

Transid	items	ordered items
1	MONKEY	KOEMY
2	DONKEY	KOEY
3	MAKE	KEM
4	MOCKY	KOMY
5	COOKIE	KOE

FP-tree :-



K:1,2,3,4,5
O:1,2,3,4
E:1,2,3
M:1
Y:1
Y:1
E:1
M:1
M:1
Y:1

Conditional item sets :-

K	NULL	
O	<K:4>	→ K:4
E	<KO:3>, <K:1>	→ K:4, O:3
M	<KOE:1>, <KO:1>, <KE:1>	→ K:3, ~~O:2~~, ~~E:2~~
Y	<KOEM:1> <KOM:1> <KOE:1>	→ K:3, O:3, ~~M:2~~, ~~E:1~~

~~Frequency pattern~~

~~E → K~~ ~~M → K~~ ~~Y → O~~ ~~Y → K~~
~~K → E~~ ~~K → M~~ ~~O → Y~~ ~~K → Y~~

Frequency pattern

K → Y, O → Y, KO → Y, K → M, K → O, K → E, O → E, KO → E

Ex-3

Trans-id	items
1	I_1, I_3, I_4
2	I_2, I_3, I_4
3	I_1, I_2, I_3, I_4

Support - 2

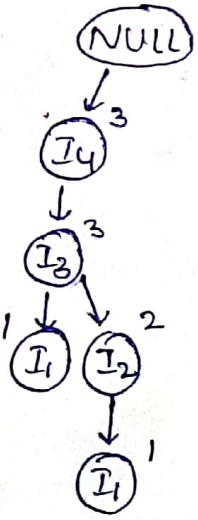
item	frequency	priority
I_1	2	4
I_2	2	3
I_3	3	2
I_4	3	1

Order

I_4, I_3, I_2, I_1

Trans id	items	ordered items
1	$I_1 I_3 I_4$	$I_4 I_3 I_1$
2	$I_2 I_3 I_4$	$I_4, I_3 I_2$
3	$I_1 I_2 I_3 I_4$	$I_4 I_3 I_2 I_1$

FP-tree



I_4 : 1, 2, 3

I_3 : 1, 2, 3

I_1 : 1

I_2 : 1, 2

I_1 : 1

Conditional item sets :-

I_4	NULL	
I_3	$\langle I_4 : 3 \rangle$	$I_4 = 3$
I_2	$\langle I_4 I_3 : 2 \rangle$	$I_4 : 2, I_3 : 2$
I_1	$\langle I_4 I_3 : 1 \rangle$ $\langle I_4 I_3 I_2 : 1 \rangle$	$I_4 : 2$ $I_3 : 2$ $I_2 : 1$

Frequency pattern

$I_4 \rightarrow I_3$

$I_4 \rightarrow I_2$ $I_3 \rightarrow I_2$, $I_4 I_3 \rightarrow I_2$

$I_4 \rightarrow I_1$, $I_3 \rightarrow I_1$, $I_4 I_3 \rightarrow I_1$

2. Sampling Algorithm :-

It is used to find out the frequency itemset from given large dataset. it is difficult by using apriori algo. Instead of working on entire population we will take some sample and apply the apriori algorithm which results indicates the frequency items in entire populatio

potentially large frequency itemset (PL) :- It indicates the large itemset in sample.

Negative Border (BD) :-

* It is generalization of apriori gen algorithm.
* The minimal set of itemsets which are not in pl, But the samples of items are present in pl.

Algorithm :-

DS D_S = Sample of database

P_l = large itemsets in D_S using "smalls"

C = $pl \cup BD^{-}(Pl)$

count c in database using S

ML = large itemsets in $BD^{-}(Pl)$

if ML = $\emptyset$ done

else

C = repeated application of BD^{-} count c in db.

Example :-

Trans Id	Itemset
1	Bread, Jelly, PB
2	Bread, PB
3	Bread, milk, PB
4	Beer, Bread
5	Beer, milk

Support = 20%.

Smalls = 10%.

PB - peanut Butter.

DS = $\{T_1, T_2\}$

Data set - select any of two transactions

T_1 = { Bread, Jelly, PB }

T_2 = { Bread, PB }

PL = { { Bread } { Jelly } { PB } { Bread, Jelly } { Bread, PB } { Jelly, PB }
{ Bread, Jelly, PB } }

$BO^-(PL)$ = { { Beer }, { milk } } → remaining items not repeated in PL

C^1 = PL ∪ $BO^-(PL)$

Canded set = { { Bread } { Jelly } { PB } { Bread, Jelly } { Bread, PB } { Jelly, PB }
{ Bread, Jelly, PB } { Beer } { milk } }

Support = $\frac{20}{100} \times 5$ → No. of Transactions

= 1 (remove the items that are not repeated atleast once)

$C^1(BO^-)$ = { { Beer, Bread } { Beer, Jelly } { Beer, PB } { Beer, milk }
{ Bread, milk } { Jelly, milk } { milk, PB } }

Support = $\frac{20}{100} \times 5 = 1$ (remove items that are not repeted atleast once as a above combo)

= { { Beer, Bread } { Beer, milk } { Bread, milk } { milk, PB } }

(A, B) ∪ (B, C)

(A, B, C)

{ { Beer, Bread, milk } { Beer, milk, PB } { Bread, milk, PB } }

Again check in support = 1

C^1 = { Bread, milk, PB }

partition Algorithm :-

To find out the frequency itemset for Entire db in apriori Algorithm is difficult because of memory restrictions.

* By using partition Algo, we can divide Entire dataset into diff partitions & allocate each partition to the processor to find out frequency itemset of each partition. At the last combine all (partition) items generated by each partition and apply apriori gen algo to find out large frequency itemset.

Ex:-

		A_1	A_2	A_3	A_4	A_5
S=20% P_1	0	1	1	0	0	0
	1	0	1	0	0	0
	2	1	0	1	0	1
	3	0	1	0	0	0
S=20% P_2	4	0	0	1	0	0
	5	0	1	0	0	0
	6	0	1	0	1	0
	7	0	0	1	0	1

Support = 20%

Support $(P_1) = \frac{20}{100} \times 4 = 0.8$ (Every state must enter atleast once)

$L^1 = \{\{A_1\}\{A_2\}\{A_3\}\{A_5\}\{A_1,A_2\}\{A_1,A_3\}\{A_1,A_5\}\{A_3A_5\}\{A_1,A_3A_5\}\}$

check in single combo whether it Enter one time

If 1 present in both A_1, A_2

$P_2 : L^2 = \{\{A_2\}\{A_3\}\{A_4\}\{A_5\}\{A_2A_4\}\{A_3,A_5\}\}$

$L = P_1 \cup P_2$

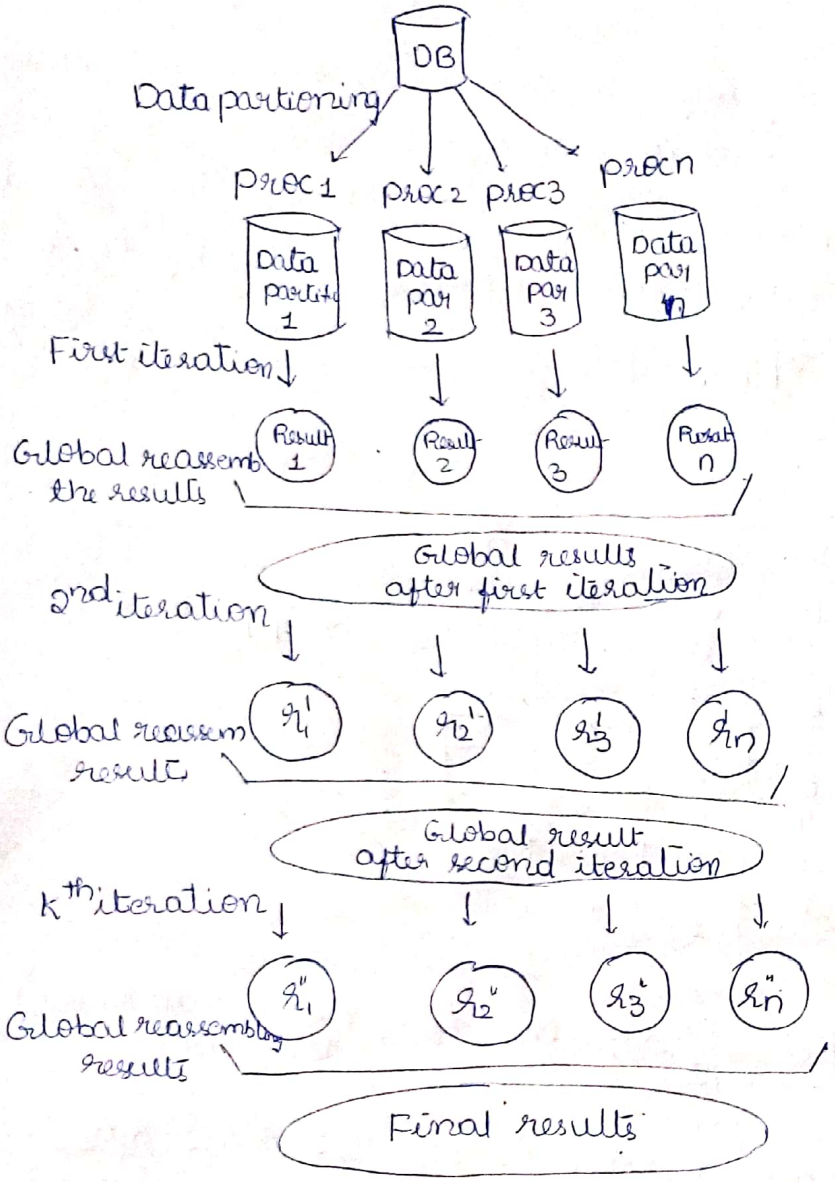
check :- $(8) \times \frac{20}{100} = 1.6 \approx 2$

(8) Total no of Transactions.

$\therefore L = \{\{A_1\}\{A_2\}\{A_3\}\{A_5\}\{A_3,A_5\}\}$

By using check, As check = 2 we need to check in above L^1, L^2 whether the inside Attributes repeat 2 times or more than 2 then write to L.

Parallel & Distributed Association Algo :- (Count Distribution Algorithm)



In most Association rules, the parallelism mechanism will be taken by 2 ways :- → Data parallelism
→ Task parallelism

* In Data parallelism, the data is distributed into dif processors and in Task parallelism, the candidate sets are distributed to the dif processors.
* In Data parallelism, the Algorithms reduce communication cost over the tasks.
* While task parallelism, not only candidates but also the local set of transactions must be broadcast to all other sites

* In data parallelism, Algo requires memory at each processor will large enough to store all the candidates at each scan but in task parallelism, It can be avoid because the subsets of the candidates that are assign to the processor during each scan must fit into memory.

Ex:

Transid	Itemset
1	A, B, C
2	A, C, B
3	B, A, A
4	C, B, D

Processor 1

1	ABC
2	ACB

Processor 2

3	B A (A)
4	C B D

consider only one in one set

A → 2
B → 2
C → 2
D → 0

A → 1
B → 2
D → 1
C → 1

Support = 40%

$\frac{2}{5} \times 4$

= 2

merging both

Processor 1

A → 3
B → 4
C → 3
D → 1 X

Processor 2

A → 3
B → 4
C → 3
D → 1 X

A → 3
B → 4
C → 3

Proc 1 ↙ Proc 2 ↘

Proc 1	Proc 2
A	C
B	
check in support → AB	C
AB	C
ABC	

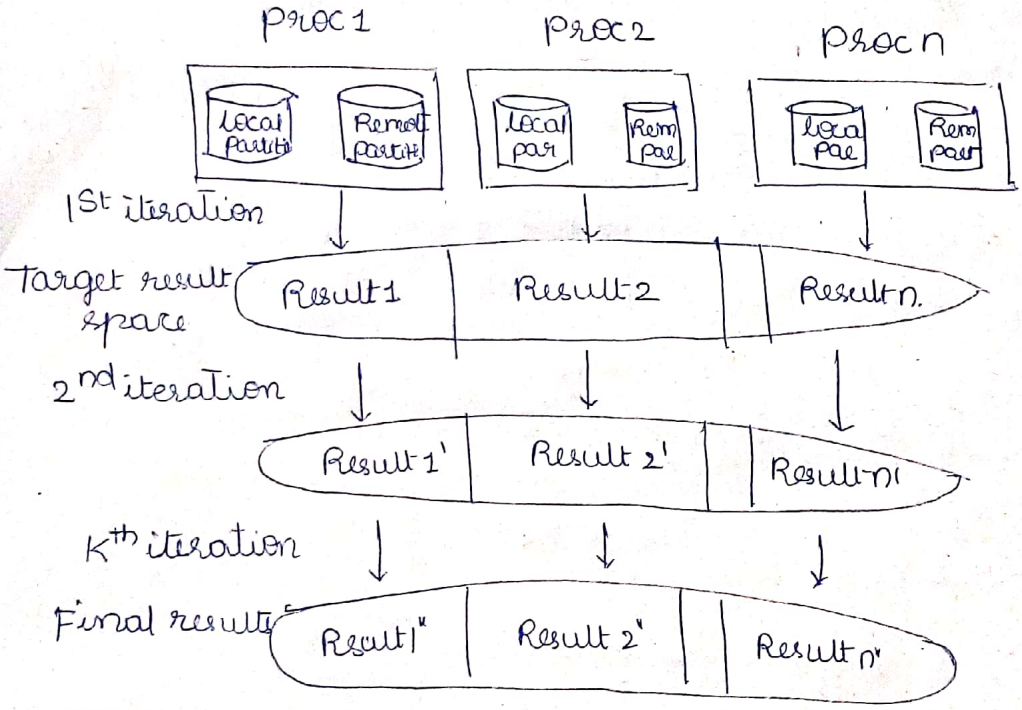
Each partition

Each processor will have a disjoint data partition to work with. Each processor however will have a complete candidate itemset, although with partial support or support count.

- At the End of Each iteration, since support of Each candidate itemset in each processor is incomplete, Each processor will need to redistribute the count to all processors. Hence term count distribution is used.

Task parallelism :-

* It is also called as data distribution Algo.
* The candidates in task parallelism is partitioned among the processors
* C_k^l indicates the candidates of size k, find out at processor P_l
* L_k^l indicates the local large k items at processor L.



Ex:-

TranId	Items
t_1	Br, Jelly, PB
t_2	Br, PB
t_3	Br, milk, PB
t_4	Beer, Br
t_5	Beer, milk

Support = 2

P_1 [Bread, Beer

P_2 [Jelly, milk

P_3 - PB.

P^1: $t_1 t_2$ P^2: $t_3 t_4$ P^3: t_5

Bread : 2 Jelly = 0 PB = 0

Beer = 0 milk = 1

Broadcast data partition to diff processors.

P^1 : $t_3 t_4$ P^2: t_5 P^3 : $t_1 t_5$

Br : 2+2 Jelly : 0+0 PB = 0+2

Be : 0+1 milk : 1+1

P^1 : t_5 P^2 : t_1, t_2 P^3 : $t_3 t_4$

Br : 4 milk : 2 PB : 3

Be : 2 ~~Jelly~~ : 1

support < 2

2-itemsets :-

{Br, milk} {Be, milk} {Br, B} {m, PB} {Br, PB} {Be, PB}

Advance Association rule mining / Techniques :—

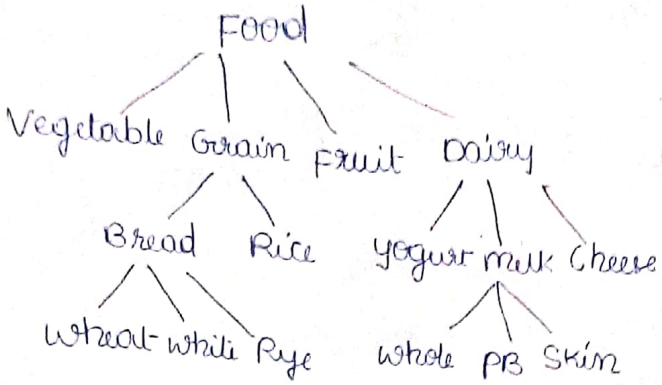
In advance association rule technique we are addressing -

- How to handle the quantitative dataset
- How to define handle the support value for the given dataset
- How to find out the frequency itemsets in Transactional db

To Address the above problems we use following association rule Techniques:-

- Generalized association rules
- Multilevel association rule
- Quantitative association rule
- Multiple minimum Support Techniques
- correlation rules.

Generalized Association rules :-



By using the concept of hierarchy that shows the relationship b/w the diff items, generalized association rules.

- A generalized association rules $x \Rightarrow y$ indicates that no item in 'y' may be above any item in x

$x \Rightarrow y$ means we can't generate association rules from lower level to higher level

Ex:- Bread $\Rightarrow$ PB

Grain $\Rightarrow$ PB

Out of these two association rules Grain $\Rightarrow$ PB has highest support value compared to Bread $\Rightarrow$ PB because Grain having higher level than bread.

Multilevel Association rules :-

A variation of generalised association rules is called as multilevel Association rules

In multilevel association rules the itemsets may occur in any level in hierarchy.

- using a variation of apriori algo the concept hierarchy is traversed in top down manner to generate large itemsets.
- The minimum support in all nodes in hierarchy at same level is identical.
- If α_i is min support for level i in hierarchy α_{i-1} is min support for level i-1 then $\alpha_{i-1} > \alpha_i$

Vegetable $\rightarrow$ Grain

Grain $\not\rightarrow$ Yogurt

Quantitative Association rules :-

In Quantitative Association rule used to find the Association rules when the Database contains the Quantity data.

By using the Discritization method called as Equal width, we divide the given Quantity data into Equal width.

- Apply the support after discritization to find out Association rules.

Ex:-

C.id	Age	married	cars
1	21	No	0
2	23	Yes	1
3	29	NO	1
4	30	Y	2
5	34	Y	2

↙

Age	
21	20:24
23	20:24
29	25:29
30	30:34
34	30:34

→

CId	Age	married	cars	
1	20:24	N	0	0:1
2	20:24	Y	1	0:1
3	25:29	N	1	0:1
4	30:34	Y	2	2
5	30:34	Y	2	2

itemset	Support
~~Age~~	~~Count~~
<Age: 20:24>	② → 20:24 appears two times Support = 2 in above
~~<Age 25:29>~~	1 — X
<Age: 30:34>	2
<married: Yes>	3
<married: No>	2
<Cars 0:1>	3
<cars 2:3>	2

✓ <Age 30:34> <married yes>
⇓
<Cars: 2:3>

X <married: Yes> ⇒ <cars: 2>

→ Multiple min support —

Rare item problem :-

If the support is too low, then most no. of items consider for frequent itemset which are not useful.

- If support value is too high then less no. of frequent itemset in this we can loose frequent items.
- TO solve the problm, we use two techniques.
  - partion (clustering) data sets & Apply Association rule.
  - MIS Apriori.

Ex:- MIS Apriori

Consider itemsets {A, B, C}

mis(A) = 20%.

mis(B) = 3%.

mis(C) = 4%.

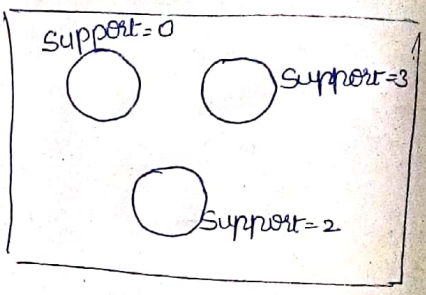
for 2 combo's

AB = min{ mis(A), mis(B)}

AC = 4%. = min{ 20%. , 3%.}

= 3%.

clustering



Correlation Techniques :-

The correlation used to find out the relation b/w two frequent itemsets.

If the value of correlation = 1, Then there is a strong association b/w two objects.

$$COR(A \rightarrow B) = \frac{P(A,B)}{P(A)P(B)}$$

Incremental Algorithm :-

This Algorithm is used to find out the frequent itemset, when the db is dynamic.

Rules to find out Association rule :-

* Find out the frequency of each item in new db, If it is already existed in old db
* The frequent item is not presented in T-old (old db) but present in new dataset, then ignore the item.
* Find out the promoted border set

promoted border set :-

If an (frequent) item is not frequent in old db but after adding the new transaction to the db, the item is frequent, Then the set of frequent items is called as promoted border set.

	TranId	Itemset
Told	1	B, J, P
	2	B, m, P
	3	m, P
	4	C, D
Tnew	5	C, J
	6	B, C, E X
	7	C, B

Support = 2

$T_{whole} = T_{old} \cup T_{new}$

Item	frequency
B	2+2 = 4
J	1+1 = 2
P	3
m	2
C	1+3 = 4
D	1 X

<2 add new value

Tnew

B	4
J	2
P	3
m	2
C	4

promoted border set

promoted border set :-

{ {J}, {C} {X}, {J, C} {J, X} {J, C, X} {C, X} }

After new set, D combinations are removed as it doesn't satisfy support.

promoted set

{J} {C} {J, C}

Measuring the Quality of rules :-

* The first support & confidence are basic methods to find out Quality of Association techniques.

1. $$S(x \to y) = \frac{S(x \cup y)}{n}$$

$$\alpha(x \to y) = \frac{S(x \cup y)}{S(x)}$$

drawback :-
we are not considering no. of times y had appproached.

2. Lift (or) Interest

$$\text{interest } (A \to B) = \frac{P(A, B)}{P(A)\,P(B)}$$

⇓ or correlation

drawback :-
Symmetric
$I(A \to B) = I(B \to A)$

3. conviction

$$(A \Rightarrow B) = \frac{P(A)(P(\sim B)}{P(A, \sim B)}$$

$A \to B \equiv \sim(A \wedge \sim B)$

If conviction = 1, then A & B are not related

4. Chi-square method

$$\leq \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

If value is less than chi-square value it is strongly less it is not accepted.