19/01/21 · Dec-09 → Data Integration

→ Entity Identification problem
→ Redundancy and correlation analysis
→ Tuple duplication
→ detection and resolution of data value conflicts.

Data Reduction

a. dimension reduction
1. wavelet transform
2. PCA
3. Attribute subset seln.

b. numerosity reduction
parametric
non-parametric

c. data compression
lossless
lossy

(1) Entity Identification Problem:

Eg: marks: -/. age
cgpa (1-10)
Grade (A, A+ - - -)

Focus on the semantic of data ie maintain meta data (for each attribute)
This includes name, meaning, data type, range of values permitted for the attribute and null rules for handling blank, zero/or null values.

→ The meta data can be used to avoid errors in schema integration.

2) Redundancy and Correlation analysis -

It is an imp. issue in data integration.

→ The redundancies can be detected using correlation analysis.

→ This measures, given two attributes, how strongly one attribute implies the other, based on available data.

for nominal attribute : χ^2 (chi-square test)

for numerical ,, : → Pearson Correln. coefficient

→ Covariance Test.

a) Chi-square Test :

	A_1	A_2	...	A_c
B_1		$(A_i\ B_j)$		
B_2				
⋮				
B_r				

$r \times c$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{Count(A = a_i) \times Count(B = b_j)}{N}$$

Null Hyp: A & B are independent

$\chi^2 \rightarrow 25$

$\alpha \rightarrow$ 0.1, 0.01, 0.001, 0.0001

DOF → $(r-1) \times (c-1)$

$(2-1) \times (2-1)$

10   01

0.001

α

1 → 10

O_{ij} → observed frequency (actual count) of joint event $(A_i\ B_j)$

e_{ij} → expected ,, of (A_i, B_j)

N = no. of data tuples

Count $(A = a_i)$ is the no. of tuples having value a_i for A

Count $(B = b_j)$,, ,, ,, ,, ,, ,, ,, b_j for B.

	Male	female	
fiction	250 (90)	200	→ 450
non-fiction	50	1000	→ 1050
Total	300	1200	1500

χ^2 test hypothesis (Null hyp.) assumes A & B are independent (Gender & Types of book are independent)

$$e_{11} = \frac{\text{count (male)} \times \text{count (fiction)}}{N}$$

$$= \frac{300 \times 450}{1500} = 90$$

$$e_{12} = \frac{1200 \times 450}{1500} = 360$$

$$e_{21} = \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{1200 \times 1050}{1500} = 840$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= 507.93$$

$$DOF = (r-1) \times (c-1) = (2-1) \times (2-1) = 01$$

For 1 DOF, the χ^2 value needed to reject the hypothesis at 0.001 significance level is 10.828 (refer χ^2 distribn. Table). Since, the computed value is above this, we reject the hypothesis that gender and book reading are independent.

Therefore, we conclude that the two attributes are dependent or strongly correlated to each other.

b) **Pearson Correlation Coefficient:**

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$-1 \leq r_{A,B} \leq +1$$

If $r_{A,B}$ is > 0, then A and B are positively correlated (ie A ↑ B ↑)

If $r_{A,B} = 0$, then A and B are independent and there is no correlation b/w them.

If $r_{A,B} < 0$, then A and B are negatively correlated

i.e A ↑ B ↓

Eg:

serial no.	Age (x)	Glucose level (y)	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

$r = 0.5298$ (This means 52.98% variables have a moderate positive correlation)

c) Covariance of numeric data:

(Next class)