THE SEARCH FOR THE PERFECT ASSOCIATION RULE !

NULL INVARIANCE MEASURE

DATA MINING

Harisaipravin Sv 17pw13

DATA-MINE
NULL INVARIANCE

Contents

The need for Mr. Perfect Association factor

RECAP

NULL Invariance - Value does not change with the # of NULL transactions

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

χ^2 tells also better than s & c

	B	¬B	Σ_{row}
C	400 (450)	350 (300)	750
¬C	200 (150)	50 (100)	250
Σ_{col}	600	400	1000

Expected value

Observed value

Lift **is more telling than s & c**

	B	¬B	Σ_{row}
C	400	350	750
¬C	200	50	250
$\Sigma_{col.}$	600	400	1000

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

$$\text{Confidence } (A \rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Measure	Definition	Range	Null-Invariant
$\chi^2(A, B)$	$\sum_{i,j=0,1} \frac{(e(a_i b_j)-o(a_i b_j))^2}{e(a_i b_j)}$	$[0, \infty]$	No
$Lift(A, B)$	$\frac{s(A \cup B)}{s(A) \times s(B)}$	$[0, \infty]$	No
$AllConf(A, B)$	$\frac{s(A \cup B)}{max\{s(A), s(B)\}}$	$[0, 1]$	Yes
$Jaccard(A, B)$	$\frac{s(A \cup B)}{s(A)+s(B)-s(A \cup B)}$	$[0, 1]$	Yes
$Cosine(A, B)$	$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}$	$[0, 1]$	Yes
$Kulczynski(A, B)$	$\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)})$	$[0, 1]$	Yes
$MaxConf(\mathrm{A}, \mathrm{B})$	$max\{\frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)}\}$	$[0, 1]$	Yes

	$milk$	$\neg milk$	Σ_{row}
$coffee$	mc	$\neg mc$	c
$\neg coffee$	$m\neg c$	$\neg m \neg c$	$\neg c$
Σ_{col}	m	$\neg m$	Σ

- Milk vs Coffee Contingency table
- Very low and very hight Null values can be fatal and misleading !

Data set	mc	$\neg mc$	$m\neg c$	$\neg m \neg c$	χ^2	$Lift$
D_1	10,000	1,000	1,000	100,000	90557	9.26
D_2	10,000	1,000	1,000	100	0	1
D_3	100	1,000	1,000	100,000	670	8.44
D_4	1,000	1,000	1,000	100,000	24740	25.75
D_5	1,000	100	10,000	100,000	8173	9.18
D_6	1,000	10	100,000	100,000	965	1.97

Confidence

MAX conf - Piking the biggest value of confidence.

ALL conf - Minimum value of confidence

C

Cosine

Lift function in disguise, but is null invariant.

Overcomes the deficiency of Lift function.

C

Jaccard

J

Does not consider the NULL Values.

The formula is such that, it only considers the used area and eliminates the unused area.

Kulczynski

K

The approach of averaging the 2 confidence together

A -> B (&) B -> A

Confidence & Kulczynski

Variation of native Confidence

Conf (A->B) =

S(A inter B) / S (A)

Confidence Measure	Definition
All Confidence	$\frac{s(A \cap B)}{\max\{s(A), s(B)\}}$
Kulczynski	$\frac{P(A\|B)+P(B\|A)}{2}$
Max Confidence	$\max\{P(A\|B), P(B\|A)\}$

“ Biggest value of confidence between the 2 ways of looking at it”

MAX Confidence

“Average of 2 confidences”

Kulczynski

“ By dividing by the maximum value of the support for A int B and B int A,

We're really just finding the minimum value of confidence between the 2 options.“

ALL Confidence

DATA-MINE

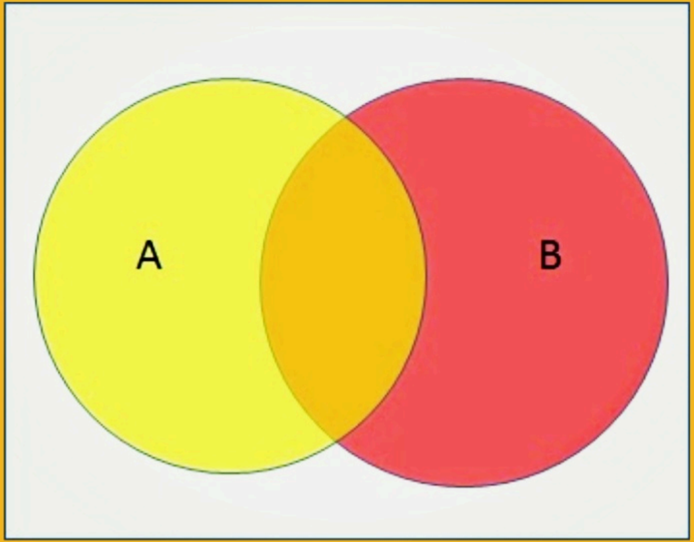
NULL INVARIANCE

Cosine

- Considering the grand total is a cause of concern.
- Lift can be simplified :- (AB)/(A*B/G)
- Cosine simplifies to :- AB/sqrt(A*B)

Confidence Measure	Definition
Lift	$\frac{s(A \cap B)}{s(A) \times s(B)}$
Cosine	$\frac{s(A \cap B)}{\sqrt{s(A) \times s(B)}}$

Confidence Measure	Fraction Explanation
Lift	$\frac{\text{Count of records with A\&B}/\text{grand total}}{\left(\text{count of records with A}/\text{grand total}\right) \times \left(\text{count of records with B}/\text{grand total}\right)}$
Cosine	$\frac{\text{Count of records with A\&B}/\text{grand total}}{\sqrt{\left(\text{count of records with A}/\text{grand total}\right) \times \left(\text{count of records with B}/\text{grand total}\right)}}$



Jaccard

The Jaccard function is defined as :-

|A int B|/|A union B|

- The numerator of the Jaccard is the Orange area
- The denominator is the area of the yellow-orange-red area
- It doesn't use any data from the blank space

Metrics

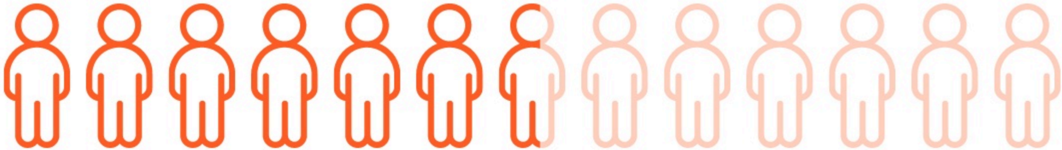
D4 is neutral and balanced; D5 is neutral but imbalanced; D6 is neutral but very imbalanced.

Data set	mc	$\neg mc$	$m \neg c$	$\neg m \neg c$	$AllConf$	Jaccard	$Cosine$	$Kulc$	$MaxConf$
D_1	10,000	1,000	1,000	100,000	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	0.01	0.01	0.10	0.5	0.99

Comparison of measures

Neutral but very imbalanced

50%



Kulc - It is proportinal !

99%

Max Conf - Always shows max !

Case Study

ID	Author A	Author B	$s(A \cup B)$	$s(A)$	$s(B)$	Jaccard	*Cosine*	*Kulc*
1	Hans-Peter Kriegel	Martin Ester	28	146	54	0.163 (2)	0.315 (7)	0.355 (9)
2	Michael Carey	Miron Livny	26	104	58	0.191 (1)	0.335 (4)	0.349 (10)
3	Hans-Peter Kriegel	Joerg Sander	24	146	36	0.152 (3)	0.331 (5)	0.416 (8)
4	Christos Faloutsos	Spiros Papadimitriou	20	162	26	0.119 (7)	0.308 (10)	0.446 (7)
5	Hans-Peter Kriegel	Martin Pfeifle	18	146	18	0.123 (6)	0.351 (2)	0.562 (2)
6	Hector Garcia-Molina	Wilburt Labio	16	144	18	0.110 (9)	0.314 (8)	0.500 (4)
7	Divyakant Agrawal	Wang Hsiung	16	120	16	0.133 (5)	0.365 (1)	0.567 (1)
8	Elke Rundensteiner	Murali Mani	16	104	20	0.148 (4)	0.351 (3)	0.477 (6)
9	Divyakant Agrawal	Oliver Po	12	120	12	0.100 (10)	0.316 (6)	0.550 (3)
10	Gerhard Weikum	Martin Theobald	12	106	14	0.111 (8)	0.312 (9)	0.485 (5)

Best NULL-invariance measure ?

You decide !

DATA-MINE
NULL INVARIANCE