



# Chapter 6. Frequent Pattern Mining: Pattern Evaluation

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

# How to Judge if a Rule/Pattern Is Interesting?

- Pattern-mining will generate a large set of patterns/rules
  - Not all the generated patterns/rules are interesting
- Interestingness measures: Objective vs. Subjective
  - Objective interestingness measures
    - Support, confidence, correlation, ...
  - Subjective interestingness measures: One man's trash could be another man's treasure
    - Query-based: Relevant to a user's particular request
    - Against one's knowledge-base: unexpected, freshness, timeliness
    - Visualization tools: Multi-dimensional, interactive examination

# Limitation of the Support-Confidence Framework

- Are s and c interesting in association rules: " $A \Rightarrow B$ " [s, c]?
- Example: Suppose one school may have the following statistics on # of students who may play basketball and/or eat cereal:

Be careful!

2-way contingency table

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

- Association rule mining may generate the following:
  - $\text{play-basketball} \Rightarrow \text{eat-cereal}$  [40%, 66.7%] (higher s & c)
- But this strong association rule is misleading: The overall % of students eating cereal is 75% > 66.7%, a more telling rule:
  - $\neg \text{play-basketball} \Rightarrow \text{eat-cereal}$  [35%, 87.5%] (high s & higher c)

# Interestingness Measure: Lift

- Measure of dependent/correlated events: **lift**

$$\text{lift}(B, C) = \frac{c(B \rightarrow C)}{s(C)} = \frac{s(B \cup C)}{s(B) \times s(C)}$$

*Lift* is more telling than s & c

- Lift(B, C) may tell how B and C are correlated

- Lift(B, C) = 1: B and C are independent
- > 1: positively correlated
- < 1: negatively correlated

	B	¬B	Σ <sub>row</sub>
C	400	350	750
¬C	200	50	250
Σ <sub>col.</sub>	600	400	1000

- For our example,  $\text{lift}(B, C) = \frac{400/1000}{600/1000 \times 750/1000} = 0.89$

$$\text{lift}(B, \neg C) = \frac{200/1000}{600/1000 \times 250/1000} = 1.33$$

- Thus, B and C are negatively correlated since  $\text{lift}(B, C) < 1$ ;
  - B and ¬C are positively correlated since  $\text{lift}(B, \neg C) > 1$

# Interestingness Measure: $\chi^2$

- Another measure to test correlated events:  $\chi^2$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- General rules

- $\chi^2 = 0$ : independent

- $\chi^2 > 0$ : correlated, either positive or negative, so it needs additional test

- Now,  $\chi^2 = \frac{(400 - 450)^2}{450} + \frac{(350 - 300)^2}{300} + \frac{(200 - 150)^2}{150} + \frac{(50 - 100)^2}{100} = 55.56$

- $\chi^2$  shows B and C are negatively correlated since the expected value is 450 but the observed is only 400
- $\chi^2$  is also more telling than the support-confidence framework

Observed value      Expected value

	B	$\neg B$	$\Sigma_{row}$
C	400 (450)	350 (300)	750
$\neg C$	200 (150)	50 (100)	250
$\Sigma_{col}$	600	400	1000

# Lift and $\chi^2$ : Are They Always Good Measures?

- Null transactions: Transactions that contain neither B nor C
- Let's examine the dataset D
  - BC (100) is much rarer than B¬C (1000) and ¬BC (1000), but there are many ¬B¬C (100000)
  - Unlikely B & C will happen together!
- But,  $\text{Lift}(B, C) = 8.44 \gg 1$  (Lift shows B and C are strongly positively correlated!)
- $\chi^2 = 670$ : Observed(BC)  $\gg$  expected value (11.85)
- Too many null transactions may “spoil the soup”!

	B	¬B	$\Sigma_{\text{row}}$
C	100	1000	1100
¬C	1000	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

*null transactions*

**Contingency table with expected values added**

	B	¬B	$\Sigma_{\text{row}}$
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

# Interestingness Measures & Null-Invariance

- *Null invariance*: Value does not change with the # of null-transactions
- A few interestingness measures: Some are null invariant

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
$Kulczynski(A, B)$	$\frac{1}{2} \left( \frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)} \right)$	$[0, 1]$	Yes
$MaxConf(A, B)$	$\max\left\{ \frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)} \right\}$	$[0, 1]$	Yes

$\chi^2$  and lift are not null-invariant

Jaccard, cosine, AllConf, MaxConf, and Kulczynski are null-invariant measures

$\max\{s(A \cup B) / s(A), s(A \cup B) / s(B)\}$

# Null Invariance: An Important Property

- Why is null invariance crucial for the analysis of massive transaction data?
  - Many transactions may contain neither milk nor coffee!

milk vs. coffee contingency table

	<i>milk</i>	$\neg milk$	$\Sigma_{row}$
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	$\neg c$
$\Sigma_{col}$	<i>m</i>	$\neg m$	$\Sigma$

- Lift and  $\chi^2$  are not null-invariant: not good to evaluate data that contain too many or too few null transactions!
- Many measures are not null-invariant!

Null-transactions  
w.r.t. m and c

Data set	<i>mc</i>	$\neg mc$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	$\chi^2$	<i>Lift</i>
$D_1$	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



# Comparison of Null-Invariant Measures

- Not all null-invariant measures are created equal
- Which one is better?
  - $D_4$ — $D_6$  differentiate the null-invariant measures
  - Kulc (Kulczynski 1927) holds firm and is in balance of both directional implications

2-variable contingency table

	<i>milk</i>	$\neg milk$	$\Sigma_{row}$
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	$m \neg c$	$\neg m \neg c$	$\neg c$
$\Sigma_{col}$	<i>m</i>	$\neg m$	$\Sigma$

All 5 are null-invariant

Data set	<i>mc</i>	$\neg mc$	$m \neg c$	$\neg m \neg c$	<i>AllConf</i>	Jaccard	<i>Cosine</i>	<i>Kulc</i>	<i>MaxConf</i>
$D_1$	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

Subtle: They disagree on those cases

# Analysis of DBLP Coauthor Relationships

- Recent DB conferences, removing balanced associations, low sup, etc.

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)

Advisor-advisee relation: Kulc: high,  
Jaccard: low, cosine: middle

- Which pairs of authors are strongly related?
  - Use Kulc to find Advisor-advisee, close collaborators

# Imbalance Ratio with Kulczynski Measure

- IR (Imbalance Ratio): measure the imbalance of two itemsets A and B in rule implications:

$$IR(A, B) = \frac{|s(A) - s(B)|}{s(A) + s(B) - s(A \cup B)}$$

- Kulczynski and Imbalance Ratio (IR) together present a clear picture for all the three datasets  $D_4$  through  $D_6$ 
  - $D_4$  is neutral & balanced;  $D_5$  is neutral but imbalanced
  - $D_6$  is neutral but very imbalanced

Data set	$mc$	$\neg mc$	$m\neg c$	$\neg m\neg c$	Jaccard	Cosine	Kulc	IR
$D_1$	10,000	1,000	1,000	100,000	0.83	0.91	0.91	0
$D_2$	10,000	1,000	1,000	100	0.83	0.91	0.91	0
$D_3$	100	1,000	1,000	100,000	0.05	0.09	0.09	0
$D_4$	1,000	1,000	1,000	100,000	0.33	0.5	0.5	0
$D_5$	1,000	100	10,000	100,000	0.09	0.29	0.5	0.89
$D_6$	1,000	10	100,000	100,000	0.01	0.10	0.5	0.99

# What Measures to Choose for Effective Pattern Evaluation?

- Null value cases are predominant in many large datasets
  - Neither milk nor coffee is in most of the baskets; neither Mike nor Jim is an author in most of the papers; .....
- Null-invariance is an important property
- Lift,  $\chi^2$  and cosine are good measures if null transactions are not predominant
  - Otherwise, Kulczynski + Imbalance Ratio should be used to judge the interestingness of a pattern

# Discussion

- Where do you want to use them?

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
$Kulczynski(A, B)$	$\frac{1}{2} \left( \frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)} \right)$	$[0, 1]$	Yes
$MaxConf(A, B)$	$\max\left\{ \frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)} \right\}$	$[0, 1]$	Yes

$$\max\{s(A \cup B) / s(A), s(A \cup B) / s(B)\}$$

# Summary

- Basic Concepts:
  - Frequent Patterns, Association Rules, Closed Patterns and Max-Patterns
- Frequent Itemset Mining Methods
  - The Downward Closure Property and The Apriori Algorithm
  - Extensions or Improvements of Apriori
  - Mining Frequent Patterns by Exploring Vertical Data Format
  - FPGrowth: A Frequent Pattern-Growth Approach
  - Mining Closed Patterns
- Which Patterns Are Interesting?—Pattern Evaluation Methods
  - Interestingness Measures: Lift and  $\chi^2$
  - Null-Invariant Measures
  - Comparison of Interestingness Measures

# References

- R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", in Proc. of SIGMOD'93
- R. J. Bayardo, "Efficiently mining long patterns from databases", in Proc. of SIGMOD'98
- N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules", in Proc. of ICDT'99
- J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent Pattern Mining: Current Status and Future Directions", Data Mining and Knowledge Discovery, 15(1): 55-86, 2007
- R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", VLDB'94
- A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large databases", VLDB'95
- J. S. Park, M. S. Chen, and P. S. Yu, "An effective hash-based algorithm for mining association rules", SIGMOD'95
- S. Sarawagi, S. Thomas, and R. Agrawal, "Integrating association rule mining with relational database systems: Alternatives and implications", SIGMOD'98
- M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li, "Parallel algorithm for discovery of association rules", Data Mining and Knowledge Discovery, 1997
- J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation", SIGMOD'00

# References (cont.)

- M. J. Zaki and Hsiao, "CHARM: An Efficient Algorithm for Closed Itemset Mining", SDM'02
- J. Wang, J. Han, and J. Pei, "CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets", KDD'03
- C. C. Aggarwal, M.A., Bhuiyan, M. A. Hasan, "Frequent Pattern Mining Algorithms: A Survey", in Aggarwal and Han (eds.): Frequent Pattern Mining, Springer, 2014
- C. C. Aggarwal and P. S. Yu. A New Framework for Itemset Generation. PODS'98
- S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD'97
- M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM'94
- E. Omiecinski. Alternative Interest Measures for Mining Associations. TKDE'03
- P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. KDD'02
- T. Wu, Y. Chen and J. Han, Re-Examination of Interestingness Measures in Pattern Mining: A Unified Framework, Data Mining and Knowledge Discovery, 21(3):371-397, 2010