



# WHICH PATTERNS ARE INTERESTING?—PATTERN EVALUATION METHODS

## INTRODUCTION

- Most association rule mining algorithms employ a support–confidence framework.
- Although minimum support and confidence thresholds help weed out or exclude the exploration of a good number of uninteresting rules, many of the rules generated are still not interesting to the users.
- Unfortunately, this is especially true when mining at low support thresholds or mining for long patterns.
- This has been a major bottleneck for successful application of association rule mining



# STRONG RULES ARE NOT NECESSARILY INTERESTING

- Whether or not a rule is interesting can be assessed either subjectively or objectively.
- Ultimately, only the user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another.
- However, objective interestingness measures, based on the statistics “behind” the data, can be used as one step toward the goal of weeding out uninteresting rules that would otherwise be presented to the user.



## A MISLEADING “STRONG” ASSOCIATION RULE

- Given minimum support = 30% and a minimum confidence = 60%.
- The following association rule is discovered:
- $\text{buys}(X, \text{“computer games”}) \Rightarrow \text{buys}(X, \text{“videos”})$
- [support = 40%, confidence = 66%].

2 × 2 Contingency Table Summarizing the Transactions with Respect to Game and Video Purchases

	<i>game</i>	$\overline{\text{game}}$	$\Sigma_{\text{row}}$
<i>video</i>	4000	3500	7500
$\overline{\text{video}}$	2000	500	2500
$\Sigma_{\text{col}}$	6000	4000	10,000



## A MISLEADING “STRONG” ASSOCIATION RULE

- The above Rule is a strong association rule and would therefore be reported, since its support value of  $4000/10,000 = 40\%$  and confidence value of  $4000/6000 = 66\%$  satisfy the minimum support and minimum confidence thresholds.
- The above Rule is misleading because the probability of purchasing videos is 75%, which is even larger than 66%.
- The computer games and videos are negatively associated because the purchase of one of these items actually decreases the likelihood of purchasing the other.



# FROM ASSOCIATION ANALYSIS TO CORRELATION ANALYSIS

- The support and confidence measures are insufficient at filtering out uninteresting association rules.
- To tackle this weakness, a correlation measure can be used to augment the support–confidence framework for association rules.
- This leads to correlation rules of the form

$A \Rightarrow B$  [support, confidence, correlation]



## CORRELATION MEASURES: LIFT

- **Lift** is the occurrence of itemset A is independent of the occurrence of itemset B  
if  $P(A \cup B) = P(A)P(B)$
- otherwise, itemsets A and B are dependent and correlated as events.
- The lift between the occurrence of A and B can be measured by computing

$$\text{lift}(A, B) = P(A \cup B) / P(A)P(B)$$

If the resulting value  $< 1$ , then A & B is negatively correlated

If the resulting value  $> 1$ , then A & B is positively correlated



## CORRELATION ANALYSIS USING LIFT

- Probability of purchasing a computer game is

$$P(\{\text{game}\}) = 0.60$$

- Probability of purchasing a video is

$$P(\{\text{video}\}) = 0.75$$

- Probability of purchasing both is

$$P(\{\text{game}, \text{video}\}) = 0.40.$$

- The lift of Rule is

$$P(\{\text{game}, \text{video}\}) / (P(\{\text{game}\}) \times P(\{\text{video}\}))$$

$$= 0.40 / (0.60 \times 0.75) = 0.89$$

As  $0.89 < 1$ , there is negative correlation

2 × 2 Contingency Table Summarizing the Transactions with Respect to Game and Video Purchases

	<i>game</i>	$\overline{\text{game}}$	$\Sigma_{\text{row}}$
<i>video</i>	4000	3500	7500
$\overline{\text{video}}$	2000	500	2500
$\Sigma_{\text{col}}$	6000	4000	10,000





## CORRELATION MEASURES: $\chi^2$ MEASURE

- The  $\chi^2$  value (also known as the Pearson  $\chi^2$  statistic) is computed as:
- where  $O_{ij}$  is the observed frequency (i.e., actual count) of the joint event  $(A_i, B_j)$
- $E_{ij}$  is the expected frequency of  $(A_i, B_j)$ , which can be computed as

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

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n},$$


$$e_{11} = \frac{\text{count}(\text{game}) \times \text{count}(\text{video})}{n}$$


Table Contingency Table, Now with the Expected Values

	<i>game</i>	$\overline{\text{game}}$	$\Sigma_{\text{row}}$
<i>video</i>	4000 (4500)	3500 (3000)	7500
$\overline{\text{video}}$	2000 (1500)	500 (1000)	2500
$\Sigma_{\text{col}}$	6000	4000	10,000

$$\chi^2 = \Sigma \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(4000 - 4500)^2}{4500} + \frac{(3500 - 3000)^2}{3000} + \frac{(2000 - 1500)^2}{1500} + \frac{(500 - 1000)^2}{1000} = 555.6.$$

- Because the  $\chi^2$  value is greater than 1, and the observed value of the slot (game, video) = 4000, which is less than the expected value of 4500, buying game and buying video are negatively correlated

