Mining Negative Correlations

Rare Patterns vs. Negative Patterns

❑ Rare patterns

    ❑ Very low support but interesting (e.g., buying Rolex watches)

    ❑ How to mine them? Setting individualized, group-based min-support thresholds for different groups of items

❑ Negative patterns

    ❑ Negatively correlated: Unlikely to happen together

    ❑ Ex.: Since it is unlikely that the same customer buys both a Ford Expedition (an SUV car) and a Ford Fusion (a hybrid car), buying a Ford Expedition and buying a Ford Fusion are likely negatively correlated patterns

    ❑ How to define negative patterns?

Defining Negative Correlated Patterns

- ❑ A support-based definition
  - ❑ If itemsets A and B are both frequent but rarely occur together, i.e.,
    $sup(A \cup B) \ll sup(A) \times sup(B)$
  - ❑ Then A and B are negatively correlated

    Does this remind you the definition of *lift*?
- ❑ Is this a good definition for large transaction datasets?
- ❑ Ex.: Suppose a store sold two needle packages A and B 100 times each, but only one transaction contained both A and B
  - ❑ When there are in total 200 transactions, we have
    - ❑ $s(A \cup B) = 0.005$, $s(A) \times s(B) = 0.25$, $s(A \cup B) \ll s(A) \times s(B)$
  - ❑ But when there are 10^5 transactions, we have
    - ❑ $s(A \cup B) = 1/10^5$, $s(A) \times s(B) = 1/10^3 \times 1/10^3$, $s(A \cup B) > s(A) \times s(B)$
  - ❑ What is the problem?—Null transactions: The support-based definition is not null-invariant!

3

Defining Negative Correlation: Need Null-Invariance in Definition

- ❏ A good definition on negative correlation should take care of the null-invariance problem
  - ❏ Whether two itemsets A and B are negatively correlated should not be influenced by the number of null-transactions
  - ❏ A Kulczynski measure-based definition
    - ❏ If itemsets A and B are frequent but $(P(A|B) + P(B|A))/2 < \epsilon$, where ϵ is a negative pattern threshold, then A and B are negatively correlated
- ❏ For the same needle package problem:
  - ❏ No matter there are in total 200 or 10^5 transactions
  - ❏ If $\epsilon = 0.01$, we have $(P(A|B) + P(B|A))/2 = (0.01 + 0.01)/2 < \epsilon$

4