

Chapter 7: Condensed representations

1

Condensed representations

- Motivation
- Closed sets
- Free sets
- Bounds
- Non-derivable sets
- k -Free sets

2

Motivation

- Too many patterns
- Incomprehensible
- A lot of redundant patterns
- Less patterns might be easier and faster to compute!

3

Example

- $fr(\{A, B\}) = fr(\{A\})$, i.e., $conf(\{A\} \Rightarrow \{B\}) = 1$
- $\Rightarrow fr(X \cup \{A, B\}) = fr(X \cup \{A\})$
- no need to count the frequencies of sets $X \cup \{A, B\}$ from the database!
- If there are lots of rules with confidence 1, then a significant amount of work can be saved
- \rightarrow useful with strong correlations and in dense 0/1 relations

4

Example

- $fr(\{C\}) = 0.6$
 $fr(\{A\}) = fr(\{A, C\}) = 0.5$
 $fr(\{B\}) = fr(\{E\}) = fr(\{B, E\}) = 0.3$
 $fr(\{B, C\}) = fr(\{C, E\}) = fr(\{B, C, E\}) = 0.2$
 $fr(\{A, B\}) = fr(\{A, E\}) = fr(\{A, B, C\}) = fr(\{A, B, E\}) =$
 $fr(\{A, C, E\}) = fr(\{A, B, C, E\}) = 0.1$

5

Closures of item sets

- The *closure* of $X \subseteq R$ in r is

$$X^+ = \{A \in R \mid conf(X \Rightarrow \{A\}, r) = 1\}$$

- in other words

$$X^+ = \bigcap_{t \in \mathcal{M}(X)} t$$

- general properties of closures:

- $X \subseteq X^+$
- $(X^+)^+ = X^+$
- $Y \subseteq X \Rightarrow Y^+ \subseteq X^+$

6

Closed sets

- item set X is *closed* iff $X^+ = X$
- the collection of all closed sets:

$$\mathcal{C}\ell = \{X^+ \mid X \subseteq R\}$$

Closed sets as condensed representation

- closed sets and their frequencies alone are a sufficient representation for the frequencies of all sets:
- either X is itself closed or some of its supersets is — in any case X^+ is closed and so its frequency is known
- which of the closed supersets of X is the closure X^+ ?
the one with the greatest frequency (why?)
- thus: $fr(X) = \max\{fr(Y) \mid Y \in \mathcal{C}\ell \text{ and } X \subseteq Y\}$

Free sets

- also called *generators* or *key patterns*
- complementary to closed sets
- set X is *free* iff there is no proper subset $Y \subset X$ such that $Y^+ = X^+$
- the collection of all free sets:

$$\mathcal{Free} = \{X \subseteq R \mid X^+ \neq Y^+ \text{ for all } Y \subset X\}$$

- \mathcal{Free} is a downward closed set!

Free sets as condensed representation

- Frequent Free sets are not a sufficient representation for all frequent sets! (why?)
- $\mathcal{Bd}^-(\mathcal{Free} \cap \mathcal{F}(r)) \cap \mathcal{Free} = \mathcal{Bd}^-(\mathcal{F}(r))$ is needed
- $fr(X) = \min\{fr(Y) \mid Y \in \mathcal{Free} \text{ and } Y \subseteq X\}$ or infrequent

Example

- Closed sets:
 $\{C\}, \{A, C\}, \{B, E\}, \{B, C, E\}, \{A, B, C, E\}$
- Generators:
 $\{C\}$
 $\{A\}$
 $\{B\}, \{E\}$
 $\{B, C\}, \{C, E\}$
 $\{A, B\}, \{A, E\}$
- frequency of $\{A, B, E\}$?
- frequency of $\{B\}$?

11

Some properties of closed and free sets

- discovery of only frequent closed sets or frequent free sets can be much more efficient than explicit discovery of all frequent sets
- Free sets are easiest to compute (downward closure!)
- Each row is a closed set: $X \in r \Rightarrow X \in \mathcal{Cl}$
- The collection of closed sets is obtained as intersections of rows:
$$\mathcal{Cl} = \{\bigcap_{X \in P} X \mid P \subseteq r\}$$
- $|\mathcal{Free}| \geq |\mathcal{Cl}| \geq |r'|$ where r' is the (non multi) set of rows in r

12

Bounds

- Can we estimate or bound the frequency of an itemset?

- Yes:

$$0 \leq fr(\{A, B\}) \leq fr(\{A\})$$

$$0 \leq fr(\{A, B\}) \leq fr(\{B\})$$

- We can do better ...

13

Inclusion - Exclusion principle

An itemset $X \cup \overline{Y}$, is contained in a transaction, if X is contained in that transaction and non of the items in Y are contained in that transaction.

$$fr(\{A, \overline{B}, \overline{C}\}) = fr(\{A\}) - fr(\{A, B\}) - fr(\{A, C\}) + fr(\{A, B, C\})$$

$$fr(\{A, \overline{B}, \overline{C}\}) \geq 0$$

$$fr(\{A, B, C\}) \geq fr(\{A, B\}) + fr(\{A, C\}) - fr(\{A\})$$

14

Inclusion - Exclusion principle

- Given an itemset X , for every possible $Y \subseteq X$, we can create such a formula, based on the inclusion-exclusion formula of

$$fr((X \setminus Y) \cup \bar{Y})$$

$$fr(\overline{abcd})$$

$$fr(abcd) \geq fr(abc) + fr(abd) + fr(acd) + fr(bcd) - fr(ab) - fr(ac) - fr(ad) \\ - fr(bc) - fr(bd) - fr(cd) + fr(a) + fr(b) + fr(c) + fr(d) - fr(\emptyset)$$

$$fr(\overline{abcd}), fr(\overline{bacd}), fr(\overline{cabd}), fr(\overline{dabc})$$

$$fr(abcd) \leq fr(a) - fr(ab) - fr(ac) - fr(ad) + fr(abc) + fr(abd) + fr(acd)$$

$$fr(abcd) \leq fr(b) - fr(ab) - fr(bc) - fr(bd) + fr(abc) + fr(abd) + fr(bcd)$$

$$fr(abcd) \leq fr(c) - fr(ac) - fr(bc) - fr(cd) + fr(abc) + fr(acd) + fr(bcd)$$

$$fr(abcd) \leq fr(d) - fr(ad) - fr(bd) - fr(cd) + fr(abd) + fr(acd) + fr(bcd)$$

$$fr(\overline{abcd}), fr(\overline{acbd}), \dots, fr(\overline{cdab})$$

$$fr(abcd) \geq fr(abc) + fr(abd) - fr(ab)$$

$$fr(abcd) \geq fr(abc) + fr(acd) - fr(ac)$$

$$fr(abcd) \geq fr(abd) + fr(acd) - fr(ad)$$

$$fr(abcd) \geq fr(abc) + fr(bcd) - fr(bc)$$

$$fr(abcd) \geq fr(abd) + fr(bcd) - fr(bd)$$

$$fr(abcd) \geq fr(acd) + fr(bcd) - fr(cd)$$

$fr(ab\bar{c}\bar{d}), fr(ab\bar{d}\bar{c}), fr(ac\bar{d}\bar{b}), fr(bc\bar{d}\bar{a}),$

$$fr(abcd) \leq fr(abc)$$

$$fr(abcd) \leq fr(abd)$$

$$fr(abcd) \leq fr(acd)$$

$$fr(abcd) \leq fr(bcd)$$

$fr(abcd)$

$$fr(abcd) \geq 0$$

In general

$$fr(I) \leq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} fr(J) \quad \text{If } |I \setminus J| \text{ odd}$$

$$fr(I) \geq \sum_{X \subseteq J \subset I} (-1)^{|I \setminus J|+1} fr(J) \quad \text{If } |I \setminus J| \text{ even}$$

Lower and Upper bounds

- for any candidate itemset, we can compute a lower and an upper bound on its frequency, given the frequencies of all of its subsets!
- What if lower bound equals upper bound?
- \rightarrow frequency is derivable, and we do not have to compute it anymore using an additional pass through the data!

Non-derivable sets

- The collection of all non-derivable frequent sets is a sufficient representation of all frequent sets!
- The collection of all non-derivable frequent sets is downward closed!
- And thus, easy to compute
- BUT, an exponential number of formulas need to be computed
- each formula contains an exponential number of terms
- BUT, numbers are small ...

23

Properties of non-derivable sets

- The width of the interval (given by lower and upper bound) shrinks exponentially with the size of the set!
- **Proof**
- \rightarrow the size of the largest non-derivable set is at most $\log |r|$.

24

k -Free sets

- Given an itemset and all of its subsets, together with their supports, compute lower and upper bound using Inclusion-Exclusion for the formulas obtained by using at most k negated items!
- If the itemset is non-derivable, compute its actual support.
- If the actual support is not equal to the lower or the upper bound, then the itemset is called k -Free.
- For $k = 1$, k -Free gives exactly the Free sets (as before)
- For $k_1 \geq k_2 \geq 1$, k_1 -Free \subseteq k_2 -Free

k -Free sets

- The set of k -Free sets is not a sufficient representation of all frequent itemsets!
- We need part of the negative border.