Closed sets, generators, condensed representations

Closed sets, generators, condensed representations

- Closure, closed set, generator
- Algorithms
- Condensed representations
- Experimental results
- Literature for this part

Example

- $\mathit{fr}(\{A, B\}) = \mathit{fr}(\{A\})$, i.e., $\mathit{conf}(\{A\} \Rightarrow \{B\}) = 1$
- $\Rightarrow \mathit{fr}(X \cup \{A, B\}) = \mathit{fr}(X \cup \{A\})$
- 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 signicant amount of work can be saved
- $\rightarrow$ useful with strong correlations and in dense 0/1 relations

Example

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

Closures of item sets

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

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

- i.e., the closure of X is the greatest set that occurs on all the rows in r on which X occurs
- general properties of closures:
  - $X \subseteq X^+$
  - $(X^+)^+ = X^+$
  - $Y \subseteq X \Rightarrow Y^+ \subseteq X^+$

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 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
- but which of the closed supersets of X is the closure X^+? the one with the greatest frequency (why?)
- thus: $\mathit{fr}(X) = \max\{\mathit{fr}(Y) \mid Y \in \mathcal{C}\ell \text{ and } X \subseteq Y\}$

Generators

- generators (also called *key patterns*) are a complementary concept
- item set X is a *generator* of X^+ iff there is no proper subset $Y \subset X$ such that $Y^+ = X^+$
- the collection of all generators:

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

- generators, too, are a sufficient representation for all sets:
- $\mathit{fr}(X) = \min\{\mathit{fr}(Y) \mid Y \in \mathcal{G}en \text{ and } Y \subseteq X\}$
- discovery of only frequent closed sets or frequent generators can be much more efficient than explicit discovery of all frequent sets

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 $\{B\}$?

Some properties of closed sets

- Each row is a closed set:
  $X \in r \Rightarrow X \in \mathcal{C}\ell$
- The collection of closed sets is obtained as intersections of rows:
  $\mathcal{C}\ell = \{\bigcap_{X \in P} X \mid P \subseteq r\}$
- $|\mathcal{G}en| \geq |\mathcal{C}\ell| \geq |r'|$ where r' is the (non multi) set of rows in r

Discovery of all frequent generators

- **Lemma** If $X \in \mathcal{G}en$ then $Y \in \mathcal{G}en$ for all subsets $Y \subseteq X$
- thus: being a generator is a downwards monotone property, just like being a frequent set
  $\Rightarrow$ the levelwise algorithm and Apriori in special are directly applicable

- recall Apriori algorithm:

  1. $\mathcal{C}_1 := \{\{A\} \mid A \in R\}$;
  2. $l := 1$;
  3. **while** $\mathcal{C}_l \neq \emptyset$ **do**
  4. compute $\mathcal{F}_l(r) := \{X \in \mathcal{C}_l \mid \mathit{fr}(X, r) \geq \mathit{min\_fr}\}$;
  5. $l := l + 1$;
  6. compute $\mathcal{C}_l := \mathcal{C}(\mathcal{F}_{l-1}(r))$;
  7. **for** all l and **for** all $X \in \mathcal{F}_l(r)$ **do** output X and $\mathit{fr}(X, r)$;

- refine $\mathcal{F}_l(r)$ and Step 4 to select frequent generators:

  4. compute $\mathcal{F}_l(r) := \{X \in \mathcal{C}_l \mid \mathit{fr}(X, r) \geq \mathit{min\_fr}$ and $\mathit{fr}(X, r) \neq \mathit{fr}(Y, r)$ for all $Y \subset X\}$;

- add a step that outputs generators in the negative border:

  8. **for** all l and **for** all $X \in \mathcal{C}_l \setminus \mathcal{F}_l(r)$ such that $\mathit{fr}(X, r) < \mathit{min\_fr}$ **do** output "X is in $\mathcal{B}d^-(\mathcal{G}en \cap \mathcal{F}(r, \mathit{min\_fr}))$";

- the negative border is needed above to determine (border of) the collection of frequent sets:
- X is frequent iff there is no Y in the border such that $Y \subseteq X$
- otherwise the frequency of X is the minimum of the frequencies of its subsets in the output of the algorithm
- frequent generators and the negative border are a *condensed representation* of frequent sets

Discovery of all frequent closures

- the easy way:
  1. find all frequent generators
  2. compute closures of the generators from the database

Condensed representations: a formulation

- a class of structures $\mathcal{S}tr = \{s_i \mid i \in I\}$, where the index set I can be finite or infinite
- examples:
  - $\mathcal{S}tr_{R,01}$, the class of all 0/1 relations over the attributes R
  - $\mathcal{S}tr_{R,D}$, the class of all relations over the domain D and attributes R
  - $\mathcal{S}tr_E$, the set of all event sequences with event types from the set E.
- $\mathcal{Q} = \{Q_1, \ldots, Q_p\}$, a finite class of queries for $\mathcal{S}tr$
- $Q(s) \in [0, 1]$ for all $Q \in \mathcal{Q}$, $s \in \mathcal{S}tr$

- example query classes for $\mathcal{S}tr_{R,01}$:
  - the *conjunctive* queries $\mathcal{Q}_\wedge = \{Q_X : r \mapsto \mathit{fr}(X, r) \mid X \subseteq R\}$
  - the *disjunctive* queries

    $\mathcal{Q}_\vee = \{Q'_X : r \mapsto \frac{|\{t \in r \mid t[A]=1 \text{ for some } A \in X\}|}{|r|} \mid X \subseteq R\}$

- an ε-*adequate representation* for $\mathcal{S}tr$ with respect to $\mathcal{Q}$:
  - $\mathcal{R}ep = \{r_i \mid i \in I\}$, a class of structures
  - $m : \mathcal{Q} \times \mathcal{R}ep \to [0,1]$, a query evaluation function
  - $|Q(s_i) - m(Q, r_i)| \leq \varepsilon$ for all $Q \in \mathcal{Q}$ and $s_i \in \mathcal{S}tr$
- *min_fr*/2-adequate representation for frequent sets:
  - original class of structures: $\mathcal{S}tr_{R,01}$
  - query class: the conjunctive frequency queries $\mathcal{Q}_\wedge$
  - condensed class of structures: frequent closed sets $\mathcal{F}(r, \mathit{min\_fr}) \cap \mathcal{C}\ell$
  - query evaluation function of $Q_X \in \mathcal{Q}_\wedge$: $r \mapsto \max(\{\mathit{fr}(Y,r) \mid Y \in \mathcal{F}(r, \mathit{min\_fr}) \cap \mathcal{C}\ell \text{ and } X \subseteq Y\} \cup \{\mathit{min\_fr}/2\})$

- Lossless (0-adequate) condensed representations for frequent sets and their frequencies
  - frequent generators and their negative border
  - frequent closed sets
- Lossless condensed representations for the collection of frequent sets (not frequencies)
  - positive border
  - negative border

- Approximate (ε-adequate, $\varepsilon \geq 0$) condensed representations:
  - a random sample
  - δ-free sets, almost closures
  - disjunction-free sets, disjunction-free generators
  - ...

δ-free sets/almost closures

- idea: relax the definition of closure
- B is "almost" in the closure of A if
  $|\mathcal{M}(\{A\}, r)| - |\mathcal{M}(\{A, B\}, r)| \leq \delta$
- do not output X if it is almost in the closure of some other set
- allows limited approximation error; can reduce the size of output and running time considerably

Experimental results

Dataset, *min_fr*	$\lvert\mathcal{F}(r, \textit{min\_fr})\rvert$	db scans	$\lvert\mathcal{F}(r, \textit{min\_fr}) \cap \mathcal{C}\ell\rvert$	db scans	$\lvert\mathcal{F}(r, \textit{min\_fr}) \cap 4 - \mathcal{C}\ell\rvert$	db scans
ANPE, 0.005			412092	11	182829	10
ANPE, 0.05	25781	11	11125	9	10931	9
ANPE, 0.1	6370	10	2898	8		
ANPE, 0.2	1516	9	638	7		
census, 0.005			85950	9	39036	8
census, 0.05	90755	13	10513	9	5090	8
census, 0.1	26307	12	4041	9		
census, 0.2	5771	11	1064	9		

[Boulicaut & Bykowski, PAKDD 2000]

Literature

- Closed sets and generators:
  N. Pasquir et al.: Discovering frequent closed itemsets for association rules, ICDT 1999.
- δ-free sets/almost closures:
  J-F. Boulicaut et al.: Approximation of frequency queries by means of free-sets. PKDD 2000.
- (Condensed representations:
  H. Mannila and H. Toivonen: Multiple uses for frequent sets and condensed representations, KDD 1996.)
- (Original work on closed sets also by M. Zaki et al.)