Association Analysis

Part 2

Limitations of the Support/Confidence framework

1. Redundancy: many of the returned patterns may refer to the same piece of information
2. Difficult control of output size: it is hard to predict how many patterns will be returned for given support/confidence thresholds
3. Significance: are the returned patterns significant, interestings?

In what follows we will address the above issues, limiting the discussion to frequent itemsets. Some of the ideas can be extended to association rules.

Closed itemsets

GOAL: Devise a lossless succint representation of the frequent itemsets.

Consider a dataset T of N transactions over the set of items I, and a support threshold minsup.

Definition (Closed Itemset)

An itemset $X \subseteq I$ is *closed* w.r.t. T if for each superset $Y \supset X$ we have $\mathrm{Supp}(Y) < \mathrm{Supp}(X)$.

Notation

- $\mathrm{CLO}_T = \{X \subseteq I \ : \ X \text{ is closed w.r.t } T\}$
- $\text{CLO-F}_{T,\mathrm{minsup}} = \{X \in \mathrm{CLO}_T \ : \ \mathrm{Supp}(X) \geq \text{minsup}\}$

Observation: the empty itemset, whose support is 1, is closed if and only if it is the only itemset of support 1.

Maximal itemsets

Definition (Maximal Itemset)

An itemset $X \subseteq I$ is *maximal* w.r.t. T and minsup if $\mathrm{Supp}(X) \geq$ minsup and for each superset $Y \supset X$ we have $\mathrm{Supp}(Y) <$ minsup.

Notation

- $\mathrm{MAX}_{T,\mathrm{minsup}} = \{X \subseteq I \ : \ X \text{ is maximal w.r.t. } T\}$

When clear from the context, the subscripts will be omitted

Exercise

Dataset T	
TID	ITEMs
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

For minsup $= 2/5$, identify:

(a) A maximal itemset

(b) A frequent closed itemset which is not maximal

(c) A closed itemset which is not frequent

Answer

(a) ACD (Support $= 2/5$)

(b) AC (Support $= 3/5$)

(c) ACDE (Support $= 1/5$)

Exercise (cont'd)

Dataset T	
TID	ITEMs
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

For minsup $= 2/5$, identify:

(d) Set F (frequent itemsets)

(e) Set CLO-F

(f) Set MAX

Answer

(d) F = A,B,C,D,E,AB,AC,AD,BC,CD,CE,DE,ABC,ACD

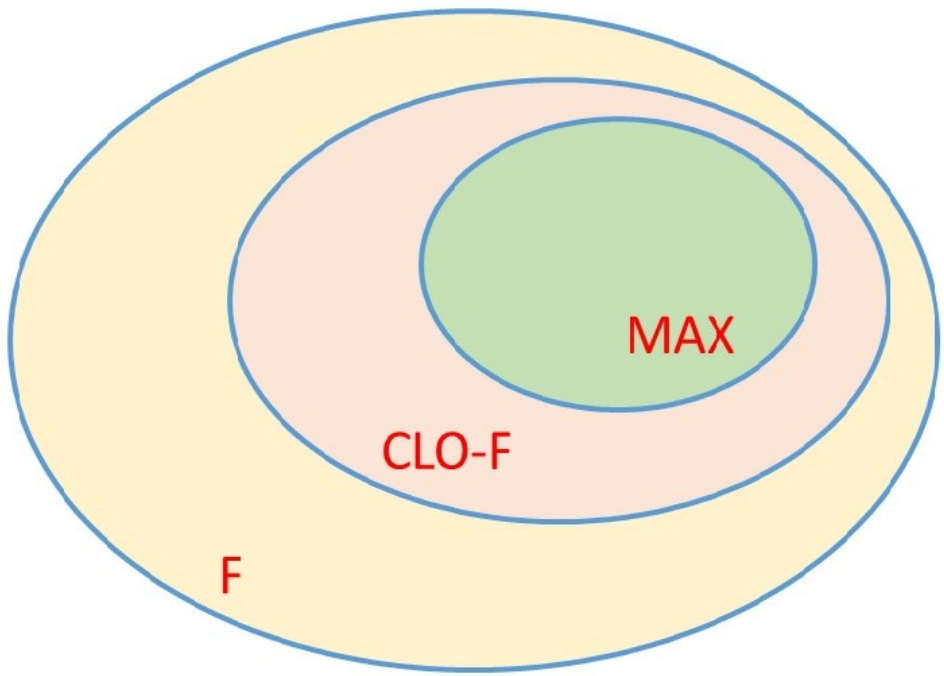
(e) CLO-F = C,D,E,AC,BC,CE,DE,ABC,ACD

(f) MAX = CE, DE, ABC,ACD

Closed/maximal itemsets

The following properties can be easily shown (exercise):

- For each itemset $X \subseteq I$ there exists $X' \in \text{CLO}$ such that $X' \supseteq X$ and $\text{Supp}(X') = \text{Supp}(X)$
- For each frequent itemset $X \in \text{F}$ there exists $X' \in \text{MAX}$ such that $X' \supseteq X$
- $\text{MAX} \subseteq \text{CLO-F} \subseteq \text{F}$.



Observations

- MAX and CLO-F provide succint representations of F: from the above properties we immediately conclude that the set of frequent itemsets *coincides* with the set of all subsets of the maximal itemsets, or, equivalently, with the set of all subsets of the frequent closed itemsets.
- MAX provides a lossy representation of F: in general, the support of a frequent itemset *cannot be derived* from the maximal itemsets and their supports.
- CLO-F provides a lossless representation of F: the support of any frequent itemset *can be derived* from the frequent closed itemsets and their supports. See next few slides.

Representativity of closed itemsets

For $X \subseteq I$, let T_X denote the set of transactions where X occurs.

Definition (Closure)

$$\text{Closure}(X) = \bigcap_{t \in T_X} t.$$

Example:

Dataset T	
TID	ITEMs
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

- $X = AB$
- $\text{Closure}(X) = ABC$

Representativity of closed itemsets (cont'd)

Theorem

Let $X \subseteq I$. We have:

1. $X \subseteq \text{Closure}(X)$
2. $\text{Supp}(\text{Closure}(X)) = \text{Supp}(X)$
3. $\text{Closure}(X)$ *is closed.*

Proof

1. Immediate, since $X \subseteq t$, for each $t \in T_X$.
2. The first property implies $\text{Supp}(X) \geq \text{Supp}(\text{Closure}(X))$. Let N_X be the number of transactions in T_X. Hence, $\text{Supp}(X) = N_X/N$. Now, by construction $\text{Closure}(X)$ is contained in each transaction of T_X, therefore $\text{Supp}(\text{Closure}(X)) \geq N_X/N = \text{Supp}(X)$.

Representativity of closed itemsets (cont'd)

Proof (cont'd).

3. By contradiction, suppose $\exists Y \supset \mathrm{Closure}(X)$, such that $\mathrm{Supp}(Y) = \mathrm{Supp}(\mathrm{Closure}(X))$. Hence, by the second property, $\mathrm{Supp}(Y) = \mathrm{Supp}(X)$. Let T_Y be the set of transactions that contain Y. Now, the relation $Y \supset \mathrm{Closure}(X) \supseteq X$ implies $T_Y \subseteq T_X$. On the other hand, since $\mathrm{Supp}(Y) = \mathrm{Supp}(X)$ we conclude that $T_Y = T_X$. Thus, Y must be contained in each transaction $t \in T_X$, hence, it is contained in $\mathrm{Closure}(X)$, which gives the contradiction.

□

Corollary

For each $X \subseteq I$, $\mathrm{Supp}(X) = \max\{\mathrm{Supp}(Y) \ : \ Y \supseteq X \wedge Y \in \mathrm{CLO}\}$.

Exercise

Prove the corollary.

Observations

- A consequence of the previous corollary is that from the frequent closed itemsets and their supports one can derive all frequent itemsets and their supports. In this sense, frequent closed itemsets and their supports provide a lossless representation of the frequent itemsets and their supports.
- Each (frequent) closed itemset Y can be regarded as a representative of all those (possibly many) itemsets X such that $\mathrm{Closure}(X) = Y$.
- There exist efficient algorithms for mining maximal or frequent closed itemsets.
- Notions of closure similar to the ones used for itemsets are employed in other mining contexts (e.g., graph mining, dna sequence analysis, etc.).

Exponentiality of maximal/frequent closed itemsets

Although maximal and frequent closed itemsets provide in practice succint representations of the frequent itemsets, still there are pathological instances where the number of maximal itemsets, hence the number of frequent closed itemsets is exponential in the input size. The following exercise provides an example.

Exercise

Let $I = \{1, 2, \ldots, 2n\}$ for some integer $n > 0$, and let $T = \{t_1, t_2, \ldots, t_{2n}\}$ be a set of $2n$ transactions, where $t_j = I - \{j\}$, for every $1 \leq j \leq 2n$.

1. For every $X \subseteq I$ determine $\mathrm{Supp}(X)$ as a function of n and of the length of X.
2. Using the result of the first point, determine the number of maximal itemsets w.r.t. minsup=1/2 and argue that it is exponential in the input size.
3. What can you say about frequent closed itemsets?

Top-K frequent (closed) itemsets

How about if we impose explicitly a limit on the output size?

Let T be a dataset of N transactions over a set I of d items. Let $\mathrm{F}_{T,s}$ and $\text{CLO-F}_{T,s}$ denote, respectively, the sets of frequent itemsets and frequent closed itemsets w.r.t. threshold s. For $K > 0$ define

$$\begin{aligned} s(K) &= \max\{s : |\mathrm{F}_{T,s}| \geq K\} \\ sc(K) &= \max\{s : |\text{CLO-F}_{T,s}| \geq K\} \end{aligned}$$

Then

- Top-K frequent itemsets w.r.t. $T = \mathrm{F}_{T,s(K)}$
- Top-K frequent closed itemsets w.r.t. $T = \text{CLO-F}_{T,sc(K)}$

Example

Top-K frequent itemsets (with supports)

Dataset T	
TID	ITEMs
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

- $K = 1$: C(4/5)
- $K = 2 \div 7$: C(4/5), A(3/5), B(3/5), D(3/5), E(3/5), AC(3/5), BC(3/5)
- $K = 8 \div 13$: C(4/5), A(3/5), B(3/5), D(3/5), E(3/5), AC(3/5), BC(3/5), AB(2/5), AD(2/5), CE(2/5), DE(2/5), ABC(2/5), ACD(2/5)
- etc.

Example (cont'd)

Top-K frequent closed itemsets (with supports)

Dataset T	
TID	ITEMs
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE

- $K = 1$: C(4/5)
- $K = 2 \div 5$: C(4/5), D(3/5), E(3/5), AC(3/5), BC(3/5)
- $K = 6 \div 9$: C(4/5), D(3/5), E(3/5), AC(3/5), BC(3/5) CE(2/5), DE(2/5), ABC(2/5), ACD(2/5)
- etc.

Observations

- K is the target number of patterns, but the acutal number of Top-K frequent (closed) itemsets could be larger than K (not much larger in practice).
- How well does the parameter K control the output size?
  - The next theorem (without proof) shows that for Top-K frequent closed itemsets, K provides a somewhat tight control on the output size.
  - The exercise in the next slide, shows that this is not the case for Top-K frequent itemsets

Theorem

For $K > 0$, the Top-K frequent closed itemsets are $O(d \cdot K)$, where d is the number of items.

Esercise

Let d be an even integer, and define T as the set of the following $N = (3/2)d$ transactions over $I = \{1, 2, \ldots, d\}$

$$\begin{aligned} t_i &= \{i\} \quad 1 \leq i \leq d \\ t_{d+i} &= I - \{i\} \quad 1 \leq i \leq d/2. \end{aligned}$$

1. Identify the itemsets of support $> 1/3$ and the itemsets of support $= 1/3$.
2. Using the result of the previous point, show that the number of Top-K frequent itemsets, with $K = d$, is exponential in d.

Significance

How do we measure the significance/interest of itemsets/rules?

- Subjective measures: user-defined criteria based on domain knowledge.
- Objective measures: quantitative criteria, often based on statistics, such as *support* and *confidence*, for which the user fixes suitable thresholds

Are support/confidence adequate to capture significance? In general, the answer is "NO", but with some amendments their effectiveness can be improved.

Beyond Confidence

Consider a dataset with 1000 transactions from a supermarket and let the following contingency table.

	coffee	$\overline{\text{coffee}}$	
tea	150	50	200
$\overline{\text{tea}}$	650	150	800
	800	200	1000

The table should be read as follows:

- 150 transactions contain tea and coffee;
- 50 transactions contain tea but not coffee;
- 650 transactions contain coffee but not tea;
- 150 transactions contain neither tea nor coffee;
- Altogether, of the 1000 transactions: 200 contain tea, 800 do not contain tea, 800 contain coffee, and 200 do not contain coffee (*marginal totals*).

Beyond Confidence (cont'd)

	coffee	$\overline{\text{coffee}}$	
tea	150	50	200
$\overline{\text{tea}}$	650	150	800
	800	200	1000

Consider rule

$$r \ : \ \text{tea} \rightarrow \text{coffee}.$$

We have:

- $\text{Supp}(r) = 0.15$ and $\text{Conf}(r) = 0.75$.
- $\text{Supp}(\text{coffee}) = 0.8$.

Observation: While $\text{Conf}(r)$ seems relatively high, in fact a random customer is more likely to buy coffee than a customer who bought tea.

Lift

The following measure is often used to assess the signficiance of high-confience rules.

Definition (Lift)

Given a dataset T and an association rule $r : X \to Y$, define

$$\text{Lift}(r) = \frac{\text{Conf}(r)}{\text{Supp}(Y)} = \frac{\text{Supp}(X \cup Y)}{\text{Supp}(X) \cdot \text{Supp}(Y)}.$$

- $\text{Lift}(r)$ is sometimes referred to as the *Interest Factor* of the pair of itemsets X, Y.
- The denominator represent the expected support of $X \cup Y$ would X and Y occur independently in the transactions.

Lift (cont'd)

- Lift(r) $\simeq 1 \Rightarrow X$ and Y are uncorrelated
- Lift(r) $\gg 1 \Rightarrow X$ and Y are positively correlated
- Lift(r) $\ll 1 \Rightarrow X$ and Y are negatively correlated

Usually, association rules are mined with the traditional support-confidence framework and then they are sorted by decreasing lift. Those rules with high lift (much larger than 1) are selected as significant.

In the previous example, Lift(tea $\rightarrow$ coffee) $= 0.9375$, hence rule tea $\rightarrow$ coffee, although it has high confidence, cannot be considered significant.

Lift is symmetric with respect to the two sides of the rule. There exists asymmmetric measures (e.g., conviction) to assess the significance of association rules.

Exercises

Exercise 1

Let T be a dataset of transactions over I. Let $X, Y \subseteq I$ be two closed itemsets and define $Z = X \cap Y$.

1. Find a relation among T_X, T_Y and T_Z (i.e., the sets of transactions containing X, Y, and Z, respectively). Justify your answer.
2. Show that Z is also closed.

Exercise 2

Let $I = \{a_1, a_2, \ldots, a_n\} \cup \{b_1, b_2, \ldots, b_n\}$ be a set of $2n$ item, e let $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n transactions over I, where

$$t_i = \{a_1, a_2, \ldots a_n, b_i\} \qquad \text{per } 1 \leq i \leq n.$$

For $\text{minsup} = 1/n$, determine the number of frequent closed itemsets and the number of maximal itemsets.

Exercises

Exercise 3

Consider the mining di association rules from a dataset T of transactions. Call *standard* the rules extracted with the classical framework. We say that a standard rule $r: X \rightarrow Y$ is also *essential* if $|X| = 1$ or for each non-empty subset $X' \subset X$, $\mathrm{Conf}(X' \rightarrow Y \cup (X - X')) < \mathrm{Conf}(r)$.

1. Let T consists of the following 5 transactions: $(ABCD)$, $(ABCE)$, (ABC), (ABE), (BCD). Using minsup=0.5 and minconf=0.5, identify a standard rule $X \rightarrow Y$ with $|X| > 1$ which is not essential.
2. Each essential rule can be regarded as *representative* of a set of non-essential standard rule. Which subset? Justify your answer.

Theory questions

- Explain how the anti-monotonicity of support is exploited by algorithm A-Priori to run efficiently
- Define what anti-monotonicity property of confidence is used when generating interesting association rules from frequent itemsets
- Let T a dataset of transactions over I and let minsup be a suitable support threshold. Define the notion of maximal itemset and argue that for every frequent itemset $X \subseteq I$ there exists at least one maximal itemset Y such that $X \subseteq Y$
- Let T be a dataset of transactions over the set of items I and let $X \subseteq I$. Define the itemset $\mathrm{Closure}(X)$ and show that $\mathrm{Supp}(\mathrm{Closure}(X)) = \mathrm{Supp}(X)$.
- Let T a dataset of transactions over I. Define the set of Top-K frequent itemsets.

References

TSK06 P.N.Tan, M.Steinbach, V.Kumar. Introduction to Data Mining. Addison Wesley, 2006. Chapter 6.

LRU14 J.Leskovec, A.Rajaraman and J.Ullman. Mining Massive Datasets. Cambridge University Press, 2014. Chapter 6.