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 patterns.

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

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

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

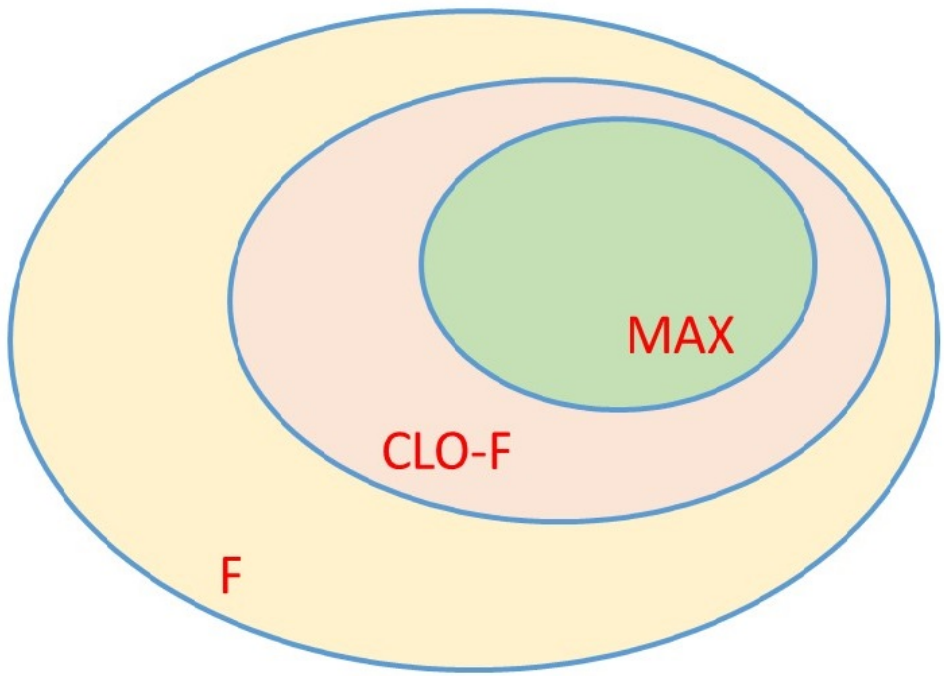
(e) CLO-F = C,D,E,AC,BC,CE,DE,ABC,ACD

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



Closed/maximal itemsets (cont'd)

Observations

- An immediate consequence of the above properties is that from the maximal itemsets (even more so from the frequent closed itemsets) *all frequent itemsets can be derived*. In this sense, MAX and CLO-F provide succint representations of F.
- In general, however, the support of frequent itemsets cannot be derived from the maximal itemsets and their supports
  $\Rightarrow$ the maximal itemsets and their supports provide a lossy representation of the frequent itemsets and their supports.
- Instead, frequent closed itemsets and their supports provide a lossless representation of the frequent itemsets and their supports.

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

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

Representativity of closed itemsets (cont'd)

Theorem

Let $X \subseteq I$. We have

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

Proof

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

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

The proof of the corollary is left as an exercise.

Representativity of closed itemsets (cont'd)

The corollary provides the following simple procedure to generate all frequent itemsets and their supports from the frequent closed itemsets and their supports. Let $\mathcal{M}$ be a map initially storing all pairs $(Y, \mathrm{Supp}(Y))$ with $Y \in$ CLO-F.

$\mathcal{M}' \leftarrow$ `empty map`
for each $(Y, s) \in \mathcal{M}$ **do**
  for each $X \subseteq Y$ **do**
    if ($\mathcal{M}'$ `contains an entry` (X, s'))
    then `substitute` s' `with` $\max\{s, s'\}$ `in` $\mathcal{M}'$
    else `add` (X, s) `to` $\mathcal{M}'$

At the end of the two for-loops $\mathcal{M}'$ contains all frequent itemsets and their supports.

Observations

- Redundancy: Frequent closed itemsets, and, to a lesser extent, maximal itemsets maintain the same information content as the frequent itemsets but remove a good deal of redundancy. In particular, each (frequent) closed itemset Y can be regarded as a representative of all those (possiby many) itemsets X such that $\mathrm{Closure}(X) = Y$.
- Control of output size: There exist pathological instances (see next slide) where even for reasonably large frequency thresholds, the number of maximal itemsets is exponential in $|\mathrm{input}|$. Even more so for frequent closed itemsets.
- 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.)

Exercise

Exponentiality of maximal/frequent closed itemsets

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. $T = \text{CLO-F}_{T,sc(K)}$

Top-K frequent (closed) itemsets (cont'd)

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)
- For closed itemsets, K provides a somewhat tight control on the output size (see next theorem)

Theorem

For $K > 0$, $CLO\text{-}F_{T,sc(K)}$ contains $O(d \cdot K)$ itemsets.

Top-K frequent (closed) itemsets (cont'd)

Proof of Theorem.

- There is only 1 closed itemset of support=1, namely the intersection of all transactions (i.e., $\mathrm{Closure}(\emptyset)$).
  $\Rightarrow$ if $sc(K) = 1$ then $|\text{CLO-F}_{T,sc(K)}| = 1$.
- Assume now $sc(K) < 1$ and let Φ be the set of closed itemsets of support $> sc(K)$, including $\mathrm{Closure}(\emptyset)$ (which may be empty). Clearly, Φ contains $\leq K$ itemsets.
- For each closed itemset X of support $sc(K)$ there must exists $X' \in \Phi$ such that $X' \subset X$. Consider one largest such X' and take an arbitrary item $a \in X - X'$. It is easy to see that $X = \mathrm{Closure}(X' \cup \{a\})$.
- The theorem follows since there are at most $d \cdot K$ itemsets $\mathrm{Closure}(Y \cup \{a\})$ with $Y \in \Phi$ and $a \in I$.

□

Example

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.

Observe that ACD $= \mathrm{Closure}(\mathrm{AC} \cup \{\mathrm{D}\}) = \mathrm{Closure}(\mathrm{D} \cup \{\mathrm{A}\})$

Top-K frequent (closed) itemsets (cont'd)

The following exercise shows that K does not always provide a tight control of the output size if the closure requirement is lifted

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: the user fixes the criteria to assess the interestingness of a pattern, based on his/her 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[1].

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

The bar is used to denote the absence of an item. For example, 200 transactions contain tea, and 150 of them contain also coffee, while the other 50 do not contain coffee.

Consider rule

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

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

[1] Learn autonomously what a contingency table is

Lift

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

Let su define a better measure alternative to confidence

Definition (Lift)

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

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

- $\mathrm{Lift}(r)$ 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) $\gg \Rightarrow X$ and Y are positively correlated
- Lift(r) $\ll 1 \Rightarrow X$ and Y are negatively correlated

In the previous example, Lift(tea $\rightarrow$ coffee) $= 0.9375$, hence tea and coffee a slightly negatively correlated.

However, Lift is symmetric with respect to the two sides of the rule. In some cases, asymmetric measures can be more adequate.

Conviction

Definition (Conviction)

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

$$\text{Conviction}(r) = \frac{\text{Supp}(X) \cdot \text{Supp}(\bar{Y})}{\text{Supp}(X \cup \bar{Y})},$$

where $\text{Supp}(\bar{Y})$ is the fraction of transactions that do not contain Y, and $\text{Supp}(X \cup \bar{Y})$ is the fraction of transactions that contain X but do not contain Y.

- Conviction is asymmetric with respect to the two sides of the rule
- In first-order logic, $X \to Y$ is equivalent to $\overline{X \wedge \bar{Y}}$.
- Maximizing the strength of the implication $X \to Y$ requires minimizing the occurrence of X and $\bar{Y}$.
- Conviction(r) is high when X and $\bar{Y}$ occurr together less frequently than expected if X and $\bar{Y}$ were independent.

Examples

(A)

	Y	$\bar{Y}$	
X	8	2	10
$\bar{X}$	72	18	90
	80	20	100

$$\begin{aligned} \text{Conf}(X \to Y) &= 0.8 \\ \text{Lift}(X \to Y) &= 1 \\ \text{Conviction}(X \to Y) &= 1 \end{aligned}$$

(B)

	Y	$\bar{Y}$	
X	40	10	50
$\bar{X}$	10	40	50
	50	50	100

$$\begin{aligned} \text{Conf}(X \to Y) &= 0.8 \\ \text{Lift}(X \to Y) &= 40/25 \\ \text{Conviction}(X \to Y) &= 25/10 \end{aligned}$$

Examples (cont'd)

(C)

	Y	$\bar{Y}$	
X	10	10	20
$\bar{X}$	1	79	80
	11	89	100

$$\begin{aligned} \text{Conf}(X \to Y) &= 0.5 \\ \text{Lift}(X \to Y) &= 100/22 \\ \text{Conviction}(X \to Y) &= 178/100 \end{aligned}$$

$$\begin{aligned} \text{Conf}(Y \to X) &= 10/11 \\ \text{Lift}(Y \to X) &= 100/22 \\ \text{Conviction}(Y \to X) &= 880/100 \end{aligned}$$

Observation: In all three cases, Conviction seem to provide a good estimate of the relevance of the rule

Beyond transactions

The mining of frequent itemsets and interesting association rules can be employed in more general contexts than transactional datasets

- Dataset T of N records over k features, $A_1, A_2, \ldots, A_k$.
- Features can be *categorical* (e.g., color, state), *binary* (e.g., yes/no, male/female), or *numerical* (e.g., age, income).
- We can search for high-support and high-confidence rules such as

  $$\text{age} \in [15, 60] \rightarrow \text{uses facebook},$$

  or

  $$\text{democrat} \wedge \text{midwest} \rightarrow \text{reads newspapers}.$$

  In general, we want rules whose LHS and RHS are *conjunctions of predicates*.

Beyond transactions (cont'd)

The standard framework can be applied after transforming the dataset T into a dataset of transactions as follows:

- Binarization. For each categorical/binary feature A over $D = \{d_1, d_2 \ldots, d_\ell\}$, define ℓ items $(A = d_i)$, for $1 \leq i \leq \ell$.
- Binning. For each numerical feature A over D, partition D into ℓ disjoint intervals $D_1, D_2 \ldots, D_\ell$ and define ℓ items $(A \in D_i)$, for $1 \leq i \leq \ell$.
- Records $\rightarrow$ Transactions. Transform each record into a transaction by mapping each feature to one of the above items, based on its value.

Observations:

- In the binarization, infrequent values of a categorical domain can be grouped
- The choice of the partition for a numerical domain is delicate and can be done using different strategies: e.g., equal width, equal frequency, clustering.

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 a dataset of transactions over I and let $X \subseteq I$. Define the itemset $\mathrm{Closure}(X)$ and show that if X is closed $X = \mathrm{Closure}(X)$.
- Let T a dataset of transactions over I. Define the set of Top-K frequent itemsets.

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.

Case Study

Year 2000 US Election Survey Data

- Source: M.MacDougall. Shopping for Voters: Using Association Rules to Discover Relationships in Election Survey Data
- Dataset: Survey on 2000 US Elections (1st George W. Bush term)
  - 1800 interviews pre- and post-election
  - Several questions per interview
- Objective: Discovery of interesting patterns
- Tool: (Generalized) Association rules
  - Interviews $\rightarrow$ Transactions
  - Questions $\rightarrow$ Items

Case Study (cont'd)

- Preprocessing:
  - Eliminate interviews with missing data (down to 1550)
  - Reduce number of items
    - Reduce the number of questions (hence, less items). E.g.: 12 questions on racial issues $\rightarrow$ 1 binary attribute RACE-BIAS
    - Reduce the size of some attribute domains. E.g.,: attribute STATE (50 values) become attribute REGION (4 values). Possible answers such as very bad, bad, good, very good, etc., become binary: bad, good.
  - Subdivide remaining items into categories: demographics, party affiliation,
- Metrics of interest: Support, confidence, lift
- Choice of parameters: Support threshold $= 5\%$, restriction tu rules with at most 4 items

Case Study (cont'd)

- Analysis:
  - Search for rules such as

$$\begin{aligned} \text{REPUBLICAN} &\rightarrow \cdots \\ \text{DEMOCRAT} &\rightarrow \cdots \end{aligned}$$

  - List the top-10 rules in decreasing order ot lift, showing support, confidence and lift for each rule
- Outcomes: (see paper for more details)
  - Republicans prefer to use surplus for reducing taxes and giving school vouchers, while democrats prefer to use it for social security
  - Gun control is more important for democrats
  - Republicans seem to show higher consensus than democrats (top rules have higher confidence)

References

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