Association Pattern Mining

Procheta Sen

UNIVERSITY OF LIVERPOOL

Applications

- **Supermarket data**
  - which items are frequently bought together
  - useful insights about target marketing and shelf placement of the items
- **Text mining**
  - identifying co-occurring terms and keywords
- **Generalization to dependency-oriented data types**
  - Web log analysis
  - software bug detection
  - spatio-temporal event detection

Terminology

- **Borrowed from the supermarket analogy**
    - Dataset objects are called **transactions**
    - Output: **large itemsets** (frequent itemsets or frequent patterns)

Usage

- Frequent itemises can be used to generate **association rules** $X \Rightarrow Y$, where X and Y are sets of items (e.g. $\{\text{Eggs, Milk}\} \Rightarrow \{\text{Yougurt}\}$)
  - promote yogurt to customers who often buy eggs and milk
  - place yogurt on shelves that are located in proximity to eggs and milk.

The Frequent Pattern Mining Model

- The **universe** of d items: U
- **Itemset** is a set of items
- The **dataset** $\mathscr{D}$ consists of n transactions $\overline{T}_1, \ldots, \overline{T}_n$, each of which is an itemset
- Each transaction can be represented as a d-dimensional binary vector
- Each binary attribute in a transaction represents a particular item from U
- **Support** of an itemset I: the fraction of the transactions in the dataset $\mathscr{D}$ that contain I as a subset (denoted by $\text{sup}(I)$)

The Frequent Pattern Mining Model

Frequent Itemset Mining Problem

Given a **dataset** $\mathscr{D}$ of transactions and a **frequency threshold** f, determine all itemsets that occur as a subset of at least fraction f of the transactions in $\mathscr{D}$.

Remarks on the frequency threshold:

- lower frequency threshold yields a larger number of large itemsets
- too high frequency threshold may lead to no large itemsets

The Frequent Pattern Mining Model

Example (let $f = 0.65$)

Transaction	Milk	Butter	Bread	Mushrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	0	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

{Milk, Butter, Bread}
is a large itemset

{Mushrooms, Onion, Carrot}
is **not** is a large itemset

Monotonicity of support

Support Monotonicity Property

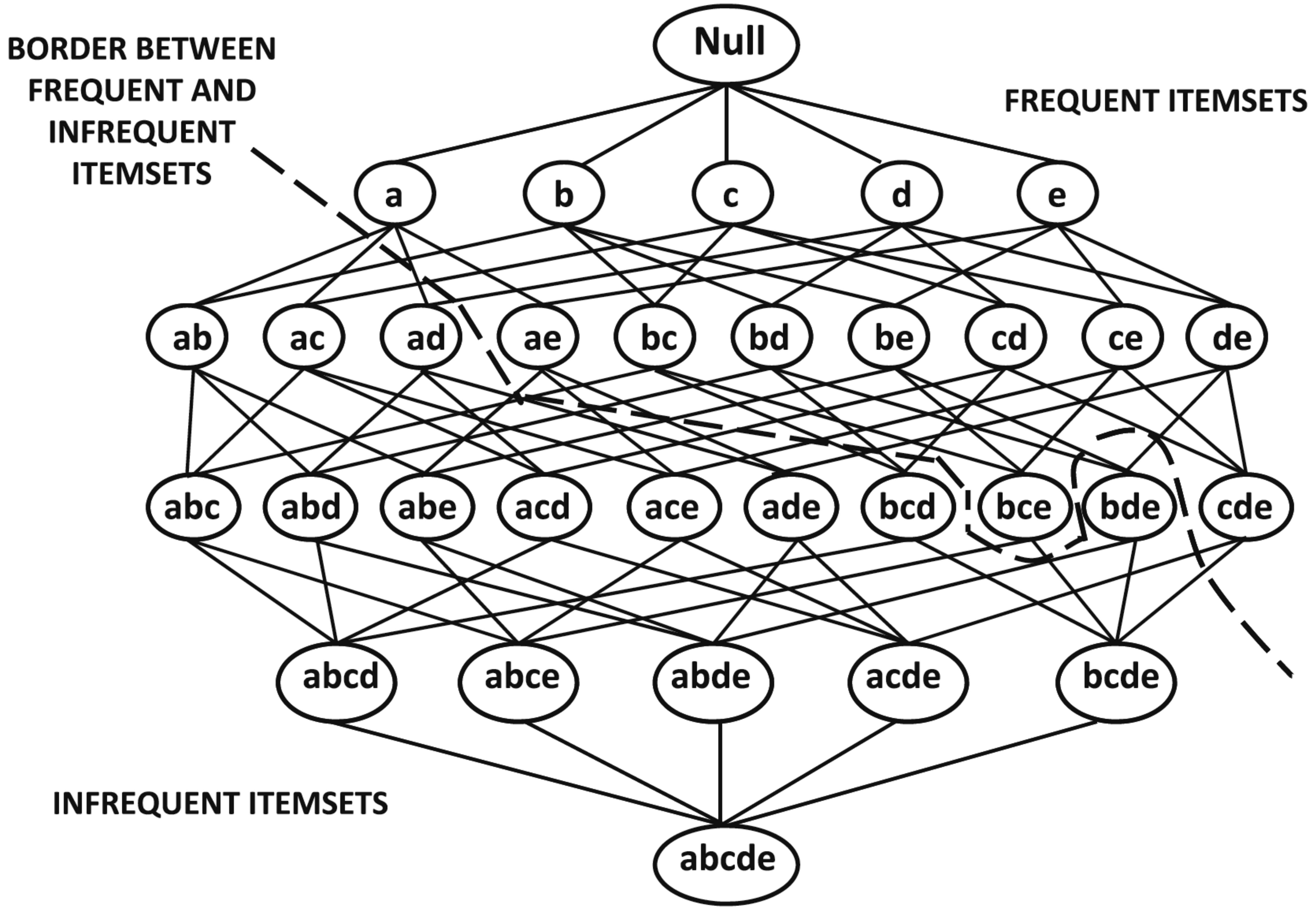
The support of ever subset J of I is at least as large as the support of itemset I, i.e. $\text{sup}(J) \geq \text{sup}(I)$ for every $J \subseteq I$.

Downward Closure Property

Every subset of a frequent itemset is also frequent.

A frequent itemset I is **maximal** at a given frequency threshold, if it is frequent and no superset of I is frequent.

Downward Closure Property



The Frequent Pattern Mining Model

Example (let $f = 0.65$)

Transaction	Milk	Butter	Bread	Mushrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	0	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

{**Milk, Butter, Bread**} is a **maximal** frequent itemset (at frequency threshold 0.65)

{**Butter, Bread**} is frequent itemset, but **not** maximal

The Frequent Pattern Mining Model

Example (let $f = 0.65$)

Transaction	Milk	Butter	Bread	Mushrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	0	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

There are 3 maximal frequent itemsets: {**Milk, Butter, Bread**} and {**Milk, Onion**} and {**Mushrooms**}, but there 10 frequent itemsets in the dataset:

{**Milk**}, {**Butter**}, {**Bread**}, {**Onion**}, {**Mushrooms**}

{**Milk, Butter**}, {**Milk, Bread**}, {**Butter, Bread**}, {**Milk, Onion**}

{**Milk, Butter, Bread**}

Representation of frequent itemsets

- All frequent itemsets can be derived from the maximal frequent itemsets
- Hence the **maximal** frequent itemsets can be considered as a compact representation of the frequent itemsets
- However, such such representation does not store information about the support values of the itemsets.

Association Rules

- We want to generate association rules of the form $X \Rightarrow Y$ meaning that if a transaction contains the set of items X, then it is "**likely**" to contain the set of items Y.

- To measure the likelihood of the association rule we use the **confidence** of the rule, which is the conditional probability that a transaction contains the the set of items Y, given that it contains the set X.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{sup}(X \cup Y)}{\text{sup}(X)}$$

By definition, the support of rule $X \Rightarrow Y$ denoted by $\text{sup}(X \Rightarrow Y)$ is $\text{sup}(X \cup Y)$.

Example: $\text{conf}(\{\text{Milk}\} \Rightarrow \{\text{Butter, Bread}\})$

Transaction	Milk	Butter	Bread	Mushrooms	Onion	Carrot
1234	1	1	1	0	1	0
324	0	0	0	1	1	1
234	1	1	1	0	1	0
2125	1	1	1	1	0	1
113	1	0	0	1	1	0
5653	1	1	1	1	1	0

$$\text{sup}(\{\text{Butter, Bread, Milk}\}) = \frac{2}{3}$$

$$\text{conf}(\{\text{Milk}\} \Rightarrow \{\text{Butter, Bread}\}) = \frac{2}{3} \cdot \frac{6}{5} = \frac{4}{5}$$

$$\text{sup}(\{\text{Milk}\}) = \frac{5}{6}$$

Association Rules

Definition

Let X and Y be two sets of items. Then the rule $X \Rightarrow Y$ is an association rule at a **frequency threshold** f and a **confidence threshold** c if

1. the support of $X \Rightarrow Y$ (i.e. the support of the itemset $X \cup Y$) is at least f, and
2. the confidence of the rule $X \Rightarrow Y$ is at least c.

The first condition ensures that there are sufficiently many transactions relevant to the rule.

The second condition ensures that the rule has sufficient strength in terms of conditional probabilities.

Association rule generation framework

Phase 1: generate all frequent itemsets for the given frequency threshold f

- Bruteforce algorithm
- Apriori algorithm

Phase 2: from the frequent itemsets, generate the association rules at the given confidence threshold c

- For each frequent item set I:
  - partition I into all possible pairs of subsets (X, Y) such that $Y = I - X$ and $X \cup Y = I$;
  - compute the confidence of the rule $X \Rightarrow Y$. If it is at least c, store the rule $X \Rightarrow Y$.

Association rule generation framework

To optimise Phase 2 one can use

Confidence Monotonicity property

Let $X_1, X_2,$ and I be itemsets such that $X_1 \subset X_2 \subset I$

$$\text{conf}(X_2 \Rightarrow I - X_2) \geq \text{conf}(X_1 \Rightarrow I - X_1)$$

Example

If we have association rules $\{\text{Butter}\} \Rightarrow \{\text{Milk, Bread}\}$ and $\{\text{Butter, Bread}\} \Rightarrow \{\text{Milk}\}$, then the second one is redundant as it has the same support as the first one, but its confidence is no less than that of the first rule.