Vytěžování dat, přednáška 6: Vyhledávání častých množin

Filip Železný

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Asociace

- Pokračujeme v metodách vytěžování bez učitele
  - tedy modelování rozdělení P_X, z něhož jsou generována data
- Připomínka: $X = X_1 \times X_2 \times \ldots \times X_n$, kde X_i je množina hodnot příznaku i
- Úloha: najít hodnoty nějaké podmnožiny příznaků, které se často objevují společně tj. jsou *asociovány*
- Příklad: $X =$ příjem $\times$ bydliště $\times$ pohlaví $\times$ úvěr
- Častá asociace např.

$$\text{příjem} = \text{vysoký} \ \& \ \text{bydliště} = \text{Praha}$$

- "častá" znamená, že pravděpodobnost $P_{\text{příjem,bydliště}}(\text{vysoký}, \text{Praha})$ je vysoká

Podpora asociace

Podpora: podíl instancí, v nichž asociace platí, mezi všemi instancemi

příjem	**bydliště**	**pohlaví**	**úvěr**
vysoký	Praha	M	splácí
vysoký	Plzeň	M	splácí
nízký	Praha	M	nesplácí
vysoký	Praha	Ž	splácí
střední	Brno	M	splácí

- Asociace $A =$

  $$\text{příjem} = \text{vysoký} \;\&\; \text{bydliště} = \text{Praha}$$

  má podporu $2/5 = 0.4$. Zapisujeme $podp(A) = 0.4$
  - $podp(A)$ je odhadem $P_{\text{příjem,bydliště}}(\text{vysoký}, \text{Praha})$

Podpora asociace

Všechny asociace s podporou alespoň 0.4

příjem	bydliště
vysoký	Praha
vysoký	Plzeň
nízký	Praha
vysoký	Praha
střední	Brno

příjem	bydliště
vysoký	Praha
vysoký	Plzeň
nízký	Praha
vysoký	Praha
střední	Brno

příjem	bydliště

1. *“true” (prázdná asociace)*

Asociace jako množiny

- Hledání asociací se uplatňuje zejm. v *"analýze transakcí"* (*"analýze nákupních košíků"*)
- Příznaky: položky sortimentu. Instance: obsah nákupního košíku. Hodnoty příznaků jsou **binární** $\{\text{ano}, \text{ne}\}$.

pivo	**párky**	**horčice**	**pleny**
ano	ne	ne	ano
ne	ano	ano	ne

- Zde místo např.

$$\text{pivo} = \text{ano} \;\&\; \text{pleny} = \text{ano}$$

- zapisujeme (a chápeme) asociaci jako *množinu položek*, např.

$$\{\text{pivo}, \text{pleny}\}$$

Princip monotonicity

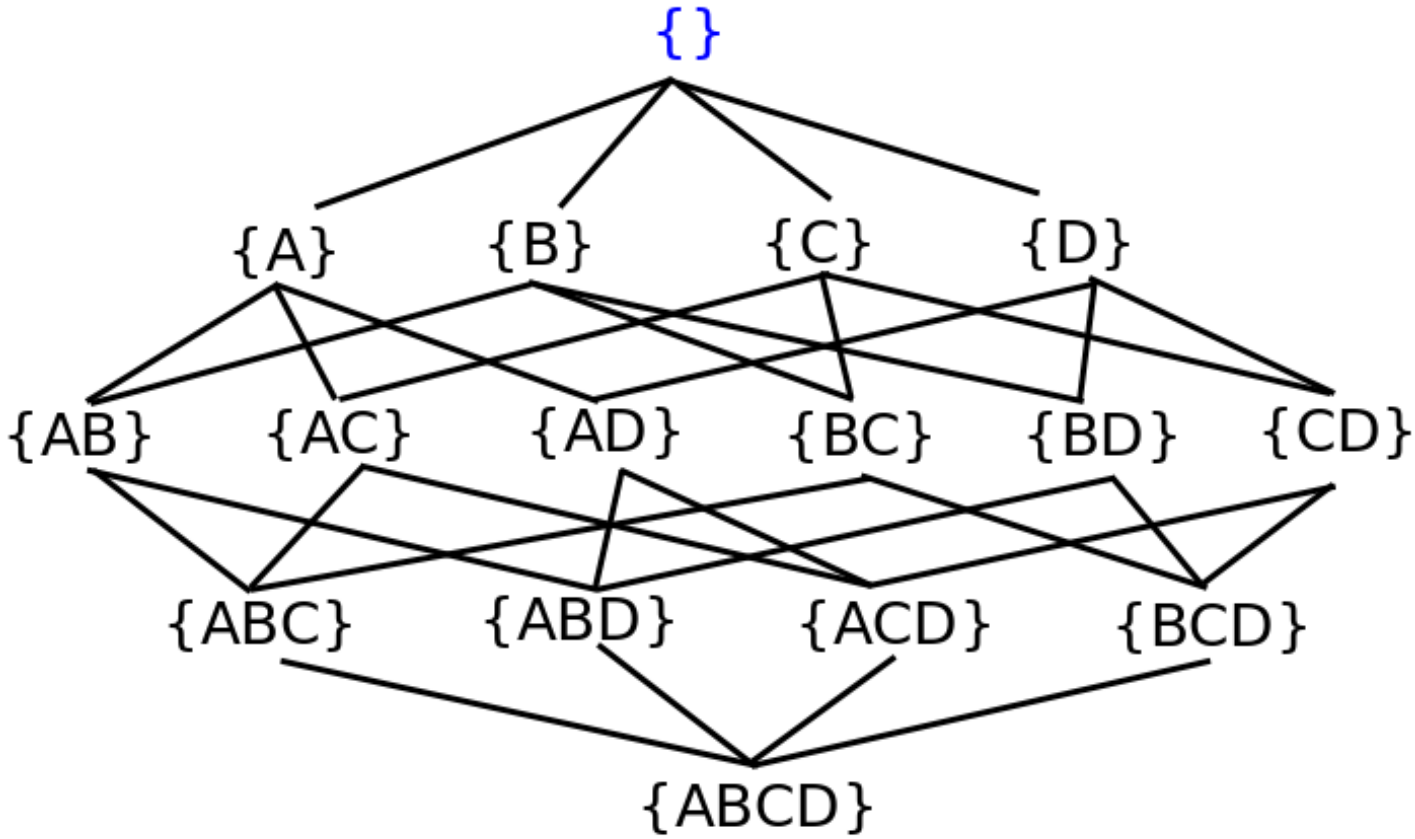
pivo	párky	horčice	pleny
ano	ne	ne	ano
ano	ne	ano	ano
ne	ano	ano	ne
ano	ano	ano	ne
ano	ano	ano	ano

pivo	párky	horčice	pleny
ano	ne	ne	ano
ano	ne	ano	ano
ne	ano	ano	ne
ano	ano	ano	ne
ano	ano	ano	ano

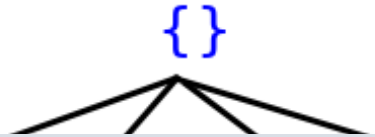
pivo	párky	horčice	pleny
ano	ne	ne	ano
ano	ne	ano	ano

Hledání častých množin položek algoritmem APRIORI

Postupujeme z vyšších pater do nižších



Prázdná množina je vždy častá. Kandidáti o patro níž:
$\{A\}, \{B\}, \{C\}, \{D\}$.



Algoritmus APRIORI

Apriori:

$C_1 = \forall$ množiny položek velikosti 1
$L_1 = \forall$ časté množiny z C_1 (podpora $\geq min\_podp$)
$i = 1$
repeat
  $i := i + 1$
  $C_i =$ Apriori-Gen(L_{i-1})
  $L_i :=$ všechny časté množiny z C_i
  {Vyžaduje průchod databází}
until $L_i = \emptyset$
$L = \bigcup L_i$,$\forall i$
return L

Apriori-Gen(L_{i-1}):

$C_i = \emptyset$
for $\forall$ dvojice množin položek $M_p, M_q \in L_{i-1}$ **do**
  if se shodují v $i-2$ položkách **then**
    přidej $M_p \cup M_q$ do C_i
  end if
end for
for $\forall$ množiny položek $M \in C_i$ **do**
  if jakákoliv podmnožina M o délce $i-1$ není v L_{i-1} **then**
    odstraň M z C_i
  end if
end for
return C_i

Maximální množiny položek

- Častých množin může být velmi mnoho
- Výstup algoritmu je pak nepřehledý
- Možné zjednodušení výstupu: zachovat pouze *maximální* množiny
- Maximální množina:
  - Je častá a *žádná z jejích vlastních nadmnožin není častá*

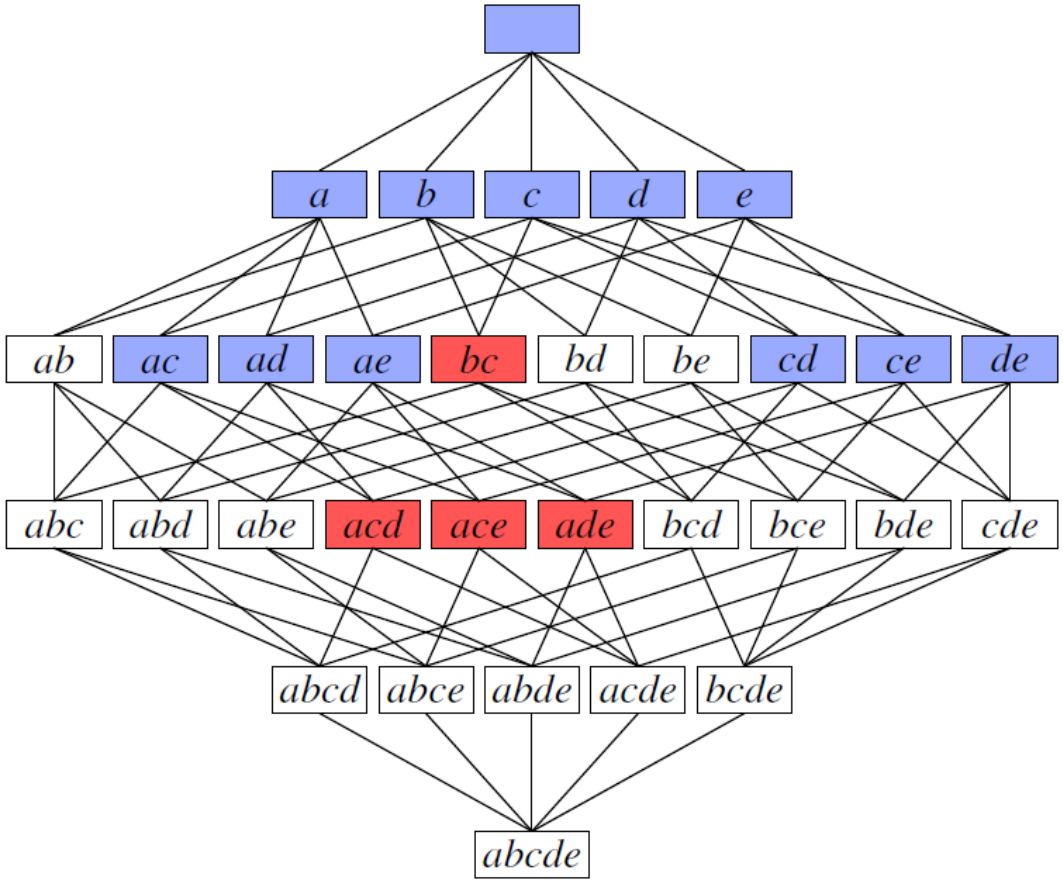
Uzavřené množiny položek

- Množina častých množin může být redundantní
- Informace o všech častých množinách je implicitně obsažena v množině *uzavřených* častých množin
- Uzavřená množina:
  - žádná její vlastní nadmnožina nemá stejnou podporu
- Každá maximální množina je uzavřená
- (obráceně nemusí platit)

Maximální a uzavřené množiny

- časté množiny položek barevně (podpora alespoň 3 instance), maximální množiny červeně

Transakce	Položky
t_1	a, d, e
t_2	b, c, d
t_3	a, c, e
t_4	a, c, d, e
t_5	a, e
t_6	a, c, d
t_7	b, c
t_8	a, c, d, e
t_9	b, c, e
t_{10}	a, d, e



Asociační pravidlo

- Pravidlo ve tvaru

$$\textbf{if } Ant \textbf{ then } Suc$$

kde Ant a Suc jsou množiny položek nazývané *antecedent* resp. *sukcedent*. Pravidlo též zapisujeme jako

$$Ant \rightarrow Suc$$

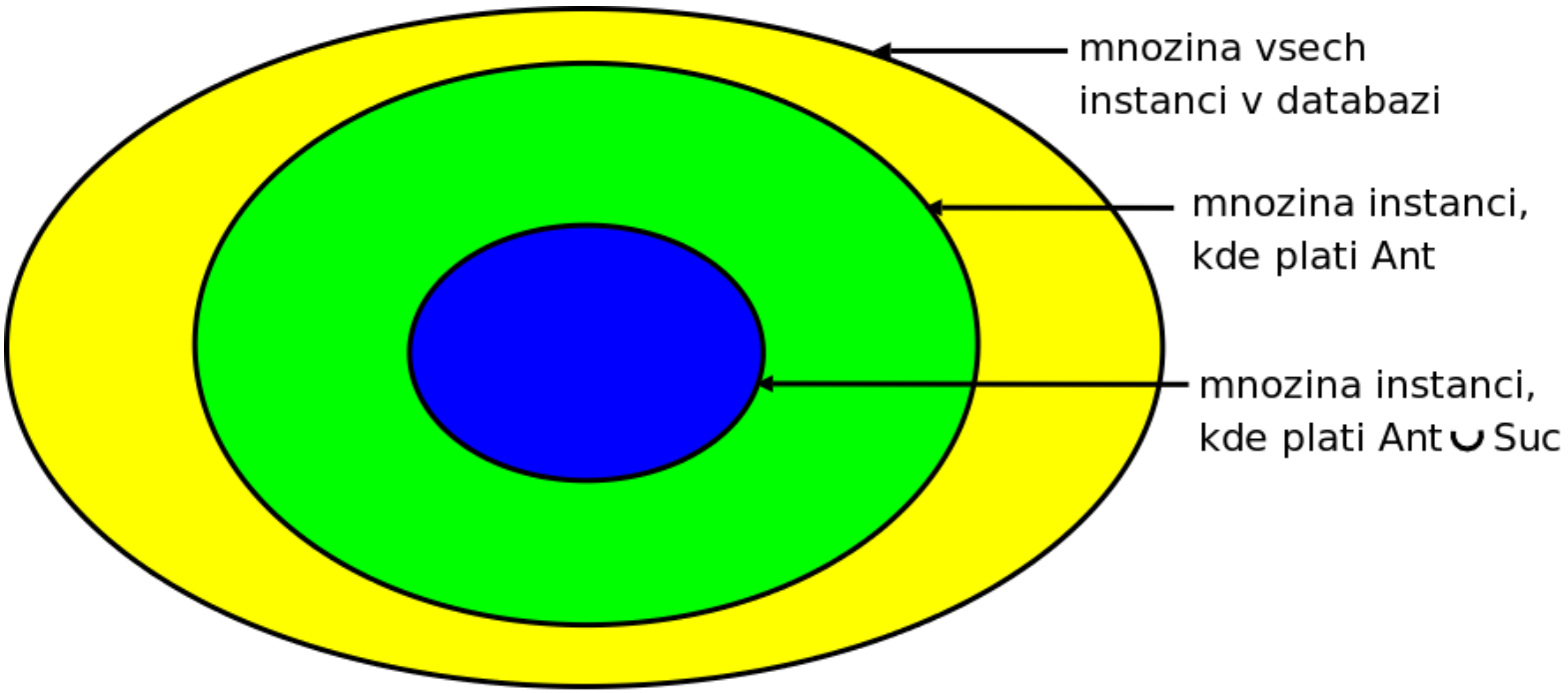
- **Podpora asociačního pravidla** R definována jako

$$podp(Ant \rightarrow Suc) = podp(Ant \cup Suc)$$

- **Spolehlivost** (confidence) **asociačního pravidla** R definována jako

$$spol(Ant \rightarrow Suc) = \frac{podp(Ant \cup Suc)}{podp(Ant)}$$

Asociační pravidlo



Hledání asociačních pravidel algoritmem APRIORI

- Hledáme pravidla $Ant \rightarrow Suc$, která jsou častá (tj. s podporou alespoň p) a spolehlivá (tj. se spolehlivostí alespoň s).
- Je-li pravidlo $Ant \rightarrow Suc$ časté, tak množina položek (asociace) $Ant \cup Suc$ je častá.
- Nejprve tedy najdeme všechny časté množiny položek.
- Pro každou častou množinu položek M a každou její neprázdnou vlastní podmnožinu $P \subset M$ zkusíme, zda

$$\frac{podp(M)}{podp(P)}$$

- Pokud ano, tak pravidlo

$$P \rightarrow M \setminus P$$

je časté a spolehlivé.

Hledání asociačních pravidel: příklad

- Hledáme asociační pravidla s podporou alespoň 0.1 a spolehlivostí alespoň 0.6. Právě 13% nákupních košíků obsahuje každou položku z množiny

$$\{\text{párky}, \text{hořčice}, \text{pivo}, \text{otvírák}\}$$

- Množina je tedy častá. Vyberme nějakou její vlastní neprázdnou podmnožinu, např.

$$\{\text{párky}, \text{hořčice}\}$$

- Pravidlo

$$\{\text{párky}, \text{hořčice}\} \rightarrow \{\text{pivo}, \text{otvírák}\}$$

má podporu 13%. Pokud navíc právě 19% nákupních košíků obsahuje párky i horčici, má pravidlo spolehlivost

$$\frac{13}{19} > 0.6$$

a je tedy časté i spolehlivé.

Algoritmus AR-Gen

Vytvoří množinu asociačních pravidel ze zadané množiny L častých množin položek a zadaného parametru min_spol minimální spolehlivosti.

AR-Gen:

$R := \emptyset$
for $\forall M \in L$ **do**
  for $\forall P \subset M,\ P \neq \emptyset$ **do**
    if $podp(M)/podp(P) \geq min\_spol$ **then**
      $R := R \cup \{P \rightarrow M \setminus P\}$
    end if
  end for
end for
return R