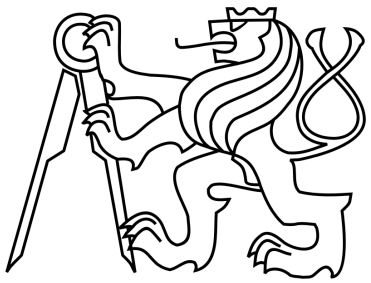
Tvorba asociačních pravidel a hledání častých skupin položek



Nature Inspired Technologies Group

Osnova

- **Asociace**
- **Transakce**
- Časté skupiny položek
- Apriori vlastnost podmnožin
- Asociační pravidla
- Aplikace

Asociace

Nechť $\boldsymbol{I}$ je množina položek. Množinu $\boldsymbol{Z} \subseteq \boldsymbol{I}$ označujeme jako **asociaci**, pokud frekvence, s jakou se $\boldsymbol{Z}$ vyskytuje, je výrazně odlišná od toho, co bychom očekávali na základě frekvencí výskytů jednotlivých prvků $X \in Z$.

Příklad:

- Nechť párky P a hořčice H se vyskytují v nákupním košíku u 10% a u 4% zákazníků.
- Očekávali bychom tedy, ze $\{P, H\}$ se bude vyskytovat u 0,4 % zákazníků.
- Ovšem frekvence $\{P, H\}$ je ve skutečnosti 1 %.
- Tedy $\{P, H\}$ je asociace.

Příklad transakcí

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Databáze transakcí příkladu

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

POLOŽKY:
A = milk
B= bread
C= cereal
D= sugar
E= eggs

Jednotlivé instance = transakce položek

Databáze transakcí příkladů

Uniformní reprezentace: položky převedeny na binární údaje

TID	Products
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Definice

- **Položka** (item): pár tvaru *atribut* = *hodnota* nebo jen hodnota
  - Obvykle se z každého atributu vytvoří několik binárních *údajů* odpovídajících možným hodnotám, např. **Produkt = "A"** se píše jako **"A"**
- **Množina položek *I*** (itemset) : podmnožina všech uvažovaných položek
  - Příklad: $I = \{A,B,E\}$ (pořadí není podstatné)
- **Transakce**: (TID, množina položek)
  - TID je identifikátor transakce (její ID)

Osnova

- Transakce
- **Časté skupiny položek**
- Apriori vlastnost podmnožin
- Asociační pravidla
- Aplikace

Podpora a časté množiny položek

- **Podpora** (support) množiny položek ***I***
  - sup(***I***) = počet transakcí ***t***, které svědčí pro (tj. obsahují) ***I***
- Pro náš příklad:
  - sup ({A,B,E}) = 2, sup ({B,C}) = 4

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

- **Množina častých položek** *I* je ta množina, jejíž podpora je vyšší než předem stanovené minimum ***minsup*** :
  sup(*I*) >= ***minsup***
- *Můžeme najít všechny množiny častých položek tak, že postupně prohlédneme všechny podmnožiny položek?*
- *Je-li všech položek **n**, pak možných podmnožin je **2^n**. Je to problém?*

Jak rostou zajímavé funkce?

Funkce\ N	10	50	100	300	1000
$N * \log N$	33	282	665	2469	9966
N^2	100	2500	10 000	90 000	10^6
N^3	1000	125 000	10^6	$27 * 10^6$	10^9
2^N	1024	10^{14}	10^{29}	10^{89}	10^{300}
$N!$	$3,6 * 10^6$	10^{63}	10^{159}	10^{621}	
N^N	10^{10}	10^{83}	10^{200}	10^{742}	

- Počet protonů ve známém vesmíru ≈ 10^{77}
- Vzdálenost „Big bang“ - dnešek? ≈ 10^{22} μs (tj.10^{-6}s)

Úlohy exp. složitosti nelze řešit hrubou silou!

Osnova

- Transakce
- Časté skupiny položek
- **Apriori vlastnost podmnožin**
- Asociační pravidla
- Aplikace

Apriori vlastnost podmnožin

Každá podmnožina množiny častých položek je také také množinou častých položek!

- Otázka: Proč?
- **Příklad**: Předpokládejme, že {A,B} je častý pár položek. Protože každý výskyt svědčící pro {A,B} musí obsahovat i položku A, musí být i položka A častá (obdobně pro B)
- Podobný argument platí i pro větší množiny položek
- *Takřka všechny algoritmy pro konstrukci asociačních pravidel využívají této vlastnosti množin častých položek !*

Hledání častých položek

- Postup zdola nahoru: vyhledáme nejdřív všechny *jednoprvkové množiny častých položek* (to je lehké)
- ***JAK?***
- Stačí spočítat frekvence jednotlivých položek

- *Princip pro další kroky:*
  - 1-prvkové množiny častých položek lze použít k vytvoření kandidátů na 2-prvkové množiny častých položek,
  - 2-prvkové k vytvoření 3-prvkových množin, …

Hledání množin častých položek

- Pokud {A,B} je častá množina položek, pak {A} i {B} musí být rovněž častými množinami položek!
- *Obecně*: pokud ***X*** je množina častých položek s *k*-prvky, pak všechny její podmnožiny četnosti (*k*-1), tj. ty obsahující právě (*k*-1) položek, musí být rovněž množinami častých položek.

Apriori algoritmus (Agrawal & Srikant)

Najdi množiny častých položek o 1 prvku a pokračuj na četnosti vždy o 1 vyšší pomocí cyklu, který obsahuje oba následující kroky:

- **Krok „spoj"**: Kandidáty na množiny častých položek o *k*-prvcích lze zkonstruovat jako „rozšíření" množin častých prvků o (*k*-1)-prvcích.

⇒ **Krok „prořež"**: Z množiny kandidátů se pak vyberou jen ty podmnožiny, které mají v databázi transakcí dostatečnou podporu.

Příklad

❖ ***Předpokládejme, že máme* 5 množin častých položek** o 3 prvcích

(A B C), (A B D), (A C D), (A C E), (B C D)

❖ *Které množiny o 4 prvcích jsou vhodnými kandidáty ?*

(A B C D) **Otázka: Je OK?**

Odpověď: *Ano* , protože všechny její 3-prvkové podmožiny jsou časté !

(A C D E) **Otázka: Je OK?**

Odpověď: *Ne* , protože (C D E) není množinou častých položek

❖ *Hledání odpovědí usnadňuje, když položky v množinách uvádíme v* ***lexikografickém uspořádání !***

Osnova

- Transakce
- Časté skupiny položek
- Apriori vlastnost podmnožin
- **Asociační pravidla a jejich tvorba**
- Aplikace

Asociační pravidla

- Asociační pravidlo ***R*** : ***Množina_pol1 => Množina_pol2***
- *Množina_pol1* i *Množina_pol2* jsou disjunktní (mají prázdný průnik) a *Množina_pol2* je neprázdná

Význam: "Pokud transakce obsahuje prvky z *Množina_pol1* , pak také obsahuje prvky z *Množina_pol2"*

- Příklady

A,B => C

A,B => C,E

A => B,C

A,B =>D

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Pravdila klasifikační X asociační

Klasifikační pravidla (~ stromy)

- Úlohou je určit cílovou třídu (hodnota specifického atributu)
- Třída musí být předem známa pro všechny zpracovávané příklady
- ***Používaná míra*** **: přesnost klasifikace** na testovací množině dat

Asociační pravidla

- Neomezují se na jediný cílový atribut
- Má smysl použít v některých příkladech
- ***Používané míry*** **: podpora, spolehlivost, zdvih**

Od množin častých položek k asociačním pravidlům

❖ *Otázka: Jaká asociační pravidla lze sestrojit z množiny častých položek {A,B,E} ?*

- **A => B, E**
- **A, B => E**
- **A, E => B**
- **B => A, E**
- **B, E => A**
- **E => A, B**
- **__ => A,B,E** (prázdné pravidlo), zapisované také jako **true => A,B,E**

Asociační pravidlo $R: I => J$: jeho podpora a spolehlivost

Definujeme základní míry významu pravidla ***R,*** kterými jsou

- **podpora** (*support*) : **sup (*R*)** = sup ($I \cup J$)

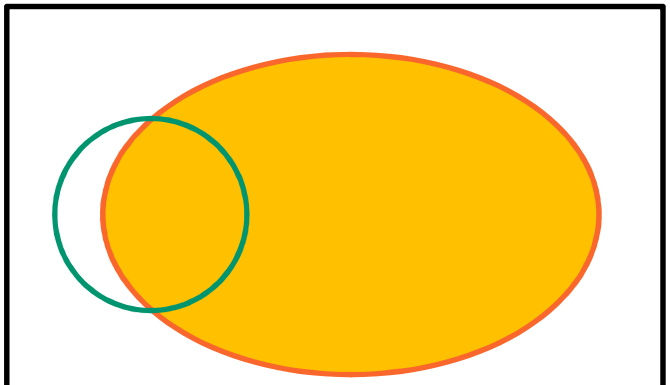
odpovídající podpoře množiny položek, které se v ***R*** vyskytují, tj. $I \cup J$

- **spolehlivost** (*confidence*)**: conf (*R*)** = sup($I \cup J$) / sup(I)

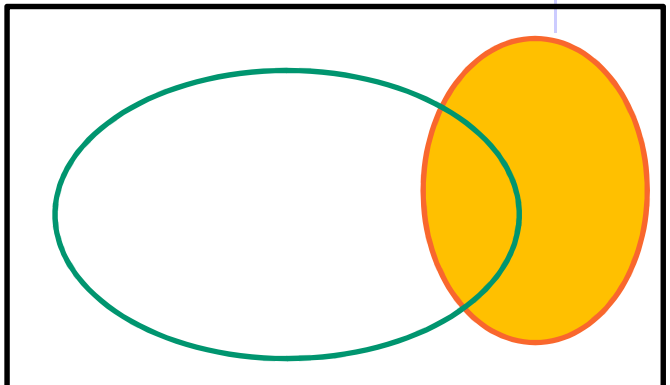
vyjadřující informaci o tom, jaká část z transakcí, ve kterých se vyskytují všechny položky tvořící množinu předpokladů ***I*** pravidla ***R***, obsahuje také položky ***J*** tvořící závěr tohoto pravidla

Jaká je spolehlivost pravidla *A* => *B* ?

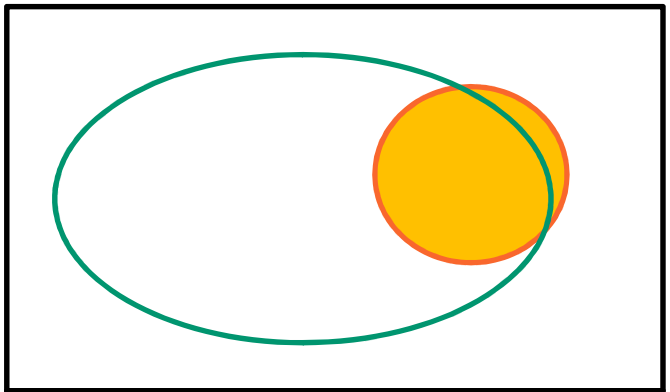
*Nechť **A** ∪ **B** mají dostatečnou podporu: výčet všech možných situací*



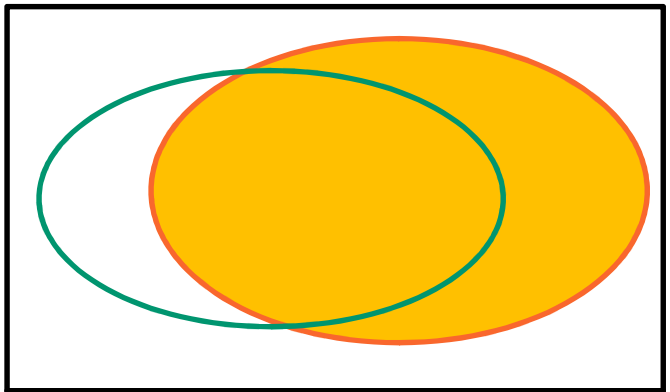
Conf(A => B) je vysoká
Conf(B => A) je nízká



Conf(A => B) je nízká
Conf(B => A) je nízká



Conf(A => B) je nízká
Conf(B => A) je vysoká



Conf(A => B) je vysoká
Conf(B => A) je vysoká

Nejčastěji používané míry ***Ant => Suc***

a je mohutnost množiny všech trans-Akcí s položkami *Ant* ∪ *Suc*

podpora(*support*) = **a/n**

spolehlivost(*confidence*) = **a/r**

pokrytí(*cover*) = **a/k**

	Suc	Non (***Suc***)	∑
Ant	**a**	**b**	**r** = a+b
Non (***Ant***)	**c**	**d**	**s** = c+d
∑	**k** = a+c	**l** = b+d	**n**= r+s

Zdvih (*lift*, „*above average*“)= **an/rk = p**

Poměr spolehlivosti uvažovaného pravidla **a/r** (tedy nad transakcemi, které splňují ***Ant***) a spolehlivosti pravidla ∅ **=>** ***Suc*** (nad všemi transakcemi) určuje hodnotu **p**, která říká kolikrát se díky předpokladu ***Ant*** spolehlivost zvýší, tj. **a/r = p*** (**k/n**)

Příklad:

Úloha: Najděte všechna pravidla s *minsup = 2* ***a*** *minconf= 50%* ***pro množinu častých položek {A,B,E }.***

conf (I => J) =
sup(I ∪ J) / sup(I)

A, B => E : conf=2/4 = 50%

A, E => B : conf=2/2 = 100%

B, E => A : conf=2/2 = 100%

E => A, B : conf=2/2 = 100%

Naopak následující pravidla požadovanou podmínku nesplňují

A =>B, E : conf= 2/6 =33%< 50%

B => A, E : conf= 2/7 = 28% < 50%

__ => A,B,E : conf= 2/9 = 22% < 50%

TID	List of items
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Generování asociačních pravidel

Proces postupující **zdola nahoru** (od pravidel vytvořených z množin častých položek s nejmenší mohutností směrem k množinám s více prvky) rozdělený do 2 fází:

1. ***Výstup*** := Ø, ***k*** := 1
2. Najděte množiny častých položek dané mohutnosti ***k***, např. algoritmem Apriori. Pokud žádná taková množina není, pak KONEC
3. Pro každou množinu ***I*** častých položek nalezněte asociační pravidla platná ve studovaných datech
   - Pro každou vlastní podmnožinu ***J*** (≠ ***I***) množiny ***I***
   - Prověřte platnost asociačního pravidla: ***I*** - ***J*** => ***J*** a platné pravidlo přidejte do množiny ***Výstup***
4. ***k*** := ***k*** +1
5. Jdi na bod **2**.

Apriori vlastnost přispívá k efektivitě celého procesu v bodě **2** i **3** !
(díky tomu, že některé mohutnosti byly na datech spočítány už v předchozích krocích výpočtu)

Příklad na generování pravidel

Množina častých položek z dat „golf/tenis“:

Je toto množina častých položek?
{Humidity = normal, Windy = False, Play = Yes }

Podpora je 4

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Určete **spolehlivost** pro všechna pravidla ze zvolené množiny častých položek:

{Humidity = Normal,Windy = False, Play =Yes}

7 potenciálních pravidel (dva typy zápisu), např.

If Humidity = Normal & Play = Yes then Windy = False

{Humidity = Normal,Play = Yes} => {Windy = False} **4/6**

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	**Normal**	**False**	**Yes**
Rainy	Cool	**Normal**	True	No
Overcast	Cool	**Normal**	True	**Yes**
Sunny	Mild	High	False	No
Sunny	Cool	**Normal**	**False**	**Yes**
Rainy	Mild	**Normal**	**False**	**Yes**
Sunny	Mild	**Normal**	True	**Yes**
Overcast	Mild	High	True	Yes
Overcast	Hot	**Normal**	**False**	**Yes**
Rainy	Mild	High	True	No

If Humidity = Normal and Play = Yes then Windy = False **4/6**
If Windy = False and Play = Yes then Humidity = Normal **4/6**
If Humidity = Normal then Windy = False and Play = Yes **4/7**
If Windy = False then Humidity = Normal and Play = Yes **4/8**
If Play = Yes then Humidity = Normal and Windy = False **4/9**
If True then Humidity = Normal and Windy = False and Play = Yes **4/14**

Pravidla pro data „golf/tenis“

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

- **Jaká jsou pravidla mající sup > 1 a conf = 100% ?**

	Association rule		Sup.	Conf.
1	Humidity=Normal & Windy=False	⇒Play=Yes	4	100%
2	Temperature=Cool	⇒Humidity=Normal	4	100%
3	Outlook=Overcast	⇒Play=Yes	4	100%
4	Temperature=Cold & Play=Yes	⇒Humidity=Normal	3	100%
...	...	...	...	...
58	Outlook=Sunny & Temperature=Hot	⇒Humidity=High	2	100%

- **Celkem**: 3 pravidla se sup=4, 5 s sup=3 a 51 se sup=2

Filtrování asociačních pravidel

- **Problém:** Každý rozsáhlý datový soubor vede ke vzniku obrovského množství různých asociačních pravidel, a to i v případě, že požadujeme rozumnou minimální spolehlivost a podporu. Asociační pravidla, která mají vyšší hodnotu podpory a spolehlivosti, než je předem zvolená hodnota pro min_sup a min_conf, se nazývají "***silná (strong)***" pravidla

- *Spolehlivost sama o sobě nestačí k výběru opravdu zajímavých pravidel !*

  - Např. pokud všechny transakce v souboru obsahuji množinu položek **Z**, pak libovolné pravidlo tvaru

    I => Z bude mít spolehlivost 100%.

- Pak je třeba využít i další míry pro filtrování pravidel

Další míry
pro pravidlo *Ant* => *Suc*

	Suc	Non (***Suc***)	∑
Ant	**a**	**b**	**r** = a+b
Non (***Ant***)	**c**	**d**	**s** = c+d
∑	**k** = a+c	**l** = b+d	**n**= r+s

podpora(*support*) = **a/n**

spolehlivost (*confidence*) = **a/r**

pokrytí (*cover*) = **a/k**

lift (*zdvih*) = (a/r)/(k/n)= **a*n/(r*k)**
"Zdvih určuje stupeň, o nějž pravidlo zlepšuje přesnost defautní predikce svého důsledku ***Suc*** na výchozích datech"

páka (*leverage*) = **(a-r*k/n)/n** = [a- (r/n)*(k/n) * n]/n "Hodnota udávající '*procento*' transakcí, které pravidlo pokývá navíc oproti odhadu, který bychom udělali za předpokladu, že ***Ant*** a ***Suc*** jsou nezávislé"

přesvědčivost (*conviction*) = **r*l/(b*n)=** (l/n)/(b/r) "Podobá se zdvihu, ale soustředí se na transakce, které nejsou pokryty ***Suc***. Proto využívá převrácený poměr četností!"

ZDVIH asoc. pravidla: **interpretace**

- Zdvih asociačního pravidla $I => J$ je definován takto:
  - Zdvih: $\text{lift}(I => J) = P(J|I) / P(J)$
  - Připomenutí: $P(J)$ = (podpora J) / (počet všech transakcí)
  - Poměr mezi spolehlivosti pravidla $I => J$ a předpokládanou spolehlivostí tak, jak platí na výchozích datech

- **Interpretace**:
  - Pokud $\text{lift}(I => J) > 1$, pak jsou I a J positivně korelované

    pokud $\text{lift}(I => J) < 1$, pak jsou I a J negativně korelované

    je-li $\text{lift}(I => J) = 1$, pak jsou I a J nezávislé

Osnova

- Transakce
- Časté skupiny položek
- Apriori vlastnost podmnožin
- Asociační pravidla
- **Aplikace**

Aplikace

- Analýza spotřebního koše může být využita pro
  - Změnu způsobu prezentace zboží (např. uspořádání zboří v prostoru)
  - Tvorbu cílené nabídky dalšího zboží (např. v internet. Knihkupectví typu Amazon)
  - Identifikaci překvapivých jevů jako např. „*Mléko se obvykle kupuje současně s chlebem. Toto však neplatí o sojovém mléce.* " Tyto jevy mohou být významné nebo mohou vypovídat třeba o organizaci výchozího experimentu, např. „*Pokud dlouhodobý kuřák přestává kouřit, jeho zdravotní stav se zhoršuje*."
  - …
- Při analýze dat o pojistných událostech mohou asoc.pravidla upozornit na jevy, které si zaslouží uvažovat o novém typu služby
  - if (Car=Porsche & Gender=Male & Age < 20) then
    (Risk=high & Insurance = high)
- …, např. identifikace změny (WSARE – What is Strange About Recent Events), …