

# Dolování asociačních pravidel

Miloš Trávníček  
UIFS FIT VUT v Brně

# Obsah přednášky

1. Proces získávání znalostí
2. Asociační pravidla
3. Dolování asociačních pravidel
4. Algoritmy pro dolování asociačních pravidel
5. Analýza získaných výsledků
6. Ukázková aplikace v SAS Enterprise Miner

# Proces získávání znalostí

- Netriviální získávání implicitních, dosud neznámých, pochopitelných a potenciálně užitečných znalostí
- **Znalost** = pravidla, omezení, pravidelnost
- **Několik kroků celého procesu:**
  - Stanovení cílů
  - Integrace, výběr a předzpracování dat
  - Výběr dolovacích prostředků
  - Získání znalostí
  - Interpretace výsledků

# Definice pojmů

## ■ Dolování v relačních datech

### – Relační tabulka

$$R=(H, R_B),$$

kde  $H$  je záhlaví relační tabulky a  $B$  je její tělo

### – **Tělo relační tabulky** = množina $n$ -tic řádků

$$R_B(t) = \{r_1, r_2, \dots, r_m(t)\},$$

### – **Doména atributu** = množina skalárních hodnot téhož typu, jichž může atribut nabýt

# Definice pojmů

- **Typy atributu** (podle velikosti domény  $D_i$ )
  - **Kategorický** = doména atributu je konečná
  - **Kvantitativní** = doména atributu je nekonečná
  - **Booleovský** = atribut nabývá hodnot *false* x *true*
- **Transakční databáze**
  - Pokud pro domény všech atributů platí, že jsou booleovského typu

# Asociační pravidla

- Snaha zjistit mezi položkami takový vztah, že přítomnost jedné nebo více položek implikuje přítomnost jiných položek v téže transakci
- **Definice:**

$$A \Rightarrow B$$

kde  $A, B$  jsou vzájemně disjunktní množiny položek z množiny  $I$  (= databáze)

# Asociační pravidla

- **Motivace získávání asociačních pravidel**
  - Úloha „Analýza nákupního košíku“

Snažíme se zjistit závislosti mezi jednotlivými prodejními položkami (př. asoc. pravidla: „pokud si zákazník kupuje mléko, zpravidla si koupí chleba“)
  - Využití výsledků pro rozmístování zboží, vytváření nabídkových katalogů, marketingové rozhodování apod.

# Asociační pravidla

## ■ Metriky pro asociační pravidla

- Podpora (support) – pravidlo  $A \Rightarrow B$  platí s podporou  $\text{supp}$ , pokud  $\text{supp} * 100\%$  řádků v relační tabulce obsahuje položky reprezentované predikáty z obou stran asociačního pravidla = **frekvence výskytu pravidla v databázi**
- Spolehlivost (confidence) - pravidlo  $A \Rightarrow B$  má spolehlivost  $c$ , pokud  $c * 100\%$  řádků v relační tabulce obsahující položky z  $A$  obsahuje také položky z  $B$  = **síla implikace v asociačním pravidle**



# Nalezení asociačních pravidel

- **Silná pravidla** – pravidla s vysokou (předem určenou) hodnotou a spolehlivostí
- **Cíl** – nalézt taková pravidla, která jsou „silná“
- **Dolování asociačních pravidel** = nalezení silných pravidel

# Nalezení asociačních pravidel

## Dva kroky procesu nalezení asociačních pravidel:

### 1. Generování frekventovaných vzorů

- = nalezení predikátů, které mají podporu vyšší, než je zadaná minimální podpora
- = nalezení *silných množin*

### 2. Generování asociačních pravidel

- = vygenerování asoc. pravidel s využitím silných množin
- = odstranění pravidel, jejichž *spolehlivost* (confidence) nedosahuje předem určené minimální hodnoty (minconf)

# Nalezení asociačních pravidel

## Definice frekventovaného vzoru

- $fp$  .... frekventovaný vzor, konjunkce predikátů tvaru  $a_1 \wedge a_2 \wedge \dots \wedge a_n$  (konjunkce predikátů), kde jednotlivé predikáty odpovídají hodnotám určitého počtu řádků, v případě kvantitativních atributů musí být hodnota uvnitř určitého intervalu
- $FP$  .... množina frekventovaných vzorů
- $s(x)$  .. podpora

$$FP = \{fp \mid s(fp) \geq \text{minsup}\}$$

# Nalezení asociačních pravidel

## Definice silných asociačních pravidel

- $ar$  ... asociační pravidlo tvaru  $A \Rightarrow B$ , kde  $A, B$  jsou konjunkce predikátů tvaru  $a_1 \wedge a_2 \wedge \dots \wedge a_n$
- $c(ar)$  ... spolehlivost pravidla
- $s(ar)$  ... podpora pravidla

Poté je *množina silných pravidel*  $AR$  definována jako:

$$AR = \{ar \mid c(ar) \geq minconf \wedge s(ar) \geq minsup\}$$

# Nalezení asociačních pravidel

## Faktory, určující výkon dolovacího algoritmu:

- Efektivita *generování frekventovaných vzorů*
- Časová a paměťová náročnost *generování frekventovaných vzorů*
- Druhý krok (*generování asociačních pravidel*) je jednoduchý a nemá v konečném důsledku větší vliv na výkon dolovacího algoritmu

# Algoritmy pro dolování asociačních pravidel

- Základem algoritmus nalezení velkých množin – tzv. *kandidátů*, založený na průchodu databází, určování možných kandidátů a počítání jejich podpory (*support*)
- **Apriori**
- **AprioriTID** – ukládají se kandidáti na frekventované množiny, kteří jsou v transakci obsaženi
- **Aprioritemset** – ukládají se i transakce, v nichž jsou kandidáti obsaženi (formou vektoru binárních čísel, vyjadřujících přítomnost/nepřítomnost kandidáta v transakci)

# Algoritmus APRIORI

- Algoritmus „prochází“ postupně databázi a počítá podporu pro kandidáty
- V každém kroku *generování tzv. kandidátů a poté kontrola minimální podpory.*
- V dalším kroku vznik kandidátů o velikosti o jednu větší, než v předchozím kroku atd.
- Konec při nenalezení kandidátů dané velikosti
- Funkce *AprioriGen* generuje všechny možné k-množiny (kandidáty) z frekventovaných množin a poté vylučuje z výběru ty množiny, jejichž některá podmnožina není frekventovaná („podpora k-množiny nemůže být větší, než podpora její podmnožiny“)

# Algoritmus APRIORI

- Výpočet podpory pro všechny prvky z množiny  $C_1$   
for (k=2;;k++) begin  
     $L_{k-1}$ ={ kandidáti z  $C_{k-1}$ , kteří mají podporu vyšší než minimální}  
    if (množina  $L_{k-1}$  je prázdná) break  
     $C_k$ =AprioriGen( $L_{k-1}$ );  
    for each (transakce  $t$ ) begin  
         $C_t$ = subset( $C_k$ ,  $t$ );  
        for each (kandidáti  $c \in C_t$ ) zvýš o 1 podporu kandidáta  $c$   
    end;  
end.



# Další typy asociačních pravidel

- **Víceúrovňová asociační pravidla**
  - nad položkami v transakcích je definována konceptuální hierarchie, která je sdružuje do tzv. *konceptů* (jablko, pomeranč = ovoce)
- **Asociační pravidla založená na omezeních**
- **Zobecněná asociační pravidla**

# Ukázka v aplikaci SAS Enterprise Miner

- Sociologický průzkum
- Výběr cvičných dat, úprava souboru dat, definování proměnných
- Určení požadovaného výsledku
- Sestavení „flow diagramu“
- Generování výstupů
- Určení vhodných asociačních pravidel

# Definice problému

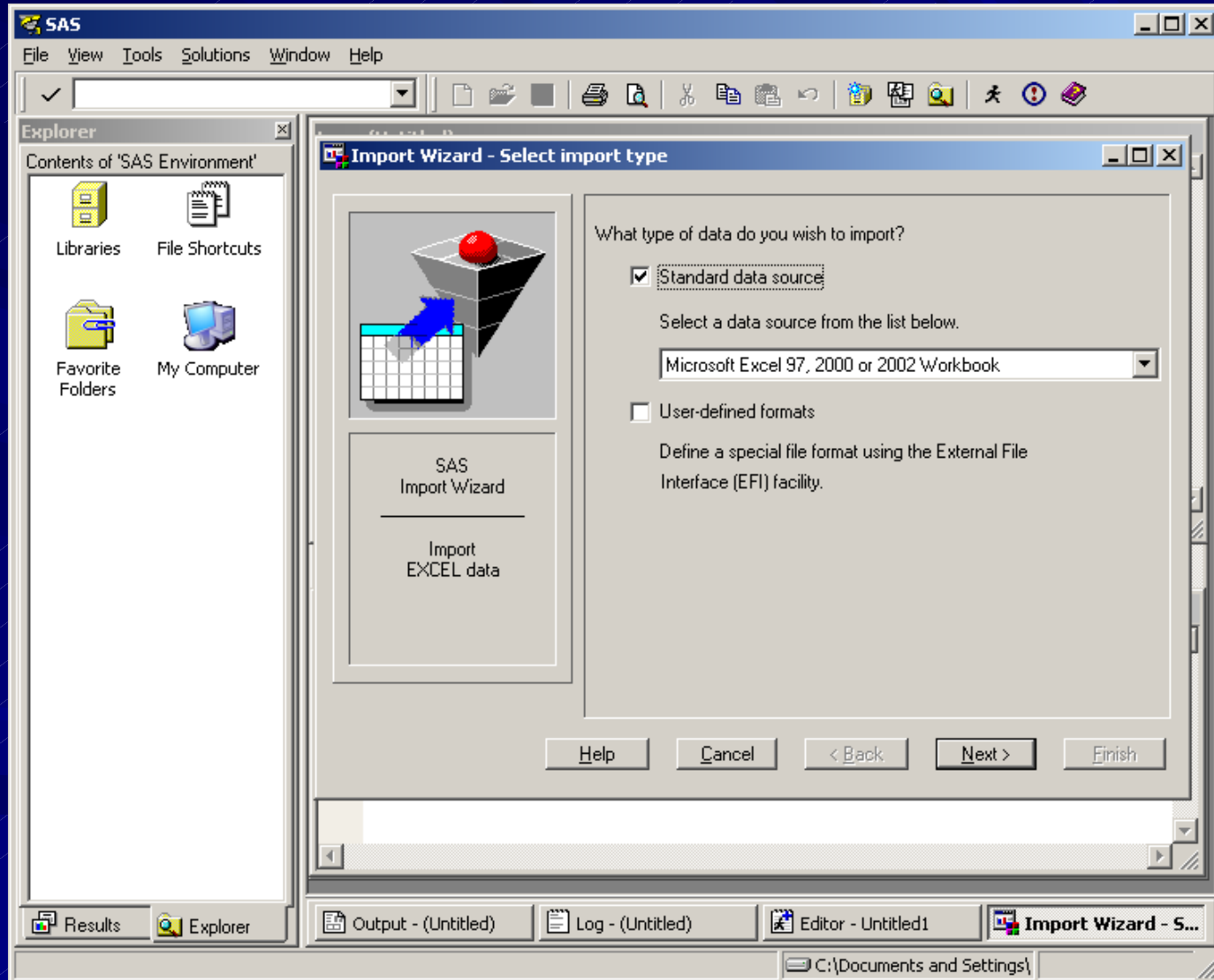
- Máme rozsáhlý soubor statistických dat, zobrazující údaje o mužích středního věku (získáno ze studie STULONG, viz. <http://euromise.vse.cz/stulong/index.php>)
- Chceme zjistit, které faktory nejvíce ovlivňují zvýšenou hladinu cholesterolu u některých jedinců v tomto věku
- Prosté zobrazení dat např. do grafu by v tomto případě nebylo dostatečně vypovídající
- Jako „vedlejší produkt“ můžeme také zjistit například souvislost se sociálním postavením a životním stylem, věkem apod.

# Úprava dat

- Vstupní data je nutno **upravit** do podoby „nákupního košíku“, jeden záznam tedy bude vypadat následujícím způsobem:

| ID    | Položka              |
|-------|----------------------|
| 10001 | kouri 21 a vice let  |
| 10001 | stari 46-50 let      |
| 10001 | 15-20 cigaret        |
| 10001 | cholesterol normalni |
| 10001 | mirna aktivita       |
| 10001 | pije prilezitostne   |
| 10001 | sedi v praci         |
| 10001 | vedouci              |
| 10001 | zenaty               |

# Import dat



# Nastavení uzlu Association

The screenshot displays the SAS software interface with the 'Association' node configuration window open. The window is titled 'Association' and has several tabs: 'Selected Output', 'Notes', 'Data', 'Variables', 'General', 'Sequences', 'Time Constraints', 'Sort', and 'Output'. The 'General' tab is active, showing the following settings:

- Analysis mode:  By Context  Association  Sequences
- Minimum Transaction Frequency to Support Associations:
  - 5% of largest single item frequency
  - Specify as a percentage:  \*
  - Specify a count:
- Maximum number of items in an association:
- Minimum confidence for rule generation:

At the bottom of the window, there is a 'Replacement' dropdown menu, a 'Diagrams' tab, and a 'Reports' tab. The status bar at the bottom indicates 'Editing association settings'.

The background shows the SAS Explorer window with the contents of the 'Work' directory, including folders like 'Dm', 'Em', 'Formats', 'Report', 'Stulong', 'Temp', and files like 'Tablemeta', '\_classpctmiss\_', '\_emtrain', and '\_listds\_'.

# Flow diagram

The screenshot displays the SAS Enterprise Miner interface. On the left, the Explorer pane shows the contents of the 'Work' folder, including folders like 'Dm' and 'Formats', and files like 'Stulong', 'Tablemeta', '\_classpctmiss\_', and '\_emtrain'. The main workspace is titled 'SAS Enterprise Miner - travnice [my\_first]' and contains a flow diagram. The diagram starts with an 'Input Data Source' icon labeled 'WORK.STULONG'. An arrow points to a 'Data Set Attributes' icon. From there, an arrow points to an 'Association' icon. Below the 'Input Data Source' icon, an arrow points to an 'Insight' icon. Below the 'Association' icon, an arrow points to a 'Reporter' icon. The diagram is titled 'Diagrams' at the bottom. The status bar at the bottom shows 'Output - (Untitled)', 'Log - (Untitled)', 'Editor - Untitled1', and 'SAS Enterprise Mi...'. The system tray shows 'C:\Documents and Settings\'.

# Výsledná pravidla

The screenshot displays the SAS interface with the 'Results - Association' window open. The window shows a table of association rules with columns for Relations, Lift, Support(%), Confidence(%), Transaction Count, and Rule. The rules are numbered 1 through 14. The 'Rule' column contains various logical expressions such as 'zenaty ==> mira aktivita' and 'mira aktivita ==> zena'.

|    | Relations | Lift | Support(%) | Confidence(%) | Transaction Count | Rule                      |
|----|-----------|------|------------|---------------|-------------------|---------------------------|
| 1  | 2         | 1.03 | 63.59      | 74.65         | 901.00            | zenaty ==> mira aktivita  |
| 2  | 2         | 1.03 | 63.59      | 87.65         | 901.00            | mira aktivita ==> zena    |
| 3  | 2         | 1.00 | 52.51      | 72.37         | 744.00            | mira aktivita ==> chol    |
| 4  | 2         | 1.00 | 52.51      | 72.87         | 744.00            | cholesterol normalni ==   |
| 5  | 2         | 1.03 | 45.94      | 53.94         | 651.00            | zenaty ==> sedi v praci   |
| 6  | 2         | 1.03 | 45.94      | 88.09         | 651.00            | sedi v praci ==> zenaty   |
| 7  | 2         | 1.01 | 45.38      | 53.27         | 643.00            | zenaty ==> pije prilezito |
| 8  | 2         | 1.01 | 45.38      | 85.96         | 643.00            | pije prilezitostne ==> ze |
| 9  | 2         | 1.04 | 40.01      | 75.80         | 567.00            | pije prilezitostne ==> mi |
| 10 | 2         | 1.04 | 40.01      | 55.16         | 567.00            | mira aktivita ==> pije    |
| 11 | 2         | 1.01 | 38.18      | 73.21         | 541.00            | sedi v praci ==> mira     |
| 12 | 2         | 1.01 | 38.18      | 52.63         | 541.00            | mira aktivita ==> sedi    |
| 13 | 2         | 1.01 | 33.38      | 46.01         | 473.00            | mira aktivita ==> kouri   |
| 14 | 2         | 1.01 | 33.38      | 73.11         | 473.00            | kouri 21 a vice let ==>   |

Below the table, there are tabs for 'Diagrams', 'Tools', and 'Reports'. A status bar at the bottom of the window indicates 'Data set with one way frequency created.' The taskbar at the bottom shows several open windows: 'Results', 'Explorer', 'Output - (Untitled)', 'Log - (Untitled)', 'Editor - Untitled1', and 'Results - Associat...'. The system tray shows the path 'C:\Documents and Settings\...'.



# Analýza výsledků

| ZE | EXP_CON | CONF  | SUPP  | LIFT | COUNT | RULE  |
|----|---------|-------|-------|------|-------|---|
| 3  | 72,05   | 76,25 | 17,22 | 1,06 | 244   | mirna aktivita & castecne nezavisly pracovn<br>==> cholestrol normalni          |
| 4  | 72,05   | 75,32 | 8,19  | 1,05 | 116   | nekurak & kouri 0 roku & jiny pracovník ==><br>cholestrol normalni              |
| 4  | 26,89   | 34,69 | 4,80  | 1,29 | 68    | mirna aktivita & kouri 21 a vice let & 15-20<br>cigaret ==> zvyse ny cholestrol |
| 4  | 26,89   | 35,53 | 5,72  | 1,32 | 81    | zenaty & kouri 21 a vice let & 15-20 cigaret<br>==> zvyse ny cholestrol         |
| 4  | 26,89   | 35,41 | 6,42  | 1,32 | 91    | zenaty & mirna aktivita & 15-20 cigaret ==><br>zvyse ny cholestrol              |
| 4  | 26,89   | 34,29 | 4,23  | 1,28 | 60    | zenaty & pije prilezitostne & 15-20 cigaret<br>==> zvyse ny cholestrol          |
| 4  | 26,89   | 37,50 | 4,23  | 1,39 | 60    | zenaty & sedi v praci & 15-20 cigaret ==><br>zvyse ny cholestrol                |
| 4  | 26,89   | 31,75 | 4,73  | 1,18 | 67    | zenaty & kouri 21 a vice let & 21 a vice<br>cigaret ==> zvyse ny cholestrol     |

# Dotazy