9. NĚKTERÉ DALŠÍ ANALÝZY

Čas ke studiu: 3 hodiny

Cíl Po prostudování této kapitoly budete umět

- popsat metodu pro studium závislosti dvou reálných veličin a na praktických příkladech ji použít
- popsat metodu pro studium závislosti dvou kategoriálních veličin a na praktických příkladech ji použít
- popsat analýzu výjimek a na praktických příkladech ji použít

Výklad

9.1. Metoda CORREL

❑ Studie závislosti dvou reálných atributů

Pro určení lineární závislosti dvou reálných veličin používáme obvykle koeficient korelace. Víme, že pro nezávislé veličiny je výsledná hodnota koeficientu korelace blízká nule. Pro lineárně závislé veličiny se blíží některé z hodnot ±1. Graf závislosti obou veličin se pak blíží přímce.

Mějme opět datovou matici **X** s objekty {O1, O2, …} v řádcích a atributy {A1, A2, …} ve sloupcích. Otestujeme dvě veličiny – dva reálné atributy A1, A2 na celé množině objektů a dostaneme korelaci velmi blízkou nule, tedy považujeme obě veličiny za nezávislé.

X

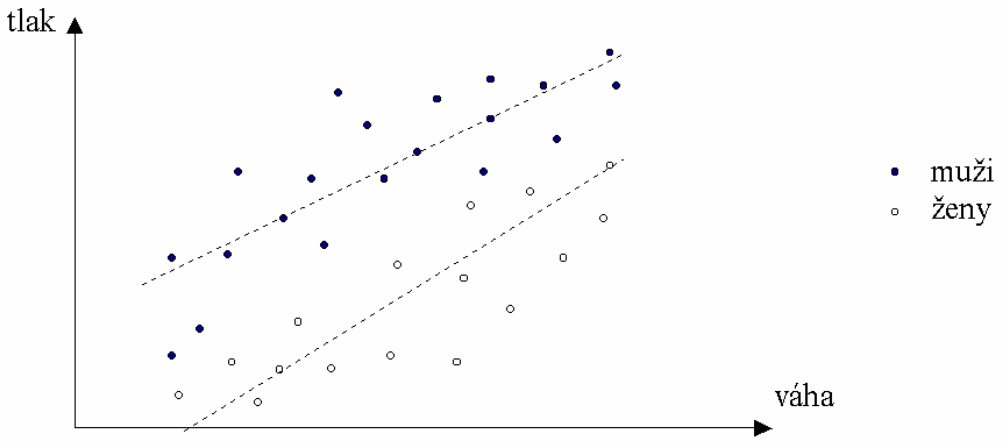
	A1	**A2**	**…**	**Aj**	**An**
O1	x11	x12			
	…	…			
Oi	xi1	xi2			
	…	…			
Om	xm1	xm2			

Může se však stát, že pro některé podmnožiny objektů jsou obě veličiny lineárně závislé a hodnota korelace se blíží ±1. Podmnožiny obvykle vybíráme pomocí hodnot některých ostatních atributů – jednoho (Aj = J) nebo konjunkce několika (Aj=J ∧ Ak= K ∧ …).

Pak můžeme konstatovat, že za podmínky P (odpovídající některému tvaru výše uvedené elementární konjunkce) jsou veličiny A1, A2 lineárně závislé.

Příklad 9.1.

Mezi atributy sledovaných lidí jsou také váha a tlak. *Jejich graf je na následujícím obrázku. Vidíme, že všechny body rozhodně nepřipomínají přímku. Ovšem když vybereme zvlášť muže a zvlášť ženy, obě podmnožiny se již přímce blíží výrazně více.*



Obrázek 9.1. Závislost váhy a tlaku dle pohlaví

♦

❑ Metoda CORREL

Jako u všech metod dolování znalostí i následující metoda netestuje jen předem formulované tvary hypotéz o možných závislostech, ale systematicky je generuje, testuje a vydává zprávu o daty podporovaných hypotézách.

Procedura CORREL generuje zajímavé pravdivé sentence tvaru

(A1 corr A2) / KONJ

kde A1, A2 jsou reálné veličiny, atributy,
KONJ je elementární konjunkce - podmínka
corr je některý z korelačních kvantifikátorů.

Při jednom výpočtu jsou veličiny A1, A2 pevné, podmínka se mění.

Vstupní parametry procedury obsahují :

- reálné veličiny A1 a A2, jejichž vztah chceme zkoumat,
- použitý korelační kvantifikátor a pro něj upřesňující parametry výpočtu:

  pro Kendallův kvantifikátor k-corr α kde α = hladina významnosti
  Spearmannův kvantifikátor s-corr$_\alpha$ α jako výše
  pořadově ekvivalenční e-corr s s = minimální frekvence pro podmínku

- povolený tvar podmínky obdobně jako u ASSOC:

  množina predikátů důležitých B (v elementární konjunkci alespoň jeden) a ostatních C,
  pro množiny B a C ke každému predikátu povolený tvar pozitivní, negativní nebo obojí,
  maximální povolenou délku podmínky.

Shrnutí pojmů 9.1.

Analýza vztahu dvou atributů reálných.

Výpočet korelace pro generované podmnožiny objektů za podmínky tvaru elementární konjunkce.

Otázky 9.1.

1. Kterými metodami je možno analyzovat vzájemný vztah mezi dvěma atributy, pokud nabývají hodnot reálných?
2. Popište princip metody CORREL a rozdíl této metody proti statistickému výpočtu korelace dvou veličin.

Úlohy k řešení 9.1.

1. Formulujte algoritmus pro metodu CORREL.
2. Napište program, realizující tento algoritmus.
3. Najděte alespoň 3 praktické úlohy na využití metody CORREL.
4. Jsou dána data AUTA03 z Torontského statistického úřadu obsahují informace o nových typech aut uvedených na trh v roce 2003. Soubor obsahuje 93 nových typů aut od 25 výrobců uvedených na trh. Každé typ vozidlo je charakterizováno 19 atributy.
   1. **Výrobce** - kateg

0 = Acura	10 = Geo	20 = Subaru
1 = Audi	11 = Hyundai	21 = Toyota
2 = BMW	12 = Lexus	22 = Volkswagen
3 = Buick	13 = Lincoln	23 = Volvo
4 = Caddilac	14 = Mazda	24 = Saturn
5 = Chevrolet	15 = Mercedes-Benz	
6 = Chrysler	16 = Mitsubishi	
7 = Dodge	17 = Nissan	
8 = Eagle	18 = Oldsmobile	
9 = Ford	19 = Pontiac	

   2. **Třída** - dělení typů bylo převzato od výrobců aut
      0 = malé auto
      1 = sportovní auto
      2 = auto střední třídy
      3 = auto vyšší třídy
      4 = nákladní vozidlo
   3. **Minimální cena** = cena za základní model (bez doplňků)
   4. **Střední cena** = průměr minimální a maximální ceny vozu
   5. **Maximální cena** = cena za nadstandartní model (s maximálními doplňky)
   6. **Spotřeba ve městě** (v mílích na galon)
   7. **Spotřeba na dálnici** (v mílích na galon)
   8. **Vybavenost airbagy**: 0 = žádný, 1 = jeden (řidič), 2 = dva (řidič i spolujezdec)
   9. **Náhon**: počet kol, na které je náhon
   10. **Počet válců**
   11. **Objem motoru** (v litrech)

12. **Maximální počet koňských sil**
13. **Vybavenost automatickou převodovkou**: 0 = Ne, 1 = Ano
14. **Kapacita nádrže** (v galonech)
15. **Maximální počet pasažérů**
16. **Délka** (v palcích)
17. **Šířka** (v palcích)
18. **Hmotnost** (v librách)
19. **Vozidlo je výroby:** 0 = neamerické, 1 = americké

Najděte možné analýzy těchto dat pomocí metody CORREL a formulujte, jaký typ výsledku může metoda nalézt.

9.2. Metoda COLLAPS

Cíl Po prostudování této kapitoly budete umět

- popsat metodu pro studium závislosti dvou kategoriálních veličin a na praktických příkladech ji použít

Výklad

Studie závislosti dvou kategoriálních atributů

Pro určení závislosti dvou kategoriálních atributů používáme obvykle frekvenční tabulku, případně z ní vypočtené koeficienty asociace. Ovšem jediný koeficient asociace opět není dostatečnou charakteristikou vztahu dvou kategoriálních veličin. Jednou z metod, která systematicky vyhledává možné vztahy mezi podmnožinami hodnot obou atributů, je metoda kolapsování. Tato metoda

- analyzuje podrobně vztah dvou kategoriálních atributů, vyhledává zdroje závislosti,
- provádí kolapsování (sečítání) hodnot několika řádků a sloupců původní kontigenční tabulky, aby našla nejsilnější vztahy mezi dvěma kategoriálními atributy
- generuje zajímavé pravdivé sentence tvaru

$$A1[Ki] \sim A_2[Kj]$$

kde A1, A2 jsou zvolené atributy,

K1, K2 jsou koeficienty vzniklé z podmnožin hodnot kategorií atributů A1, A2 takové, že jejich numerické charakteristiky vyhovují určitým podmínkám, které říkají, že veličiny spolu v datech souvisejí.

Vstupem pro proceduru jsou složené dvouhodnotové veličiny A1[K1] a A2[K2] a z nich vypočtené frekvence a, b, c, d. Na ně se aplikuje některý asociační koeficient.

Název procedury plyne z toho, že frekvence a, b, c, d dostaneme sečtením - kolapsováním řádků a sloupců úplné frekvenční tabulky obou kategoriálních veličin.

Frekvenční (kontigenční) tabulka atributů A1[s hodnotami H1 ... Hh] a A2 [s hodnotami G1 ... Gg]:

A1\A2	G1	G2	…	Gg	
H1	a_{11}	a_{12}		a_{1g}	r1
H2	a_{21}	a_{22}		a_{2g}	r2
…					…
Hh	a_{h1}	a_{h2}		a_{hg}	rh
	k1	k2	…	kg	m

⇒

K1\K2	ano	ne	
ano	a	b	k
ne	c	d	l
	r	s	m

Metoda vyšetřuje **všechny dvojice neprázdných koeficientů K1 (podmnožin hodnot H1 ... Hh) a K2 (podmnožin hodnot G1 ... Gg)** a jako výsledek vydává všechny ty dvojice, které prošly testem

pro zvolený kvantifikátor a zvolené parametry. Výsledky uspořádá podle hodnoty kvantifikátoru, od největšího.

Vstupní parametry upřesňující, co je pro řešitele zajímavé, a jaké požaduje výstupy:

- Atributy A1, A2, z nichž jsou odvozeny binární veličiny A1[K1] a A2[K2], pro něž se má procedura provést.
- Použitý asociační kvantifikátor a parametry upřesňující jeho použití:

  N = maximální povolená délka řešení, maximální počet hypotéz,
  C = χ^2, kritická hladina
  Q = omezení podílu charakteristiky nejlepšího a nejhoršího prvku řešení (o kolik může být poslední hypotéza horší, než první).
- Zda řešíme úlohu nehierarchickou nebo hierarchickou (viz níže).
- Informaci o požadovaných výsledcích. Základním výsledkem je blokový rozklad frekvenční tabulky s absolutními frekvencemi. Na požadavek řešitele je možno počítat další charakteristiky jako

řádkové relativní frekvence	$(a_{ij} / r_i) * 100$
sloupcové relativní frekvence	$(a_{ij} / k_j) * 100$
relativní frekvence	$(a_{ij} / m) * 100$
očekávané frekvence	$o_{ij} = r_i * k_j / m$
odchylky	$a_{ij} - o_{ij}$
standardizované odchylky	$(a_{ij} - o_{ij}) / \sqrt{o_{ij}}$

❑ Příklad použití metody COLLAPS

Příklad 9.3.

V datech o 592 sledovaných lidech jsou mj. atributy barva očí *a* barva vlasů. *Frekvenční tabulka těchto dvou atributů je*:

OČI \ VLASY	1 = černá	2 = brunet	3 = zrzavá	4 = blond	SUMA
1 = hnědé	68	119	26	7	220
2 = šedé	15	54	14	10	93
3 = zelené	5	29	14	16	64
4 = modré	20	80	17	94	215
SUMA	108	286	71	127	592

Kvantifikátor CHIQ (N=3, C=0, Q=0) se vypočte pro koeficienty.

OČI [1] ~ VLASY [1]OČI [1,2] ~ VLASY [1] OČI [1] ~ VLASY [1,2]
OČI [1] ~ VLASY [2]OČI [1,2] ~ VLASY [2] OČI [1] ~ VLASY [1,3]
...
OČI [2] ~ VLASY [1] OČI [1,3] ~ VLASY [1] OČI [1] ~ VLASY [2,4]
OČI [2] ~ VLASY [2] OČI [1,3] ~ VLASY [2] OČI [1] ~ VLASY [3,4]
...
OČI [3,4] ~ VLASY [3,4]

(ostatní kombinace jsou doplňkové a není je třeba počítat znovu).

Pak posloupnost výsledků uspořádaná podle výsledné hodnoty CHIQ je:

1. OČI [3, 4] ~ VLASY [4] — tj. zelené nebo modré oči souvisí s blond vlasy
   OČI [1, 2] ~ VLASY [1,2,3] — doplňkový tvar téže hypotézy, pro obě je v([3, 4], [4]) = **101.17**
2. OČI [4] ~ VLASY [4] — v([4], [4]) = **99.35**
3. OČI [1] ~ VLASY [1,2,3] — v([1], [1, 2, 3]) = **69.36**

Všechny tři hypotézy nejsou v rozporu, obecně s tmavšími vlasy souvisí tmavší oči.

Volbou první hypotézy dostáváme výsledný rozklad frekvenční tabulky:

OČI\VLASY	4=blond	1=čer	2=bru	3=zrz	
3 = zelené	16	5	29	14	
4 = modré	94	20	80	17	
1 = hnědé	7	68	119	26	
2 = šedé	10	15	54	14	
					592

⇒

O\V	4	1,2,3	
3,4	110	165	
1,2	17	296	
			592

♦

Hierarchická varianta metody COLLAPS

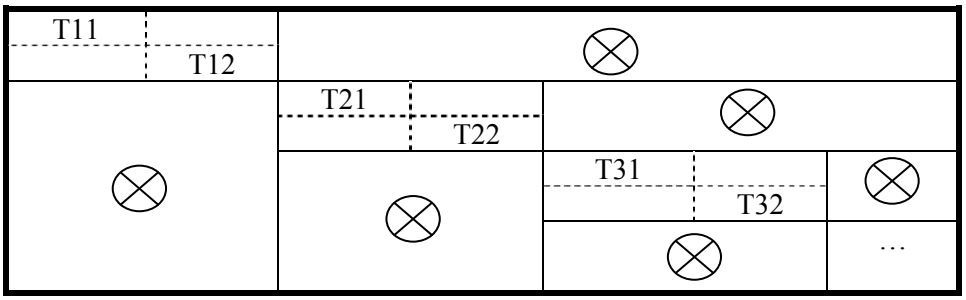
Procedura COLLAPS může řešit i hierarchickou úlohu.

Nehierarchickou úlohou rozumíme popsané vyhodnocení vztahu dvou atributů.

Hierarchická úloha znamená rekurzivní opakování této úlohy na podtabulkách četností, dokud podtabulky obsahují počet prvků > 1. V každém kroku se 2 nové podtabulky T1 a T2 tvoří z předcházející tak, že první obsahuje řádky a sloupce příslušné zpracovávaným veličinám K1 a K2, druhá řádky a sloupce ostatní.

Po prvním kroku (vyhledání K1, K2 s největší hodnotou kvantifikátoru) vytvoříme novou frekvenční tabulku tak, že vyškrtneme řádky s indexy nepatřícími do K1 a sloupce s indexy nepatřícími do K2. Takové podtabulky jsou dvě, základní T1 a doplňková T2.

Na tyto podtabulky aplikujeme znovu proceduru COLLAPS, dokud nejsou podtabulky jednořádkové nebo jednosloupcové a tedy je již nelze dělit.



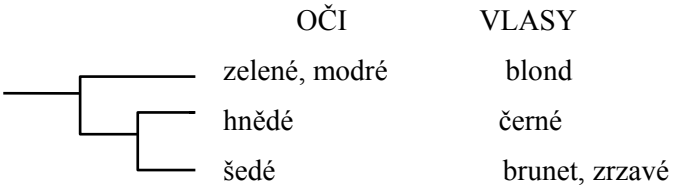
Výsledkem je pak binární strom nejlepších hypotéz, který vyjadřuje strukturu závislosti v tabulce.

Příklad 9.4.

V našem příkladě odpovídá plná čára rozkladu v 1. kroku. Krok 2 se neprovede, protože podtabulka T1 = [blond] je jednosloupcová. Krok 3 s rozkladem tabulky T2 = [černá, brunet, zrzavá] se provede (má 2 řádky a 3 sloupce), jeho výsledkem je rozklad označený přerušovanou čarou. Další kroky se neprovedou, protože další podtabulky jsou jednoprvkové.

OČI \ VLASY	4 = blond	1 = černá	2 = brunet	3 = zrzavá
3 = zelené	16	5	29	14
4 = modré	94	20	80	17
1 = hnědé	7	68	119	26
2 = šedé	10	15	54	14

Výsledek hierarchického rozkladu je vhodné vyznačit do stromové struktury, výsledkem je binární strom nejlepších hypotéz, vyjadřující strukturu závislostí v tabulce.



♦

❑ Uživatelem řízená hierarchie

Někdy může být vhodné ponechat na uživateli - analytikovi rozhodnutí, která hypotéza je nejzajímavější (nejsilnější) a které tabulky se mají dále rozkládat rekurzivní procedurou.

Příklad 9.5.

V našem případě může analytik například rozhodnout o výběru 2. hypotézy v 1. kroku. Obecná zkušenost totiž říká, že tmavší vlasy souvisí s tmavšíma očima a obě hypotézy mají jen velmi malý rozdíl hodnot CHIQ.

Pak by se rozkládala jiná podtabulka:

OČI \ VLASY	4 = blond	1 = černá	2 = brunet	3 = zrzavá
4 = modré	94	20	80	17
1 = hnědé	7	68	119	26
2 = šedé	10	15	54	14
3 = zelené	16	5	29	14

=

O\V	4	1,2,3
4	94	117
1,2,3	33	344

Pak z tabulky T2 vyjdou hypotézy:

1. hnědé oči černé vlasy 11.82
2. zelené oči zrzavé vlasy 7.64
3. zelené oči hnědé nebo zrzavé vlasy 6.73

První hypotéza má dostatečný odstup, zvolíme tedy ji a dostáváme rozklad:

OČI \ VLASY	4 = blond	1 = černá	2 = brunet	3 = zrzavá
4 = modré	94	20	80	17
1 = hnědé	7	68	119	26
2 = šedé	10	15	54	14
3 = zelené	16	5	29	14

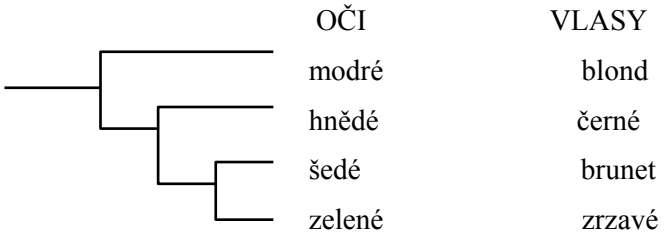
=

O\V	4	1	2,3
4	…	x	
1		68	145
2,3		20	111

Zbývá rozložit tabulku:

OČI \ VLASY	2 = brunet	3 = zrzavá
2 = šedé	54	14
3 = zelené	29	14

Celkově pak dostáváme výsledek, které odpovídá i naší intuitivní představě:



♦

Shrnutí pojmů 9.2.

Analýza vztahu dvou atributů kategoriálních.

Výpočet asociace pro generované podmnožiny hodnot atributů.

Hierarchie hodnot asociací dvojice atributů.

Otázky 9.2.

1. Kterými metodami je možno analyzovat vzájemný vztah mezi dvěma atributy, pokud nabývají hodnot kategoriálních?
2. Jaký je rozdíl mezi statistickým výpočtem CHIQ dvou veličin a metodou COLLAPS:
3. Jaký význam a využití má hierarchická metoda COLLAPS?

Úlohy k řešení 9.2.

1. Pro data AUTA03 (viz úloha 9.2./4) najděte zadání pro využití metody COLLAPS a formulujte význam případných výsledků.

9.3. Analýza výjimek

Cíl Po prostudování této kapitoly budete umět

- popsat úlohu analýza výjimek pro kategoriální atributy a na praktických příkladech ji použít

Výklad

Testování hypotéz prostřednictvím p-value

Úskalí klasických testů spočívá především v možnosti libovolného zvolení hladiny významnosti α. Nulovou hypotézu, kterou můžeme na jisté hladině významnosti zamítnout, lze při jinak zvoleném α přijmout. Libovolně zvolená hladina významnosti α tedy vede k libovolnému rozhodnutí. Tomuto problému se lze vyhnout, užijeme-li metodu testování hypotéz prostřednictvím p-value.

Konstrukce testu

1) stanovení nulové (H_0) a alternativní (H_A) hypotézy
2) volba výběrové statistiky a nulového rozdělení
3) určení hodnoty p-value
4) rozhodnutí o přijetí či zamítnutí nulové hypotézy

P-value

P-value je hodnota, která ukazuje jaká je shoda mezi daty a nulovou hypotézou. P-value lze definovat třemi různými způsoby, podle aktuálního tvaru H_A.

Jednostranné p-value

H_A: parametr populace < nulová hypotéza

p-value: pravděpodobnost, že výběrová statistika bude alespoň tak malá, jako skutečně zjištěná hodnota, za předpokladu, že H_0 je pravdivá

$$p - value = \int_{-\infty}^{x} f(t)dt = F(x) \qquad (1)$$

kde *f(t)* je hustota pravděpodobnosti a *F(x)* distribuční funkce

HA: parametr populace > nulová hypotéza

p-value: pravděpodobnost, že výběrová statistika bude alespoň tak velká, jako skutečně zjištěná hodnota, za předpokladu, že H_0 je pravdivá

$$p - value = \int_{x}^{\infty} f(t)dt = 1 - F(x) \qquad (2)$$

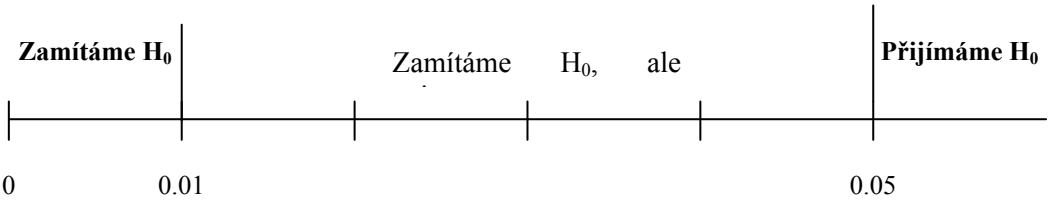
kde *f(t)* je hustota pravděpodobnosti a *F(x)* distribuční funkce

Oboustranné p-value

p-value: pravděpodobnost, že výběrová statistika bude alespoň tak extrémní (tj. velká nebo malá) jako skutečně zjištěná hodnota, za předpokladu, že H_0 je pravdivá

$$p - value = 2 * \min\{F(x),\ 1 - F(x)\} \qquad (3)$$

Rozhodovací proces



Obr. 1: Rozhodovací proces p-value

V případě, že se p-value nachází v intervalu (0.01 až 0.05), doporučuje se opakování testu se zvětšeným rozsahem výběru (obecně – čím menší p-value, tím menší je platnost H_0).

❑ Konstrukce testu použitá pro analýzu výjimek

1) Stanovení nulové hypotézy H_0, a alternativní hypotézy H_A

   a) pro jednovýběrový test

   $H_0 : p_T = p_Z$

   $H_A : p_T \neq p_Z$

   kde p_T je relativní četnost v testovaném (výběrovém) souboru, a p_Z relativní četnost v základním souboru

   b) pro dvouvýběrový test

   $H_0 : p_1 = p_2$

   $H_A : p_1 \neq p_2$

   kde p_1 relativní četnost v prvním testovaném (výběrovém) souboru, a p_2 relativní četnost ve druhém testovaném souboru

2) Určení testovacího kriteria (testovací statistiky)

   Jako testovací statistiku zvolíme

   a) pro jednovýběrový test

$$T(\overline{X}) = \frac{p_T - p_Z}{\sqrt{p_Z(1 - p_Z)}}\sqrt{n} \qquad (4)$$

   kde *n* značí rozsah výběru (počet prvků testovaného souboru).

b) pro dvouvýběrový test

$$T_2(\overline{X}) = \frac{p_1 - p_2}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad p_1 = \frac{x_1}{n_1};\ p_2 = \frac{x_2}{n_2};\ p = \frac{x_1 + x_2}{n_1 + n_2} \qquad \textbf{(5)}$$

kde x_1, resp. $x2$, značí počet jednotek výběru mající specifikovanou vlastnost (nazývaný často jako „počet úspěchů“ a pak $n_{1(2)} - x_{1(2)}$ jako „počet neúspěchů“) v prvním, resp. druhém testovaném souboru. Dále n_1, resp. n_2, značí rozsah výběru u prvního, resp. druhého testovaného souboru.

Obě testovací statistiky mají dle [4] normální normované rozdělení *N(0, 1)*. Jedná se o jeden z jednovýběrových, resp. dvouvýběrových testů, používaných pro testování hypotéz o četnostech kategorií [5]. Testujeme, zda se statisticky liší střední hodnoty některého z prvků v testovaném souboru a základním souboru při jednovýběrovém testu, resp. mezi dvěma testovanými soubory (výběry) při dvouvýběrovém testu.

3) Nulové rozdělení F_0

$$F_0(x) = P(T(\overline{x}) < x \mid H_0) = \Theta(x) \qquad \textbf{(6)}$$

4) Určení hodnoty p-value

$$p - value = 2 * \min\{F_0(x),\ 1 - F_0(x)\}$$

5) Rozhodovací proces

Při zvolené hladině významnosti α=0.05 (jinak řečeno pravděpodobnost chyby 1.druhu je 5 %, což znamená že v 95 případech ze 100 se zamítla nulová hypotéza oprávněně. U zbylých nejvýše 5 % případů byla zamítnuta neoprávněně a ve skutečnosti platí. Se spolehlivostí 95 % nebo větší bylo přijato správné rozhodnutí) rozhodneme podle hodnoty p-value o přijetí či zamítnutí nulové hypotézy. Pokud tedy p-value > 0.05, pak přijímáme nulovou hypotézu. Pokud hodnota p-value leží v intervalu 0.01 – 0.05, jedná se o nepřesvědčivou oblast. Při hodnotě p-value < 0.01 (resp. p-value < 0.001), zamítáme nulovou hypotézu (resp. zamítáme H_0 ještě „silněji“). Výsledek výpočtu může tedy nabývat jedné ze šesti hodnot:

1. p-value < 0.001 (zamítáme H_0) a pravděpodobnost prvku základního souboru je větší než pravděpodobnost prvku testovaného souboru ($p_Z > p_T$) pro jednovýběrový test, resp. pravděpodobnost prvku druhého testovaného souboru je větší než pravděpodobnost prvku prvního testovaného souboru ($p_2 > p_1$) pro dvouvýběrový test; označíme ji znakem (--)
2. p-value < 0.01 (zamítáme H_0) a pravděpodobnost prvku základního souboru je větší než pravděpodobnost prvku testovaného souboru ($p_Z > p_T$) pro jednovýběrový test, resp. pravděpodobnost prvku druhého testovaného souboru je větší než pravděpodobnost prvku prvního testovaného souboru ($p_2 > p_1$) pro dvouvýběrový test; označíme ji znakem (-)
3. p-value < 0.01 (zamítáme H_0) a pravděpodobnost prvku základního souboru je menší než pravděpodobnost prvku testovaného souboru ($p_Z < p_T$) pro jednovýběrový test, resp. pravděpodobnost prvku druhého testovaného souboru je menší než pravděpodobnost prvku prvního testovaného souboru ($p_2 < p_1$) pro dvouvýběrový test; označíme ji znakem (+)
4. p-value < 0.001 (zamítáme H_0) a pravděpodobnost prvku základního souboru je menší než pravděpodobnost prvku testovaného souboru ($p_Z < p_T$) pro jednovýběrový test, resp. pravděpodobnost prvku druhého testovaného souboru je menší než pravděpodobnost prvku prvního testovaného souboru ($p_2 < p_1$) pro dvouvýběrový test; označíme ji znakem (++)

První čtyři hodnoty tedy představují oblast zamítnutí nulové hypotézy, přičemž první dvě (1, 2) hodnoty odpovídají statisticky významně méně častému výskytu „úspěchu“ (jednotky výběru, mající specifickou vlastnost) v datech, a hodnoty 4 a 5 odpovídají statisticky významně častému výskytu „úspěchu“ v datech.

5. p-value leží mezi 0.01 a 0.05. Jedná se o nepřesvědčivou oblast viz. výše; označíme ji znakem (~)
6. p-value > 0.05 (přijímáme H_0), označíme ji znakem (.)

❑ Využití statistiky při dolování znalostí

Analýza výjimek je postupem, kdy se na základě porovnávání množiny základního a testovaného souboru hledají statisticky významné výskyty (ne)úspěchů, výjimek.

Základním souborem pro dolování dat, tedy i pro analýzu výjimek, rozumíme velké množství strukturovaných údajů o nějakém výseku reality. Tato množina dat bývá konečná a můžeme o ní tvrdit, že je pořízena statisticky náhodným výběrem. Nevíme tedy předem nic o pravděpodobnostech, rozděleních atd., a ve výpočtech budeme především používat četnosti sledovaných znaků.

Příkladem analýzy výjimek je hledání důvodů přerušení léčebného cyklu v databázi klinického registru léčebných cyklů centra asistované reprodukce (blíže viz. kapitola 7).

Vlastností této metody je to, že s využitím matematické statistiky, konkrétně testování hypotéz prostřednictvím p-value, automaticky vyhledává výjimky, na základě kterých pak vyslovíme určité hypotézy. Můžeme pak říct, že takovéto hypotézy buď přijímáme nebo zamítáme. Hypotézy vytvářené metodou analýzy výjimek jsou ve tvaru:

„statisticky významně častější výskyt výjimky v datech byl zaznamenán u“

- *1. nalezená výjimka*
- *2. nalezená výjimka*
- *...*

Použití analýzy výjimek je vhodné pro řešení výzkumných problémů na základě rozsáhlých dat. Analyzujeme data, protože se s jejich pomocí chceme něco zajímavého dozvědět. Zkoumáme problémy, které jsou definovány specificky, kdy na začátku máme položenu určitou otázku: *„co je příčinou výjimky v datech“*, a data jsou cíleně rozdělena tak, aby tato otázka byla zodpovězena.

Příklad 9.6.

Dlouhodobě sbíraná data o asistované reprodukci (umělém oplodnění) obsahují 8500 záznamů o provedených léčebných cyklech. Z nich jen 170 případů má v binárním atributu „hyperstimulační syndrom“ hodnotu 1. Jde tedy jen o 2% případů a v rámci předzpracování by se tento atribut za normálních okolností vyřadil ze zpracování. Pro lékaře by však bylo velmi důležité poznat příčiny vzniku tohoto syndromu. V rámci generování asociací nevyšly žádné výsledky.

Použijeme analýzy výjimek. První výběrový soubor tvoří záznamy bez syndromu,druhý soubor oněch 170 záznamů se syndromem. Analyzují se všechny atributy (kategoriální nebo kategorizované), které teoreticky mohou mít vliv na vznik syndromu – antecedenty.

Pro oba soubory se spočítají relativní četnosti všech hodnot antecedentových atributů, pro každou si odpovídající dvojici se spočítá hodnota p-value. Hodnoty vykazující statistickou odchylku se generují jako výsledky spolu s vyznačení nad- nebo podprůměrnosti hodnoty v testovaném souboru.
Výsledky si shrneme v tabulce, z níž je patrno, které atributy a s kterými hodnotami vykazují statistickou odchylku a tedy mohou být potencionální příčinou syndromu (nevysvětlujeme zde některé lékařské termíny, jako typ stimulace apod., jsou určeny k posouzení odborníkům).

Výskyt ovariálního hyperstimulačního syndromu (OHSS)

Atribut	OHSS (++)	OHSS (--)
Věk pacientky	mladší 25 let	Starší 30 let
OHSS dříve	1 a vícekrát	Dosud ne
Délka cyklu	> 30 dnů	< 21 dnů
Délka menstruace	7 – 9	
Operace dříve	0	1
Idiopat.faktor	Ano	Ne
Imunolog.faktor	Ano	Ne
Stimulace-typ	FSH/FSH+GnRH, HMG+GnRH	
Hodnota E2	> 10	< 10
Den cyklu AS	> 15	10 – 14
Získaných oocytů	> 8	0,1

♦

Shrnutí pojmů 9.3.

Analýza výjimek.

T test, hodnota p-value.

Generování hypotéz o hodnotách kategoriálních atributů, vykazujících statistické odchylky od základního souboru.

Otázky 9.3.

1. Který typ „zajímavosti“ v datech vyhledává analýza výjimek?
2. Popište princip analýzy výjimek a uveďte, pro která data je vhodná.

Úlohy k řešení 9.3.

1. Najděte alespoň 3 praktické úlohy na využití analýzy výjimek, definujte k nim výjimku a atributy zkoumané.

10. EXPERTNÍ SYSTÉM NAD DATOVOU MATICÍ

Čas ke studiu: 3 hodiny

Cíl Po prostudování této kapitoly budete umět

- popsat podstatu expertního systému, založeného na datové matici objektů a jejich konstruktů
- analyzovat, ladit a řešit praktické úlohy pomocí tohoto typu expertního systému

Výklad

10.1. Psychologický prostor člověka

❑ Psychologický prostor, objekty a konstrukty

Psychologové zkoumají, jak lidé rozpoznávají a třídí zkušenosti a klasifikují své okolí a jak na základě toho předvídají budoucí jevy a řídí své jednání. Americký psycholog G. A. Kelly popsal postupy, kdy člověk vytváří svou strukturu pojmů (tzv. **psychologický prostor)** a model svých znalostí jako tzv. systém osobních konstruktů.

Osobní konstrukty jsou šablony, které si člověk vytváří pro klasifikaci objektů reality a potom se do nich snaží zasadit skutečnosti, ze kterých se skládá svět.

Používá k tomu dva pojmy: **objekty** jako oblasti zájmu (osoby, zvířata, věci, jevy, akce) a **konstrukty** jako vlastnosti objektů (atributy, znaky, ...):

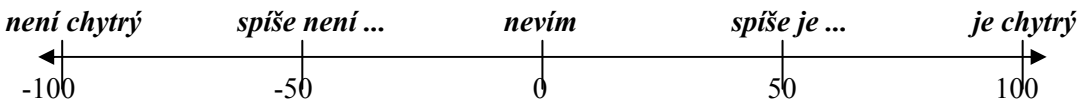
Konstrukt je tvořen dvěma protikladnými vlastnostmi, póly, které tvoří opačné mezní hodnoty atributu. Tím konstrukt svým způsobem normalizuje.

Psychologická vzdálenost mezi póly je rozdělena na stupně vyjadřující míru příslušnosti objektu k jednomu či druhému pólu. Střed škály znamená nevím, netýká se, proto je vhodné používat vícehodnotové stupnice s lichým počtem stupňů.

Rozdíl mezi atributem a konstruktem je právě v normalizaci konstruktu do škály <-100, 100>. Konstrukt je tedy zvláštním případem obecného atributu.

Příklad 10.1.

Objekty jsou studenti,
konstrukty např. chytrý, pracovitý, *..., nejsou zde podstatné atributy jako* rodné číslo, ...



♦

Pro jemnější rozlišení informací pro uživatele se zavádí další parametr, **důležitost konstruktu**, kterou uživatel modifikuje vliv jednotlivých konstruktů na rozhodování při provádění konzultace:

Příklad 10.2.

1. *vybíráme-li studenty pro brigádu na tenisových kurtech, pak konstrukt* je chytrý *má důležitost 0*
2. *vybíráme-li studenty pro SVOČ, pak. konstrukt* je chytrý *má důležitost 100.*

♦

❑ Expertní systémy založené na pravidlech a založené na datech

Expertem v nějaké oblasti reality nazýváme člověka, který tyto oblast zná podstatně lépe, než je běžný průměr. Z člověka se stane expert tak, že teoretickým studiem nebo praktickým zkoumáním světa nebo získáváním znalostí dalšími způsoby pozná část reality dostatečně důkladně.

Aby se expertovy znalosti využily i ku prospěchu ostatních lidí, uděluje expert svou práci nebo rady ostatním (lékař rozpozná nemoc a léčí, učitel rozpozná neznalost a naučí apod.). Jedna z oblastí umělé inteligence – konstrukce expertních systémů – se snaží znalost experta vložit vhodným způsobem do počítače a řešení problémů či rady uživatelům pak nacházet automatizovaně.

Zjednodušeně řečeno, nejčastější způsob uložení znalostí experta do báze znalostí je formou pravidel tvaru „jestliže platí … pak …“

kde pravidla jsou právě získána od experta. Podmínky na sebe mohou navazovat, v programu je zabudován inferenční mechanizmus. Vytváří tak rozvětvené stromy, které modelují expertův způsob myšlení.

Na rozdíl od těchto klasických expertních systémů, které se snaží modelovat expertovy myšlenkové pochody, náš popsaný přístup modeluje expertovu paměť (modelují zkoumanou tématickou oblast jako psychologický prostor objektů a konstruktů).

Σ Shrnutí pojmů 10.1.

Psychologický prostor člověka.

Objekty a konstrukty. Póly konstruktu, psychologická vzdálenost.

Důležitost konstruktu.

Expert a jeho osobní psychologický prostor.

10.2. Metodologie konstrukce expertního systému

Výklad

❑ Realizace psychologického prostoru repertoárovou tabulkou

Uvedli jsme si, že expertní znalosti do databáze se přenáší pomocí **interview s expertem.** Expert tím pomáhá budovat bázi znalostí. Experta chápeme jako učitele, který předává systému své znalosti v podobě modelu svého psychologického prostoru objektů a konstruktů a prostřednictvím repertoárové tabulky průběžně sleduje, jak jeho „žák“ (báze znalostí) „chápe“ poskytnuté expertní zkušenosti. Učení je ukončeno, když báze znalostí poskytuje řešení podle jeho představ.

Zobrazením psychologického prostoru je **repertoárová tabulka**, kde v řádcích jsou objekty {O1,O2,…} a ve sloupcích konstrukty {K1, K2, …}

Objekt \ Konstrukt	*K1*	*K2*	*. . .*	*Kn*
O 1	*50*	*-50*		*-100*
O 2	*75*	*0*		*0*
. . .				
O m	*100*	*-100*		*-75*

Hodnoty v tabulce znamenají, do jaké míry přísluší odpovídajícímu objektu jeden nebo druhý pól příslušného konstruktu.

Cílem je vytvořit z tabulky **bázi znalostí**, která bude řešit analytické úlohy podobně jako expert. To znamená vyplnit tabulku tak, aby dobře pokrývala celou modelovanou oblast zájmu – obsahovala dostatek příslušných objektů, charakterizovaných vhodnými konstrukty.

Příklad 10.3.

Uveďme jako příklad „expertní znalosti“ pracovníka cestovní kanceláře, který doporučuje svým klientům, kam se mají vypravit na dovolenou v závislosti na jejich požadavcích, jako jsou: roční období, finanční náročnost, sportovní a společenské požadavky apod. Souhrn takových požadavků – konstruktů je vstupem pro zadání, výstupem pak jsou konkrétní doporučená místa – objekty. Přitom každé z míst není charakterizováno konstruktem jen binárně ano/ne, ale mírou příslušnosti (Alpy jsou spíše drahé, Beskydy spíše levné).

♦

Bude to báze s jednou vrstvou pravidel, pravidla směřují od dotazovacích uzlů (předpokladů), kterými jsou osobní konstrukty, k cílovým uzlům (závěrům), kterými jsou objekty. V našem případě se bude báze skládat z pravidel typu:

Jestliže se vyskytuje vlastnost K, pak jde o objekt O s vahou V

Automatizace budování expertního systému

Všechny fáze práce s psychologickým prostorem člověka je možné automatizovat, realizovat pomocí vhodného programového systému. Jeho úkolem je

1. zprostředkovat dialog mezi expertem s jeho osobním psychologickým prostorem a repertoárovou tabulkou
2. vytvořit z repertoárové tabulky bázi znalostí ve formě soustavy pravidel,
3. pomocí báze znalostí podávat konzultace dalším uživatelům.

Postup při budování expertního systému

1. Zadavatel vybere oblast z reality, které se bude expertní systém týkat, vybere experta na zvolenou oblast.
2. Expert zadá seznam všech možných **řešení** problému, kterého se bude vytvářená báze znalostí týkat. Prvky seznamu reprezentují závěry, které by měl poskytovat budoucí expertní systém (cíle konzultace, objekty).
3. Expert předem provede analýzu problémové oblasti, **vybere objekty**, které budou zkoumanou oblast reality dobře popisovat. Objekty mají být téhož typu, stejné úrovně složitosti a pokrývat téma co nejúplněji.
4. Expert vybere možné **konstrukty** objektů, jejichž póly by objekty dobře rozlišovaly.
5. Program vhodnými dotazy získává od experta informace o jeho psychologickém prostoru a sestavuje podle něho repertoárovou tabulku. Průběžně ji testuje a o výsledcích podává expertovi informace. Program analyzuje tabulku ze tří hledisek:
   - testuje postačitelnost informací pro popis objektů,
   - testuje vzájemné vazby mezi konstrukty,
   - vyhledává ekvivalentní nebo nedůležité informace.
6. Expert na každé upozornění dále upřesňuje poskytnuté informace a tak dále zjemňuje a rozšiřuje repertoárovou tabulku.
7. Program vytváří z každého políčka tabulky dvě pravidla, protože spojuje pravidlem každý pól každého konstruktu s každým objektem (mimo políčka „nevím“).

Postup při využívání expertního systému

1. Uživatel vybere nebo zadá typ konzultace, vybere existující nebo zadá nový dotaz na expertní systém.
2. Při konzultaci uživatele s programem prochází systém všechny konstrukty a ptá se na jejich důležitost pro uživatele. Z hodnot políček repertoárové tabulky vypočítává váhu jednotlivých jednoduchých pravidel.
3. Příspěvky všech pravidel pro každý objekt sloučí a dospěje k závěru o vhodnosti jednotlivých objektů pro danou soustavu vlastností.
4. Výsledkem je seznam všech objektů báze s doporučením vhodnosti pro zadaný dotaz - pro každý objekt určí hodnocení z intervalu <-100,100>, kde -100 znamená naprosto nevyhovuje, 100 naprosto vyhovuje.

Tvorba báze znalostí tedy probíhá v následujících krocích

1. Expert zadá seznam všech možných řešení problému, kterého se bude vytvářená báze znalostí týkat, prvky seznamu reprezentují závěry, které by měl poskytovat budoucí expertní systém (cíle konzultace, objekty);
2. pro dané cíle systém vytváří jejich trojice a žádá experta, aby pro každou trojici určil znak odlišující dva prvky od třetího; znak je určen pomocí dvojice opačných hodnot, pólů; vzniká tabulka ohodnocení, ve které je každý cíl popsán pomocí znaků, konstruktů; hodnota znaku se zadává s možností neurčitosti (ano, spíše ano, ..., nevím, ..., spíše ne, ne)
3. pro vytvořenou tabulku ohodnocení systém konstruuje znalosti ve formě implikací mezi jednotlivými póly různých znaků;
4. generují se pravidla, nejprve cílová (na levé straně se vyskytuje některý ze závěrů), potom mezilehlá (pro každou implikaci nalezenou v bodě 3); z každého pole tabulky ohodnocení se vygeneruje jedno pravidlo spolu s faktorem jistoty;
5. po vytvoření všech pravidel začíná testování báze znalostí; pokud expert nesouhlasí s výsledky konzultace, provádí se další zjemňování báze znalostí - přidáním nových znaků, nových objektů, změnami v tabulce ohodnocení ap.).

Využití programového systému

- podstatně zkracuje dobu vytváření báze znalostí,
- po vytvoření slouží jako expertní systém vhodný pro výběr optimální varianty z množiny zaznamenaných variant nebo
- k automatizaci řízení systémů, u nichž je předem známo, co způsobí určitý řídicí zásah na výstupu systému (dopředné řízení).

Shrnutí pojmů 10.2.

Repertoárová tabulka.

Vyplnění a ladění repertoárové tabulky, zadání objektů, zadání konstruktů.

Analýza implikací. Shoda objektů, shoda konstruktů.

Konzultace experta s bází, zlepšování chování báze.

Konzultace uživatele s bází.

10.3. Automatizované získávání expertíz SAZZE

Výklad

Popsané principy realizace expertního systému nad datovou maticí byly využity pro implementaci programového systému, nazvaného SAZZE (**S**ystém **A**utomatizovaného **Z**ískávání **Z**nalostí od Experta).

❑ Fáze definování a naplňování repertoárové tabulky

Systém sérií otázek postupně od experta získává a ukládá tyto informace.

Název: NÁZEV BÁZE
Cíl: Určení a využití báze znalostí
Objekty: Konkrétní typy objektů
Konstrukty: Charakteristika konstruktů

Následuje dialog, ve kterém se systém postupně ptá na

1. Zadané objekty (vyplní levý sloupec repertoárové tabulky).

2. Zadané konstrukty (vyplní hlavičku – první řádek repertoárové tabulky).

3. Zadání hodnot konstruktů pro všechny objekty (vyplní sloupce a řádky repertoárové tabulky).

Výsledkem tohoto dialogu je repertoárová tabulka s hodnotami:

Potom z každého pólu každého konstruktu zformuluje systém cílová pravidla.

V rámci budování báze znalostí systém dále umožňuje provádět analýzu implikací a test shody objektů nebo konstruktů.

4. Analýza implikací

znamená nalezení všech dvojic logicky totožných implikací, které podle informací v bázi mohou platit. Analýzu provede systém a výsledky zobrazí. Expert pak potvrdí všechny skutečně platné. Ty jsou uloženy a budou využívány při konzultacích.

Expertem potvrzené implikace tvoří mezilehlá pravidla expertního systému.

5. Shoda objektù

znamená nalezení objektù, které mají ve všech konstruktech stejné hodnocení, porovnání se provádí procentuelně. Ve výsledku se uvádějí shody dvojic objektů nad 80%. Na dotaz systému může expert zadat nový konstrukt, který oba objekty rozliší.

Procentuelní shoda dvou objektů Oi a Oj (z celkového počtu n objektů popsaných m konstrukty) se vypočítá podle vztahu

$$S_{ij} = \sum_{a=1}^{n} \left(1 - \frac{|V_{ia} - V_{ja}|}{2 \times V_{\max}} \right) \times \frac{100}{n}.$$

kde x_{ik} je hodnota konstruktu k u objektu Oi.

6. Shoda konstruktů

Obdobně se mohou objevit podobné póly některých konstruktů, **shoda konstruktů**. Řešení je obdobné, jako u shody objektů, za hranici se bere 70%.

Sloučení konstruktů: pokud je shoda konstruktů dána nevhodně zvolenou skladbou konstruktů, je vhodné tyto konstrukty sloučit do jednoho. Jeho jméno mùže být shodné s jedním z původních nebo je možno zadat nové.

7. Konzultace s expertem

Po vytvoření báze znalostí expert konzultacemi ověřuje, jak se shodují doporučení expertního systému s jeho vlastními a v případě neshody modifikuje bázi tak dlouho, dokud systém nedává odpovědi stejné, jako on.

Každá konzultace má svůj název a je uloženo její zadání - požadavky, důležitost a hodnota konstruktů. Je možno vybrat z již existujících nebo definovat novou.

Uživatel vlastně zadává požadované hodnoty vlastností hledaného objektu, program vyhodnotí podobnost každého objektu tabulky se zadáním.

8. Zlepšování chování báze

Celkově má expert ladit bázi znalostí těmito nástroji:

- Přidáváním nových objektů
    - podle nápovědy, dotazem na nový objekt s hodnotami konstruktů, jejichž kombinace v bázi nejsou,
    - bez nápovědy
- Přidáváním nových konstruktů
    - bez nápovědy
    - podle nápovědy využitím mezilehlých pravidel (výsledku analýzy implikací), pro neplatná doplnit konstrukty s hodnotou různou u doporučených objektů a u nedoporučených objektů,
    - podle nápovědy s využitím výsledku konzultace, dotazem na nový konstrukt s hodnotami různými pro expertem doporučené a nedoporučené objekty, přičemž výsledek konzultace dopadl jinak,
- Zjemněním stupnice hodnot konstruktu z 5 základních hodnot na celou stupnici <-100, 100>, aby výsledek konzultace odpovídal představě experta.
- Dodáním hodnoty důležitost konstruktu ke každému konstruktu v %.

❑ Fáze využívání expertního systému

Je.li báze znalostí odladěna, to znamená, že se ve všech typech konzultací s expertem chová podle jeho představ, může být nabídnuta uživatelům.

Uživatel buď vybere konzultaci z již existujících, vytvořených expertem nebo některým předcházejícím uživatelem, a to podle názvu, cíle a zadání.

Pokud si nevybere, zadá svou novou konzultaci: zadá název a její zadání - požadavky, důležitost a hodnoty konstruktů. Uživatel vlastně zadává požadované hodnoty vlastností hledaného objektu, program vyhodnotí podobnost každého objektu tabulky se zadáním.

10.4. Příklad využití systému SAZZE

Výklad

Teoreticky popsané etapy vývoje systému SAZE i jeho využití pro konzultace ukážeme na podrobně popsaném reálném příkladě. Celý vývoj, ladění expertní báze i využití budeme prezentovat formou dialogu uživatele se systémem (**SYS**). V první fázi bude uživatelem expert (**EXP**), po odladění báze to bude běžný uživatel (**UŽIV**), požadující konzultace.

Příklad 10.5.

Báze znalostí o rekreačních oblastech.

Název: REKREACE
Cíl: Doporučení nejvhodnější rekreační oblasti požadovaných vlastností.
Objekty: Konkrétní rekreační oblasti.
Konstrukty: Umístění, roční doba, cena.

I. fáze - dialog při definování a naplňování repertoárové tabulky

1. Zadání objektů

SYS Zadejte konkrétní objekty (minimálně 5):
EXP Špindlerův Mlýn, Zemplínská Šírava, Balaton, Karlovy Vary, Benátky.

2. Zadání konstruktů

SYS Uveďte 1.významnou vlastnost, kterou dva z objektů Špindlerův Mlýn, Zemplínská Šírava, Balaton mají a jeden nemá:
EXP je na horách
SYS Jmenuje se opačný pól konstruktu takto? není na horách
EXP [potvrzení nebo oprava]
SYS Uveďte 2.významnou vlastnost, kterou dva z objektů Balaton, Zemplínská Šírava, Karlovy Vary mají a jeden nemá:
EXP je v cizině
SYS Jmenuje se opačný pól konstruktu takto? není v cizině
EXP [potvrzení nebo oprava]
SYS Uveďte 3. významnou vlastnost, kterou dva z objektů Balaton, Karlovy Vary, Benátky mají a jeden nemá:
EXP je u moře
SYS Jmenuje se opačný pól konstruktu takto? není u moře
EXP [potvrzení nebo oprava]
atd.

3. Zadání hodnot konstruktů pro všechny objekty

Po zadání objektů a konstruktů popisuje expert jednotlivé objekty.

SYS Popište každý objekt pomocí zadaných konstruktů.
Objekt Špindlerův Mlýn rozhodně je na horách
poněkud je na horách
nevím / ani jedno
poněkud není na horách
rozhodně není na horách

EXP rozhodně je na horách
atd.

Jednotlivé volby se transformují na hodnoty repertoárové tabulky. Výsledkem tohoto dialogu je repertoárová tabulka s hodnotami:

Objekt \ Konstrukt	je na horách	je v cizině	je u moře
Špindlerův Mlýn	100	-100	-100
Zemplínská Šírava	-100	100	-100
Balaton	-100	100	100
Karlovy Vary	50	-100	-100
Benátky	-100	100	100

Z každého pólu každého konstruktu zformuluje systém cílová pravidla. Příklad několika pravidel našeho příkladu:

SYS Zadaná pravidla jsou
Když je v cizině, pak
Zemplínská Šírava s váhou 100
Balaton s váhou 100
Benátky s váhou 100
Karlovy Vary s váhou -100
Špindlerův Mlýn s váhou -100
Když není v cizině. pak
Karlovy Vary s váhou 100
Špindlerův Mlýn s váhou 100
Zemplínská Šírava s váhou -100
Balaton s váhou -100
Benátky s váhou -100

ostatní pravidla pro jejich velké množství neuvádíme.

V rámci budování báze znalostí systém dále umožňuje provádět analýzu implikací a test shody objektů nebo konstruktů.

4. Analýza implikací

znamená nalezení všech dvojic logicky totožných implikací, které podle informací v bázi mohou platit. Analýzu provede systém a výsledky zobrazí. Expert pak potvrdí všechny skutečně platné. Ty jsou uloženy a budou využívány při konzultacích.

V našem příkladě se po provedení analýzy implikací objeví:

Analýza implikací

SYS Zdá se, že mezi vlastnostmi mohou platit následující vazby. Potvrďte platné z nich:

1. je na horách → není v cizině
   je v cizině → není na horách
2. je na horách → není u moře
   je u moře → není na horách
3. není v cizině → není u moře
   je u moře → je v cizině

EXP [platné jsou implikace číslo 3, neplatné číslo 1]
SYS Znáte nějaký objekt, pro který neplatí implikace číslo 1?
EXP Insbruck
[zadá objekt ... a k němu hodnoty všech konstruktů], výsledkem je tabulka:

Objekt \ Konstrukt	je na horách	je v cizině	je u moře
Špindlerův Mlýn	100	-100	-100
Zemplínská Šírava	-100	100	-100
Balaton	-100	100	100
Karlovy Vary	50	-100	-100
Benátky	-100	100	100
Insbruck	100	100	-100

Expertem potvrzené implikace tvoří mezilehlá pravidla expertního systému.

5. Shoda objektù

znamená nalezení objektù, které mají ve všech konstruktech stejné hodnocení, porovnání se provádí procentuelně. Ve výsledku se uvádějí shody dvojic objektů nad 80%. Expert může zadat nový konstrukt, který oba objekty rozliší.
Procentuelní shoda dvou objektů Oi a Oj (z celkového počtu n objektů popsaných m konstrukty) se vypočítá podle vztahu

$$S_{ij} = \sum_{k=1}^{n} ((x_{ik} - x_{jk}) / 200) * 100 / m$$

kde x_{ik} je hodnota konstruktu k u objektu Oi.

Shoda objektů

SYS Podobnost mezi objekty Špindlerův Mlýn a Karlovy Vary je 92%. Znáte nějakou vlastnost, která by je rozlišila?
EXP možnost zimních sportů
SYS Jmenuje se opačný pól konstruktu takto? nemožnost zimních sportů
atd.

Následuje doplnění hodnot nového konstruktu u všech objektů, výsledkem je tabulka:

Objekt \ Konstrukt	je na horách	je v cizině	je u moře	zimní sporty
Špindlerův Mlýn	100	-100	-100	100
Zemplínská Šírava	-100	100	-100	-100
Balaton	-100	100	100	-100
Karlovy Vary	50	-100	-100	0
Benátky	-100	100	100	-100
Insbruck	100	100	-100	100

6. Shoda konstruktů

Obdobně se mohou objevit podobné póly některých konstruktů, **shoda konstruktů**. Řešení je obdobné, jako u shody objektů, za hranici se bere 70%.

Shoda konstruktů

SYS Podobnost mezi konstrukty je na horách a možnost zimních sportů je 83%. Pokuste se najít objekt, který by měl zároveň vlastnosti
je na horách a nemožnost zimních sportů nebo
není na horách a možnost zimních sportů.
EXP neznám
[zadá nový objekt a k němu hodnoty všech konstruktů, nebo takový objekt není znám]

Sloučení konstruktů: pokud je shoda konstruktů dána nevhodně zvolenou skladbou konstruktů, je vhodné tyto konstrukty sloučit do jednoho. Jeho jméno mùže být shodné s jedním z původních nebo je možno zadat nové.
SYS Mají se tedy konstrukty je na horách a možnost zimních sportů spojit do jediné vlastnosti?
EXP ano
SYS Jak se bude nový konstrukt jmenovat:
je na horách / není na horách
možnost zimních sportů / nemožnost zimních sportů
jinak
zrušit oba konstrukty
EXP jinak : vhodný pro zimní rekreaci

atd.,

systém se ptá na název opačného pólu, zruší 2 původní konstrukty, vytvoří nový, expert zadá hodnoty nového konstruktu pro všechny objekty (případnou modifikací hodnot zrušených konstruktů).

Výsledkem je tabulka:

Objekt \ Konstrukt		je v cizině	je u moře	zimní rekreace
Špindlerův Mlýn		-100	-100	100
Zemplínská Šírava		100	-100	-100
Balaton		100	100	-100
Karlovy Vary		-100	-100	-50
Benátky		100	100	-50
Insbruck		100	-100	100

7. Konzultace s expertem

Po vytvoření báze znalostí expert konzultacemi ověřuje, jak se shodují doporučení expertního systému s jeho vlastními a v případě neshody modifikuje bázi tak dlouho, dokud systém nedává odpovědi stejné, jako on.

Každá konzultace má svůj název a je uloženo její zadání - požadavky, důležitost a hodnota konstruktů. Je možno vybrat z již existujících nebo definovat novou.

Konzultace s expertem

SYS Vyberte si konzultaci:
Vánoce a Nový rok
nová
EXP nová: Letní rodinná s dětmi

SYS [pro všechny konstrukty se zadává jeho důležitost pro danou konzultaci]
Jak důležitý je konstrukt Je v cizině / není v cizině
velmi
trochu
vůbec ne
EXP trochu

atd. pro všechny konstrukty.

SYS U těchto konstruktů jste zadal,že pro konzultaci nejsou důležité:

vhodné pro zimní sporty
Pokud tomu tak není, doplňte zadání konzultace.
EXP [případné doplnění zadání nebo potvrzení platnosti zadání]
SYS [dále se systém ptá na požadované hodnoty konstruktů označených za velmi nebo trochu důležité]
EXP [zadá požadované hodnoty]

Pro náš příklad je celkem zadáno:
Je trochu důležité, aby platilo rozhodně není v cizině.
Je trochu důležité, aby platilo rozhodně není u moře.
Je velmi důležité, aby platilo rozhodně možnost zimních sportů.
Je velmi důležité, aby platilo poněkud laciné.

Uživatel vlastně zadává požadované hodnoty vlastností hledaného objektu, program vyhodnotí podobnost každého objektu tabulky se zadáním.

Po ukončení zadání se zobrazí výsledek konzultace ve tvaru:
SYS Pro konzultaci Rodinná s dětmi systém doporučuje objekty
83 % Zemplínská Šírava
55 % Balaton
27% Karlovy Vary
Systém nedoporučuje
14 % Špindlerùv Mlýn
-14 % Benátky
-46 % Insbruck
[doporučuje objekty s váhou > 20, nedoporučuje ostatní]
Pokud s některým zařazením nesouhlasíte, označte příslušný objekt.

8. Zlepšování chování báze

- Přidáváním nových objektů
  - podle nápovědy, dotazem na nový objekt s hodnotami konstruktů, jejichž kombinace v bázi nejsou,
  - bez nápovědy
- Přidáváním nových konstruktů
  - bez nápovědy
  - podle nápovědy využitím mezilehlých pravidel (výsledku analýzy implikací), pro neplatná doplnit konstrukty s hodnotou různou u doporučených objektů a u nedoporučených objektů,
  - podle nápovědy s využitím výsledku konzultace, dotazem na nový konstrukt s hodnotami různými pro expertem doporučené a nedoporučené objekty, přičemž výsledek konzultace dopadl jinak,

- Zjemněním stupnice hodnot konstruktu z 5 základních hodnot na celou stupnici <-100, 100>, aby výsledek konzultace odpovídal představě experta.
- Dodáním hodnoty důležitost konstruktu ke každému konstruktu v %.

Výsledná upřesněná a doplněná repertoárová tabulka našeho příkladu:

Objekt \ Konstrukt	v cizině	u moře	zimní rekreace	vodní sporty	drahá	stanování
Špindlerův Mlýn	-100	-100	100	-100	-100	-100
Zemplínská Šírava	100	-100	-100	100	-100	100
Balaton	100	-100	-100	100	50	100
Karlovy Vary	-100	-100	-50	-50	50	-100
Benátky	100	100	-50	100	100	100
Insbruck	100	-100	100	-50	100	-100

II. Konzultace uživatele s expertním systémem

Uživatel expertního systému konzultuje s hotovým naplněným systémem obdobným způsobem, jako expert při ověřování báze znalostí, pouze bez možnosti změn báze.

Výsledkem je výpis konzultace obsahující název konzultace, zadání a výsledná doporučení s váhou.

Mimo to může uživatel požádat o výpisy

- pravidel cílových i mezilehlých (výsledků expertem uznaných implikací),
- báze znalostí (repertoárové tabulky),
- výsledku uložených konzultací.

10.5. Implementace systému SAZZE

Výklad

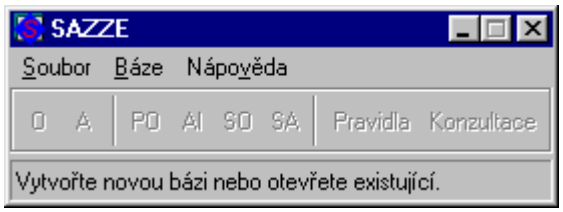
Na katedře informatiky VŠB-TUO byl v rámci diplomové práce implementován programový systém SAZZE, který realizuje všechny popsané funkce. Ukážeme si jeho ovládání opět na příkladu z praxe.

Příklad 10.6.

Vytápění

S pomocí experta z oblasti návrhu druhu a způsobu vytápění bude ukázkově sestavována báze znalostí expertního systému, který bude schopen z daného hlediska doporučit druh topení. Vytápění se týká běžných objektů pro bydlení, tj. objektů, které mají topení s výkonem kotle do 50 kW. Není zde zahrnuto vytápění velkých budov (panelové domy, školy, úřady, atd.) a dálkové způsoby vytápění (dostupnost pouze v nevelké vzdálenosti od poskytovatelů odpadního tepla). Netýká se ani výrobních objektů.

Po spuštění aplikace SAZZE se zobrazí následující okno:

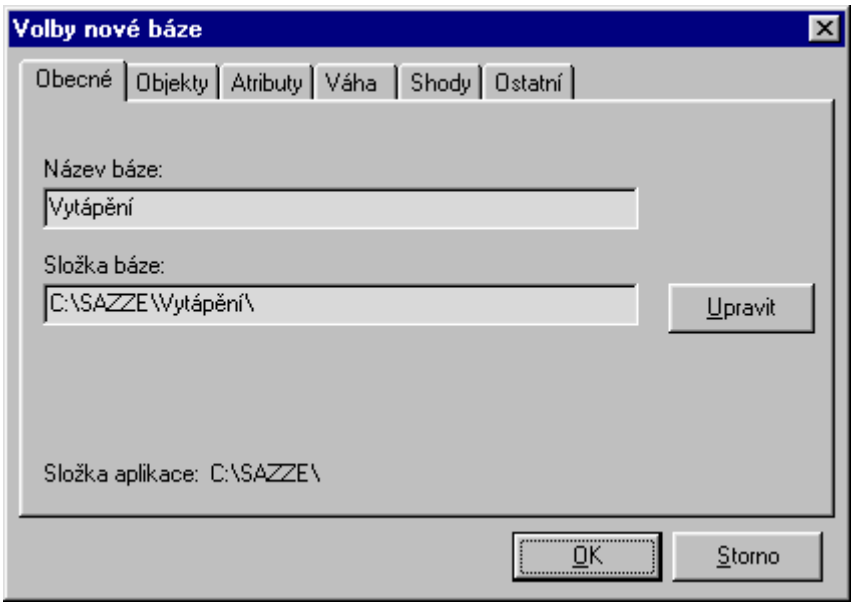


Na nástrojové liště nyní není dostupná žádná ikona (tak jako v menu Báze žádná položka), protože není otevřena žádná báze dat a znalostí. Jelikož vytváříme novou bázi, nebudeme otevírat žádnou již existující bázi.

Než začneme analyzovat problém vytápění, vytvoříme tedy prázdnou bázi výběrem položky Nová v hlavním menu aplikace Soubor. Zobrazí se formulář, do kterého vyplníme v záložce Obecné název nové báze „Vytápění“, v záložce Shody nastavíme hranici uvádění shod objektů na 90 % a v záložce Ostatní vyplníme jméno experta. Ostatní údaje použijeme standardní.

Navržené objekty, jejich konstrukty a váhy jsou zřejmé z repertoárové tabulky, pro stručnost tedy neuvádíme kompletní dialog.

Aby se dala určit hranice, kdy je instalace (nebo provoz) daného objektu drahá a kdy levná, bereme za rozhodně drahou nejdražší instalaci a za rozhodně levnou nejlevnější instalaci.



Většina objektů využívá při provozu také jiný druh energie, i když její spotřeba není významná. Výjimku tvoří objekty č. 6 – Solární energie a č. 7 – Tepelné čerpadlo, které by bez elektrické energie nemohly z technologického hlediska správně fungovat. Při popisu těchto (i ostatních) objektů byl na tento fakt brán zřetel a váha atributu určujícího finanční náročnost provozu upravena.

Do atributu „je ekologické“ byla zahrnuta také obnovitelnost a neobnovitelnost zdrojů energií. Dřevo jako obnovitelný zdroj energie má větší váhu než elektrická energie. Pro ni, ačkoliv je obnovitelným zdrojem, se využívají neobnovitelné zdroje, jako je uhlí a plyn.

Atribut „je rychlý ohřev topného média“ byl zvolen kvůli potřebě energie při přerušovaném provozu, kde je spotřeba energie ovlivňována akumulační schopností budov. Je-li schopnost budovy akumulovat energii nízká, je výhodná pro přerušovaný provoz oproti stejně zateplené budově ale s větší akumulací tepla.

***Vytváření báze znalostí**:*

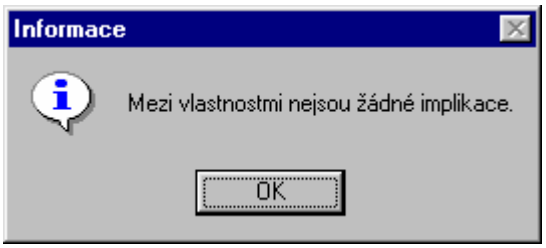
1. *Zadáme všechny objekty.*
2. *Pokračujeme zadáním navržených atributů.*
3. *Pomocí formuláře pro popis objektů vytvoříme repertoárovou tabulku. Použijeme jemnou stupnici, takže můžeme také zadávat hodnoty jiné, než na pětibodové stupnici. Ohodnocení atributů je následující:*

Repertoárová tabulka

	instalace je drahá	provoz je drahý	je ekologické	rychlost ohřevu
dřevo	-100	-50	70	60
elektřina	30	100	50	100
hnědé uhlí	-50	-20	-50	40
kaly	-50	-50	-100	20
koks	-50	10	20	10
solární energie	80	-80	80	-90
tepelné čerpadlo	100	-100	100	-50
zemní plyn	60	40	70	80
černé uhlí	-50	-10	-30	50

4. *Provedeme analýzu implikací (AI).*

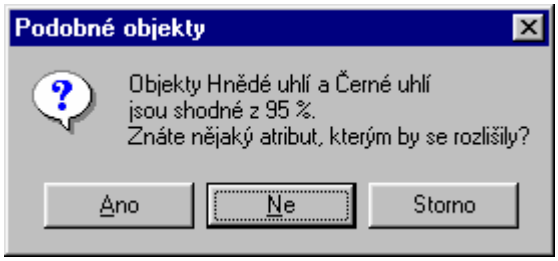
   V našem případě nebyla nalezena žádná implikace:



5. *Analýza SO, SA*

- *Po provedené analýze shod objektů se objeví okno s informací, že objekty „hnědé uhlí“ a „černé uhlí“ jsou shodné z více než 90 %, a to z celých 95 %. Protože tyto zdroje energie jsou z principu podobné (přibližně stejně drahé, stejně dostupné a mají podobný způsob spalování), nebudeme vkládat žádný nový atribut, kterým by se rozlišily.*

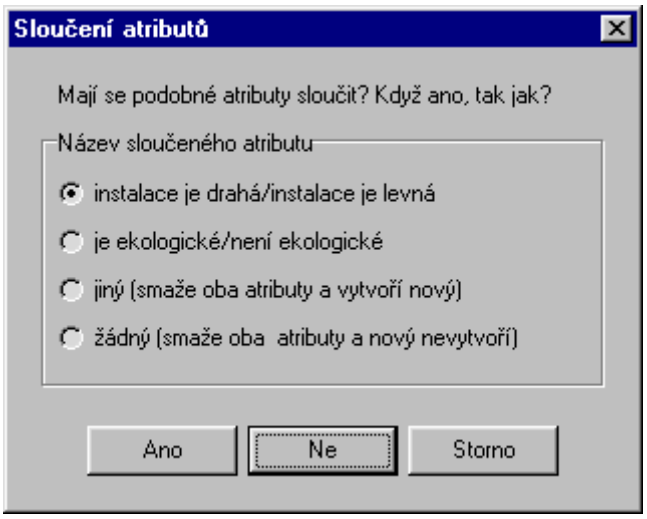
  Na následující dotaz odpovíme stisknutím tlačítka Ne.



- *Následně provedeme analýzu shod atributů. Je zobrazeno okno s informací, že atributy „instalace je drahá“ a „je ekologické“ jsou shodné z 81 %, a že atributy „provoz je drahý“ a „ohřev topného média je rychlý“ jsou shodné ze 78 %. Po výběru první možnosti je po nás požadováno vložení nového objektu uvedených vlastností. Protože platí, že co je ekologické bývá i drahé, nebudeme shodu těchto atributů vyvracet.*

Na dotaz odpovíme ne a zobrazí se okno s výběrem způsobu sloučení těchto podobných atributů:

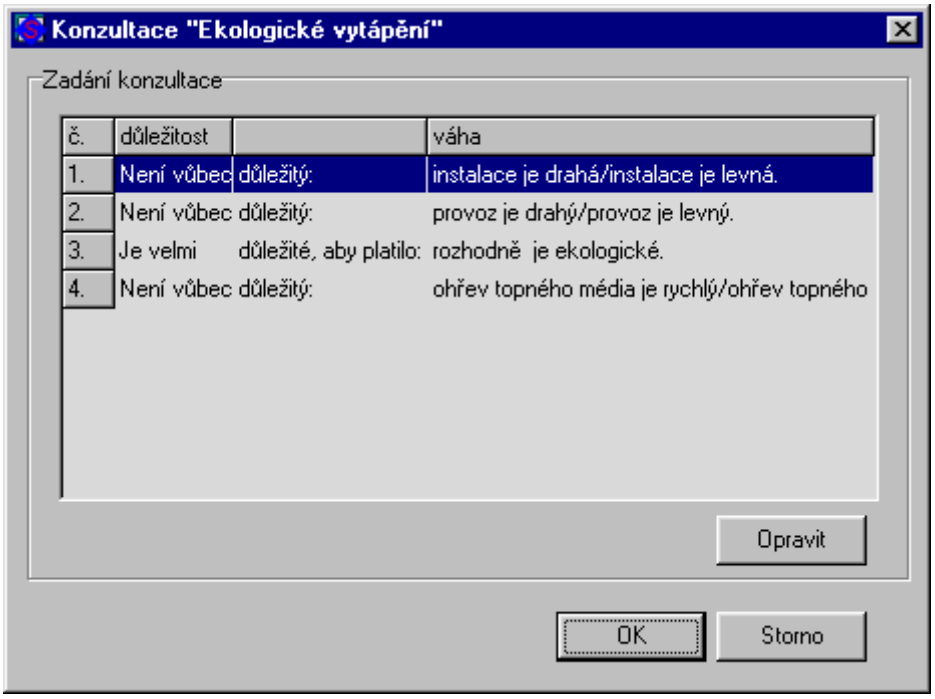
Protože atributy nebudeme slučovat ani mazat, odpovíme rovněž ne. Touto poslední odpovědí jsme získali přístup ke zbývajícím dvěma položkám lišty, kterými jsou Pravidla a Konzultace.



6. *Pravidla platící v bázi jsou buď cílová nebo mezilehlá. Mezilehlá pravidla v naši bázi nejsou, protože zde neplatí žádná implikace. Protože pro ostatní atributy mají pravidla stejný tvar, stačí zde uvést příklad výpisu koncových pravidel pro jeden atribut. Koncová pravidla totiž vedou od každého pólu každého atributu ke všem objektům a jejich kompletní výpis by byl rozsáhlý.*

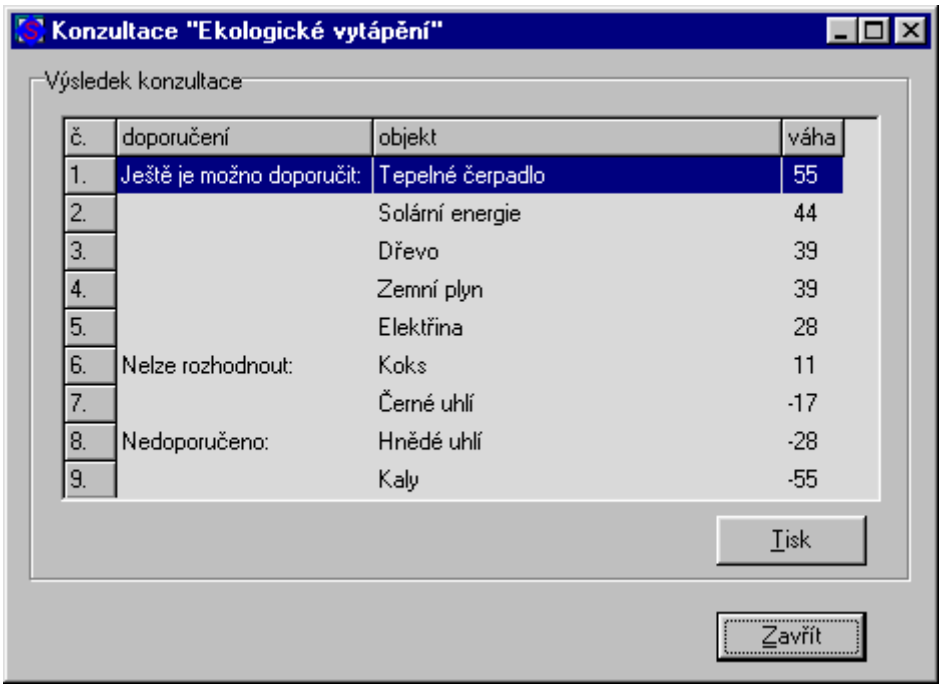
- Když je ekologické, pak:
  Tepelné čerpadlo s váhou 55.
  Solární energie s váhou 44.
  Dřevo s váhou 39.
  Zemní plyn s váhou 39.
  Elektřina s váhou 28.
  Koks s váhou 11.
  Černé uhlí s váhou -17.
  Hnědé uhlí s váhou -28.
  Kaly s váhou -55.
- Když není ekologické, pak:
  Tepelné čerpadlo s váhou -55.
  Solární energie s váhou -44.
  Dřevo s váhou -39.
  Zemní plyn s váhou -39.
  Elektřina s váhou -28.
  Koks s váhou -11.
  Černé uhlí s váhou 17.
  Hnědé uhlí s váhou 28.
  Kaly s váhou 55.

7. *Nyní přistoupíme k testování báze znalostí – konzultacím. Po vložení názvu nové konzultace „Ekologické vytápění“ zadáme důležitosti a váhy jednotlivých atributů pro tuto konzultaci. Zadání je v následujícím formuláři:*



Pokud nás zajímá pouze ekologie a nevíme nic o nákladech na instalaci a provoz, považujeme za jediný velmi důležitý (100 %) atribut „je ekologické“. Jeho váhu zadáme „rozhodně ano“, tedy +100 (váha je v intervalu –100 až 100). Ostatní atributy nejsou vůbec důležité (0 %) a váhu tudíž zadávat nemusíme.

Výsledkem konzultace je následující okno, ve kterém jsou uvedeny všechny objekty s rozdělením do doporučujících intervalů a s uvedením váhy:



Konzultace

Jako příklad využití báze je uveden výpis konzultace, která doporučí co nejlevnější vytápění s ohledem na životní prostředí.

Konzultace "Levné vytápění"

A) ZADÁNÍ:

1. Je velmi důležité, aby platilo: rozhodně instalace je levná.
2. Je velmi důležité, aby platilo: rozhodně provoz je levný.
3. Je trochu důležité, aby platilo: spíše je ekologické.
4. Je trochu důležité, aby platilo: spíše ohřev topného média je rychlý.

B) VÝSLEDKY *(váha doporučení je z intervalu <-100, 100>.*

Rozhodně doporučeno:
- Dřevo (váha 74)

Ještě je možno doporučit:
- Kaly (váha 41)
- Tepelné čerpadlo (váha 37)
- Hnědé uhlí (váha 35)
- Černé uhlí (váha 34)
- Koks (váha 27)

Nelze rozhodnout:
- Solární energie (váha 20)
- Zemní plyn (váha -13)

Nedoporučeno:
- Elektřina (váha -25)

Shrnutí pojmů 10.

Psychologický prostor člověka. Objekty a konstrukty. Póly konstruktu, psychologická vzdálenost. Důležitost konstruktu.

Expert a jeho osobní psychologický prostor.

Repertoárová tabulka. Vyplnění a ladění repertoárové tabulky, zadání objektů, zadání konstruktů.

Analýza implikací. Shoda objektů, shoda konstruktů.

Konzultace experta s bází, zlepšování chování báze.

Konzultace uživatele s bází.

Otázky 10.

1. Co nazýváme psychologickým prostorem člověka a čím je popsán?
2. Čím se liší objekty a konstrukty v pojetí psychologického prostoru člověka od databázových objektů a atributů?
3. Co je repertoárová tabulka ve vztahu k psychologickému prostoru člověka?

4. Co je důležitost konstruktu a k čemu se využívá?
5. Které kroky a které nástroje slouží k odladění báze znalostí v repertoárové tabulce?
6. K čemu slouží konzultace nad bází znalostí a jak probíhají?

Úlohy k řešení 10.

1. Zvolte si oblast, v níž se považujete za experta a vytvořte repertoárovou tabulku pro tuto oblast: navrhněte vhodné cíle využití běžnými uživateli, k nim vhodné objekty a konstrukty. Pro ně pak naplňte repertoárovou tabulku, případně ji několika testovacími pokusy odlaďte.

11. SW PRO PODPORU DOLOVÁNÍ ZNALOSTÍ

Čas ke studiu: 1 hodina

Cíl Po prostudování této kapitoly se seznámíte s

- SW produkty, které nabízejí některé z metod dolování znalostí
- metodami, které tyto SW produkty nabízejí.

Výklad

11.1. Obecně o SW pro Data Mining

Původně (historicky) byly implementovány jednotlivé metody samostatně a využívány většinou pro výzkumné účely nebo sociální průzkumy. Později – přibližně se vznikem datových skladů se začaly využívat metody získávání znalostí i pro shromážděná data a datové sklady. Výsledky přesahující prosté agregované dimenzionální řady a jejich hierarchie, jak již víme, metody mohou objevovat i dosud netušené souvislosti a pravidla v datech ukrytá.

Firmy nabízející Minery většinou nezveřejňují podrobnější informace o metodách nebo dokonce algoritmech, použitých ve svých produktech. Často popisují metodu jako „velmi podobnou“ metodě XYZ, ale vyvinutou vlastními pracovníky firmy. Zkušený programátor a analytik někdy odhadne i ze stručných informací metodu. Pokud však chce mít jistotu o výsledku metody, je nutné ji zřejmě otestovat na známých datech nebo porovnat se známými výsledky.

Často firmy také inzerují, že grafické uživatelské rozhraní a automatizovaný rámec znamenají, že nemusíte vědět, jak nástroje dolování pracují k tomu, abyste je používali. Tvrdí například, že „obchodní technolog s malou statistickou odbornou znalostí může **rychle a snadno** používat systém“. Není to pravda. Je sice často pravdou, že uživatel bez znalostí metod může snadno a rychle naklikat zadání a dostat nějaké výsledky. Pokud ale neví, co vlastně výsledky znamenají, jak se mohou interpretovat, nejsou mu buď k ničemu, nebo – což je horší – si je může interpretovat chybně a nadělat tak i škodu. Prakticky by měli s nástroji pro dolování dat pracovat (na žádost managementu nebo při vlastní analytické práci) jen analytici. I těm snadné a rychlé ovládání systému velmi urychlí práci.

Následující stručné informace jsou převzaty z reklamních materiálů firem nabízejících jednotlivé produkty. Proto je každý z nich „nejlepší nebo jediný“. Proto jsou zde někdy přidány k takovým tvrzením uvozovky. Pravidelně systémy pro dolování dat nabízí i základní metody matematické statistiky. Často je uvádí dohromady s metodami dolování, proto i v našem přehledu budou uváděny tak, jak jsou prezentovány

11.2. Enterprise Miner

Společnost SAS Institute Inc. založena v roce 1976 v USA.

SAS Systém je integrovaný softwarový systém poskytující kompletní kontrolu nad přístupem k datům, managementem, analýzou a jejich prezentací. Řešení je vhodné především pro velké organizace, pro banky, pojišťovny, finanční a telekomunikační organizace, a to zejména v oblastech řízení vztahů se zákazníky (CRM), řízení výkonnosti organizace (BSC) a finanční konsolidace. Svým zákazníkům umožňuje transformovat data – včetně velkého objemu dat generovaného v rámci e-businessu – do podoby využitelných informací. Software od SASu je využíván ve více než 38.000 obchodních, vládních a univerzitních organizacích ve 111 zemích. Mezi produkty SASu patří statistický program JMP-IN.

Produkt **Enterprise Miner** je „první a jediné“ řešení dolování dat, které zahrnuje celý proces dolování dat, má plně intuitivní grafické uživatelské rozhraní. V kombinaci se SAS datovým skladem a OLAP technologiemi vytváří součinné řešení, které osloví plné spektrum objevování znalostí.

Pro dolování dat používá firma **metodologii SEMMA**

Sample – identifikování množiny vstupních dat

Explore – prozkoumává množinu dat statisticky a graficky

Modify – příprava dat na analýzu

Model – výběr vhodného prediktivního modelu

Assess – srovnání konkurenčních prediktivních modelů

Systém obsahuje tyto **metody dolování dat**

Filtrování umožňuje vzít náhodná data (vhodné pro extrémně velké databáze).

Asociace nalezne asociativní vztahy v datech.

Shlukování nalezne data, které jsou si nějak podobné.

Neuronové sítě umožní zkonstruovat, natrénovat a ověřit vícevrstvou neuronovou síť.

Rozhodovací stromy rozdělení databáze založené na nominálních, ordinálních a spojitých proměnných.

Self-Organizing Maps / Kohonenovy sítě

Regrese

Automatizace procesu dolování dat

Enterprise Miner pomáhá tvořit otázky, na které by se možná nikdy nikdo nezeptal. Zabudovaná metodologie SEMMA poskytuje uživateli logický, organizovaný rámec pro provádění dolování dat. Začíná se statistickým vzorkem dat, tato metodologie usnadňuje použití výzkumné statistické a vizualizační techniky, vybrat a transformovat nejvýznamnější prediktivní proměnné, modelovat proměnné k předvídání výsledků a potvrdit modelovou přesnost.

Získání výsledků

Oceňováním výsledků získaných z každého stupně procesu se může určit, jak modelovat nové otázky z předchozích výsledků a vracet se tak do fáze zkoumání pro další zjemňování dat. Kompletní vyhodnocovací vzorec pro všechny stupně vývoje modelu je automaticky zachycen ve formě SAS, C a Java jazyků pro další rozmístění modelu. Nakonec nástroj Reporter poskytuje stručné HTML reporty o

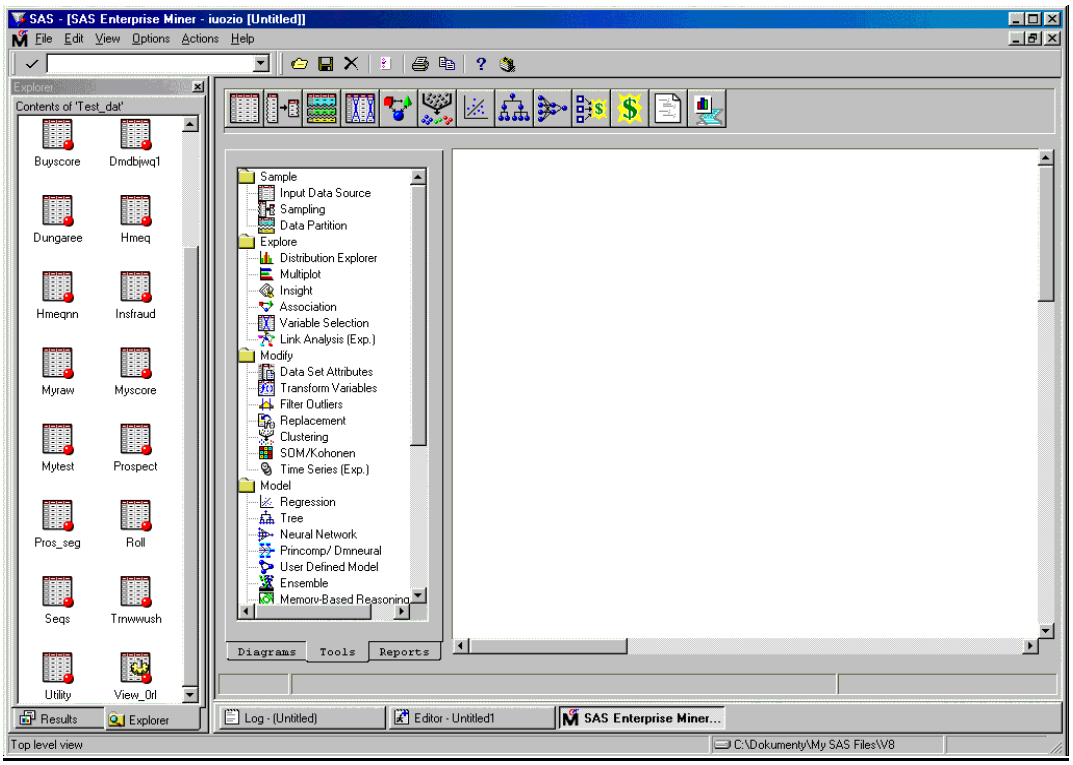
výsledcích procesního tok dolování dat, které si můžete pohlížet ve vašem oblíbeném internetovém prohlížeči.

V několika následujících ukázkách předvedeme skutečně intuitivní, snadné ovládání programu pomocí grafického rozhraní.

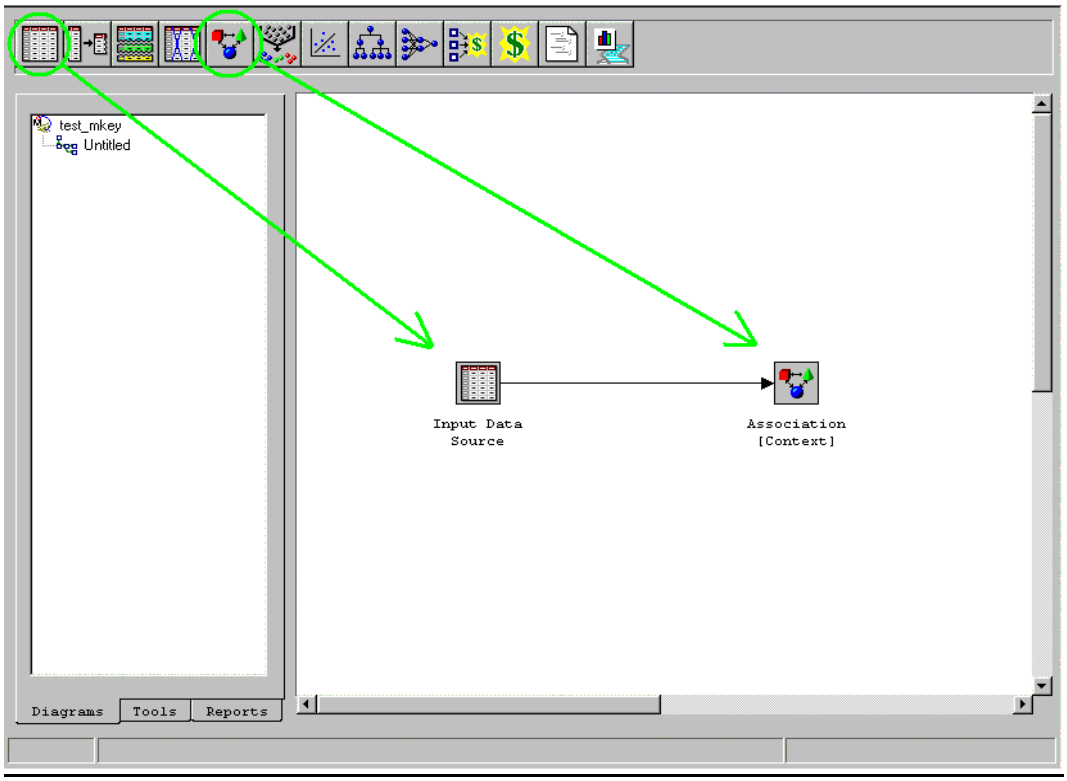
DM PRIMER - skládání částí

Každá metoda má svou ikonu a celý proces – od načtení dat přes provedení základních statistik, předzpracování, vlastní analýzy, vzájemné porovnání výsledků různých analýz až po prezentaci výsledků v grafické či tabulkové formě – se zadává vykreslením orientovaného grafu zpracování. Uzly tvoří vždy jedna z ikon příslušné metody, která se přetáhne z menu na pracovní plochu, hrany určují pořadí zpracovávaných metod.

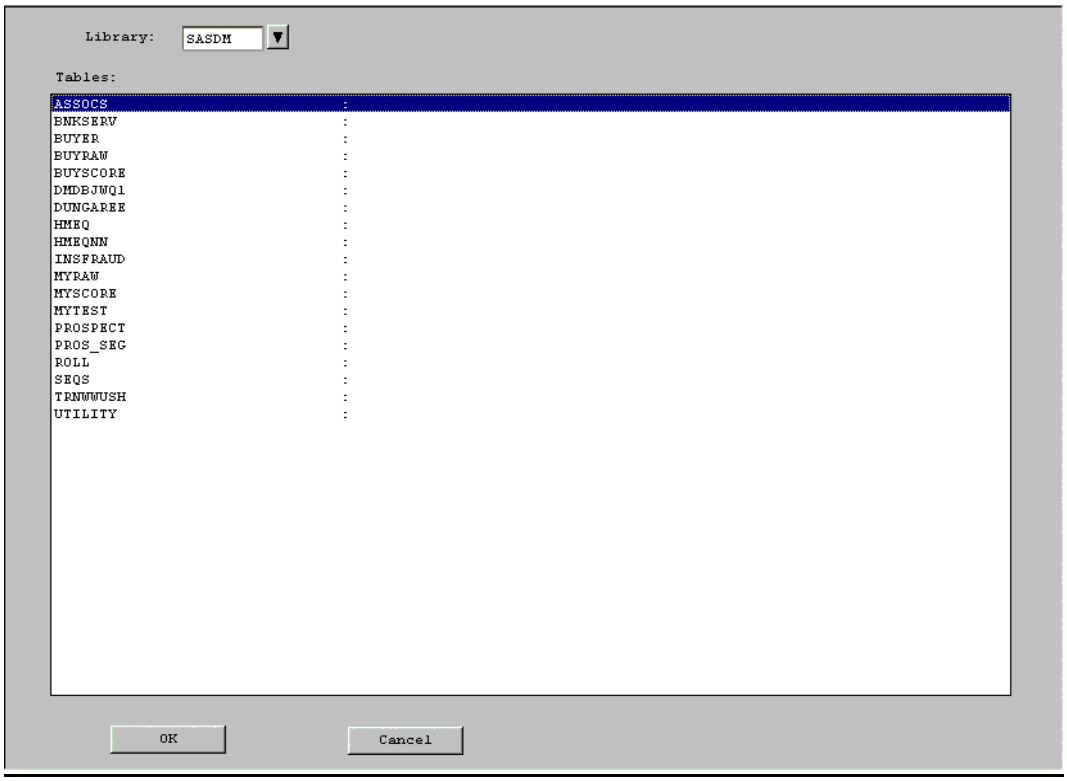
Základní obrazovka systému má tvar:



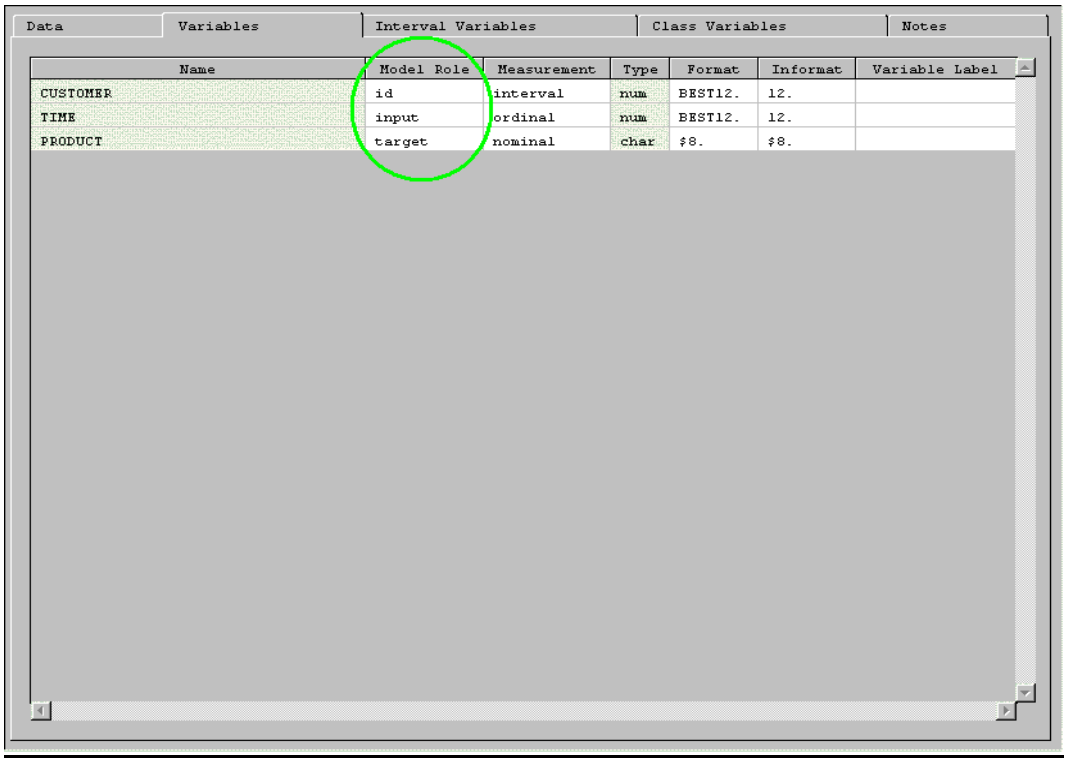
Výběr vstupů, výstupů a vazby



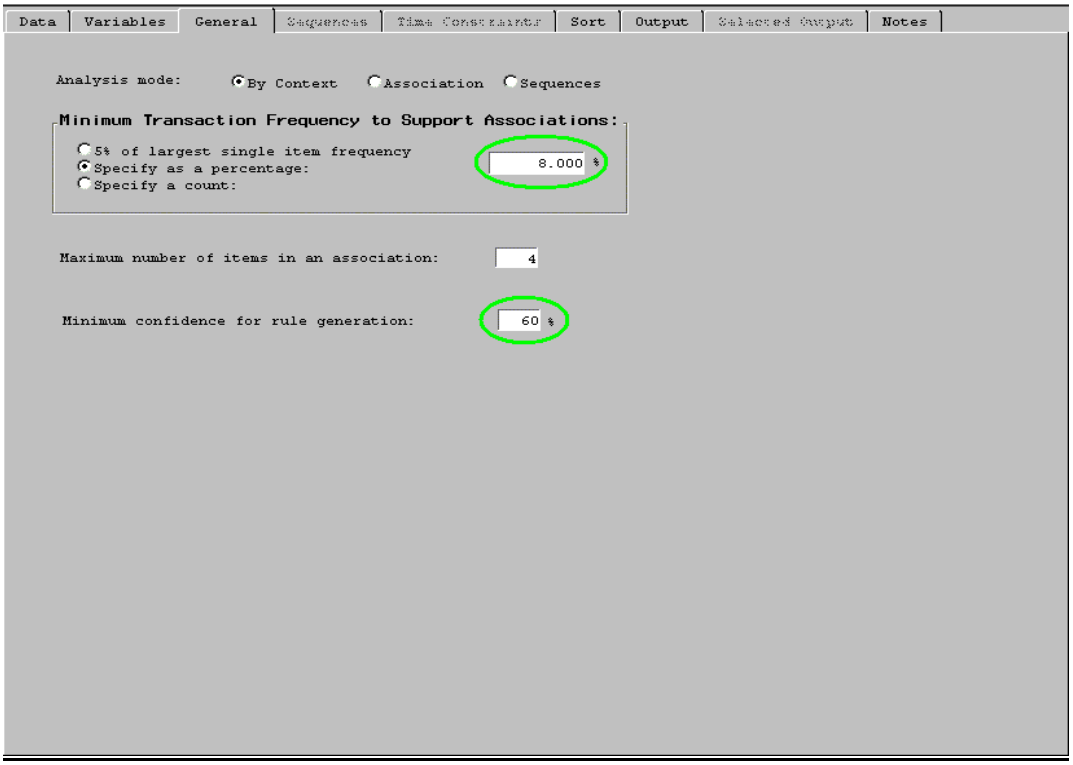
Výběr dat



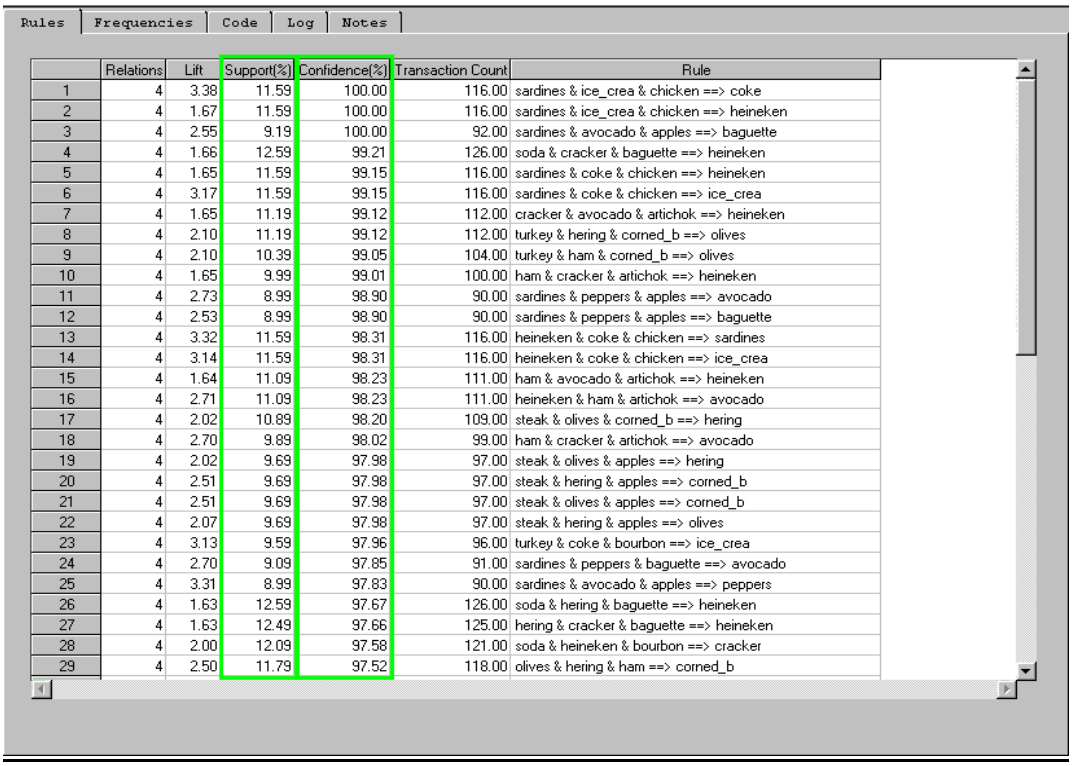
Změny v datech



Nastavení asociací



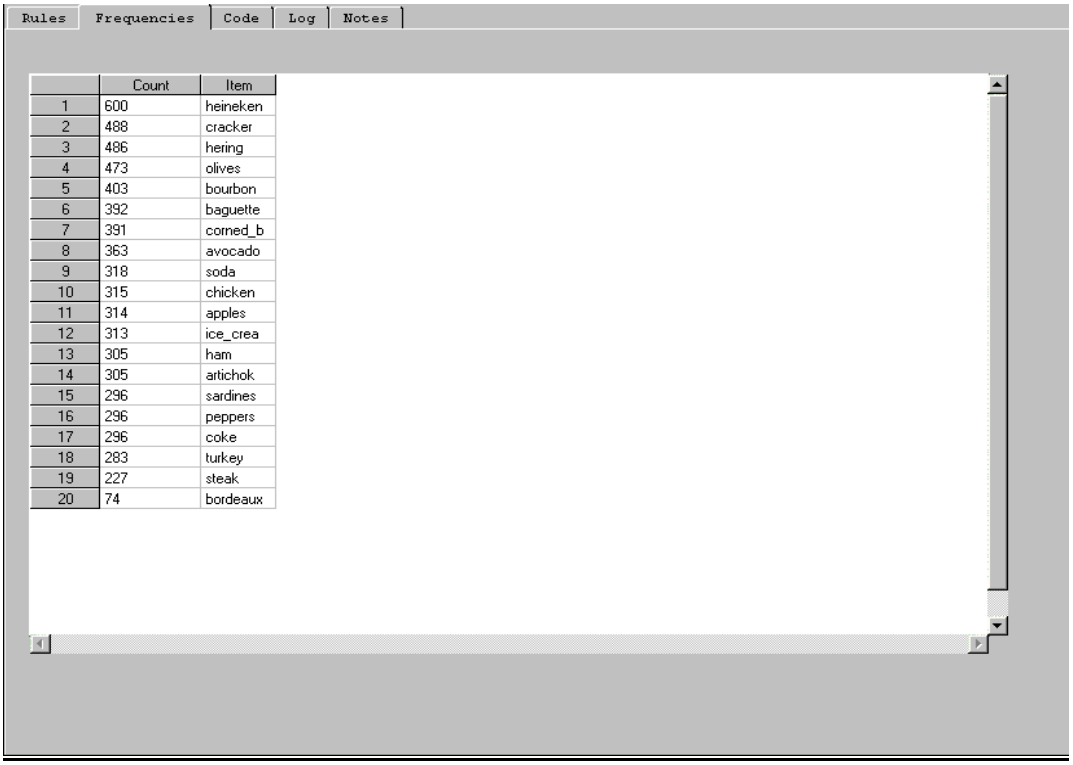
Výsledky (1. část)



Rules | Frequencies | Code | Log | Notes

	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	4	3.38	11.59	100.00	116.00	sardines & ice_crea & chicken ==> coke
2	4	1.67	11.59	100.00	116.00	sardines & ice_crea & chicken ==> heineken
3	4	2.55	9.19	100.00	92.00	sardines & avocado & apples ==> baguette
4	4	1.66	12.59	99.21	126.00	soda & cracker & baguette ==> heineken
5	4	1.65	11.59	99.15	116.00	sardines & coke & chicken ==> heineken
6	4	3.17	11.59	99.15	116.00	sardines & coke & chicken ==> ice_crea
7	4	1.65	11.19	99.12	112.00	cracker & avocado & artichok ==> heineken
8	4	2.10	11.19	99.12	112.00	turkey & hering & corned_b ==> olives
9	4	2.10	10.39	99.05	104.00	turkey & ham & corned_b ==> olives
10	4	1.65	9.99	99.01	100.00	ham & cracker & artichok ==> heineken
11	4	2.73	8.99	98.90	90.00	sardines & peppers & apples ==> avocado
12	4	2.53	8.99	98.90	90.00	sardines & peppers & apples ==> baguette
13	4	3.32	11.59	98.31	116.00	heineken & coke & chicken ==> sardines
14	4	3.14	11.59	98.31	116.00	heineken & coke & chicken ==> ice_crea
15	4	1.64	11.09	98.23	111.00	ham & avocado & artichok ==> heineken
16	4	2.71	11.09	98.23	111.00	heineken & ham & artichok ==> avocado
17	4	2.02	10.89	98.20	109.00	steak & olives & corned_b ==> hering
18	4	2.70	9.89	98.02	99.00	ham & cracker & artichok ==> avocado
19	4	2.02	9.69	97.98	97.00	steak & olives & apples ==> hering
20	4	2.51	9.69	97.98	97.00	steak & hering & apples ==> corned_b
21	4	2.51	9.69	97.98	97.00	steak & olives & apples ==> corned_b
22	4	2.07	9.69	97.98	97.00	steak & hering & apples ==> olives
23	4	3.13	9.59	97.96	96.00	turkey & coke & bourbon ==> ice_crea
24	4	2.70	9.09	97.85	91.00	sardines & peppers & baguette ==> avocado
25	4	3.31	8.99	97.83	90.00	sardines & avocado & apples ==> peppers
26	4	1.63	12.59	97.67	126.00	soda & hering & baguette ==> heineken
27	4	1.63	12.49	97.66	125.00	hering & cracker & baguette ==> heineken
28	4	2.00	12.09	97.58	121.00	soda & heineken & bourbon ==> cracker
29	4	2.50	11.79	97.52	118.00	olives & hering & ham ==> corned_b

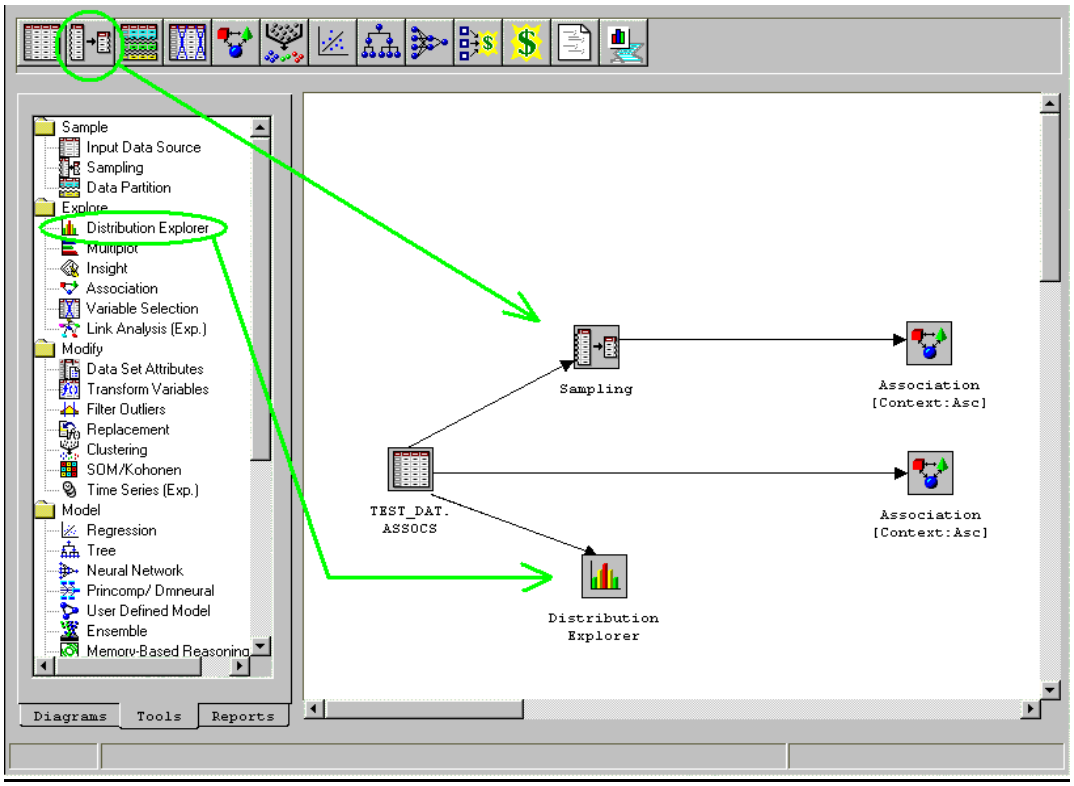
Výsledky (2. část)



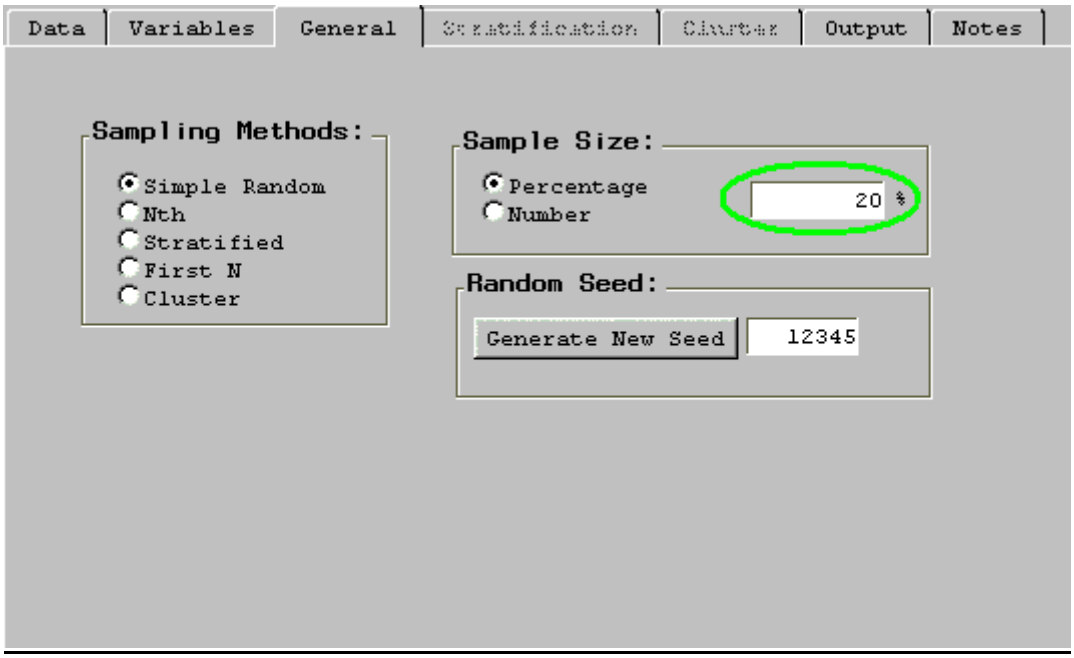
Rules | Frequencies | Code | Log | Notes

	Count	Item
1	600	heineken
2	488	cracker
3	486	hering
4	473	olives
5	403	bourbon
6	392	baguette
7	391	corned_b
8	363	avocado
9	318	soda
10	315	chicken
11	314	apples
12	313	ice_crea
13	305	ham
14	305	artichok
15	296	sardines
16	296	peppers
17	296	coke
18	283	turkey
19	227	steak
20	74	bordeaux

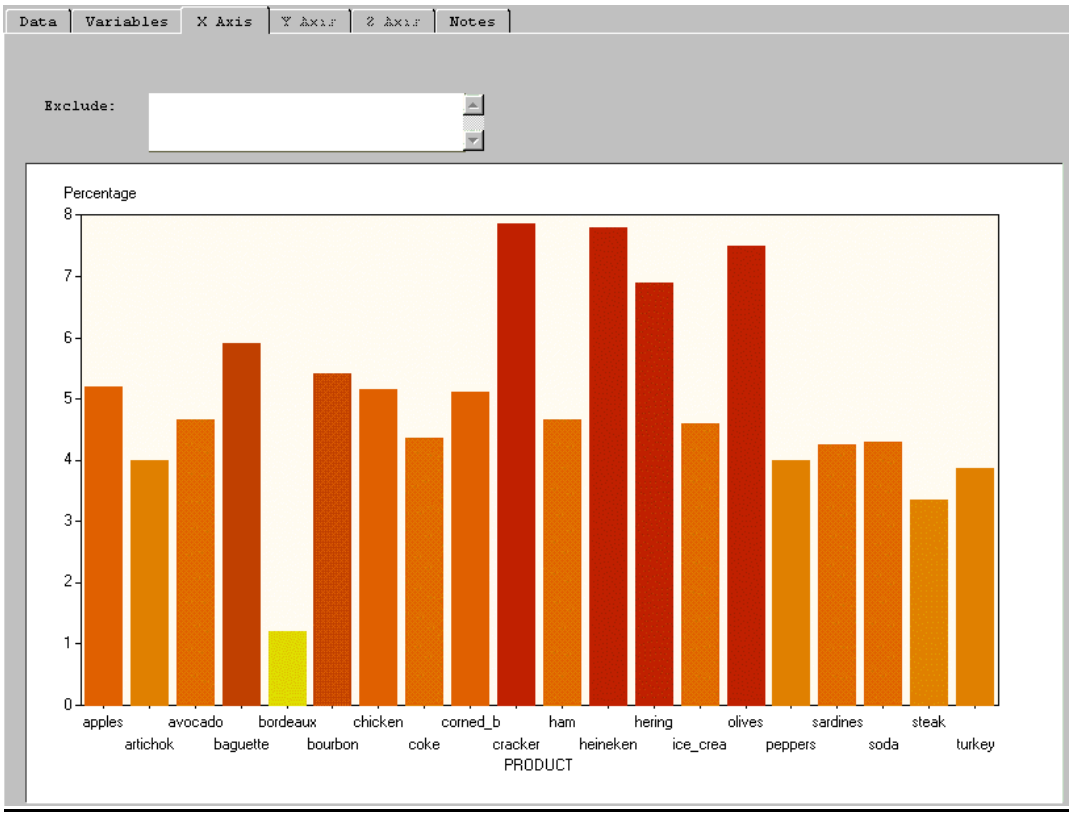
Asociace - menší počet dat



Sampling



Distribution Explorer



11.3. SPSS - Clementine

Společnost SPSS je výrobcem produktů, pokrývajících celou oblast běžně používané statistiky. Jejím největším produktem je data miningový program Clementine. Jde o produkt, orientující se téměř výhradně na data mining marketingový. Proto je jeho ovládání stavěno pro manažery a obchodníky, tedy i „naprosté laiky v oblasti metod data miningu a databází". Je kladen důraz především na kvalitu prezentace výsledků, přehlednost získaných dat a celkový uživatelský komfort. Jako většina firem na informace o používaných metodách firma již tak vstřícná a otevřená není.

Pražská pobočka firmy SPSS ČR, spol. s r.o. pořádá pravidelné reklamní akce i profesionální školení uživatelů.

Produkty jsou vyvíjeny jako standardy metodologie CRISP – DM (Cross Industry Standard Process for Data Mining). Pro Data Mining firma používá **metodologii** nazvanou **5A** (Assess, Access, Analyze, Act, Automate). Proces zahrnuje etapy:

Assess – posouzení konkrétní situace a určení jejích aspektů. Jde o vnoření procesu dolování do kontextu manažerského rozhodovacího problému a určení cílů.

Access – zajištění potřebných dat z databází a datových skladů, jejich předzpracování.

Analyze – samotné dolování závislostí a vztahů. Analytická část se dělí na dva kroky: odvození modelu a jeho validace, zjištění efektivity, přesnosti, spolehlivosti.

Act – formulace výsledků do podoby srozumitelné všem.

Automate – implementace prověřených a opakovaně využívaných modelů do běžné praxe.

Vstup a výstup dat přes ODBC umožňuje pracovat se širokou škálou vstupních formátů. Přímo lze načítat data formátu Excel, Oracle Express, SPSS, vlastním Clementine.

Nástroje pro filtrace a transformace dat: náhodný výběr případů ze souboru, filtrování dat, transformace časových řad, agregace, odvození nových proměnných apod. Funkce Balance umožňuje „vyvážit" vzorek dat tak, aby se počet příznivých a nepříznivých případů pro analýzy řádově rovnal (například v případě, že ve skutečnosti je úspěchů mizivé procento).

Analytické nástroje tvoří jádro programu. Zahrnují tradiční metody, neuronové sítě – mnohovrstvé perceptrony, radiální bazické sítě, Kohonenovy mapy, i „moderní metody". Firma uvádí následující rozdělení metod:

- Metody predikce a klasifikace
  - neuronové sítě
  - rozhodovací stromy a indukční pravidla
  - lineární regrese, logistická regrese, multinominální logistická regrese
- Shlukování a segmentace
  - Kohonenovy sítě
  - K-means
  - two step
- Detekce asociací
  - GRI (Generalized Rule Induction)
  - Apriori
  - pavučinová vizualizace
- Redukce dat

- faktorová analýza
- analýza hlavních komponent

Nástroje pro prezentaci zahrnují grafická zobrazení a tabulkové výsledky. Za hlavní trumf proti konkurentům inzeruje firma obrovské množství grafů a možných zobrazení výsledků:

- Dotazy myší k exploraci datových podmnožin v grafu
- Histogramy, distribuční a ostatní sloupcové grafy
- Čárové a bodové grafy
- Detekce vazeb pavučinovým grafem

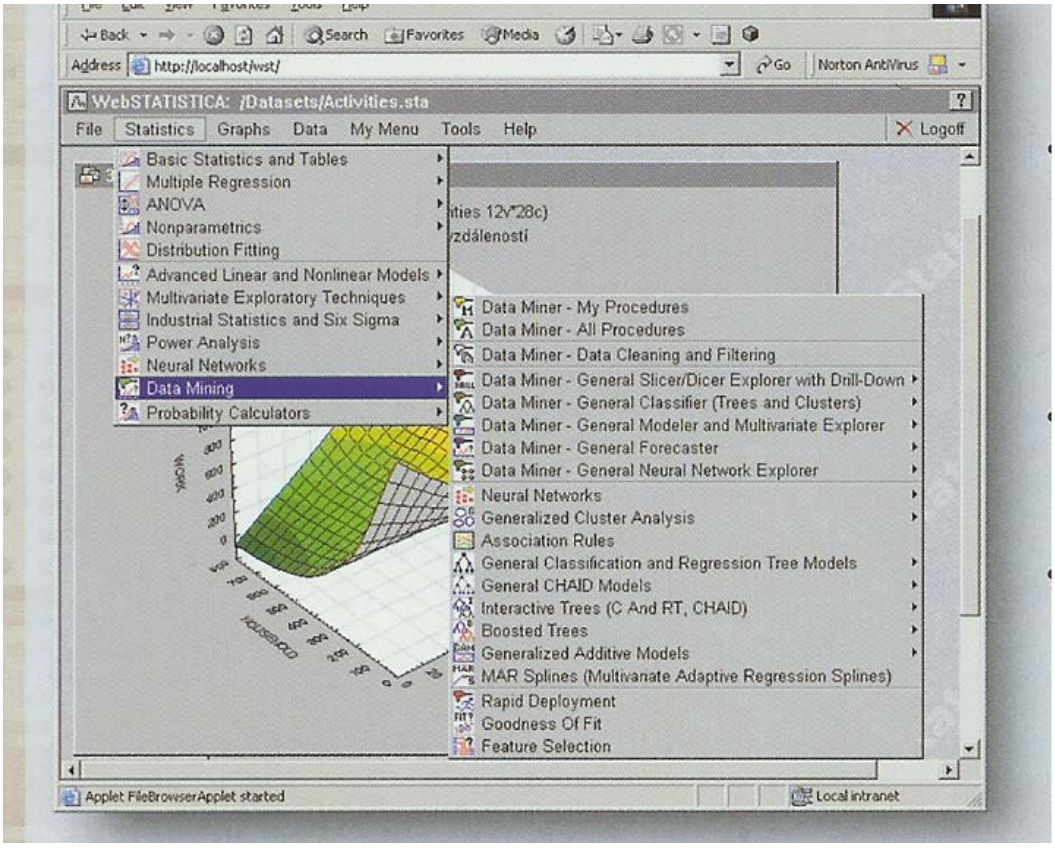
Například u histogramů je možno kliknutím určit místo rozhraní mezi dvěma skupinami případů. Vytvořený filtr je možno použít kdekoliv dále. Pavučinový graf ukazuje přehledně míru ekvivalence jednotlivých vlastností (vlastnosti jsou zobrazeny jako uzly, hrany znázorňují ekvivalenci mezi nimi).

11.4. Statistica Data Miner

Firma StatSoft, Inc. USA je producentem rozšířeného mohutného statistického systému **Statistica**. Do něj jsou postupně doplňovány i metody Data Mining. Firma příslušný subsystém Statistica Data Miner inzeruje jako „podnikový systém pro vytěžování dat (Data Miner) nabízející nejrozsáhlejší výběr technik pro vytěžování dat, který je v současnosti na trhu dostupný“. Je vybaven snadno použitelným uživatelským rozhraním založeným především na ikonách. Obsahuje kolekci plně integrovaných, automatizovaných a k činnosti připravených systémů jednotlivých řešení vytěžování dat pro nejrůznější obchodní aplikace.

Nástroje pro vytěžování dat

Analýzy jsou prováděny pomocí výkonných procedur pěti obecných nástrojů (každý se skládá z několika modulů systému Statistica). Každý se dá použít interaktivně, ale také se z nich dá vytvořit nový, samostatný postup.



Obrázek 11.1. Menu pro výběr metod

1. Obecný graficky průzkumník s technologii OLAP

(základní komponenta produktu STATISTICA Data Miner). Základ pro všechny procedury systému. Umožní sestavit vlastní aplikace, má nástroje pro týmovou práci, databázi skriptů, dotazů, dat a jiných informací, které se dají sdílet mezi uživateli. Navíc obsahuje i nástroje tvořící rozhraní databázových dotazů, které umožní zpracovávat vzdálené databáze přímo "na místě" (nemusíte si vytvářet lokální kopie části dat). Tato část Data Mineru obsahuje také velký výběr obecných průzkumných technik pro

vytěžování dat - od mocné a všestranné implementace OLAP, přes nástroje pro ověřování a očišťování dat, nástroje pro rozřezání dat "na plátky" (Slicer) a "na kostičky" (Dicer), až po flexibilní generování vícerozměrných rozkladů, křížových tabulek a protokolů. Opět tedy „nejširší nabídka grafických a průzkumných technik pro vytěžování dat, jakou dosud není vybaven žádný produkt na trhu“.

2. Obecný klasifikátor

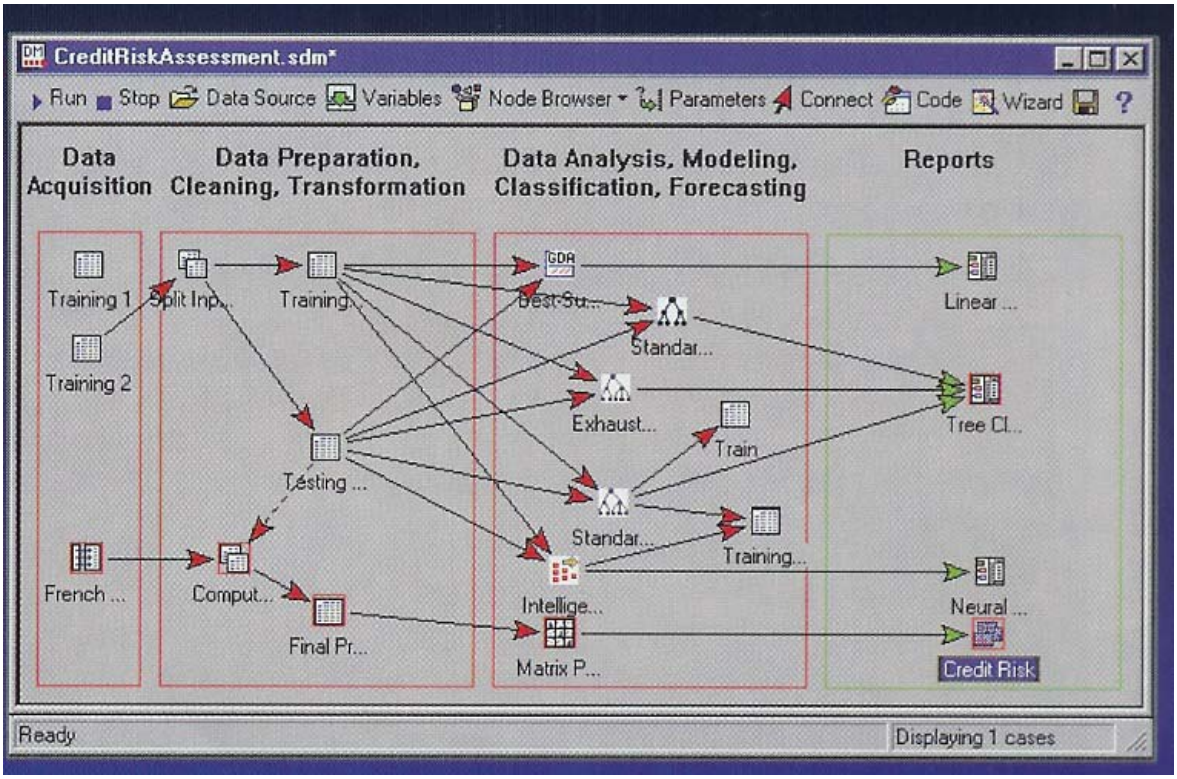
nabídka nástrojů vybavených klasifikačními technikami vytěžovaní dat a metodami pro tvorbu příslušných modelů. Jsou zde Klasifikační stromy, Obecné klasifikační a regresní stromy (GTrees), Obecné modely CHAID, Techniky shlukové analýzy, Obecné modely diskriminační analýzy.

3. Obecný modul tvorby modelů / vícerozměrný průzkumník

umožní sestavit modely lineárními a nelineárními technikami, prozkoumat data a sestavit prediktivní modely pomocí obecných vícerozměrných technik, včetně pokročilých obecných lineárních a nelineárních regresních modelů, GLM, GLZ, GPLS, log-lineárních modelů, modelů přežívání a ostatních specializovaných modelů. Nabízí techniky modelování pomocí strukturálních rovnic a opravdu rozsáhlé možnosti vícerozměrných průzkumných technik (včetně vysoce specializovaných a výkonných metod vzniklých kombinací korespondenční analýzy a technik hlavních komponent, vícerozměrného škálování nebo faktorové analýzy).

4. Obecný prediktor

široká škála tradičních predikčních technik (tzn. nezaložených na neuronových sítích). Obsahuje modely ARIMA, exponenciální vyhlazování, sezónní rozklad, a také regresní a polynomiální distribuované modely.



Obrázek 11.2. Grafický návrh procesu analýzy dat

5. Obecný průzkumník neuronových sítí

nejúplnější sada metod neuronových sítí „jakou v současnosti můžete vidět“, pomocí nichž se řeší problémy související s vytěžováním dat (klasifikace, detekce skryté struktury, predikce, atd.). Jednou z unikátních vlastností průzkumníka NS je výběr inteligentních metod pro řešení problému a automatičtí průvodci, kteří využívají technik umělé inteligence a kteří pomohou s řešením nejnáročnějších problémů pokročilé analýzy pomoci NS (např. volba nejlepší architektury sítě a nejlepší sady proměnných). Součásti je také modul pro generování vysoce optimalizovaného kódu v jazyce C. Průzkumník NS obsahuje opravdu špičkové procedury využívající NS a optimalizované algoritmy - vícevrstvé perceptrony, samoorganizující se mapy funkcí, zpětné šíření (Back Propagation), metodu konjugovaných gradientů, převzorkování, krizové ověření, analýzu citlivosti, křivky ROC, soubory sítí, a spoustu dalších.

❑ Specializované moduly pro vytěžování dat

Velká část analytických funkcí je převzata z výpočetních jader modulů jiných produktů z rodiny Statistica. Jsou zde tři moduly, které obsahují vysoce specializované modelovací techniky pro vytěžování dat, které najdete pouze ve Statistica Data Mineru.

1. Zobecněné aditivní modely (GAM)

Program umí pracovat se spojitými a kategorizovanými proměnnými prediktoru. Statistica obsahuje velké množství metod pro tvorbu nelineárních modelů odpovídajících datům, jako třeba modul nelineárních odhadů, zobecněné lineární modely, obecné klasifikační a regresní stromy, atd.

Rozdělení a funkce vlivu. Program umožňuje uživateli vybrat si z mnoha rozdělení závisle proměnné a funkce vlivu efektu proměnných prediktoru na závislou proměnnou:

Normální, Gamma a Poissonovo rozdělení:

Logaritmická funkce vlivu: $f(z) = \log(z)$
Inverzní funkce vlivu: $f(z) = 1/z$
Identická funkce vlivu: $f(z) = z$

Binomické rozdělení:

Logitová funkce vlivu: $f(z)=\log(z/(1-z))$

Vyhlazovač bodových grafů. Program využívá kubickou spline vyhlazovací metodu s volitelným počtem stupňů volnosti. Pomocí ní vyhledává optimální transformační funkci proměnných prediktoru.

Výsledkové statistiky. Program umí vytvářet rozsáhlou sadu výsledkových statistik, čímž zjednodušuje ověřování adekvátnosti modelu, přesnosti proložení a interpretaci výsledku. Mezi výsledky jsou: záznam historie iterací při určování optimálního modelu, souhrn statistik (včetně celkové hodnoty R-kvadrát počítané z rozptylu), počet stupňů volnosti modelu a detailní statistiky predikované odezvy, reziduí a vyhlazení proměnných prediktoru. Výsledkové grafy obsahují grafy pozorovaných odezev v závislosti na predikovaných odezvách, grafy predikovaných hodnot v závislosti na reziduích, histogramy pozorovaných a reziduálních hodnot, normální pravděpodobnostní grafy reziduálních hodnot a částečně reziduální grafy pro každý prediktor.

2. Obecné klasifikační a regresní stromy (GTrees). Tento modul je obsáhlou implementací metod popsaných v literatuře CART od Breiman, Friedman, Olshen a Stone (1984). Modul GTrees ale navíc obsahuje různá rozšíření a možnosti, které se obvykle v jiných implementacích tohoto algoritmu nenacházejí a které jsou velice užitečné pro aplikace vytěžování dat.

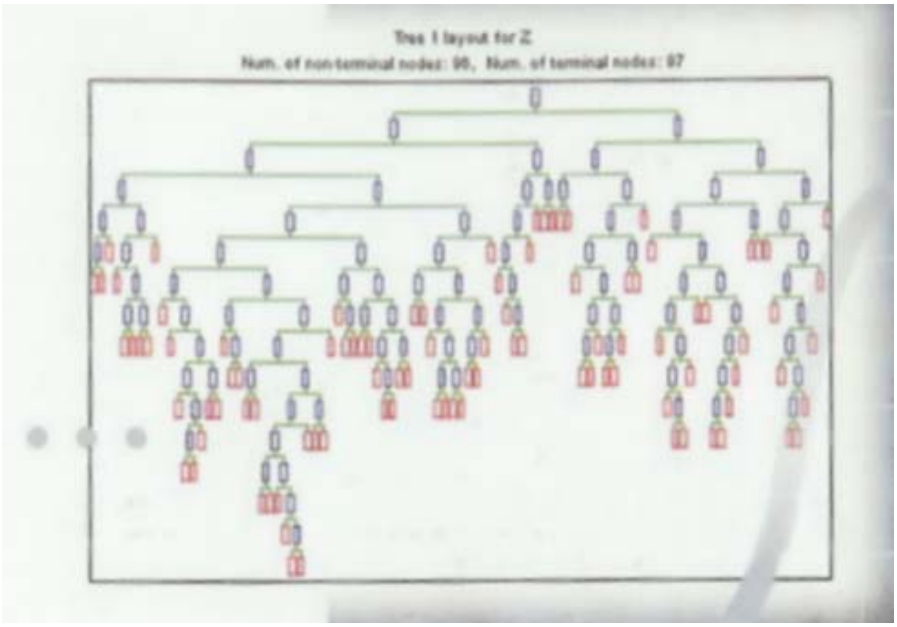
Uživatelské rozhraní; specifikace "modelu." Navíc ke standardním analýzám (které jsou popsány v Breiman a kol.) vám implementace těchto metod v systému STATISTICA umožní vytvářet modely ANOVA/ANCOVA se spojitými nebo kategorizovanými proměnnými prediktoru a jejich interakcemi.

Ve třech alternativních uživatelských rozhraních budete moci tyto modely vytvářet; návrhy modelu jsou analogické návrhům a metodám v modulech GLM (Obecné lineární modely), GLZ (Zobecněné lineární modely), GRM (Obecné regresní modely), GDA (Obecná diskriminační analýza) a PLS (Metoda částečných nejmenších čtverců) a detailně jsou popsány v příslušných sekcích. V krátkosti, návrhy prediktoru ANOVA/ANCOVA se dají určit v dialozích a průvodcích nebo ve formě syntaxe příkazu; ta je navíc kompatibilní ve všech modulech, takže různé analýzy můžete vytvořit prakticky stejným způsobem (např. můžete porovnat kvalitu klasifikace pomoci GDA proti klasifikaci pomoci GTrees).

Prořezávaní stromu, selekce a ověřeni. Program poskytuje velký výběr možnosti pro ovládaní procesu sestavování stromu, jejich prořezávaní a pro volbu nejlepšího řešení. Pro spojité závislé kriteriální proměnné může být prořezávaní založeno na rozptylu nebo se dá použít prořezávaní stylem FACT. Pro kategorizované závislé kriteriální proměnné se zase používá prořezávaní podle počtu špatně klasifikovaných případů, podle rozptylu nebo prořezávaní stylem FACT. Můžete si určit maximální počet uzlů stromu a minimální počet větví jednoho uzlu. Máte možnosti pro nalezení nejlepšího rozhodovacího stromu (např. pomocí krizového ověření nebo aplikací rozhodovacího stromu na nová pozorování ve vzorku dat). Pro kategorizované kriteriální proměnné, např. pro klasifikační problémy, můžete zvolit různé míry, jimiž lze modifikovat algoritmus a ohodnotit kvalitu konečného rozhodovacího stromu. Máte možnost určit také apriorní pravděpodobnosti jednotlivých tříd klasifikace a penalizace za chybnou klasifikaci. Míry kvality rozhodovacího stromu obsahují míru Gini, Chi-kvadrát a G-kvadrát.

Chybějící data a náhradní dělení. Chybějící hodnoty dat v prediktorech se dají zpracovat tak, že umožníte programu určit vhodná "místa" větvení pomocí náhradních proměnných, tj. podle proměnných, které jsou podobné příslušné proměnné, použité pro určité větvení (uzel stromu).

Návrhy ANOVA/ANCOVA modelu. Navíc k tradiční CART analýze, můžete zkombinovat kategorizované a spojité proměnné prediktoru do modelu typu ANOVA/ANCOVA a provádět analýzu pomoci navržené matice pro prediktory. To vám umožní ohodnotit a porovnat složité modely prediktoru a jejich efektivit při předpovídaní a klasifikaci pomocí různých analytických technik (např. obecné lineární modely, zobecněné lineární modely, obecné modely diskriminační analýzy, atd.).



Obrázek 11.3. Zobrazení výsledného rozhodovacího stromu

Prohlížeč stromů. Možnost zobrazit si jednoduchý souhrnný graf stromu je doplněna ještě funkci intuitivního interaktivního stromového prohlížeče, který vám umožní nechat zkolabovat nebo naopak expandovat libovolný uzel stromu a prohlednout si k němu příslušné informace. Například můžete kliknutím označit určitý uzel v panelu prohlížeče a okamžitě uvidíte poměr dobře a špatně klasifikovaných případů pro daný uzel. Prohlížeč stromů poskytuje velice efektivní a intuitivní možnost pro kontrolu složitých stromových struktur pomocí metod, které jsou běžně používaný v počítačových aplikacích. Několik oken prohlížeče stromových struktur může být otevřeno najednou, takže můžete zároveň sledovat celkový strom i několik jednotlivých podstromů, což vám umožní jejich snadné porovnání. STATISTICA Prohlížeč stromů je důležitou inovací, která velice pomůže uživatelům při interpretaci složitých rozhodovacích stromů.

Výsledkové statistiky. Modul STATISTICA GTrees poskytuje velké množství výstupů. Dostupné jsou souhrnné výsledky pro každý uzel, počítají se detailní statistiky klasifikace, atd. Najdete zde i unikátní grafická shrnutí, včetně histogramu (pro klasifikační problémy) každého uzlu, detailních souhrnných grafů spojitých závislých proměnných (např. normální pravděpodobnostní grafy, bodové grafy) a včetně grafu každého uzlu, které vám poskytnou efektivní souhrn vzoru odezev pro velké klasifikační problémy. Stejně jako v ostatních statistických procedurách systému STATISTICA mohou byt všechny numerické výsledky použity jako vstup pro další analýzy, díky čemuž budete moci rychle shlédnout a dále analyzovat pozorování provedena na určitém uzlu (např. můžete použít modul GTrees k vytvoření počáteční klasifikace případu a pak použít výběr nejlepší podmnožiny v GDA, což vám umožní nalézt další proměnné, které mohou vylepšit další klasifikaci).

3. Obecné modely CHAID (Chi-kvadrát automatická detekce interakcí).

Stejně jako implementace obecných klasifikačních a regresních stromů v systému STATISTICA, ani modul Obecné modely CHAID neposkytuje "jen" obsáhlou implementaci původní techniky, ale její metody rozšiřuje na analýzu modelu ANOVA/ANCOVA.

Standardní CHAID. Analýza CHAID se dá provádět jak na spojitých, tak i na kategorizovaných závislých proměnných. K dispozici je spousta možností, jak ovládat proces tvorby hierarchického stromu. Máte možnost stanovit nejnižší počet větví v každém uzlu, maximální počet uzlu a pravděpodobnosti pro větvení a slučování kategorii. Uživatel také může provádět mohutná vyhledávaní nejlepšího řešení (Exhaustive CHAID). Dají se také spočítat statistiky V-fold pro ověření a ohodnocení stability cílového řešení. Pro klasifikační problémy můžete určit i pokuty za chybnou klasifikaci.

Modely ANOVA/ANCOVA. Jako doplněk k tradiční analýze CHAID vám program umožňuje zkombinovat kategorizované a spojité proměnné prediktoru do modelu ANOVA/ANCOVA a provést analýzu pomoci navržené matice pro prediktory. To vám umožni ohodnotit a porovnat složité modely prediktoru a jejich efektivitu pro předpovědi a klasifikaci pomoci nejrůznějších analytických technik (např. obecných lineárních modelů, zobecněných lineárních modelů, obecných modelů diskriminační analýzy, obecných klasifikačních a regresních stromů, atd.). Více detailů naleznete v popisu GLM (Obecné lineární modely) a GTrees (Obecné klasifikační a regresní stromy).

Prohlížeč stromů. Stejně jako binární výsledkové stromy používané pro shrnuti klasifikačních a regresních stromu, i výsledky analýzy CHAID se dají prohlížet v STATISTICA Prohlížeči stromů. Prohlížeč stromu poskytuje velice efektivní a intuitivní možnost pro kontrolu složitých stromových struktur a pro porovnání několika řešení v podobě stromu pomocí metod, které jsou běžné používány v počítačových aplikacích. STATISTICA Prohlížeč stromů je důležitou inovací, která velice pomůže uživatelům při interpretaci složitých rozhodovacích stromů. Další detaily naleznete také v sekci GTrees (Obecné klasifikační a regresní stromy).

Výsledkové statistiky. Modul STATISTICA Obecné modely CHAID poskytuje velké množství výstupu. Dostupné jsou souhrnné výsledky pro každý uzel, počítají se detailní statistiky klasifikace, atd. Najdete zde i unikátní grafická shrnutí, včetně histogramu (pro klasifikační problémy) každého

uzlu, detailních souhrnných grafů spojitých závislých proměnných (např. normální pravděpodobnostní grafy, bodové grafy) a včetně grafu každého uzlu, které vám poskytnou efektivní souhrn vzorů odezev pro velké klasifikační problémy. Stejně jako v ostatních statistických procedurách systému STATISTICA mohou byt všechny numerické výsledky použity jako vstup pro další analýzy, díky čemuž budete moci rychle shlédnout a dále analyzovat pozorování provedena na určitém uzlu (např. můžete použít modul GTrees k vytvoření počáteční klasifikace případu a pak použít výběr nejlepší podmnožiny v GDA, což vám umožni nalézt další proměnné, které mohou vylepšit další klasifikaci).

❑ Pokročilé lineární/nelineární modely

jsou doplňkem, který obsahuje nejširší škálu pokročilých lineárních a nelineárních modelovacích nástrojů, jaké jsou dnes na trhu dostupné. Podporuje spojité prediktory i kategorizované prediktory, interakce, hierarchické modely, možnosti automatické volby modelu, komponenty rozptylu, časové řady a mnoho dalších metod. Všechny analýzy mají rozsáhlou grafickou podporu a jsou vybavené možnosti psaní skriptu ve vestavěném Visual Basicu. Obsahuje následující moduly:

Komponenty rozptylu a smíšené modely ANOVA / ANCOVA
Analýza přežívaní, analýza poruch
Obecné nelineární odhady (a rychlá logit / probit regrese)
Log-lineární analýza kontingenčních tabulek
Analýza časových řad/Predikce
Modelování pomocí strukturálních rovnic / Analýza cesty (SEPATH)
Obecné lineární modely (GLM)
Obecné regresní modely (GRM)
Zobecněné lineární modely (GLZ)
Metoda částečných nejmenších čtverců (PLS)

11.5. Oracle 9i – Data Mining

Jedna z velkých databázových firem Oracle, založená 1983 v USA, ve své verzi Oracle 9i rel2 database z roku 2002 mimo jiné nabízí také podporu Business Intelligence, což zahrnuje OLAP a Data Mining. Firma zveřejňuje na svých webových stránkách poměrně podrobně informace i o metodách.

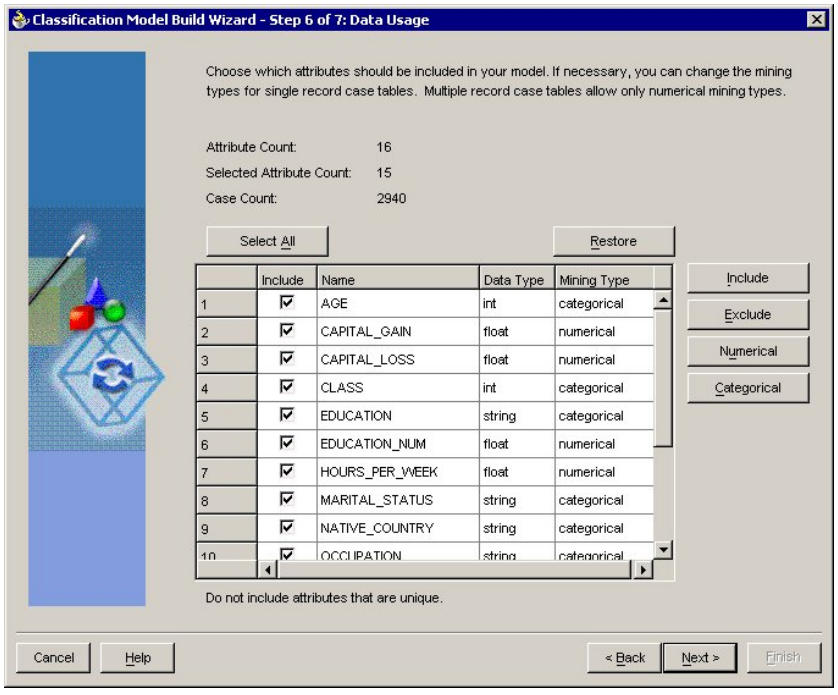
Zajímavostí je, že Data Mining probíhá přímo v databázi.

Do Data Miningu zahrnuje

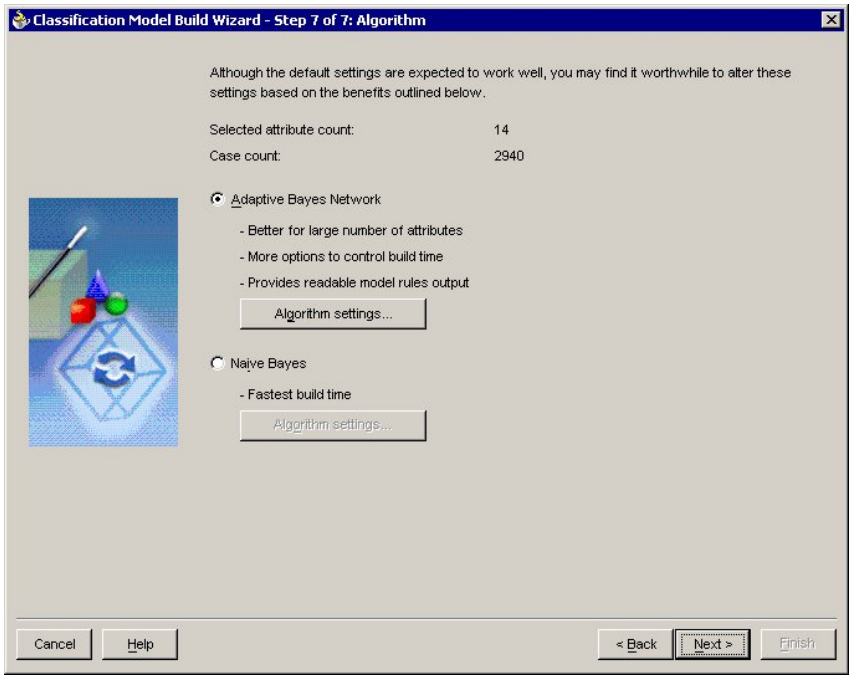
- přípravu dat
    - ošetření chybějících údajů – místa s NULL v datech jsou ignorována
    - diskretizaci dat několika technikami
- klasifikaci
    - nativní Bayesův algoritmus
    - firmou vyvinutý algoritmus pro rozhodovací strom s využitím adaptivní Bayesovy sítě
    - Model Seeker – modul využívající vztah pro vyhodnocení nejlepšího z vypočtených modelů
- shlukování
    - k-means algoritmus
    - O-Cluster, vlastní algoritmus pro rozsáhlá data
- asociační pravidla – nákupní košík s danou podporou a spolehlivostí
- korelační analýza, výpočet parciálních korelací (korelace zkoumané veličiny s cílovou veličinou s ohledem na ostatní veličiny)

Standardem je i zde grafické uživatelské prostředí se sadou grafických nástrojů (wizards), pomocí nichž se vytvářejí data miningové modely. Z nich se automaticky generuje Java kód. V následujících ukázkách si předvedeme několik klasických obrazovek:

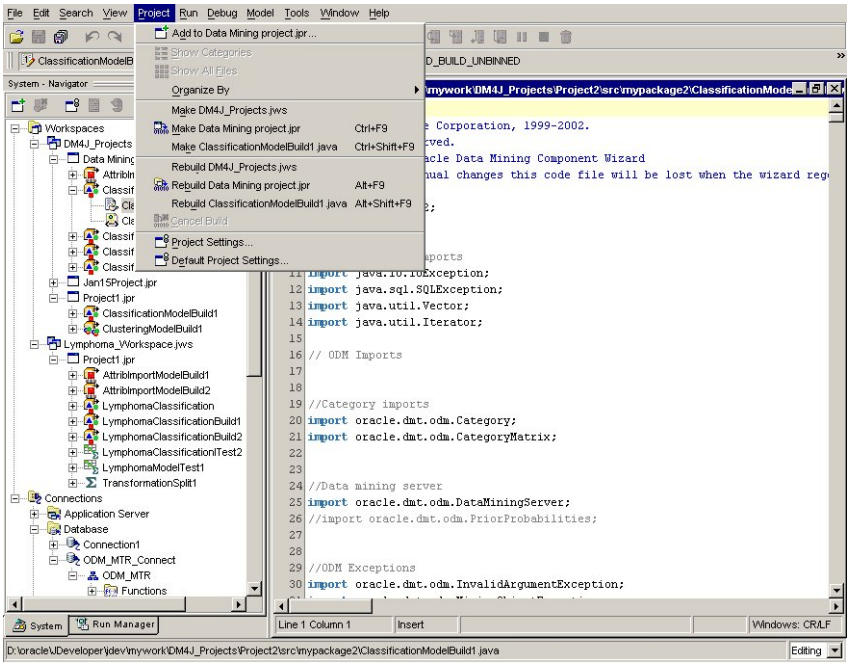
Pomocí wizardů se volí, které atributy budou zahrnuty do modelování



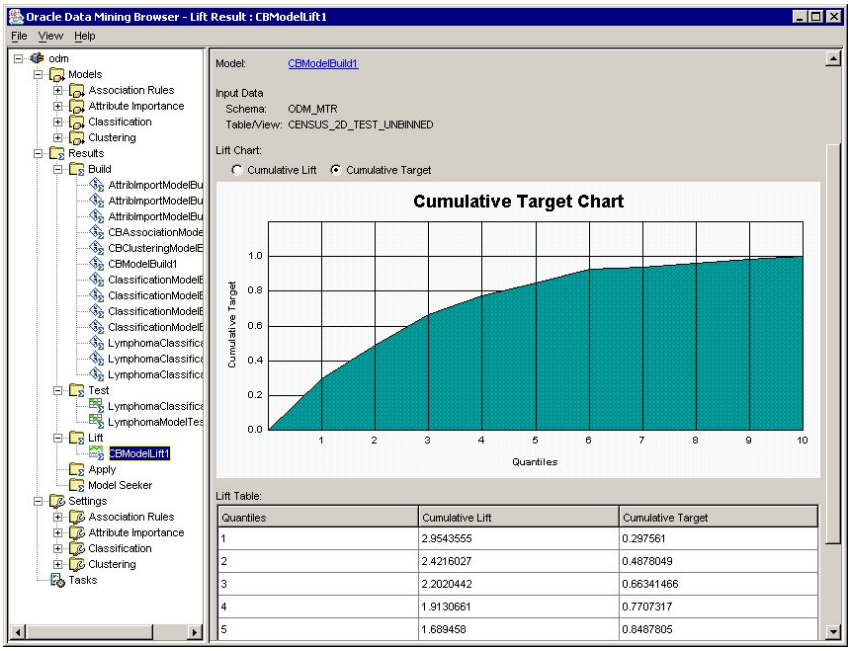
Volba algoritmu pro dolování a nastavení jejích parametrů



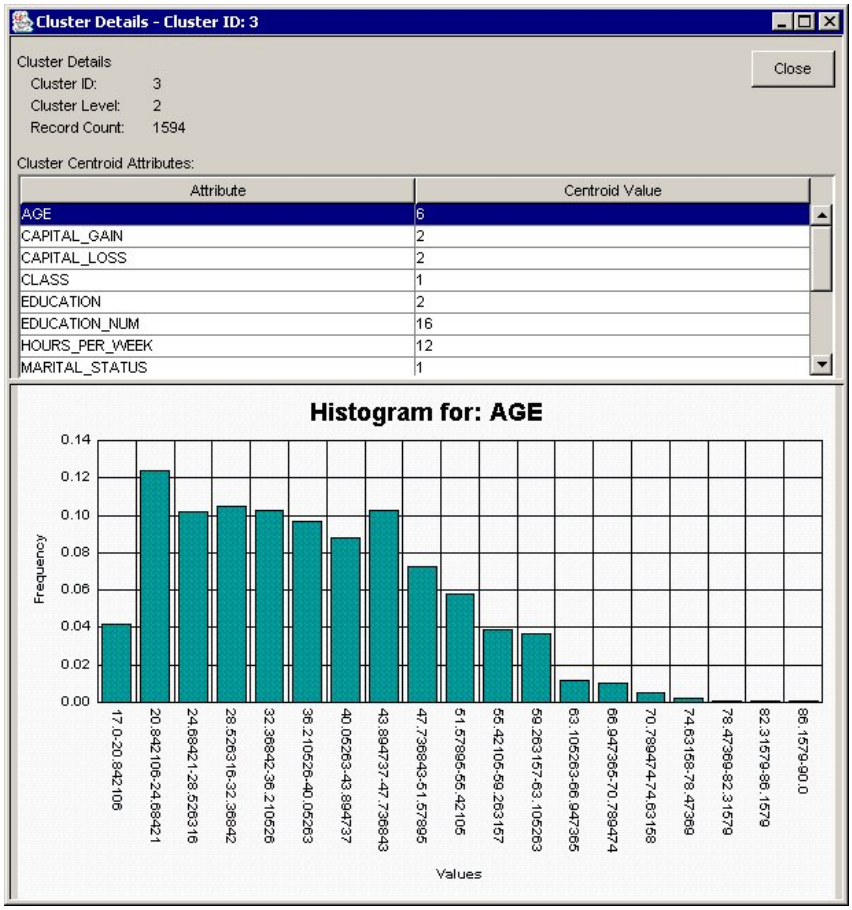
ODM komponenty automaticky generují Java kód, který může být dále využíván při tvorbě aplikací.



Prohlížeč pro vizualizaci výsledků



Výsledky „shlukovací analýzy“ pomocí histogramu



11.6. Microsoft SQL Server 2000

MS SQL Server 2000 je produkt, který je možno rozdělit do několika oddělených bloků:

- **Microsoft SQL Server** – relační databázový transakční systém (OLTP)
- **Microsoft Data transformation Services** – systém pro datovou pumpu, pro **e**xtrakci, **t**ransformace a **l**oading (ETL)
- **Microsoft OLAP Server** – systém pro On Line Analytical Processing
- **Microsoft Data Mining Services** – nástroj pro dolování znalostí z dat
- **Off-line OLAP** – otevřené prostředí pro komunikaci s uživatelem.

Analytické nástroje, které jsou jeho součástí, tedy lze rozdělit na nástroj pro „pumpování" dat ETL a pro vlastní analýzy dat (OLAP s DS a Data Mining).

Analýza v MS SQL Serveru

Součástí dodávky MS SQL Serveru 2000 je nástroj **Analysis Manager** a jeho průvodce (Wizard), pomocí kterého se vytváří tzv. *mining model* nad daty, která chceme analyzovat. Údaje pro Data Mining lze získávat jak z relačních databází, tak z multidimenzionálních datových struktur OLAP.

1. Zadání typu zdroje údajů (relační databáze nebo OLAP).
2. Po načtení dat se pomocí dialogu se seznamem tabulek databáze určuje tabulka nebo více tabulek, které se stanou podkladem dataminingu.
3. Výběr algoritmu podle typů atributů (reálná - **MS clustering**, kategoriální - **MS decision Trees**).
4. Výběr vstupních a predikovaných sloupců (atributů).
5. Za předpokladu, že byla vybrána metoda MS decision Trees, bude výsledkem analýzy přehledný diagram pro všechny hodnoty predikovaného atributu v další pomůcce — **Relation Mining Model editoru**. V této aplikaci je diagram stromem, kde se tmavou barvou zobrazí všechny hodnoty atributů, které mají největší vliv na předpovídaný atribut.
6. Pro zjištění informace, na čem závisí určitá hodnota predikovaného atributu, slouží nastavení *Tree color based on*. Pomocí tohoto prvku můžeme nastavit, ke které hodnotě predikovaného atributu se bude intezita barvy v diagramu vztahovat.
7. Naše případné závěry si lze velice jednoduše ověřit pomocí nástroje **Dependency Network Browser**. Tento nástroj názorně zobrazuje diagram atributů, na kterých predikovaná veličina závisí. Jednoduchým posouváním posuvníku postupně odpadají méně významné vlivy (atributy). Takto si jednoduše nastavíme potřebný počet atributů, na kterých predikovaná veličina závisí.

Predikce podle provedené analýzy

Pokud máme k dispozici nová (nebo jiná) data se stejnou strukturou a stejnými atributy, můžeme si vyzkoušet predikci vlastností na základě předchozí analýzy. Jako nástroj pro predikci slouží **Prediction Query Task** v aplikaci **DTS Package**. Parametry pro tuto analýzu se nastavují ve třech záložkách:

- nastavení připojení na přislušnou databázi analytického serveru
- specifikace predikované veličiny na základě vstupních veličin (pomocí SQL příkazu nebo pomocí průvodce **Prediction Query Builder**)

- určení místa (databáze) pro uložení výsledku predikce

Úspěšnost předpovědí lze pak jednoduše ověřit SQL příkazem, v kolika popř. ve kterých objektech byla či nebyla predikce správná.

Algoritmy pro analýzu dat

Microsoft clustering

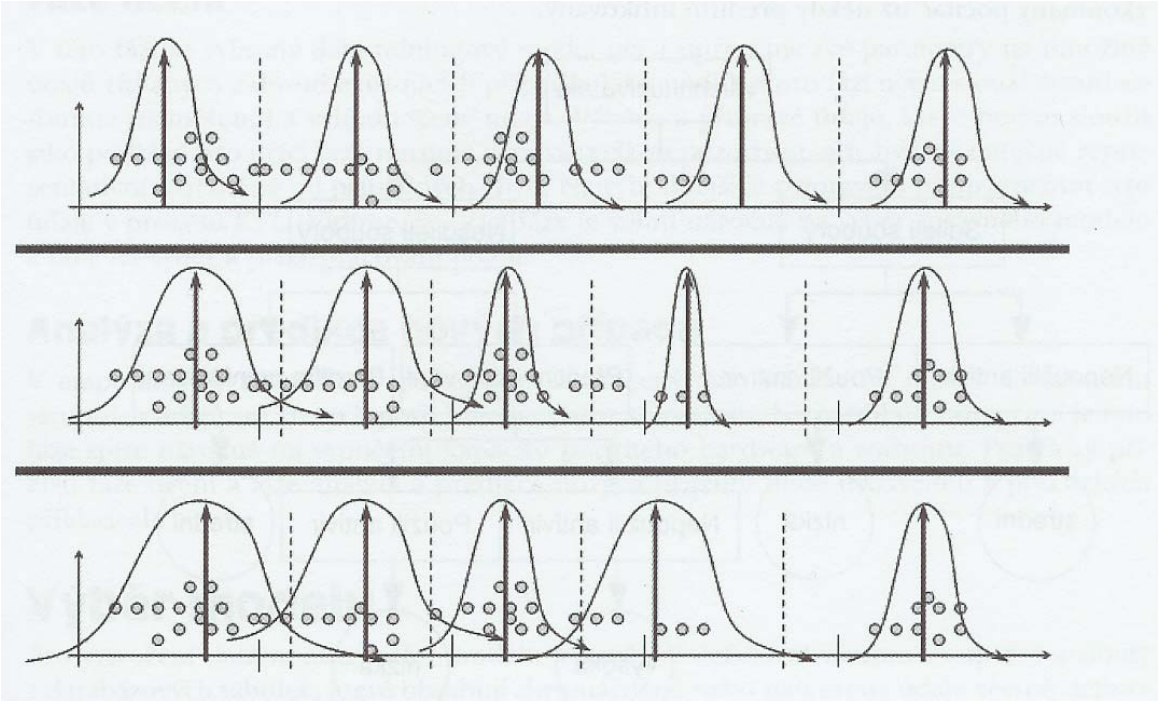
Pod tímto názvem se skrývá metoda *vícerozměrných shlukovacích diagramů*, která odhaluje shluky reálných dat v multidimenzionálních prostorech a je založena na předpokládajícím a maximalizačním (Expectation and Maximization — EM) algoritmu. Algoritmus iteruje ve dvou krocích:

1. předpovídající krok (E-step)
2. maximalizační krok (M-step)

V prvním kroku jsou vypočítány příslušnosti objektů k jednotlivým shlukům, v druhém jsou parametry modelů přehodnoceny s ohledem na členy shluku. EM je podobný K—středovým algoritmům, které mají tyto hlavní kroky:

1. Ustanovení počátečních středů shluku.
2. Přiřazení objektů k jednotlivým středům pomocí některé metriky.
3. Přepočet nových středů shluků podle přiřazených objektů shlukům
4. Vyhrazení hranic shluků na základě nových středů
5. Opakuje dokud algoritmus konverguje (mění se shluky)

EM se v několika aspektech od K- středových algoritmů odlišuje. Hlavním rozdílem je, že EM nemá žádné striktní hranice kolem shluků. Objekt je ke každému shluku připojen s určitou pravděpodobností. Následující obrázek ukazuje průběh tří iterací algoritmu, kdy se jednotlivé shluky mění (jejich hranice).



Obrázek 11.4.. Iterace průběhu shlukování

Microsoft Clustering používá několik proměnných, které se zadávají v Mining Model editoru jako seznam parametrů oddělených čárkou (např. parametr_1 = hodnota_1, parametr_2 = hodnota_2, atd.). Následuje seznam těchto parametrů s popisem.

CLUSTER_COUNT - počet shluků.

CLUSTERING_METHOD - určuje algoritmus pro přiřazování objektů do jednotlivých shluků. Některé algoritmy jsou rychlejší nebo vhodnější pro rozsáhlá data, popř. může existovat i jiné hledisko — např. kvalita. Možné hodnoty jsou: [1 | 2].

HOLDOUT_PERCENTAGE - procento úspěšných cvičných objektů, které je potřeba k výpočtu modelu.

HOLDOUT_SEED - vzorek pro udržení generátoru náhodných čísel.

MINIMUM_CLUSTER_CASES - minimální počet objektů, které mohou vytvořit shluk. Jestliže shluk obsahuje příliš hodně objektů, bude dočasně odložen a může být přerozdělen do jiného uskupení.

MODELLING_CARDINALITY - kontroluje počet iterací algoritmu, před vybráním nejlepší odpovědi ze sady pokusů o shlukování. Protože shlukovací algoritmy obsahují určitou náhodnost, různé pokusy o shlukování mohou přinášet různé výsledky.

SAMPLE_PERCENTAGE - procento cvičných objektů.

SAMPLE_SEED - původní vzorek na nasamplování generátoru náhodných čísel.

STOPPING_TOLERANCE - shlukovací algoritmy opakovaně procházejí data, kdy se s každou iterací přibližují k optimálnímu řešení. Tento parametr určuje, jaký minimální rozdíl musí být mezi jednotlivými iteracemi v souvislosti s konečným řešením. Možné hodnoty jsou v intervalu: <0,1>, kde hodnota bližší 0 způsobuje větší počet iterací.

Microsoft decision Trees

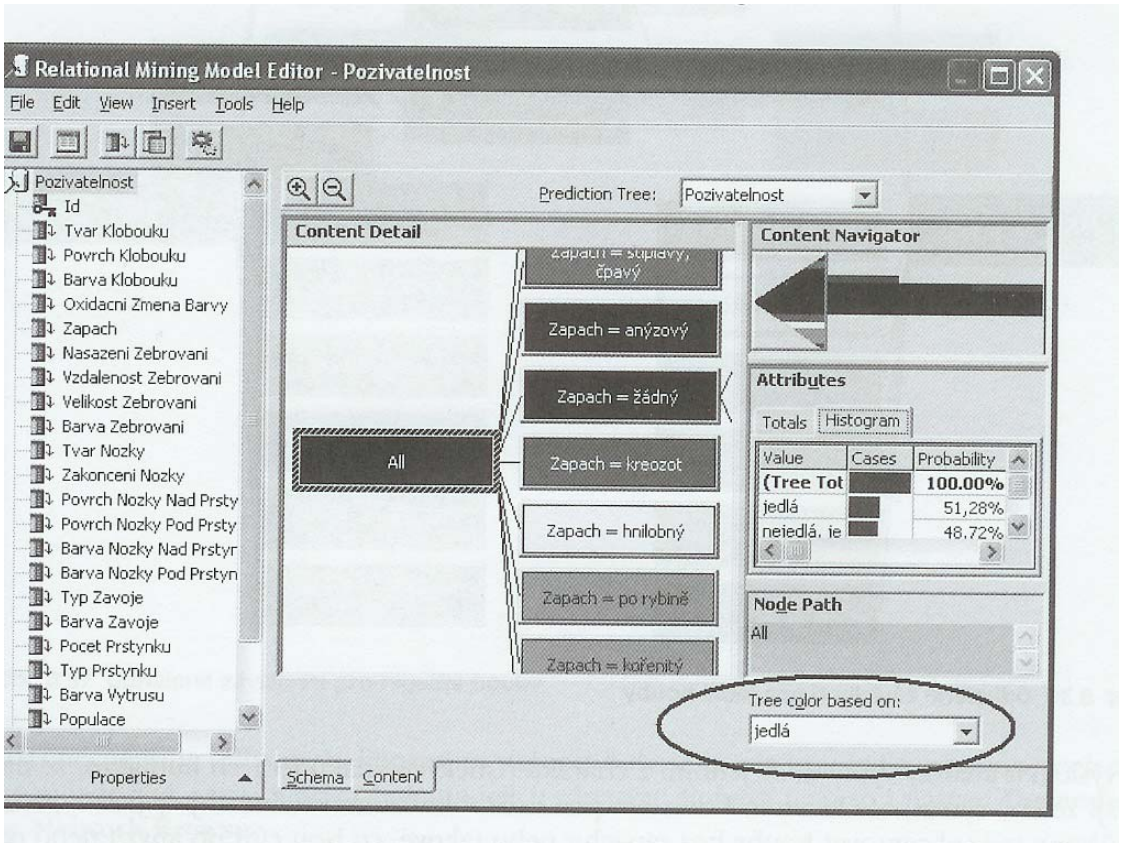
Metoda *nevyváženého rozhodovacího stromu* odhaluje závislosti a vyhledává specifické vlastnosti kategoriálních atributů dokud nejsou ustanoveny čisté korelace. Tyto vlastnosti pak mohou sloužit pro předpovídání (predikci). Rozhodovací stromy jsou prospěšné v případě, kdy chceme vytvářet určité prognózy založené na informaci ve zdrojových datech.

Microsoft Decision Trees je algoritmus je pravděpodobnostní třídění stromu (PCT). Je velmi podobný algoritmu *C4.5*, ale místo entropie používá jako default *Bayesian score*.

Také tento algoritmus má několik rozšiřujících parametrů, které se zadávají v Mining Model editoru stejným způsobem jako u předchozího algoritmu.

COMPLEXITY_PENALTY - reálné číslo z intervalu (0,1), které působí jako omezení rostoucího stromu. Parametr se aplikuje v každém kroku větvení stromu. Hodnota blížící se 0 označuje menší omezení. Použitím parametru omezení ohraničuje hloubku a složitost učících se stromů, což zamezuje přetečení v algoritmu. Nicméně použitím příliž silného omezení může nepříznivě působit na předpovídací schopnost učícího stromu.

HOLDOUT_PERCENTAGE - procento úspěšných cvičných objektů, které je potřeba k výpočtu modelu.



Obrázek 11.5. Příklad výsledku rozhodovacího stromu

HOLDOUT_SEED - vzorek pro udržení generátoru náhodných čísel.

MINIMUM_LEAF_CASES - kontroluje růst stromu za účelem prevence tvoření koncových uzlů stromu, které obsahují méně než MINIMUM listů. Například jestliže se atribut SCORE_METHOD váže na uzel, který obsahuje 30 objektů a hodnota atributu MINIMUM_LEAF_CASES je 10 a jedna potencionální větev uzlu stromu by obsahovala 23 objektů, druhá 7 objektů, pak tento uzel (rozštěpení) nebude povolen.

SAMPLE_PERCENTAGE - procento cvičných objektů.

SAMPLE_SEED - původní vzorek na nasamplování generátoru náhodných čísel.

SCORE_METHOD - určuje algoritmus, použitý k řízení růstu rozhodovacího stromu. Tento algoritmus vybírá atributy tvořící strom, postup použití jednotlivých atributů, způsob, kterým by se měly hodnoty atributů dělit a místo, kde by se měl růst (dělení) stromu zastavit. Možné hodnoty [1 | 2 | 3 | 4].

SPLIT_METHOD - popisuje různé způsoby ovlivňování. Například když atribut má 5 potencionálních hodnot, pak tyto hodnoty mohou být rozčleněny do binárních větví (např. 3 a 1,2,4,5). Tyto hodnoty mohou být také rozděleny do pěti oddělených větví nebo do jiných kombinací. Hodnota 1 říká, že rozhodovací stromy mají jen binární větve, hodnota 2 určuje vícenásobný (nebo nulový) počet větví. Default je hodnota 3, která dovoluje Analytickým službám používaní binárních nebo vícenásobných větví, podle toho jak potřebují.

Možnosti prezentace výsledků analýz

Bez správné a také pro méně zasvěceného člověka pochopitelné prezentace výsledků analýzy, by samotná analýza ztrácela svůj smysl a výhodu pro osoby, pro které je určena. Na základě přehledného

zobrazení těchto výsledků umožňuje manažerům, neorientující se v technologiích datových analýz, přijímat rychlé a kvalitní rozhodnutí a např. plánovat účinnou reklamní a marketingovou strategii.

Jednoduché výpisy

Nejjednodušším příkazem pro Data Mining je příkaz SELECT, kde výsledkem vyhledávání může být jeden nebo více záznamů databáze. Údaje se zobrazují ve formě textového výpisu, nebo formou tabulky.

Grafické zobrazení

V tomto případě se výsledky zobrazí jako dvourozměrný (2D) nebo trojrozměrný (3D) graf. Podle typu použitých zobrazovacích prostředků lze záznamy zobrazit v grafu v podobě sloupců, koláče apod.

Vizualizace údajů

Prakticky každá množina údajů, poskytuje informace k popisu nějakého objektu, jevu, zákonitosti, časové nebo jiné závislosti apod. Na základě poznání takovýchto vlastností lze zdokonalit grafické zobrazení výsledků analýzy. Např. pro zobrazení výsledku sčítání lidu je vhodné použít geografické zobrazení ve formě mapy. Různé získané informace lze pak zobrazit v trojrozměrném obraze mapy, kde oblasti, která představuje plochu kraje, přiřadíme třetí rozměr - výšku. Při správně zvolené axonometrii pohledu je možné získat přehledné a také pro laika pochopitelné zobrazení výsledku analýzy.

Shrnutí pojmů 11.

SW pro Data Mining.

Enterprise Miner. SPSS Clementine. Statistica Data Miner. Oracle 9i Data Mining. Microsoft SQL Server 2000.

Otázky 11.

1. Pokud se rozhodnete vybudovat vlastní datový sklad a v něm mj. používat i metody pro Data Mining, které systémy máte možnost použít a který byste volil? Proč?
2. Co všechno bude hrát roli při rozhodování o použitém systému?

12. ZÁVĚR

Čas ke studiu: 1/2 hodiny

Cíl Přečtením této kapitoly se seznámíte s tím, že

- nejen z numerických dat je možno dolovat
- jiné datové typy se převádějí na numerická data a pak se pro ně používají již známé metody dolování

Výklad

Jiné metody pro data numerická

Zvláště pro některé specifické úlohy se vyskytuje v literatuře řada dalších metod získávání znalostí. Mnoho snahy se věnuje zlepšování algoritmů, především v souvislosti s ohromnými rozsahy analyzovaných dat. Častým řešením u rozsáhlých dat je vzorkování. Místo celých dat se analyzují jejich části = vzorky, vybírané náhodně nebo systematicky v rozsahu, který je možno počítat bez využití vnějších pamětí pro omezení přenosů dat mezi vnitřní pamětí a diskem. Výsledek může být považován za odhad celkového výsledku, pokud dostačuje nižší spolehlivost výsledku. Jindy je po vzorcích zpracován celý soubor, z každého vzorku jsou vybráni kandidáti nebo reprezentanti pro další analýzu a s nimi je znovu proveden výpočet.

Data textová

Užitečné znalosti jsou samozřejmě rozptýleny nejen ve strukturovaných databázích, ale v mnohých dalších zdrojích, především v textech – článcích, knihách, zprávách, textových databázích. Vyhledávat informace a získat z nich zobecněné znalosti je úloha náročnější, než zpracovat strukturovaná data. Jednou z možností získávat znalosti z textů je popsat množinu textových dokumentů vhodnými numerickými parametry a na ně použít metody výše popsané.

Uveďme si jako příklad metodu pomocí strukturovaných dotazů v podobě “topiku” [7]. Princip spočívá v porovnání textu s popisem hledané informace a vypočtení pravděpodobnosti, s jakou je obsah dokumentu k informaci relevantní. Hledaná informace se zadává pomocí klíčových slov, jejich struktury, váhy a operátorů. Struktura ve formě vyhledávacího stromu (topiku), váha je přiřazena klíčovým slovům a operátory jsou nejen klasické logické AND a OR, ale i operátor Accrue (= čím více, tím lépe), řešící klasický rozpor mezi přesností a úplností vyhledávání jen pomocí AND/OR. Vstupem metody jsou analyzované dokumenty a seznam „profilů“ pomocí jednoduchého seznamu klíčových slov nebo strukturovaným topikem. Výsledkem je strukturovaná matice, obsahující v řádcích nalezené relevantní dokumenty, ve sloupcích relevanci dokumentu odpovídající jednotlivým profilům. Nabývají hodnot 0 – 100. Na tyto matici je možno použít shlukování, hledání asociací, klasifikace apod.

❑ Data grafická, zvuková, nestrukturovaná

Obdobně i pro další data multimediální jsou vyvíjeny metody, jak je převést na datovou numerickou matici s pevným počtem atributů. I zde jde o jistý druh komprese informace, účelem metod je snaha o minimalizaci její ztráty při strukturovaném popisu informace.

Jako příklad uveďme typ dat, který se podle významu informace nazývá sekvencí, signálem, časovou, dimenzionální nebo jednorozměrnou řadou apod. Sekvence přísluší zkoumané proměnné jako jeden z jejích nestrukturovaných atributů. Jde o posloupnost dvojic {t,X}, kde t je nezávislou dimenzí a X={x1, x2, …, xn} je množinou vlastností příslušné entity v konkrétním bodě nezávislé dimenze. Jako příklad uveďme naměřenou sekvenci EEG pacienta. Chceme-li taková data analyzovat, opět jednou z možností je charakterizovat sekvenci pevným vektorem charakteristik (min, max, průměr, šikmost, …), přičemž zřejmě množina takových charakteristik bude pro každý typ sekvence různá. Některé typy sekvencí (např. právě EEG, EKG, …) je možno považovat za "periodické". Pak lze rozdělit celou sekvenci na podsekvence (též grafoelementy), charakterizovat každou z nich samostatnou n-ticí atributů a analyzovat je jako množinu entit: shlukováním vyhledávat odchylky apod.

❑ Výhled do blízké budoucnosti

Mimo klasické databázové zdroje – cíleně sebraná data nebo relační, transakční a objektově-relační databáze - se postupně stále častěji objevují snahy o rozšíření této základny. Již jsme zmínili data databázová multimediální všech typů. V blízké budoucnosti se bude zřejmě stále častěji obracet pozornost na využití ohromného potenciálu informací na Internetu ve všech jeho podobách.

Shrnutí pojmů 12.

Dolování v datech nenumerických. Převod dat nenumerických na numerickou reprezentaci.

Data textová, grafická, zvuková.

LITERATURA

1. Applix, TM1 White Paper - http://www.applix.com/tm1/rescentr/tm1wppr.htm
2. Codd E.F., A Relational Model for Data for Large Shared Data Banks, CACM 13, No 6, June 1970
3. Codd E.F. Codd S.B. & Salley C.T., Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate, E.F.Codd & Associates 1993
4. Computer (magazín), Computer Press, 1/1999
5. Computer World, seriál Databázová abeceda, http://www.cw.cz/db
6. Data Warehousing Knowledge Center, http://www.datawarehousing.org
7. DB2 magazine, ročníky 1997, 1998
8. Database Programming & Design (magazín), ročník 1996DM Review magazine, http://www.dmreview.com/issues/archive.htm
9. Hájek, P. – Havránek, T. – Chytil, M.: *Metoda GUHA*. Academia Praha, 1982
10. Humphries, M. a kol: Data Warehousing. Computer Press, Praha, 2002
11. Inteligent Enterprise, The magazine, http://www.intelligententerprise.com
12. Lacko, L.: Datové sklady, analýza OLAP a dolování dat. Computer Press, Brno, 2003
13. Lukasová, A. – Šarmanová, J.: *Metody shlukové analýzy*. SNTL Praha, 1985.
14. OLAP Page - http://www.olap.cz
15. Rud, O.P.: Data Mining. Computer Press, Praha, 2001
16. Sborník přednášek konference DATASEM '96 – DATASEM '2000
17. Sborník přednášek konference DATAKON '2001 - '2003
18. Systémová integrace - ČSSI, Ročník 4, Číslo 3, Září 1997 - Datové sklady
19. Statsoft: Data Mining. Sborník k seminářům, Praha, 2002