Využíváte efektivně potenciál skrytý v datech?

Praktický příklad využití data miningu

Petra Beranová

V databázích, zejména velkých společností, bývá uloženo množství informací, které v případě efektivního zpracování mohou přinést nemalé úspory a zvýšit šance uspět v silném konkurenčním boji. A právě proces data miningu (dolování dat) umožňuje nalézt souvislosti v datech, které nejsou přímo zřejmé a které napomáhají lépe porozumět firemním procesům a obstát v konkurenčním prostředí.

Využívání data miningu v posledních letech výrazně roste a jde napříč obory. Typickými uživateli těchto nástrojů jsou společnosti z bankovního sektoru, pojišťovny nebo například telekomunikace. Data mining se čím dál více využívá také v různých obchodních společnostech, službách či státní správě. Velké datové sklady přímo vybízí k lepšímu využívání cenných informací.

V tomto článku se zaměříme na datamaningové úlohy týkající se hledání asociací a sekvencí v datech. Tyto metody mohou výrazně pomoci například při plánování marketingových strategií, tvorbě produktových balíčků, při péči o zákazníky, detekci podvodů atd.

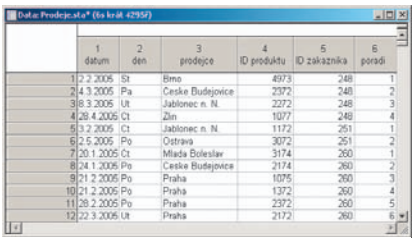
Zdroj dat

Data se týkají prodeje náhradních dílů pro „bílé zboží“ jedné nadnárodní společnosti. Sledována byla skupina zákazníků, kteří nakoupili zboží u konkrétních prodejců této společnosti v určitém období. Celkově se jedná o 898 zákazníků, kteří nakupovali u osmi prodejců v průběhu šesti měsíců (viz tab. 1). Datový zdroj obsahuje datum a den v týdnu, kdy se prodej uskutečnil, a dále místo sídla prodejce, kde bylo zboží objednáno. Každá komponenta má svůj

identifikátor (ID produktu). Jednotliví zákazníci jsou vedeni pod identifikačním číslem (ID zakaznika) a je patrné, v jakém pořadí uskutečňovali jednotlivé nákupy.

Cíl analýzy

Cílem je popsat typické chování zákazníků, tj. nalézt charakteristické skupiny náhradních dílů „bílého zboží“, které zákazníci



	1 datum	2 den	3 prodejce	4 ID produktu	5 ID zakaznika	6 poradi
1	2.2.2005	St	Brno	4973	248	1
2	4.3.2005	Pa	Ceske Budejovice	2372	248	2
3	8.3.2005	Ut	Jablonec n. N.	2272	248	3
4	28.4.2005	Ct	Zlin	1077	248	4
5	3.2.2005	Ct	Jablonec n. N.	1172	251	1
6	2.5.2005	Po	Ostrava	3072	251	2
7	20.1.2005	Ct	Mlada Boleslav	3174	260	1
8	24.1.2005	Po	Ceske Budejovice	2174	260	2
9	21.2.2005	Po	Praha	1075	260	3
10	21.2.2005	Po	Praha	1372	260	4
11	28.2.2005	Po	Praha	2372	260	5
12	22.3.2005	Ut	Praha	2172	260	6

Tab. 1: Vstupní data

odebírají. Výsledek analýzy je podkladem pro tvorbu strategie produktového managementu s cílem rozšířit sortiment zboží odebíraného jednotlivými zákazníky a podpořit obrat u prodejců.

Jak z dat získat užitečné informace?

Z hlediska plánovaného projektu je zajímavé sledovat:

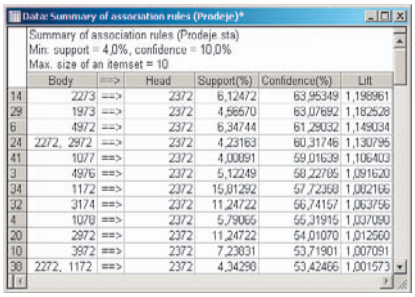
1. Které komponenty jednotliví odběratelé za sledované období nakoupili?
2. V jakém pořadí se nákupy uskutečnily?

K nalezení skrytých vzorů v datech, které popisují chování zákazníků, použijeme analýzu sekvencí a asociací. Výsledkem analýzy budou pravidla tvaru KDYŽ podmínka PAK následek. (V angličtině se používá označení IF body THEN had.) Pravidla jsou určována na základě četností, s jakými se podmínka a následek vyskytují v datech. Díky této analýze dokážeme efektivně odpovědět na otázky typu: Které produkty je dobré zákazníkovi nabízet současně? Když zákazník koupil zboží A, který další produkt je vhodné mu nabídnout?

Vlastní analýza

Na data můžeme pohlížet sekvenčně, či nesekvenčně. Podívejme se nejprve na nesekvenční přístup. V tomto případě se nejedná o klasický „nákupní košík“, kdy zákazník nakoupí více druhů zboží najednou. Vzhledem k plánovanému záměru managementu budeme za jeden „nákupní koš“ považovat nákupy jednoho zákazníka za půl roku. Výsledkem analýzy mohou být asociační pravidla uvedená v tabulce 2. Z výsledků je například patrné, že zákazník si s 16% pravděpodobností současně koupí zboží s ID 1172 a 2372. Pokud si zákazník koupí zboží s ID 2272 a 2972, je 60% pravděpodobnost, že koupí také produkt s ID 2372.

Prohlédněme si některá asociační pravidla v grafické podobě na obrázku 1. Vidíme například, že se často společně kupují komponenty s ID 2372 a 1172, a komponenty

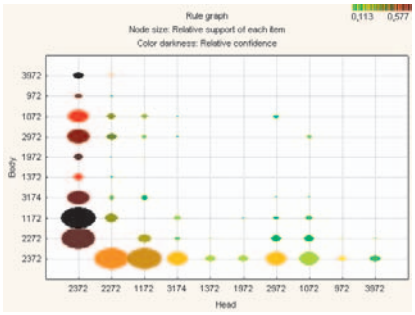


Summary of association rules (Prodeje.sta)
Min: support = 4,0%, confidence = 10,0%
Max. size of an itemset = 10

	Body	==>	Head	Support(%)	Confidence(%)	Lift
14	2273	==>	2372	6,12472	63,95349	1,196861
29	1973	==>	2372	4,56570	63,07692	1,182528
6	4972	==>	2372	6,34744	61,28032	1,149034
24	2272, 2972	==>	2372	4,23163	60,31746	1,130795
41	1077	==>	2372	4,00891	59,01639	1,106403
3	4976	==>	2372	5,12249	58,22785	1,091620
34	1172	==>	2372	15,81292	57,72358	1,082166
32	3174	==>	2372	11,24722	56,74157	1,063756
4	1078	==>	2372	5,79066	56,31915	1,037090
20	2972	==>	2372	11,24722	54,01070	1,012660
10	3972	==>	2372	7,23031	53,71901	1,007091
38	2272, 1172	==>	2372	4,34298	53,42466	1,001573

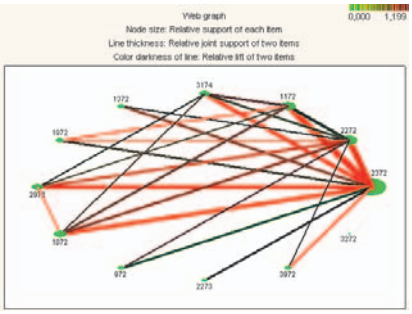
Tab. 2: Asociační pravidla

s ID 2372 a 2272. Pokud zákazník koupí komponentu s ID 1172, je velká pravděpodobnost, že pořídí také produkt s ID 2372 (nákup v obráceném sledu je méně pravděpodobný).



Obr. 1: Grafické znázornění vybraných asociačních pravidel – rule graph

Také další graf přehledně popisuje některé vybrané asociace – viz obrázek 2. „Web graph“ navíc ukazuje, že nejžádanější je

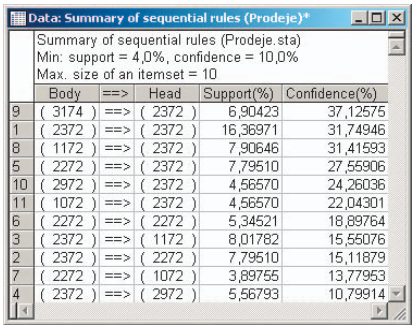


Obr. 2: Grafické znázornění vybraných asociačních pravidel – web graph

náhradní díl s ID 2372 nebo například že se častěji kupují náhradní díly s ID 2372 a 2273 dohromady než zvlášť (lift = 1,2 viz tabulka 2).

Nyní se podívejme na data sekvenčně. Zohledníme tedy pořadí, v jakém jednotliví zákazníci zboží v rámci sledovaného období nakupovali. Výsledkem mohou být sekvenční pravidla uvedená v tabulce 3. Již víme, že komponenta s ID 2372 je nejžádanějším náhradním dílem. Ze sekvenční analýzy navíc vyplývá, že pokud již zákazník komponentu s ID 2372 koupil, objedná další stejný díl s 32% pravděpodobností.

Vybraná sekvenční pravidla jsou graficky znázorněna na následujícím obrázku 3.

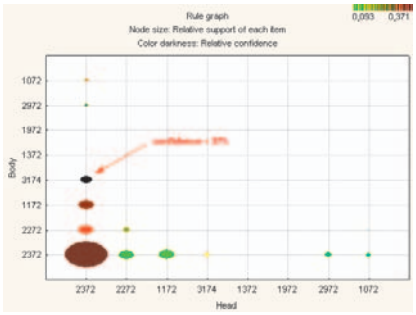


Summary of sequential rules (Prodeje.sta)
Min: support = 4,0%, confidence = 10,0%
Max. size of an itemset = 10

	Body	==>	Head	Support(%)	Confidence(%)
9	(3174)	==>	(2372)	6,90423	37,12575
1	(2372)	==>	(2372)	16,36971	31,74946
8	(1172)	==>	(2372)	7,90646	31,41593
5	(2272)	==>	(2372)	7,79510	27,55906
10	(2972)	==>	(2372)	4,56570	24,26036
11	(1072)	==>	(2372)	4,56570	22,04301
6	(2272)	==>	(2272)	5,34521	18,89764
3	(2372)	==>	(1172)	8,01782	15,55076
2	(2372)	==>	(2272)	7,79510	15,11879
7	(2272)	==>	(1072)	3,89755	13,77953
4	(2372)	==>	(2972)	5,56793	10,79914

Tab. 3: Sekvenční pravidla

Můžeme si například všimnout poměrně vysoké pravděpodobnosti (37%), že zákazník, který koupil komponentu s ID 3174, koupí následně také náhradní díl s ID 2372.



Obr. 3: Grafické znázornění vybraných sekvenčních pravidel – rule graph

Závěr

Uvedené postupy ukazují cestu, jak odhalit zajímavé vzorce chování zákazníků. Nyní dokážeme popsat, které komponenty jednotliví odběratelé za sledované období nakoupili a v jakém pořadí se nákupy uskutečnily. Abychom tato pravidla mohli použít jako podklad pro tvorbu strategie produktového managementu, je samozřejmě nutné dobře znát věcnou povahu dat. Nalezení významných rysů v chování stávajících zákazníků pak může výrazně přispět například k tvorbě dobrých predikcí chování nových zákazníků nebo ke zvýšení efektivity marketingových kampaní. ■

Autorka působí jako senior consultant ve společnosti StatSoft CR.