Možnosti extrakcie asociačných pravidiel z údajov pomocou SAS Enterprise Miner

Iveta Stankovičová
Univerzita Komenského v Bratislave, Fakulta managementu, Katedra informačných systémov
iveta.stankovicova@fm.uniba.sk

Abstract

Customer relationship management (CRM) is a process when company exploits information about customers. Company that wants to gather information about its and competitive customers needs to analyze data sources using various quantitative and data mining methods. For example - market basket analysis. Article explains how to exploit results of market basket analysis, specifically association rules and sequence analysis, in financial companies.

Abstrakt

Riadenie vzťahov so zákazníkmi (CRM) predstavuje proces, v ktorom spoločnosť maximálne využíva informácie o zákazníkoch. Aby firma získala informácie o svojich a konkurenčných zákazníkoch, potrebuje analyzovať svoje a cudzie dátové zdroje pomocou rôznych kvantitatívnych metód a metód data miningu. Medzi takéto metódy patrí aj analýza nákupného koša (market basket analysis). Cieľom článku je ukázať využitie výsledkov analýzy nákupného koša, konkrétne asociačných pravidiel a sekvenčnej analýzy, vo finančných organizáciách.

Keywords

undirected data mining, market basket analysis, associations rules, transaction data, SAS Enterprise Miner

Kľúčové slová

nepriamy data mining, analýza nákupného koša, asociačné pravidlá, transakčné dáta, SAS Enterprise Miner

Úvod

Analýza nákupného koša nie je jednoduchá technika. Názov tejto skupiny metód je odvodený od situácie, keď zákazník chodí s nákupným vozíkom po supermarkete a dáva si spolu do košíka potrebné produkty (tovary).

Pojem „analýza nákupného koša“ zahŕňa skupinu techník a metód analýzy transakčných dát o predaji produktov v sledovanej oblasti (point-of-sale transaction data). Najčastejšie používanou technikou v praxi sú asociačné pravidlá. Automatické generovanie asociačných pravidiel z transakčných dát je možné robiť pomocou data miningových softvérov (napr. SAS Enterprise Miner, SPSS Clementine, STATISTICA Data Miner a pod.).

Asociačné pravidlá reprezentujú schémy (vzory, pravidlá) v dátach bez toho, aby bola definovaná cieľová (modelovaná, závislá, target) premenná, preto sa táto technika zaraďuje medzi techniky nepriameho data miningu (undirected data mining). Medzi ďalšie techniky analýzy nákupného koša patrí sekvenčná analýza a taxonomická analýza nákupného koša. Sekvenčná analýza analyzuje postupnosť (poradie) nakupovaných produktov. Taxonomická analýza nákupného koša vyžaduje údaje o hierarchickej štruktúre položiek (tzv. taxonómiu), čiže údaje o kategóriách a subkategóriách nakupovaných produktov. Zisťuje, ktoré kategórie, resp. subkategórie produktov si zákazník kupuje spolu a v akej postupnosti.

Cieľom predkladaného príspevku je vysvetliť princípy asociačnej a sekvenčnej analýzy a aplikovať tieto techniky na transakčné dáta z oblasti poisťovníctva a bankovníctva pomocou softvéru SAS Enterprise Miner 5.3.

1. Údaje pre analýzu nákupného koša

Údaje pre techniky analýzy nákupného koša majú povahu transakčných dát. Je pre ne typické, že sa skladajú zo štyroch základných častí, resp. dátových tabuliek:

- Dátová tabuľka o zákazníkoch obsahuje ID zákazníka (*customer*), meno, adresu a pod..
- Dátová tabuľka o zákazkách (objednávkach, nákupoch) obsahuje ID zákazky (*order*), ID zákazníka, dátum objednávky (nákupu), typ platby, celkovú cenu, dátum zaslania, náklady na dopravu a pod..
- Dátová tabuľka o položkách zákazky (účtovných položkách, *line items*) obsahuje ID účtovnej položky, ID zákazky, ID produktu (tovaru, služby), množstvo, jednotkovú cenu, jednotkové náklady a pod..
- Dátová tabuľka o produktoch môže obsahovať ID produktu, produktovú kategóriu, produktovú subkategóriu, popis produktu a pod..

Dátové tabuľky predstavujú relačnú databázovú štruktúru, v ktorej základným prvkom je zákazka (*order*). Zákazka predstavuje transakciu a skladá sa z jednotlivých položiek (*items*). Položky sú zvyčajne napojené na tabuľku charakteristík produktov. Zákazka sa viaže na určitého zákazníka (ID zákazníka), ktorý môže mať aj niekoľko zákaziek a preto sa ID zákazníka v transakčných dátach opakuje na viacerých riadkoch dátovej tabuľky.

2. Asociačné pravidlá

Asociačné pravidlá sú pôvodne odvodené z nákupných dát a vyjadrujú, ktoré tovary si zákazník kúpil, resp. kupuje spolu v obchode, čiže dáva si ich zvyčajne spolu do nákupného košíka (*basket*). Dnes sa však táto technika aplikuje aj v iných oblastiach, napr. analýza položiek kúpených cez kreditnú kartu, analýza zakúpených voliteľných služieb zákazníkmi telekomunikačnej spoločnosti, analýza bankových služieb používaných retailovými zákazníkmi a podobne.

Výsledkom asociačnej analýzy sú asociačné pravidlá (*association rules*), ktoré hovoria o tendencii zoskupovania sa nakupovaných produktov alebo služieb a sú dobre zrozumiteľné. Napríklad pravidlo „*ak si zákazník kúpil pomarančový džús, tak si kúpi aj sódu*“, je každému hneď jasné. Od asociačných pravidiel požadujeme aby boli nielen jasné ale aj užitočné. Asociačná analýza produkuje tri základné druhy pravidiel:

1. použiteľné (osožné, prospešné),
2. triviálne a
3. nevysvetliteľné.

Použiteľné asociačné pravidlá obsahujú informáciu vysokej kvality. Napríklad pravidlo „*ak zákazník kúpi bábiku Barbie, tak si kúpi aj čokoládu*“ vedie obchodníka k tomu, aby tieto tovary umiestnil blízko seba v obchode a zvýši tým svoje tržby.

Triviálne asociačné pravidlá obsahujú výsledok, ktorý je všeobecne známy pre danú oblasť podnikania. Napríklad pravidlo „*ak zákazník kúpi počítač, tak si kúpi aj monitor*“ patrí medzi triviálne zistenia a zbytočne sme mrhali prostriedky na výskum.

Nevysvetliteľné asociačné pravidlá sa nedajú objasniť a nevedú k nejakému prospešnému činu. Napríklad pravidlo „*keď sa otvorilo nové oddelenie hardvéru, tak sa stal jedným z najpredávanejších tovarov WC čistič*“ patrí medzi takéto výsledky. Vysvetlenie môže byť rôzne, napr. že zľava na tento WC čistič v sledovanom období bola taká vysoká, že zákazníci kupovali tento produkt vo zvýšenej miere.

Podľa originálnej definície Agrawala [R. Agrawal a kol., 1993] pre definovanie asociačných pravidiel je potrebné zaviesť nasledujúce pojmy. Nech $I=\{i_1, i_2, \ldots, i_n\}$ je súbor binárnych atribútov o rozsahu n, ktoré nazývame položkami (*items*) a nech $D=\{t_1, t_2, \ldots, t_m\}$ je súborom transakcií o rozsahu m, nazývaným databáza. Každá transakcia v D má pridelené jedinečné identifikačné číslo ID a obsahuje podmnožinu položiek z I. Príklad databázy D je uvedený v tabuľke 1.

ID transakcie (zákazníka)	Názov produktu (položka)			
	A	B	C	D
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Tabuľka 1 Príklad databázy D pre 4 položky

Asociačné pravidlo je definované ako implikácia $A \Rightarrow B$, kde A, B sú z I a sú to disjunktné položky (prvky), čiže $A \cap B = \emptyset$. Každé asociačné pravidlo sa teda skladá z dvoch častí, z podmienky A (*condition*) a výsledku B (*result*) a zvyčajne predstavuje výrok:

Ak podmienka, tak výsledok (ak A potom B, resp. A==>B).

Asociačné pravidlá môžu byť aj zložitejšie, čiže môžu obsahovať aj viac prvkov ako dva (prvky A, B). Napríklad medzi trojprvkové pravidlá (prvky A, B, C) patrí výrok:

Ak A a B, tak potom C (resp. A&B==>C).

Kvalita asociačných pravidiel sa posudzuje troma ukazovateľmi: *support*, *confidence* a *lift*. Vzorce a charakteristiky týchto ukazovateľov pre dvojprvkové asociačné pravidlá typu $A \Rightarrow B$ sú nasledujúce:

- *Support* udáva pravdepodobnosť, že dve položky (A a B) sa vyskytnú súčasne u zákazníka. Je mierou dôležitosti (*significance, importance*) pravidla A==>B. Je to symetrický ukazovateľ.

$$support = \frac{\text{počet transakcií s obsahom položiek } A \text{ a } B}{\text{počet všetkých transakcií v databáze } |D|}$$

- *Confidence (strenght)* udáva podmienenú pravdepodobnosť, že ak má zákazník položku A, tak má aj položku B. Nie je to už symetrický ukazovateľ a je veľmi citlivý na výskyt položky B (výsledku pravidla) v databáze.

$$confidence = \frac{\text{transakcie s obsahom položiek } A \text{ a } B}{\text{transakcie s obsahom položiek } A} = \frac{P(A \cap B)}{P(A)} = P(B/A)$$

- *Lift (improvement, interest)* vypočítame ako podiel dvoch pravdepodobností. Očakávaná pravdepodobnosť (*expected confidence*) vyjadruje teoretickú pravdepodobnosť za predpokladu nezávislosti položiek A a B, čiže pre pravidlo A==>B vyjadruje pravdepodobnosť, že zákazník má produkt B. Ukazovateľ liftu je tiež symetrický.

$$lift = \frac{confidence}{expected\ confidence} = \frac{P(B/A)}{P(B)} = \frac{P(A \cap B)/P(A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)} = \frac{P(\text{podmienka a výsledok})}{P(\text{podmienka})P(\text{výsledok})}$$

Lift interpretujeme ako všeobecnú mieru závislosti (asociácie) medzi dvoma prvkami. Hodnoty vyššie ako 1 indikujú pozitívnu závislosť, hodnoty nižšie ako 1 indikujú negatívnu závislosť a hodnoty rovné 0 indikujú nezávislosť. Napr. ak je lift=2 pre pravidlo A==>B, tak zákazník, ktorý má produkt A, má produkt B s dvojnásobnou pravdepodobnosťou v porovnaní s náhodne vybratým zákazníkom.

Pre pravidlo s liftom nižším ako 1, je lepšie využiteľná negácia tohto pravidla. Napr. pravidlo „ak B a C, tak A" má lift len 0,74 a confidence 0,33. Položka A sa objavuje u zákazníkov napr. s pravdepodobnosťou 0,45 a teda jej negácia (non A) má pravdepodobnosť 0,55. Negácia tohto pravidla má tvar „ak B a C, tak non A" a má confidence 0,67 (1-0,33). Lift tohto nového pravidla je už 1,22 (0,67/0,55), čo je v praxi lepšie použiteľné.

Generovanie asociačných pravidiel je viackrokový proces. Algoritmus môžeme všeobecne popísať takto:

1. Generovanie matice výskytu (co-occurrence matrix) pre jednotlivé prvky.
2. Generovanie matice výskytu (co-occurrence matrix) pre dva prvky. Jej použitie pre hľadanie dvojprvkových asociačných pravidiel.
3. Generovanie matice výskytu (co-occurrence matrix) pre tri prvky. Jej použitie pre hľadanie trojprvkových asociačných pravidiel.
4. Atď.

Výsledkom sú asociačné pravidlá vytvorené kombináciami prvkov (produktov). Už napríklad pre 5 prvkov je to veľký počet pravidiel, lebo ich počet rastie exponenciálnym radom. Je potrebné tento proces niekde zastaviť a riešením je tzv. orezanie (*pruning*). Na orezanie sa najčastejšie používa nastavenie minimálnej hodnoty *supportu*.

3. Využitie metód analýzy nákupného koša

Na analýzu použijeme transakčné dáta o nákupe štyroch poistných produktov za určité sledované obdobie zákazníkmi nemenovanej poisťovne, teda *I={HP,PZP, PDOM, ZIP}*. Dátový súbor *ASOC_DATA.sas7bdat* obsahuje 4913 riadkov (transakcií) pre tri premenné (*Tabuľka 2*).

Kód produktu (PRODUCT)	Popis	Počet transakcií	Počet zákazníkov s produktom*
HP	Havarijné poistenie	118	113
PZP	Povinné zmluvné poistenie motorového vozidla	658	550
PDOM	Poistenie domácnosti	1882	1425
ZIP	Životné poistenie	2255	1709
Počet transakcií	Celkový počet riadkov v dátovej tabuľke	**4913**	
Počet zákazníkov	Počet rôznych zákazníkov, resp. CUSTOMER_ID	**2835**	

Tabuľka 2 Poistné produkty a ich výskyt v transakčných dátach z poisťovne (ASOC_DATA.sas7bdat)

Zoznam a charakteristika premenných dátového súboru z poisťovne *ASOC_DATA.sas7bdat* je nasledovný:

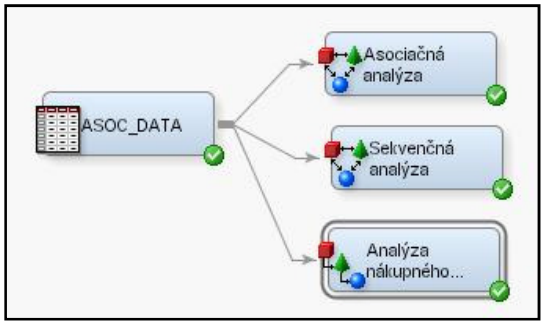
- **CUSTOMER_ID** je identifikačná premenná nominálneho typu, v ktorej je uvedené ID zákazníka. Počet rôznych ID zákazníkov v transakčnej databáze je 2835 (t.j. |D|=2835) a môžeme to zistiť z výstupných súborov softvérového riešenia z uzla *Association*.
- **PRODUCT** je cieľová (*target*) premenná nominálneho typu, ktorá obsahuje kód poistného produktu, ktorý si zákazník kúpil v sledovanom období. Z dát sme zistili, že v sledovanom období si jeden zákazník kúpil maximálne štyri rôzne produkty (HP,PZP, PDOM a ZIP) a ten istý produkt si mohol kúpiť aj viackrát (Porovnaj v tabuľke 2 stĺpce počet transakcií a počet zákazníkov). Hodnoty kódov poistných produktov, ktoré sa v transakciách vyskytovali, ich popis a výskyt je v uvedený tabuľke 2.
- **SEQUENCE** je sekvenčná (poradová) premenná, ktorá vyjadruje postupnosť nakupovania produktov u jednotlivých zákazníkov.

Na uskutočnenie asociačnej a sekvenčnej analýzy sme použili data minigový nástroj SAS Enterprise Miner 5.3 (ďalej len SAS EM). Vytvorili sme pracovný diagram (*Obrázok 1*), ktorý vyjadruje postup analýzy a skladá sa z troch uzlov (*nodes*).

Asociačná analýza

Na vykonanie asociačnej analýzy je v SAS EM určený uzol *Association (*premenovali sme ho na *Asociačná analýza*, viď *Obrázok 1*), ktorý sa nachádza v časti hlavnej ponuky *Explore*. Použili sme len dve vstupné premenné a to CUSTOMER_ID ako identifikačnú (ID) premennú a PRODUCT ako cieľovú (*target)* premennú. Premennú SEQUENCE sme nepoužili.

Využili sme preddefinované nastavenia uzla *Association* systémom SAS EM (*Obrázok 2*). Softvér tak bude vytvárať maximálne štvorprvkové pravidlá (*Obrázok 2* – časť *Association*: *Maximum Items=4*), ktorých *support* bude vyšší ako 5% (Obrázok 2 – časť *Association*: *Support Type=Percent a Support Percentage=5.0*).



Obrázok 1 Postup asociačnej a sekvenčnej analýzy v SAS Enterprise Miner 5.3

Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	1
Support Percentage	5.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	0.0
Support Type	Percent
Support Count	1
Support Percentage	2.0

Obrázok 2 Časť nastavení v uzle Association v SAS Enterprise Miner 5.3

Z asociačnej analýzy s týmito nastaveniami sme získali 10 výsledných asociačných pravidiel (*Tabuľka 3*), ktoré sú usporiadané podľa hodnoty *liftu* a ktorých *support* je vyšší ako 4%.

index	SET_SIZE	EXP_CONF	CONF	SUPPORT	LIFT	COUNT	RULE
1	3	19.400	26.400	4.656	1.361	132	ZIP & PDOM ==> PZP
2	3	17.637	24.000	4.656	1.361	132	PZP ==> ZIP & PDOM
3	3	50.265	52.590	4.656	1.046	132	ZIP & PZP ==> PDOM
4	3	60.282	54.772	4.656	0.909	132	PZP & PDOM ==> ZIP
5	2	50.265	43.818	8.501	0.872	241	PZP ==> PDOM
6	2	19.400	16.912	8.501	0.872	241	PDOM ==> PZP
7	2	19.400	14.687	8.854	0.757	251	ZIP ==> PZP
8	2	60.282	45.636	8.854	0.757	251	PZP ==> ZIP
9	2	60.282	35.088	17.637	0.582	500	PDOM ==> ZIP
10	2	50.265	29.257	17.637	0.582	500	ZIP ==> PDOM

Tabuľka 3 Asociačná analýza - 10 výsledných pravidiel pre dáta z poisťovne

Najvyššiu hodnotu *liftu* 1,36 má 3-prvkové pravidlo ZIP & PDOM ==> PZP. *Support* tohto pravidla 4,66% (132/2835 = 0,04656) znamená, že zákazníkov ktorí si kúpili 3 produkty poisťovne, konkrétne: životné poistenie, poistenie domácnosti a aj povinné zmluvné poistenie motorového vozidla, je v súbore 4,66%. *Confidence* (*CONF*) pre toto pravidlo je 26,4% (132/500 = 0,264) a vyjadruje pravdepodobnosť, že ak si zákazník kúpil poistenie domácnosti (PDOM) a životné poistenie (ZIP), tak si kúpi v tejto poisťovni aj povinné zmluvné poistenie motorového vozidla (PZP). Očakávaná pravdepodobnosť (*EXP_CONF* = *expected confidence*) pre toto pravidlo je 19,4% (550/2835 = 0,194) a vyjadruje pravdepodobnosť výskytu povinného zmluvného poistenia motorového vozidla u zákazníka za predpokladu nezávislosti od zakúpenia produktov poistenie domácnosti a životné poistenie. *Lift* 1,36 (0,264/0,194 = 1,36) znamená, že u zákazníka ktorý si kúpil poistenie domácnosti a životné poistenie, je šanca 1,36 ku 1, že si kúpi aj povinné zmluvné poistenie motorového vozidla v porovnaní s náhodne vybratým zákazníkom.

Analogicky je možné interpretovať aj ostatné asociačné pravidlá v tabuľke 3. Praktický význam majú len asociačné pravidlá s liftom vyšším ako 1. Takto získané informácie sa dajú využiť v marketingových kampaniach poisťovne pri ponuke produktov už len vhodným zákazníkom (tzv. *cross-sell* a *up-sell*).

Sekvenčná analýza

Sekvenčná analýza, na rozdiel od asociačnej analýzy, využíva aj premennú, v ktorej sú zachytené informácie o poradí nakupovania produktov firmy jednotlivými zákazníkmi. Výsledkom sú dvoj až niekoľkoprvkové pravidlá, z ktorých je zrejmá postupnosť nákupu produktov firmy u zákazníkov zoradené podľa hodnoty *supportu*. Hodnoty *supportu* sú zoradené zostupne a udávajú percento, resp. pravdepodobnosť výskytu sekvenčného pravidla v databáze.

Pre vykonanie sekvenčnej analýzy v SAS EM sme tiež využili uzol *Association (*premenovali sme ho na *Sekvenčná analýza,* viď *Obrázok 1*). Použili sme však všetky tri vstupné premenné a to CUSTOMER_ID ako identifikačnú (*ID*) premennú, PRODUCT ako cieľovú (*target)* premennú a premennú SEQUENCE ako sekvenčnú (poradovú) premennú. Ponechali sme preddefinované nastavenia systému SAS EM pre tento uzol (Obrázok 2). Softvér tak bude vytvárať maximálne trojprvkové sekvenčné pravidlá (Obrázok 2 – časť *Sequence*: *Chain Count=3*), ktorých *support* bude vyšší ako 2% (*Obrázok 2* – časť *Sequence* : *Support Type=Percent a Support Percentage=2.0*).

Zo sekvenčnej analýzy v SAS EM s týmito nastaveniami sme získali až 17 výsledných pravidiel (*Tabuľka 4*), ktoré sú usporiadané podľa hodnoty *supportu* a ich *confidence* je vyššia ako 2%. Softvér vytváral maximálne 3-prvkové sekvenčné pravidlá.

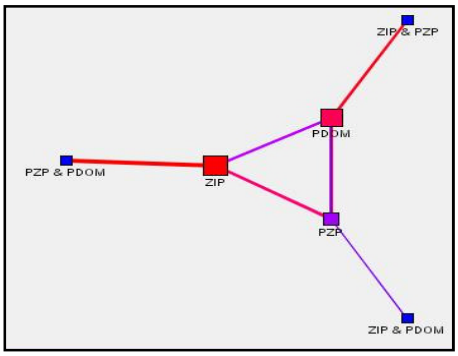
index	NITEMS	COUNT	SUPPORT	CONF	RULE
1	2	408	14.392	23.874	ZIP ==> ZIP
2	2	374	13.192	26.246	PDOM ==> PDOM
3	2	334	11.781	23.439	PDOM ==> ZIP
4	2	302	10.653	17.671	ZIP ==> PDOM
5	2	166	5.855	9.713	ZIP ==> PZP
6	2	151	5.326	10.596	PDOM ==> PZP
7	2	136	4.797	24.727	PZP ==> ZIP
8	2	133	4.691	24.182	PZP ==> PDOM
9	3	105	3.704	25.735	ZIP ==> ZIP ==> ZIP
10	2	97	3.422	17.636	PZP ==> PZP
11	3	92	3.245	30.464	ZIP ==> PDOM ==> ZIP
12	3	89	3.139	23.797	PDOM ==> PDOM ==> ZIP
14	3	83	2.928	27.483	ZIP ==> PDOM ==> PDOM
13	3	83	2.928	24.850	PDOM ==> ZIP ==> PDOM
15	3	81	2.857	24.251	PDOM ==> ZIP ==> ZIP
16	3	70	2.469	17.157	ZIP ==> ZIP ==> PDOM
17	3	64	2.257	17.112	PDOM ==> PDOM ==> PDOM

Tabuľka 4 Sekvenčná analýza - 17 výsledných pravidiel pre dáta z poisťovne

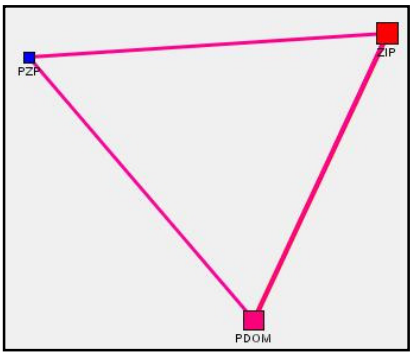
Na základe výsledkov môžeme konštatovať, že zákazníci poisťovne si najčastejšie kupovali dve životné poistenia za sebou (pravidlo ZIP ==> ZIP) a to s pravdepodobnosťou (*support*) 14,4% (408/2835 = 0,143915) a s podmienenou pravdepodobnosťou (*CONF=confidence*) 23,87% (408/1709 = 0,23873). Táto podmienená pravdepodobnosť znamená, že ak si zákazník ako prvý kúpil produkt životné poistenie (ZIP), tak si aj ako druhý kúpi nejaký produkt životného poistenia (ZIP) a to s pravdepodobnosťou 23,87%.

Trojprvkové sekvenčné pravidlá s najvyššou pravdepodobnosťou výskytu v sledovanej poisťovni sú ZIP ==> ZIP ==> ZIP (3,7%) a ZIP ==> PDOM ==> ZIP (3,2%). Trojprvkové pravidlo ZIP ==> PDOM ==> ZIP má súčasne najvyššiu hodnotu *confidence* (až 30,46%, t.j. 92/302 = 0,304635). Toto číslo znamená, že ak má klient už kúpené postupne 2 produkty a to životné poistenie (ZIP) a potom poistenie domácnosti (PDOM), tak si s pravdepodobnosťou 30,46% kúpi ako tretí produkt opäť nejakú životnú poistku. Takto sa chová však len 3,2% klientov poisťovne, na ktorých môžeme aplikovať toto pravidlo. V prípade, že oslovíme v kampani na kúpu životného poistenia len zákazníkov, ktorí už majú v našej poisťovni zakúpenú sekvenciu dvoch produktov a to ZIP ==> PDOM, tak pravdepodobnosť úspešnosti tejto kampane bude 30,46%. Rozhodnutie, či je to veľa alebo málo, závisí od peňažného vyjadrenia nákladov a ziskov z takejto kampane.

Softvérové produkty poskytujú aj širokú paletu grafických prostriedkov pre názorné zobrazenie výsledkov dataminigových metód. Aj SAS Enterprise Miner 5.3 v spomínaných uzloch pre analýzu nákupného koša (uzly *Association* a *Market Basket*) ponúka niekoľko grafov. Medzi veľmi názorné grafické pomôcky patrí tzv. linkový graf (*Link Graph*), ktorý slúži na lepšiu vizualizáciu asociačných alebo sekvenčných pravidiel (*Obrázok 3 a 4*). Veľkosť a farba uzlov v tomto grafe znázorňuje dôležitosť produktu. Čím je uzol väčší a má červenú farbu, tým sa produkt (položka, *item*) častejšie v transakčnej databáze vyskytuje. Čím je spojovacia čiara medzi položkami hrubšia, tým je väzba medzi produktmi dôležitejšia. Z linkového grafu asociačných pravidiel je hneď zrejmé (*Obrázok 3*), že najviac predávaným produktom v sledovanom období bolo životné poistenie (ZIP). Produkt havarijné poistenie (HP) patrí medzi produkty s nízkou početnosťou výskytu v databáze a preto sa na linkovom grafe ani v asociačnej a ani v sekvenčnej analýze nevyskytuje pri použitých nastaveniach v SAS EM. Asociačné pravidlá s produktom HP mali nízky *support* (nižší ako 2%).



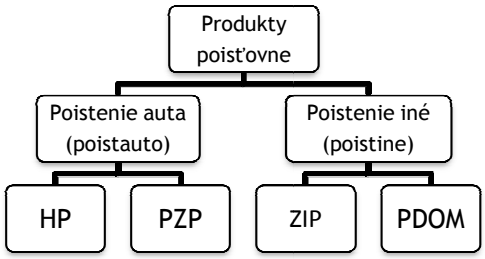
Obrázok 3 Linkový graf asociačných pravidiel pre dáta z poisťovne



Obrázok 4 Linkový graf sekvenčných pravidiel pre dáta z poisťovne

Analýza nákupného koša s hierarchickou taxonómiou produktov

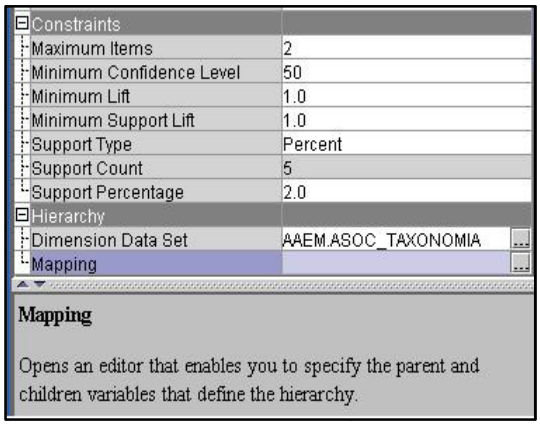
Softvér SAS EM vo svojej poslednej verzii 5.3 obsahuje uzol s názvom *Market Basket*. Tento uzol sa tiež nachádza v ponuke *Explore*, tak ako uzol *Association*. Je to špeciálny uzol na vykonanie taxonomickej asociačnej analýzy, t.j. asociačnej analýzy so zadefinovanou hierarchickou štruktúrou produktov (položiek). Napríklad pre naše údaje z poisťovne by sme štyri produkty zoskupili do dvoch skupín. Prvú skupinu by tvorili len produkty na poistenie auta (označenie: *poistauto*), čiže by obsahovala položky (produkty) HP a PZP. Druhú skupinu by tvorili ostatné poistné produkty (označenie: *poistine*), čiže ZIP a PDOM. Túto veľmi jednoduchú jednostupňovú hierarchickú štruktúru vyjadruje *Obrázok 5*. Je potrebné túto štruktúru zapísať do dátového súboru (*Tabuľka 5*, v stĺpci *Level* len číslo *1*) a použiť v uzle *Market Basket* v časti *Hierarchy (*viď *Obrázok 6)*.



Obrázok 5 Hierarchická štruktúra produktov poisťovne

Product	ParentProd	Level
HP	poistauto	1
PZP	poistauto	1
PDOM	poistine	1
ZIP	poistine	1

Tabuľka 5 Dátový súbor, ktorý mapuje hierarchickú štruktúru produktov poisťovne (ASOC_TAXONOMIA.sas7bdat)



Obrázok 6 Nastavenia uzla Market Basket

ITEM	LEVEL	COUNT	SUPPORT
HP	1	113	3.99
PZP	1	550	19.40
PDOM	1	1425	50.26
ZIP	1	1709	60.28
poistauto	2	586	20.67
poistine	2	2634	92.91

Tabuľka 6 Výstup z uzla Market Basket - Výskyt položiek 1-stupňovej hierarchickej štruktúry produktov poisťovne

RULEID	COUNT	SUPPORT	SUPLIFT	CONF	LIFT	RULE
3	1709	60.282	0.298	64.882	1.076	poistine ==> ZIP
1	1425	50.265	0.082	54.100	1.076	poistine ==> PDOM
2	550	19.400	0.877	93.857	4.838	poistauto ==> PZP
5	77	2.716	-0.274	68.142	3.512	HP ==> PZP
4	74	2.610	0.842	65.487	1.303	HP ==> PDOM

Tabuľka 7 Výstup z uzla Market Basket - Hierarchické asociačné pravidlá pre dáta z poisťovne

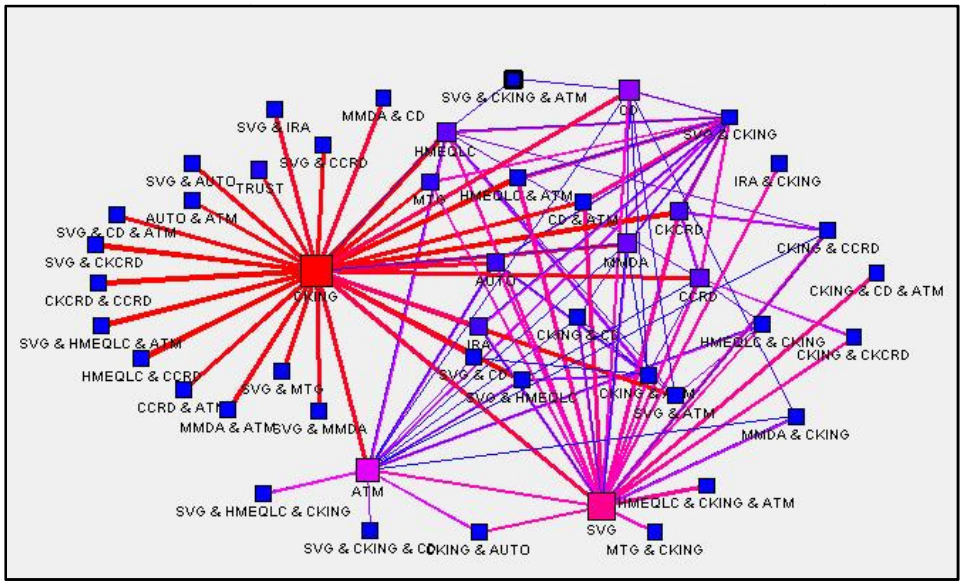
V prípade takejto jednoduchej jednostupňovej štruktúry len pre 4 produkty nie je výsledok veľmi zaujímavý a neprináša nové poznatky oproti asociačnej analýze z uzla Association. Pri nastaveniach v SAS EM, že sa majú zobraziť len pravidlá s podmienenou pravdepodobnosťou (CONF) vyššou ako 50% (viď Obrázok 6), máme vo výstupe (*Tabuľka 7*) len 5 triviálnych pravidiel.

Uvedieme preto zaujímavejšiu aplikáciu z prostredia banky, ktorá ponúka svojím klientom až 13 rôznych produktov. Výstup v podobe linkového grafu z asociačnej analýzy je pomerne neprehľadný (viď *Obrázok 7*). Je zrejmé, že najdôležitejšie produkty banky sú dva účty, konkrétne produkty CKING – bežný účet a SVG – sporiaci účet. Je vhodné navrhnúť hierarchickú taxonomickú štruktúru produktov banky, aby sme mohli lepšie pochopiť vzťahy medzi skupinami položiek. Navrhujeme použiť dvojstupňovú hierarchickú štruktúru pre týchto 13 produktov banky tak, že ich zoskupíme na prvej úrovni do šiestich podskupín a na druhej úrovni vytvoríme tri hlavné skupiny produktov: účty, karty a výpožičky (viď *Obrázok 8*).

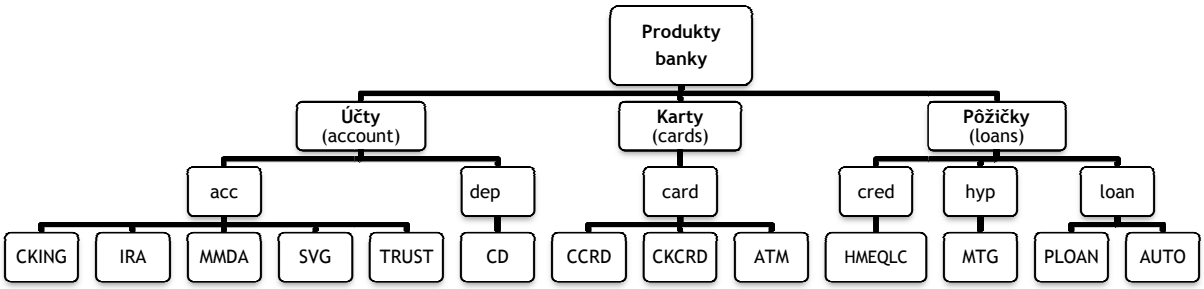
Z výstupnej tabuľky (*Tabuľka 8*) je zrejmé, že najviac produktov banky je z kategórie účtov (7384 transakcií), potom sú to rôzne výpožičky (5724) a nakoniec rôzne druhy kariet (4370). Zo subkategórií (*LEVEL 2*) má najvyšší výskyt podskupina *acc*, čo sú rôzne druhy účtov a z účtov je to konkrétne účet CKING – bežný účet. Tento produkt sa v databáze banky za sledované obdobie vyskytol

s pravdepodobnosťou 85,78% a má lift 1,12 (pozri RULEID 5 v *Tabuľke 9*). Zaujímavé zistenie ďalej je, že v banke je vysoký podiel klientov ktorí majú pôžičku a potom aj bežný účet a platí to aj opačne (pozri *RULEID* 47 46 v *Tabuľke 9*). Asi novú informáciu sa nedozvieme z pravidiel číslo 87 a 86, ktoré majú lift 1 a hovoria o tom, že keď má klient v banke nejakú pôžičku, tak má v banke aj nejaký druh účtu a opačne. To sú triviálne pravidlá.

Najvyšší lift 1,83 má pravidlo card ==> ATM. Znamená to, že keď si klienti kupujú nejakú kartu, tak je to karta ATM (*automated teller machine debit card*). Takých klientov je v banke 38,46%.



Obrázok 7 Výstup asociačnej analýzy pre bankové dáta - linkový graf pre 13 produktov



Obrázok 8 Dvojstupňová hierarchická štruktúra produktov v banke (BANK.sas7bdat)

Záver

Analýza nákupného koša patrí medzi techniky nepriameho data miningu a jej výsledkom sú jasné a pochopiteľné pravidlá o tom, aké produkty si zákazníci najčastejšie kupujú a v akej postupnosti. Ich nevýhodou je hlavne vysoká výpočtová náročnosť pri veľkom objeme transakcií a pri veľkom počte produktov. Je potrebné si stanoviť limit *supportu*, aby sme znížili nároky na softvérové výpočty. Táto metóda má však tiež problémy, keď sa niektoré produkty vyskytujú v databáze zriedkavo. Analýza nákupného koša pracuje dobre, keď sa všetky produkty (položky) vyskytujú približne s rovnakou frekvenciou a to nebol náš prípad, pretože v našich dátach z poisťovne sa dva produkty vyskytovali častejšie (PDOM a ZIP) a dva produkty zriedkavejšie (HP a PZP). Bankové dáta obsahovali viac položiek, až 13 rôznych produktov, pre ktoré je už možné vytvoriť viacstupňovú hierarchickú štruktúru a použiť taxonomickú analýzu nákupného koša.

Techniky analýzy nákupného koša sú vhodné pre organizácie, ktoré ešte nemajú vytvorené vlastné historické databázy o chovaní svojich zákazníkov, čiže na začiatku podnikania. Ak však organizácia už má takúto historickú databázu vybudovanú, tak na modelovanie chovania zákazníkov je vhodná logistická regresia alebo iné metódy tzv. prediktívneho modelovania, napr. rozhodovacie stromy a neurónové siete.

Techniky analýzy nákupného koša je možné použiť na analýzu dát s cieľom lepšie poznať chovanie svojich zákazníkov tak v obchodných reťazcoch, ako aj vo finančných inštitúciách, v telekomunikáciách, ale aj vo web a textminingu.

Literatúra

1. Agrawal, R.; Imielinski, T.; Swami, A.: Mining Association Rules Between Sets of Items in Large Databases. SIGMOD Conference 1993. p. 207-216.
2. Berry, M. J. A. ; Linoff, G.: Data mining Techniques. For Marketing, Sales, and Customer Relationship management. Second Edition. Wiley Publishing, Inc., Indianapolis, Indiana, 2004. ISBN 0-471-47064-3.
3. Munk, M.: Optimalizácia webu na základe sekvenčných pravidiel. Forum Statisticum Slovacum, č 7/2008. SŠDS Bratislava. 2008, s. 80-85. ISS 1336-7420.
4. Personal Home Page of Michael Hahsler: A Comparison of Commonly Used Interest Measures for Association Rules. Dostupné na internete: <http://michael.hahsler.net/research/association_rules/measures.html#support>
5. Řezanková, H.: Analýza dat z dotazníkových šetření SPSS Profesional Publishing. Praha. 2007. ISBN 978-80-86946-49-8.
6. SAS OnLine Doc 9.1.3. Dostupné na internete: <http://support.sas.com/onlinedoc/913/docMainpage.jsp>
7. Wikipedia. Association rule learning. Dostupné na internete: <http://en.wikipedia.org/wiki/Association_rule_learning#cite_ref-lift_10-0>

ITEM	LEVEL	COUNT	SUPPORT
TRUST	1	390	4.88
MTG	1	594	7.43
AUTO	1	742	9.29
IRA	1	866	10.84
CKCRD	1	903	11.30
CCRD	1	1237	15.48
HMEQLC	1	1316	16.47
MMDA	1	1394	17.44
CD	1	1960	24.53
ATM	1	3073	38.46
SVG	1	4944	61.87
CKING	1	6855	85.78
hyp	2	594	7.43
cred	2	1316	16.47
dep	2	1960	24.53
card	2	4370	54.69
loan	2	5214	65.25
acc	2	7159	89.59
cards	3	4370	54.69
loans	3	5724	71.63
account	3	7384	92.40

Tabuľka 8 Výstup z uzla Market Basket - Hierarchická štruktúra produktov banky a ich výskyt v databáze

RULEID	COUNT	SUPPORT	SUPLIFT	CONF	LIFT	RULE
6	7159	89.59	0.94	96.95	1.08	account ==> acc
5	6855	85.78	2.83	95.75	1.12	acc ==> CKING
3	5214	65.25	1.73	91.09	1.40	loans ==> loan
87	5132	64.22	0.01	89.66	1.00	loans ==> acc
86	5132	64.22	0.01	71.69	1.00	acc ==> loans
78	5060	63.32	0.03	73.81	1.03	CKING ==> loans
79	5060	63.32	0.03	88.40	1.03	loans ==> CKING
9	4944	61.87	0.90	94.82	1.53	loan ==> SVG
47	4552	56.96	0.02	66.40	1.02	CKING ==> loan
46	4552	56.96	0.02	87.30	1.02	loan ==> CKING
2	4370	54.69	0.00	100.00	1.83	cards ==> card
73	4329	54.17	0.03	63.15	1.02	CKING ==> SVG
74	4329	54.17	0.03	87.56	1.02	SVG ==> CKING
33	4185	52.37	0.00	95.77	1.04	card ==> account
105	4185	52.37	0.00	95.77	1.04	cards ==> account
34	4185	52.37	0.00	56.68	1.04	account ==> card
104	4185	52.37	0.00	56.68	1.04	account ==> cards
84	4173	52.22	0.03	58.29	1.07	acc ==> cards
25	4173	52.22	0.03	95.49	1.07	card ==> acc
85	4173	52.22	0.03	95.49	1.07	cards ==> acc
26	4173	52.22	0.03	58.29	1.07	acc ==> card
23	4149	51.92	0.04	94.94	1.11	card ==> CKING
77	4149	51.92	0.04	94.94	1.11	cards ==> CKING
76	4149	51.92	0.04	60.53	1.11	CKING ==> cards
24	4149	51.92	0.04	60.53	1.11	CKING ==> card
...						
1	3073	38.46	1.11	70.32	1.83	card ==> ATM

Tabuľka 9 Výstup z uzla Market Basket pre bankové dáta - vybrané asociačné pravidlá