Data, informace, nebo vědomosti?

Data mining a statistické metody a software pro získávání poznatků z databází a podporu rozhodování.

Motto: "Topíme se v datech, ale nemáme informace." *(John Naisbett)*

ing. Karel Kupka, TriloByte Pardubice
kupka@trilobyte.cz; http://www.trilobyte.cz

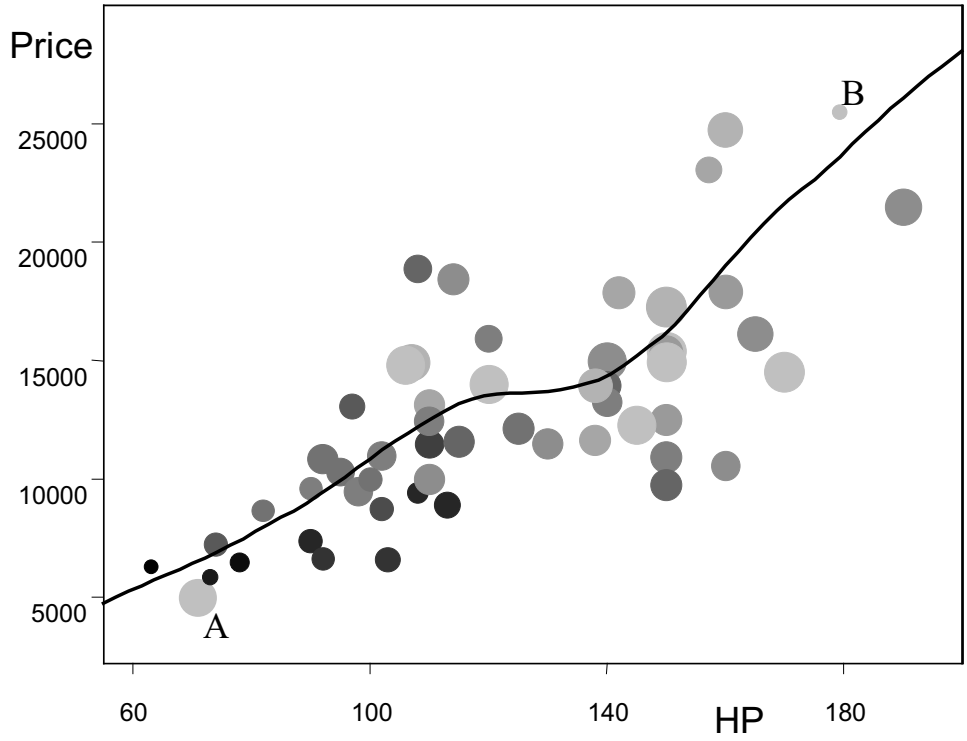
Jsme svědky prudkého rozvoje informačních technologií (LAN, Data Warehousing, Intranet, Internet) a elektronických a optických médií. Exponenciálně (podle některých zdrojů dokonce rychleji) rostoucí množství dat, které má k dispozici průměrný člověk kupodivu nepřináší lepší informovanost. Spíše přispívá k dezorientaci a neschopnosti se objektivně rozhodnout. Toho mimochodem zručně zneužívají obchodní a marketingové sítě různých supermarketů zaplavující spotřebitele "výhodnými" diskonty, slevami, soutěžemi, časovými bombami, věrnostními výhodami, koly štěstí a kdoví čím ještě, aby ho přesvědčily, že nejdražší chleba je nejvýhodnější, a že nejvíce potřebujete to, co vám k ničemu není. Podobně nepřehledná situace začíná být bohužel i ve sféře profesionálních oborů, v sociologii, ekonomice, průmyslu a výzkumu.

Množství dat a informací, které má odpovědnému pracovníku umožnit kvalifikované rozhodnutí, nevede k porozumění situaci, nýbrž k dezorientaci a časovému a informačnímu stresu. Mít mnoho dat totiž ještě neznamená něco vědět. V této situaci se objevují filozofie obecně nazývané Data Mining (DM), která vychází z předpokladu, že ve velkých databázích jsou ukryty poznatky, které lze vyjádřit jednoduchými tvrzeními vyjadřujícími příčinné vztahy, závislosti, klasifikaci. Některé takové poznatky mohou být nečekané a mohou vést k novým odhalením i objevům. Hovoří se proto o "objevování znalostí (vědomostí) v databázích" (anglicky Knowledge Discovery in Databases, KDD). Technické nástroje DM (do češtiny se někdy překládá jako dolovaní v datech) zahrnují především statistické metody (klasifikační stromy, zobecněnou regresi, shlukovou analýzu, robustní metody), dále vícerozměrnou grafickou vizualizaci dat, fuzzy metody, optimalizaci, případně neuronové sítě a samoučící se algoritmy. Tyto metody jsou obvykle integrovány do prostředí schopného vytvářet a modifikovat a pamatovat si algoritmy, spojovat výstup jedné metody se vstupem jiné, případně se (do jisté míry) samostatně učit. Použití DM vedlo již k zajímavým výsledkům v mapování kosmických objektů v Mt. Palomar Observatory, odhalení nečekaným podobností v genech savců, úspěchům v předpovědi zemětřesení, či klasifikaci sopek na Venuši podle radarové (SAR) mapy. Mimořádný zájem o DM je rovněž v oblasti financí a ekonomiky, marketingu, demografie, politologie psychiatrie a medicíny. V následujícím textu uvedeme několik příkladů využití grafických a statistických metod pro získání informace z dat v oblasti ekonomiky a marketingu.

Mezi základní a přirozené nástroje patří určitě vizualizace dat. Nebudeme hovořit o grafu jednoduché závislosti dvou veličin X-Y, na němž je vidět vše. Jak však zobrazit a vyšetřovat vztahy ve vícerozměrných datových souborech, když přímo umíme "nakreslit" nanejvýš 3-rozměrná data (tedy data se 3 sloupci)? Vícerozměrná data si můžeme představit jako tabulku s několika sloupci a mnoha řádky (obvykle bude řádků více nebo mnohem více než sloupců). Příkladem může být databáze firem, u nichž máme zjištěny třeba základní jmění, počet zaměstnanců, aktiva, pasiva, zadluženost, podíl cizího kapitálu, obrat a informaci o pozici na trhu cenných papírů. Dalším příkladem je databáze klientů s informacemi o měsíčním a ročním odběru, trvání obchodních styků, odebíraném sortimentu, způsobu komunikace (pošta, osobně, Internet), stupně spokojenosti, počtu transakcí za rok a geografické vzdálenosti. Cílem analýzy takových databází je samozřejmě zjistit co nejvíce smysluplných informací, nebo vědomostí, které bude možné využít například pro rozhodnutí či tvorbu strategie podnikatelských aktivit. Pokročilé technologie Data Mining se pokoušejí

hledat informace i v daleko složitějších databázích s rozsáhlou dynamickou strukturou. Uvedeme nyní jednoduchý příklad.

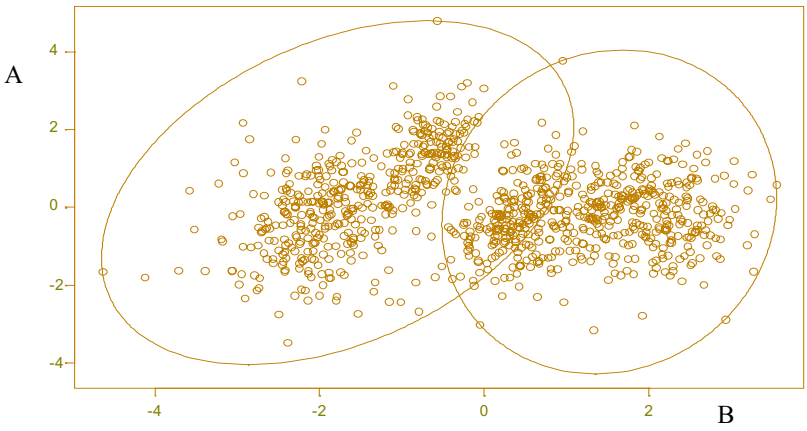
Příklad 1. Máme 4-rozměrná data, totiž údaje o amerických, japonských a západoevropských osobních autech: cenu, výkon (HP), spotřebu a hmotnost 111 automobilů (Obr. 1). Hmotnost je zde vyjádřena velikostí bodu (čím větší, tím těžší), spotřeba barvou (čím tmavší, tím menší). Body A a B se zřetelně liší svým chováním od ostatních: Bod A je v tomto místě nečekaně velký a světlý, bod B je nečekaně malý. Tato informace má soustředit naši pozornost na tyto dva případy, neboť zjištění jakékoliv neobvyklosti či odlišnosti je nová informace, která může mít rozhodující význam a pomohla-li nám nějaká metoda DM/KDD takovou odlišnost odhalit, splnila svůj úkol. Bod A je Moskvič, bod B je Porche.



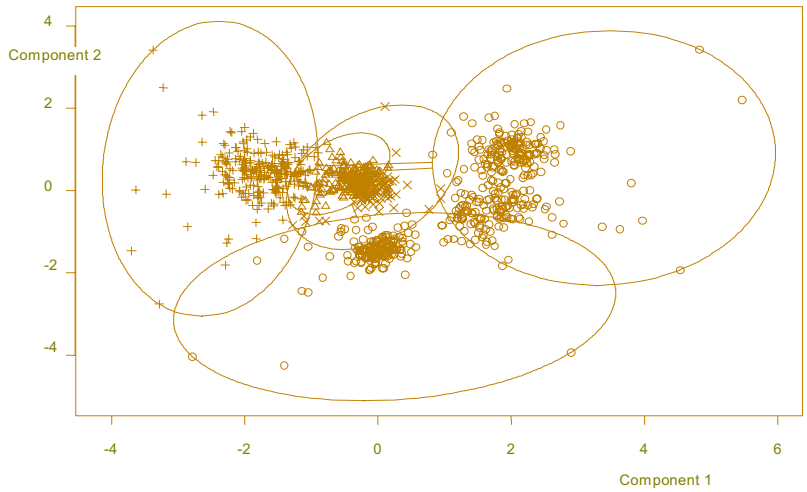
Obr. 1 Graf pro 4-rozměrná data

Mezi nejpoužívanější patří různé metody pro klasifikaci dat, jako jsou techniky shlukové analýzy nebo klasifikační a regresní stromy (často používaná zkratka CART pochází z anglického Classification And Regression Trees). Výhodou metod CART je schopnost předpovídat (predikovat). Použití shlukové analýzy ilustruje následující příklad.

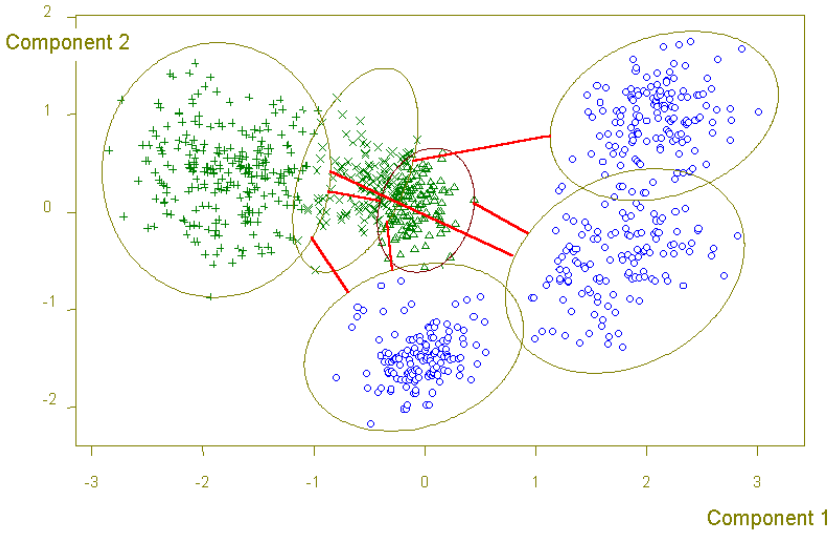
Příklad 2. Jde o ekonomická data pořízená finančním subjektem za účelem klasifikace a posouzení bonity a perspektivy klientských podniků. Jednalo se o 15 číselných ukazatelů (proměnných). Mimo jiných metod byla použita shluková analýza. Na Obr. 2 je nejlepší výsledek dosažitelný výběrem ukazatelů, identifikující dvě odlišné skupiny. Při použití hlavních komponent místo původních ukazatelů (hlavní komponenty jsou takové kombinace původních proměnných, ve kterých jsou jednotlivá data nejlépe rozlišitelná, v nichž mají největší rozptyl) se již objevilo 5 skupin klientů, avšak jejich rozlišení není příliš zřetelné, Obr. 3. Teprve současné použití robustního výpočtu "vzdáleností" a fuzzy shlukové analýzy vedlo ke zřetelnému rozlišení 6 skupin, které jsou patrné v projekci do prvních dvou komponent na Obr. 4. Toto rozlišení výrazně zefektivnilo a usnadnilo strategické zaměření a rozhodování subjektu. Nutnost použít terminologie bez podrobného vysvětlení je snad vyvážena názorností grafů. Podrobnější informace najde čtenář v doporučené literatuře.



Obr. 2 Shluková analýza, původní data



Obr. 3 Shluková analýza, projekce do roviny s maximální variabilitou (hlavní komponenty)



Obr. 4 Shluková analýza, projekce po fuzzy-shlukové analýze s robustními vzdálenostmi

Příklad 3. Další zajímavý příklad je aplikace regresního stromu při analýze zákaznické databáze. Tento postup je formálně podobný analýze shluků, ale na rozdíl od ní umožňuje predikci. Vytvoří binární strom v jehož uzlech jsou jednoduché podmínky. Statistická významnost těchto podmínek klesá shora dolů. Koncové větve obsahují buď odhad střední hodnoty v případě číselné proměnné, nebo nejpravděpodobnější hodnotu s udanou pravděpodobností výskytu v případě nečíselné proměnné (může to být například hodnota anone, barva, název dodavatele, apod.). Tabulka 1 shrnuje údaje o klientech větší obchodní společnosti za posledních 5 let a vypadá takto:

Tabulka 1 Databáze klientů

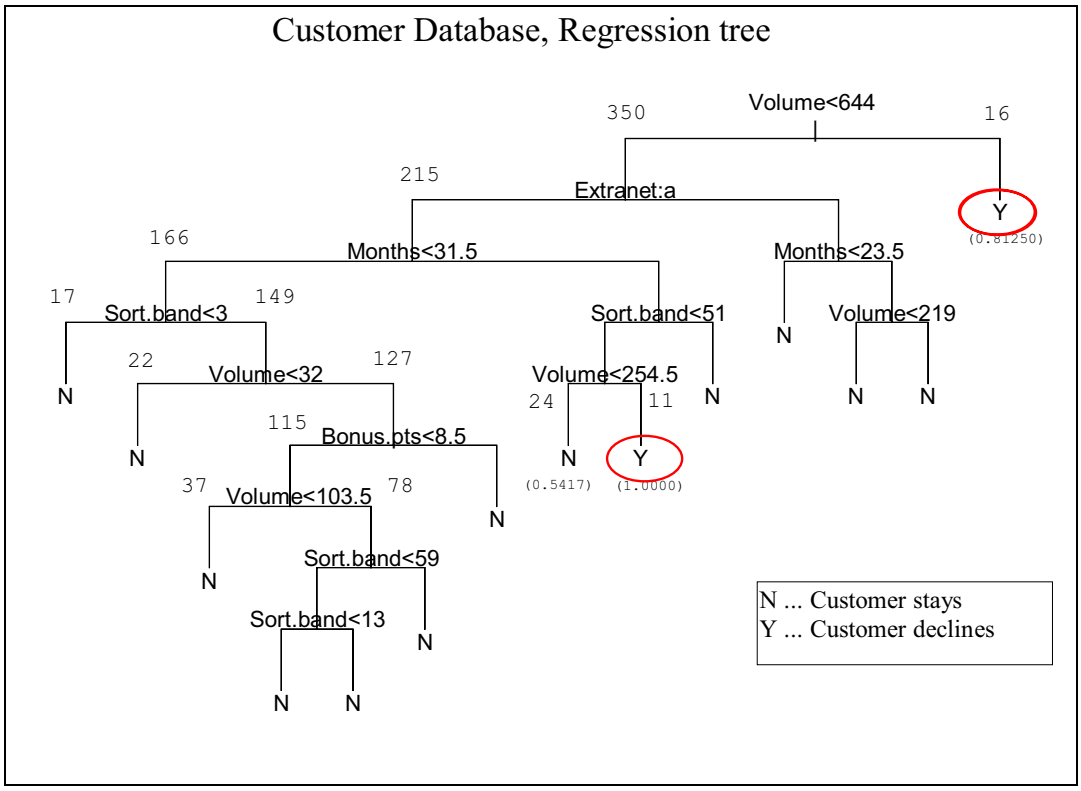
Volume	Sort band	Extranet	Bonus pts	Satisfy	Months	Losing
149	46	Y	6	2	38	N
294	8	N	7	0	18	N
55	84	N	4	5	14	N
169	14	N	5	1	37	N
66	38	N	7	5	9	Y
52	96	Y	4	5	9	N
631	60	N	8	0	23	N
17	2	N	6	3	73	Y
292	48	N	6	4	9	N
36	1	N	7	5	11	N
91	64	Y	5	5	15	N
250	152	N	4	5	3	N
33	24	N	5	5	11	Y
634	34	Y	0	4	7	N
787	58	N	11	1	11	Y
48	40	N	6	5	40	N
etc...	etc...	etc...	etc...	etc...	etc...	etc...

Volume = odebraný objem zboží (v dolarech), Sort band = šíře odebíraného sortimentu (počet položek), Extranet = zda klient obchoduje prostřednictvím Internetu, Bonus pts = interní hodnocení aktivity klienta, Satisfy = subjektivního hodnocení obchodního styku klientem, Months = jak dlouho je klientem společnosti, Losing = klient přestal obchodovat se společností (Y=ztráta klienta, N=klient zůstává).

Poslední sloupec představuje kritickou informaci o ztrátě zákazníka. Objasnit důvod této ztráty je cílem analýzy. Stromový model má za úkol vysvětlit proměnnou "Losing" pomocí ostatních proměnných. Zápis modelu zní:

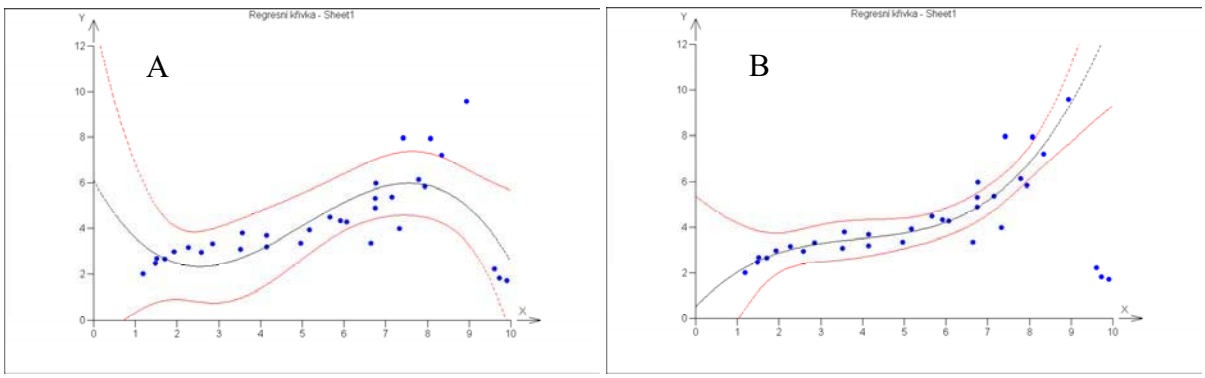
```
Classification tree model:
Losing ~ Volume + Sort.band + Extranet + Bonus.pts + Satisfy + Months
```

Výsledek v grafické podobě je na Obr. 5. Na první pohled je alarmující skutečnost, že společnost ztrácí zákazníky odebírající největší objemy, 81.2% zákazníků s odběrem větším než $644.000 odchází. Z ostatních zákazníků odchází klienti, kteří nemají přístup na Internet, obchodují déle než 31.5 měsíců, odebírají malý počet položek (<51, specializovaní klienti) a odebírají velké objemy (>254.5). Překvapující je, že v rozhodování klientů se nijak neprojevila jejich subjektivní spokojenost (proměnná Satisfy), což asi znamená, že se řídí spíše ekonomickými a pragmatickými podmínkami. Řešením v této situaci bude například poskytnutí výraznějších množstevních slev při velkých ročních odběrech a poskytnutí přístupu do obchodního systému společnosti přes Internet zdarma, případně na vlastní náklady k tomuto účelu vyškolit pracovníky klientů, protože "Internetoví" zákazníci jsou zjevně stálejší. Tento regresní strom tedy jednak odhalí možné příčiny problémů a pomůže je řešit, a na druhé straně skvěle umožní akvizici, plánování reklamy a výběr nových klientů, protože umíme předpovědět, který druh klienta bude pro společnost výhodný. Tuto zjednodušenou ukázku ovšem musela předcházet analýza kvality a spolehlivosti dat, která musela zahrnout i robustní filtry a metody, které "vyčistily" nespolehlivé údaje.



Obr. 5 Regresní strom pro klasifikaci zákazníků

Příklad 4. Zastavme se na závěr u pojmu „robustnost metody“. Diskutované metody jsou určeny především (ale ne pouze!) pro rozsáhlé databáze a automatické zpracování, často bez velkého zásahu či kontroly člověka. Je prakticky nemožné, aby v datech nebyly chyby, překlepy, nesmyslné nebo chybějící údaje a podobně. Je rovněž nemožné každý údaj kontrolovat či zkonstruovat obecný filtr, který databázi „vyčistí“ od takových dat. Chceme-li, aby výsledky byly smysluplné a schopné rozumné interpretace i za těchto reálných podmínek, musíme od výpočetních metod vyžadovat **robustnost**. Zhruba řečeno, robustnost metody (statistické, stochastické, neurální, nebo jiné) znamená její necitlivost na přítomnost chybných a nesmyslných dat (při vědomí, že zcela obecná definice „chybného“ data je nemožná). Z hlediska robustnosti se nejlépe osvědčují moderní statistické metody. Názorný příklad robustnosti je na Obr. 6. V grafu A byla data proložena křivkou (kubická parabola) pomocí klasické (nerobustní) regresní metody nejmenších čtverců. V grafu B jsou ***táž data*** proložena ***stejnou křivkou*** (kubická parabola), jenže s použitím robustní metody (konkrétně se jedná o tzv. *m-odhad*). Diametrálně odlišný výsledek a podstatně přesnější možnost predikce pomocí křivky B má na svědomí robustnost druhé metody, která správně identifikovala a ignorovala tři chybná data vpravo dole.



Obr. 6 Porovnání klasické (A) a robustní (B) metody proložení dat křivkou.

Závěr. Velmi výkonné algoritmy založené na principech statistiky, rozeznávání, neurálních sítí, učících cyklů, nelineární metrice a podobně, které plně využívají obrovský výpočetní potenciál současných osobních počítačů a pracovních stanic nabízí řadu postupů, které mohou z relativně neznámých a nepřehledných dat extrahovat užitečné informace, které mohou svému uživateli poskytnout rozhodující vodítko v konkurenčním prostředí. Vtip je v tom, že nerozhoduje pouze jak mnoho máme dat a jak jsou tato data přesná a aktuální (on-line přístup ke světovým finančním a komoditním burzám může mít kdokoliv), ale jak účinně a originálně jsme schopni z těchto dat získat smysluplnou a správnou informaci. Řada softwarových firem samozřejmě nabízí zaručené programy s komplikovanými názvy, které údajně vyřeší úplně každý problém. Při podrobnějším pohledu často zjistíme, že jde o ne příliš povedené poměrně jednoduché implementace metod neurálních sítí, či vícerozměrné analýzy zabalené do „třpytivého obalu“. Optimistická představa černé skříňky vyrábějící objevy z každé databáze naráží na několik předpokladů, především: (1) že samotná data musí být smysluplná (tj. že například neobsahují mnoho redundantních či irelevantních informací, příliš mnoho hrubých chyb a že se týkají předpokládaného fenoménu), (2) že máme alespoň hrubou představu co chceme nebo můžeme v datech nalézt, (3) že použijeme vhodné, robustní a spolehlivé metody se správně nastavenými parametry a (4) že uživatel je náležitě poučen nejen o samotném problému (ekonomická data musí analyzovat a interpretovat ekonom), ale také o vlastnostech použitých metod a je schopen správné diagnostiky a interpretace výsledků. Při porovnání různých metod se ukazuje, že v efektivitě analýzy mírně převažují moderní statistické metody (jsou obvykle stabilnější), a že automaticky se učící algoritmy představují spíše nevýhodu a větší nebezpečí chybné interpretace. Při tom je výhodné mít na mysli tzv. princip Occamovy břitvy, a nehledat za každou cenu nejsložitější modely dané situace, ale zastavit se na určité hranici komplexnosti problému. Příliš složité interpretace a závěry totiž většinou nejsou k ničemu, neboť pozbývají obecnosti a stability řešení, pro detaily přestaneme vidět celky. Je jisté, že nejlépe umí výsledky analýzy interpretovat uživatel, který dobře rozumí příslušné profesi a chápe principy použitých algoritmů. Rychle rostoucí řada publikovaných výsledků svědčí o tom, že se nepochybně vyplácí aplikace netriviálních algoritmů pro analýzu dat ale skutečný zisk není možný bez invence, kreativity a jisté dávky zdravé skepse, kteréžto vlastnosti nelze čekat od stroje.

Tipy pro datové horníky

- Buďte adaptivní, pečlivě vybírejte metody a přizpůsobujte je svým potřebám.
- Používejte programové struktury, opakované iterativní výpočty, automatizujte analýzu.
- Používejte efektivní metody a způsoby uložení dat.
- Při návrhu modelu využívejte předchozích informací o databázi.
- Používejte grafickou reprezentaci dat, modelů a výsledků, kde jen to jde.
- 70-90% dat v databázi je zbytečných (neobsahují užitečnou informaci).
- 5-30% dat v databázi je netypických (vybočuje, kazí analýzu, je třeba používat robustní metody).
- Databáze často obsahují nesmyslné údaje.
- Obrovská databáze většinou neznamená mnoho informací.
- DM je dobrý pro velké i pro malé databáze. I v malé databázi je užitečná informace.
- *Každá* matematicko-statistická metoda poskytne bohužel pro jakákoliv data *nějaký* výsledek.
- Pro jednoduché databáze použijte jednoduchou metodu a čekejte jednoduchý výsledek.
- Nepoužívejte zbytečně složité modely tam, kde stačí jednoduché (Occamova břitva).
- Nevěřte rafinovaným černým skříňkám do kterých nevidíte.
- Až do 3-rozměrného prostoru je nejlepším klasifikačním nástrojem lidské oko.
- Držte se *přirozené interpretace* výsledků.
- Nikdy nevěřte prvnímu výsledku.

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 8/15/99 - 8/18/99, San Diego, CA, USA

Sekce:
Data Reduction & Scalability
Mining Temporal Data
Clustering Techniques
Classification Techniques

Event Detection From Time Series Data
Discovering Trends and Differences
Noticing interesting changes in behavior
Statistics and Data Mining Techniques for Lifetime Value Modeling
Generalized Additive Neural Networks

Příklady uvedené v tomto článku byly zpracovány softwarem S-Plus a QC.Expert.

Doporučená literaura:

1. M. Meloun, J. Militký: Statistické zpracování experimentálních dat, EP Praha 1998
2. J. Antoch, D. Vorlíčková: Vybrané metody statistické analýzy dat, Academia Praha 1992
3. P. Hebák, J. Hustopecký: Vícerozměrní statistické metody, SNTL/ALFA 1987
4. Data Mining and Knowledge Discovery (časopis), Kulver Academic Publishers, ISSN: 1384-5810
5. P. Adriaans, D. Zantinge: Data Mining, Addison-Wesley Longman, Harlow 1998
6. R. S. Michalski et al.: Machine Learning and Data Mininng, Wiley, Chichester 1998
7. U. M. Fayyad: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press 1996
8. MathSoft: S-Plus Manual, Statistical methods

Další literatura

1. Jesus Mena: Data Mining Your Website, Digital Press 1999
2. Dorian Pyle: Data Preparation for Data Mining, Morgan Kaufmann Publishers 1999
3. Michael J. A. Berry, Gordon Linoff: Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons 1999
4. Rob Mattison, Brigitte Kilger-Mattison: Web Warehousing and Knowledge Management (Enterprise Computing Series), McGraw-Hill 1999
5. Elliot King: Data Warehousing and Data Mining: Implementing Strategic Knowledge Management, Computer Technology Research Corporation, 2000
6. Peter Cabena et al.: Discovering Data Mining from Concept to Implementation, Prentice Hall 1997
7. Boris Kovalerchuk, Evgenii Vityaev: Data Mining in Finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers 2000
8. Bhavani Thuraisingham: Data Mining : Technologies, Techniques, Tools, and Trends, CRC Press 1998
9. Alex Berson et al.: Building Data Mining Applications for CRM, McGraw-Hill Professional Publishing 1999
10. Chambers, John M., Hastie, Trevor J.: Statistical Models in S, Chapman and Hall, New York 1995
11. MathSoft: S-Plus Manual, Statistical methods