StatSoft

Úvod do data miningu

Tento článek je úvodním povídáním o data miningu, jeho vzniku, účelu a využití.

Historie data miningu

Rozvoj počítačů, výpočetní techniky a zavedení elektronického sběru dat vedlo ke vzniku velkých datových souborů. Tím vyvstala potřeba společností a podniků tyto velké objemy dat nějakým způsobem zpracovávat a využívat informaci v nich skrytou ke zvýšení výdělku, optimalizaci produktů, získání výhod oproti konkurenci nebo mnoha dalším cílům. Pro takovéto objemy dat nicméně standardní statistické metody nejsou příliš vhodné, bylo tedy potřeba nalézt metody, které dokáží nalézt i složité nelineární vztahy a to navíc bez omezujících předpokladů. Zadání tedy vzešlo přímo z potřeby mít nové metody, prostředkem bylo využití výpočetní síly počítačů k nalezení struktury (pravidla, vzory, asociace), namísto statistických parametrů (středních hodnot, vah, uzlů). Pojem data mining se začal objevovat v devadesátých letech, v roce 1991 napsal první definici data miningu pan Frawley: *„Data mining je netriviální získávání předtím neznámé a potenciálně užitečné informace ukryté v datech."* Do češtiny se občas překládá jako „dolování" či „vytěžování" dat. Na začátku nového tisíciletí se pak data mining osamostatnil jako nové odvětví statistiky.

Co je data mining

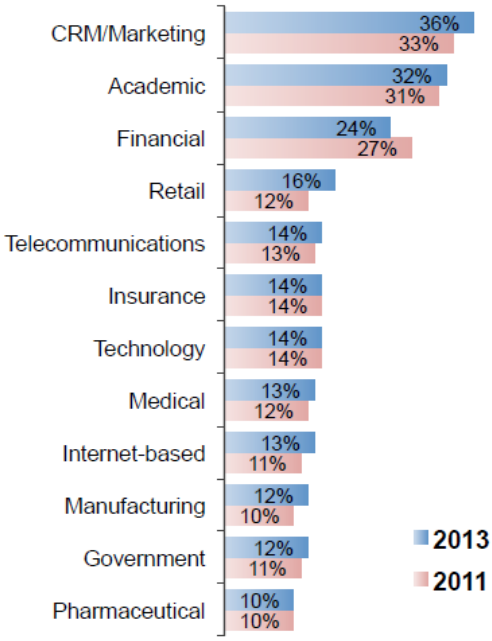
Data mining je dnes zcela určitě nejrychleji rostoucím segmentem business intelligence. S jeho pomocí se snažíme z ukládaných dat získat složitější a užitečnější informace než jen grafy a základní přehledy.

Ze statistického úhlu pohledu se jedná o vyšetřování vzájemných vztahů nebo vzorů v datech. Smyslem je analyzovat datové závislosti, určit trendy, a pokud to typ dat umožňuje, předpovědět budoucí vývoj.

Data mining pomáhá hledat odpovědi na různé otázky. Články na toto téma je možné si přečíst v sekci Publikační činnost: **http://www.statsoft.cz/o-firme/publikacni-cinnost/**.

Na co data mining použít?

Využití v praxi je celkem široké, jak se za chvíli přesvědčíme. Začněme výsledky nezávislé studie společnosti Rexer o využití dataminingových nástrojů za rok 2013 (a také 2011) v jednotlivých oblastech:

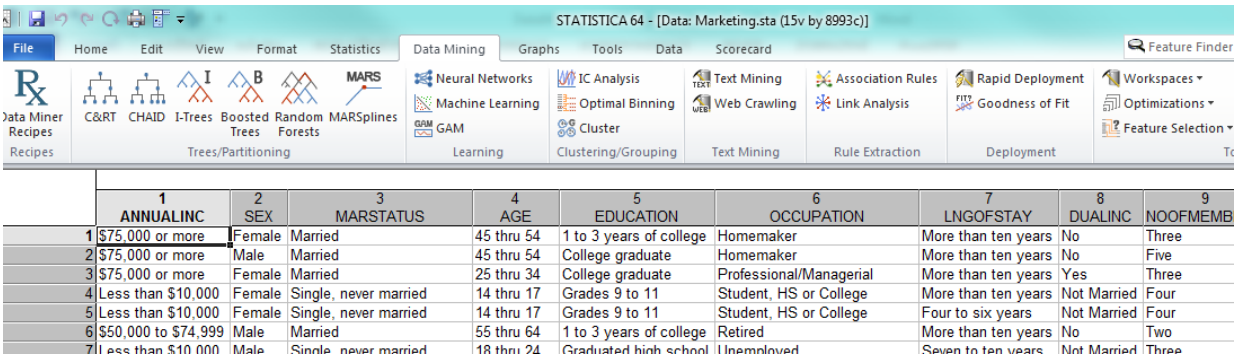


Je vidět, že data mining se používá především v oblastech, kde se sbírá velké množství dat. Typickými příklady obrovských datových souborů jsou například:

- údaje o klientech, pohyby na účtech (bankovnictví),
- údaje o volání (telefonní operátoři),
- informace o tom, jak lidé nakupují (obchodní řetězce a internetové obchody),
- pohyb uživatelů na internetu, datové informace o expresi genů (genetika),
- provozy zaznamenávají průběh provozních parametrů (průmysl).

Nicméně data mining již nějakou dobu není výsadou pouze velkých společností, ale jak ukazuje praxe, tyto postupy mají nemalý potenciál i v menších firmách, výzkumných ústavech, lékařství apod.

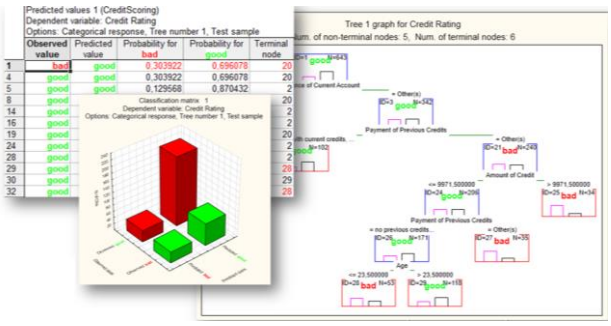
Dataminingové úlohy



	1 ANNUALINC	2 SEX	3 MARSTATUS	4 AGE	5 EDUCATION	6 OCCUPATION	7 LNGOFSTAY	8 DUALINC	9 NOOFMEMB
1	$75,000 or more	Female	Married	45 thru 54	1 to 3 years of college	Homemaker	More than ten years	No	Three
2	$75,000 or more	Male	Married	45 thru 54	College graduate	Homemaker	More than ten years	No	Five
3	$75,000 or more	Female	Married	25 thru 34	College graduate	Professional/Managerial	More than ten years	Yes	Three
4	Less than $10,000	Female	Single, never married	14 thru 17	Grades 9 to 11	Student, HS or College	More than ten years	Not Married	Four
5	Less than $10,000	Female	Single, never married	14 thru 17	Grades 9 to 11	Student, HS or College	Four to six years	Not Married	Four
6	$50,000 to $74,999	Male	Married	55 thru 64	1 to 3 years of college	Retired	More than ten years	No	Two
7	Less than $10,000	Male	Single, never married	18 thru 24	Graduated high school	Unemployed	Seven to ten years	Not Married	Three

Statistické úlohy, které stojí nad samostatnými daty, můžeme rozdělit na několik skupin:

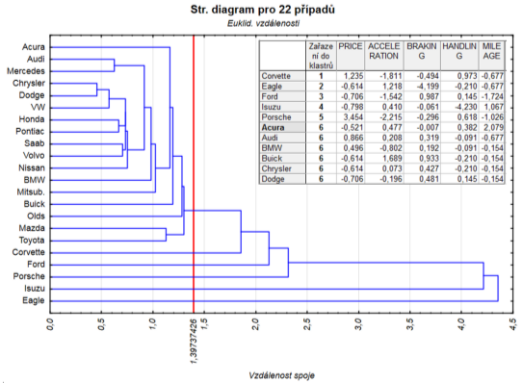
- ✓ *Klasifikace* - Klasifikační metody mají poměrně široké využití v různých oblastech, kde se shromažďuje větší množství dat. Definujeme je jako zařazování objektů (zákazníků, pacientů, dlužníků, příležitostí) do tříd, přičemž třídou rozumíme například: *Splatí/nesplatí, zdravý/nemocný, odpoví/neodpoví, registruje se/neregistruje se, koupí/nekoupí, SPAM/non SPAM.* Jde o nejčastější dataminingovou úlohu, kterou nad



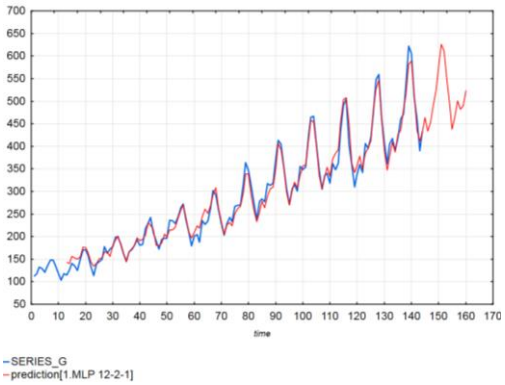
daty děláme. V těchto úlohách máme tzv. cílovou proměnnou (učitele), která definuje příslušnost konkrétního zákazníka do nějaké třídy. V tabulce níže je cílová proměnná *Credit Rating*, každý řádek reprezentuje konkrétního klienta, kterému byla v minulosti poskytnuta půjčka, a proměnná *Credit Rating* ukazuje ohodnocení konkrétních klientů. Jde tedy o historická data, nad kterými chceme vystavět model, s jehož pomocí potom budeme klasifikovat nové klienty.

	1 Credit Rating	2 Balance of Current Account	3 Duration of Credit	4 Payment of Previous Credits	5 Purpose of Credit	6 Amount of Credit	7 Value of Savings
67	good	>$300	24	no problems with current credits	other	$1 003,80	>1400
68	good	no running account	6	paid back	other	$1 204,00	no savings
69	bad	>$300	18	no previous credits	household appliances	$2 720,20	no savings
70	bad	no running account	36	no previous credits	other	$12 979,40	no savings
71	good	>$300	15	paid back	furniture	$1 842,40	140-700
72	good	>$300	24	no previous credits	new car	$4 883,20	<140
73	bad	no running account	21	no previous credits	furniture	$2 569,00	no savings

- ✓ *Shlukování/Segmentace* – Cílem této úlohy je najít objekty, které jsou si vzájemně podobné, případně skupiny vzájemně podobných objektů (zákazníků) bez znalosti či nějaké definice těchto skupin. V této úloze tedy nemáme cílovou proměnnou. Tento typ analýzy nám umožní shlukovat objekty (zákazníky) do skupin dle jejich vzájemné podobnosti, která ale není na první pohled zřejmá.



- ✓ *Predikce* – do této skupiny řadíme úlohy, které se zaměřují na předpovědi vývoje nějakého ukazatele v čase (objem poptávky, ceny a dalších ekonomických, ale také např. průmyslových ukazatelů) pomocí netriviálních statistických technik (neuronové sítě).



- ✓ *Regrese* – regresní úlohy slouží obecně pro vysvětlení a předpověď spojitých proměnných za pomoci dostupných informací z historických dat. Regresní úloha se liší od klasifikační především typem výsledku. V regresi je výsledkem spojitá číselná hodnota, nikoliv odhad dané kategorie (třídy). V některých oblastech se tyto metody nazývají úlohami typu: „Co se stane, když…".

- ✓ *Asociační pravidla* – specifické metody, které jsou vhodné pro konkrétní typ úloh. Tyto metody umožňují z velkého počtu záznamů stanovit pravidlo, které např. říká, že pokud návštěvník klikne na záložku „Pro ženy", tak s určitou pravděpodobností klikne také na „hubnutí a diety". Snahou asociačních pravidel je zjistit mezi položkami takový vztah, že přítomnost jedné nebo více položek v transakci implikuje výskyt jiných položek. Jinak řečeno, pomocí těchto metod hledáme odpovědi na otázky typu:

  - › Jaké jsou typické průchody webovou prezentací?
  - › Jaké jsou překážky při procházení stránek?
  - › Které URL navštíví uživatel obvykle před tím, než klikne na reklamní banner?

Ukázky konkrétní asociací z různých oblastí jsou následující:

- › pečivo => pátek/pondělí/středa

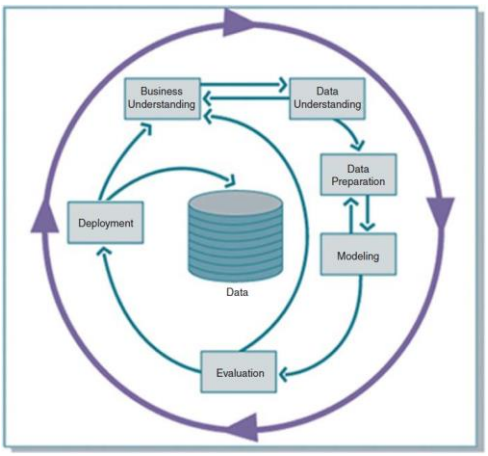
- brusle & chrániče => helma
- Apartment Rent & Credit => Full time job
- plenky & mléko => Basa piva
- Pro-ženy & *cokoliv* => hubnutí-diety
- detoxikace & pro-muže => Přírodní viagra

✓ *Text Mining* – textminingové úlohy obecně řadíme do úloh dataminingových. Text Mining pracuje s nestrukturovaným textem, lze ho tedy definovat jako proces vytěžení cenné informace z textu. V textové proměnné obvykle hledáme klíčová slova, následně děláme jejich frekvenční analýzu. Případy (konkrétní klienti, záznamy apod.), kde se tato klíčová slova vyskytla, indexujeme a následně vrátíme do souboru (databáze) jako novou číselnou proměnnou, kterou využijeme v rámci klasifikačních metod. Dalším typem úlohy je potom porovnávání dokumentů podle frekvence jednotlivých slov. Článek o této problematice si můžete přečíst např. v článku **Text Mining aneb Kladivo na nestrukturovaná data.**

Typickými metodami, které spadají do oblasti data miningu jsou klasifikační a regresní stromy, neuronové sítě a metody strojového učení.

Data miningový projekt

Nevládne jednota v tom, co znamená samotné spojení data mining. Někteří vidí za data miningem pouze analytické metody, které jsou potřeba k tvorbě modelů. Jiní berou data mining jako ucelený proces od vytažení dat, jejich zpracování přes tvorbu modelu až k nasazení modelu na nová data – tento druhý pohled je nicméně častější a vhodnější.



Nyní se pojďme podívat, co takový dataminingový projekt obnáší. Představme si, že před nás jako statistiky někdo právě položil nějaké zadání. První fází každého projektu je pochopení dat a pochopení cíle celého projektu a to nejen ze statistického, ale také z obchodního pohledu (je potřeba vědět, na co bude případně model použit, jaký je tedy jeho smysl). Poté, co se do detailu seznámíme s daty, čeká nás ta nejpracnější část a to očištění a zpracování dat, zde vyřazujeme a hledáme nesmyslné hodnoty a připravujeme si nové proměnné, které by se nám mohly hodit pro následné modelování. Pokud máme data připravena, čeká nás fáze modelování, což je proces vytváření nejrůznějších modelů, přičemž se nám často stane, že zjistíme, že by se nám například velmi hodila ještě nějaká další proměnná, musíme se tedy občas vrátit do fáze přípravy dat a vhodné proměnné si dovyrobit. Pokud již máme modely, vybereme z nich ten nejlepší (úspěšnost modelu hodnotíme například pomocí tzv. ROC, Gains a ROC křivek, které velmi pěkně popisuje **článek v starším newsletteru**). Pokud se model chová dobře a dává smysl i z obchodního hlediska, dostáváme se do fáze nasazení (deployment) - vytvořený model použijeme již v praxi na novou sadu dat. Kolečko uzavřeme tím, že pozorujeme, zda je model stále aktuální, což provedeme porovnáváním výsledků modelu, stejně jako i rozložením vstupních dat aktuálních a historických. V případě velkých odchylek poté musíme přikročit k aktualizaci modelu, tedy tvorbě modelu na základě nových poznatků.

Toto je samozřejmě pouze stručný popis dataminingového projektu.

Možné cíle dataminingového projektu:

- pochopení stávajících zákazníků,
- zlepšení spokojenosti zákazníka,
- zabránění odchodu klienta,
- eliminování rizikového klienta,
- vzbuzení zájmu potencionálního zákazníka,
- optimalizovaní pojistné sazby,
- detekce podvodného jednání,
- určení pravděpodobnosti pojistné události, nesplácení úvěru,
- Cross-up/sell analýzy
- ...

Konkrétní aplikace a zajímavá řešení najdete mezi **našimi případovými studiemi**.

StatSoft CR s. r. o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz