Data mining

Data mining ([dejta majnyn], angl. *dolování z dat* či *vytěžování dat*) je analytická metodologie získávání netriviálních skrytých a potenciálně užitečných informací z dat. Někdy se chápe jako analytická součást *dobývání znalostí z databází* (Knowledge Discovery in Databases, KDD),[1] jindy se tato dvě označení chápou jako souznačná.

Data mining se používá v komerční sféře (například v marketingu při rozhodování, které klienty oslovit dopisem s nabídkou produktu), ve vědeckém výzkumu (například při analýze genetické informace) i v jiných oblastech (například při monitorování aktivit na internetu s cílem odhalit činnost potenciálních škůdců a teroristů).

1 Historie

První náznaky aktivit, které dnes označujeme jako data mining, se objevily v 60. letech 20. století s rozvojem počítačové techniky. Šlo například o využívání regresní analýzy s automatickým výběrem proměnných a prvních rozhodovacích stromů. Většinou však šlo jen o ojedinělé nebo akademické záležitosti.

Rozvoj statistických metod, databázových aplikací a umělé inteligence spolu s rychlým růstem rychlosti a paměti počítačů byly předpoklady, které umožnily v sedmdesátých a osmdesátých letech první systematická využití data miningové metodologie v praxi. Slovní spojení *data mining* tehdy ovšem stále mělo spíše hanlivý přídech: Označovalo „vyzobávání rozinek“ z dat, hledání korelací ve velkých datových souborech, které – jak známo ze statistické teorie – je vystaveno obrovskému nebezpečí, že „objeví“ pouze nahodilé fluktuace v datech bez možnosti zobecnění a praktického využití.

Obrat přišel počátkem devadesátých let. V té době byly již vybudovány metody, umožňující vyhnout se zmíněnému nebezpečí falešných korelací (například kontrola založená na vynechaných datech nebo na metodě Monte Carlo). Navíc zejména v USA rostla poptávka ze strany komerčních organizací, disponujících již velkými objemy dat a neschopných z nich pomocí klasických tabelačních metod získat potřebné podklady pro rozhodování. To napomohlo k rychlému etablování data miningu jako svébytného oboru aplikované vědy a k jeho širokému použití v komerční praxi. Časté aplikace jsou především v oblastech přímého marketingu (výběr klientů pro oslovení), finančnictví (např. odhadování rizika, hledání podvodů), maloobchodního prodeje (analýza nákupních košíků aj.), telekomunikací (segmentace klientů, prodej programů aj.) a internetového prodeje (analýza přechodů mezi stránkami, efektivity reklamy apod.).

Nárůst aplikací v oblasti data miningu se projevil i na softwarovém a konzultačním trhu. Existuje již poměrně široká nabídka specializovaných softwarů pro tento účel. Vedoucími trhu dataminingových softwarů jsou komerční aplikace SAS Enterprise Miner, SPSS Clementine a STATISTICA Data Miner, mezi známé nekomerční softwary patří Weka a Orange.

2 Metodologie data miningu

Protože data mining zahrnuje velkou šíři metod a způsobů práce, je obtížné podat jednoznačný návod k postupu. Přesto během 90. let vykrystalizovaly dvě obecné metodologie, které alespoň v hrubých rysech popisují jednotlivé kroky: metodologie *SEMMA*, za níž stojí firma SAS, a *CRISP-DM*, vyvinutá konsorciem firem, mezi něž patřil druhý hlavní hráč na trhu, SPSS.

Společnou podstatou všech metodologií je následnost několika kroků:

- **Obchodní/praktický** – formulace úlohy a porozumění problému. Ani automatické vyhledávání znalostí nelze provádět zcela naslepo.
- **Datový** – vyhledání a příprava dat pro analýzu. Statistické algoritmy většinou potřebují data připravená v určité podobě, a proto není možné použít přímo surových dat z obchodních databází.
- **Analytický** – hledání informace v datech, vytváření statistických modelů a podobně. Využívají se nejrůznější metody od jednoduchých tabelací a vizualizací až po sofistikované přístupy jako je genetické programování. Asi nejčastěji používanými metodami však jsou logistická regrese s automatickým výběrem proměnných, rozhodovací stromy a neuronové sítě. Výstup této fáze bývá dvojí: Jednak obecnější *znalosti* (např. že svobodní klienti nejčastěji nakupují pozdě večer, zatímco ženatí po obědě), jednak matematické *modely* (např. postup, jak vytipovat potenciálního klienta pro daný produkt).
- **Aplikační** – zjištěné poznatky a modely je třeba uvést do praxe, například spuštěním reklamní kampaně nebo reorganizací webových stránek.

- **Kontrolní** – je třeba zajistit zpětnou vazbu (jak efektivní byla obchodní akce) a v případě dlouhodobě nasazovaných modelů i kontrolovat, zda model příliš nezestárl a zachovává si svoji efektivitu.

3 Potenciální nebezpečí data miningu

Protože komerční data mining představuje často masivní a inteligentní zpracování osobních údajů, vznikají často obavy ze zneužití těchto informací.

Kromě obvyklých negativ spojených se shromažďováním osobních údajů, jako je záměrný i nezáměrný únik dat a jejich využití k různým nečestným aktivitám od spamu až po vydírání, zde teoreticky hrozí i specifické zneužití statistických technik. Lze si například představit zločince, který si pomocí analýzy dat vytipovává své oběti.

Zdá se však, že toto nebezpečí je – alespoň v současném stavu data miningu – nepatrné. I kdyby se náhodou zločinci dostali k využitelným osobním datům, pravděpodobně by jim použití sofistikovaných statistických metod příliš nepomohlo, už proto, že by jim chyběla databáze „pozitivních příkladů“ úspěšných zločinů, na níž by mohli své modely postavit.

Za větší potenciální nebezpečí lze považovat technologie, k jejichž vzniku data mining přispívá v akademické sféře. Například dekódování genomu může být použito k nehumánním selekcím osob podobným eugenice, ale postaveným na vědeckém základě. Anebo pokročilé metody identifikace osob mohou být spolu s kamerovými systémy používány ke špehování pohybu občanů.

4 Používané techniky

- rozhodovací stromy
- asociační pravidla
- neuronové sítě
- regresní analýza
- shluková analýza (pro marketingovou segmentaci)

4.1 Problémy

Ze správných dat, použijeme-li správný způsob dobývání, dostaneme správné výsledky. Proto by dobývání dat mělo být založeno na správných datech. Nicméně protože se používají statistické metody, tak výsledky jsou tzv. *statisticky správně*, tj. s malou pravděpodobností odvodíme chybné výsledky (falešně pozitivní).

Dobývání typicky neodhalí všechny souvislosti schované v datech.

I z nesmyslných dat, například špatně připravených anebo náhodných dat dostaneme nějaké výsledky. Je možné dostat i výsledky, co vypadají smysluplně. *Smetí dovnitř, smetí ven* (angl. Garbage In, Garbage Out - GIGO).

5 Reference

[1] BERKA, Petr. *Dobývání znalostí z databází*. Praha : Academia, 2003. ISBN 80-200-1062-9.

6 Literatura

- BERKA, Petr. *Dobývání znalostí z databází*. Praha : Academia, 2003. ISBN 80-200-1062-9. S. 366.

7 Externí odkazy

- Obrázky, zvuky či videa k tématu Data mining ve Wikimedia Commons

7.1 Informace

- KDnuggets – on-line čtrnáctideník o data miningu (anglicky)
- Průvodce data miningem (anglicky)
- Themanagement.de – sbírka článků (německy)
- CRISP-DM: Cross Industry Standard Process for Data Mining (anglicky)
- Eruditionhome – rozcestník (anglicky)
- CRM Today (anglicky)
- Data Warehousing Review o Data miningu (anglicky)
- SIGKDD – sdružuje vědce v oblasti, pořádá konference (anglicky)
- Data Mining for Scientific and Engineering Applications – Tutorial (anglicky)
- MLnet – on-line seznam softwaru
- Poskytovatel data miningu
- Specialisté na data mining
- Open-source nástroje pro data mining - recenze programů Orange, KNIME a RapidMiner v magazínu LinuxEXPRES

7.2 Software

Stránky producentů softwaru obsahují často i řadu obecnějších informací (případové studie, zkušenosti klientů apod.)

- Bayesia – komerční
- Ferda DataMiner - společné vizuální prostředí pro procedury dataminingu, nekomerční, český produkt
- LISp-Miner – nekomerční, český produkt
- Megaputer – komerční
- Miner3D (vizualizace) – komerční, slovenský produkt
- Minitab – komerční;české stránky jsou zde
- Orange – nekomerční
- Neural Designer – komerční
- OpenNN – nekomerční
- R – free software environment for statistical computing and graphics
- Salford Systems – komerční
- SAS – komerční; české stránky jsou zde
- SPSS – komerční; české stránky jsou zde
- STATISTICA Data Miner – komerční; české stránky jsou zde
- Tanagra – nekomerční
- WEKA – nekomerční
- YALE – nekomerční

8 Zdroje textu a obrázků, přispěvatelé a licence

8.1 Text

- **Data mining** *Zdroj:* http://cs.wikipedia.org/wiki/Data_mining?oldid=12573207 *Přispěvatelé:* Ludek, Bota47, YurikBot, Jan Spousta, Zagothal, Joker Island, Porthos, FlaBot, Mercy, Harold, Dinybot, JAnDbot, Demian79, Thijs!bot, Rei-bot, OdderBot, VolkovBot, TXiKiBoT, Blek, MiroslavJosef, AlleborgoBot, Jj14, SilvonenBot, Djuke, Luckas-bot, Jana Lánová, Ptbotgourou, Nallimbot, Udufruduhu, RibotBOT, MastiBot, MauritsBot, TobeBot, Lmises, Klenot, EmausBot, JackieBot, WikitanvirBot, ChuispastonBot, Daemonicky, Addbot a Anonymové: 16

8.2 Obrázky

- **Soubor:Commons-logo.svg** *Zdroj:* https://upload.wikimedia.org/wikipedia/commons/4/4a/Commons-logo.svg *Licence:* Public domain *Přispěvatelé:* This version created by Pumbaa, using a proper partial circle and SVG geometry features. (Former versions used to be slightly warped.) *Původní autor:* SVG version was created by User:Grunt and cleaned up by 3247, based on the earlier PNG version, created by Reidab.

8.3 Licence obsahu

- Creative Commons Attribution-Share Alike 3.0