Informační Bulletin

České Statistické Společnosti číslo 3, ročník 16, prosinec 2005

STARÝ ROK KONČÍ A NOVÝ PŘICHÁZÍ

Vážené kolegyně, vážení kolegové, vážené čtenářky a vážení čtenáři,

dovolte nám popřát Vám všem vše nejlepší v nastávajícím roce 2006, mnoho pohody v osobním i pracovním životě a především hodně zdraví.

Všem, kteří letos přispěli do našeho bulletinu, velmi děkujeme za jejich příspěvky. Zároveň doufáme, že příliv příspěvků jak do bulletinu naší společnosti, tak i do časopisu *Statistika* nejenom že nevyschne, nýbrž bude stále „houstnout“.

Srdečně Vás všechny zveme na výroční schůzi České statistické společnosti, která se uskuteční již po sedmnácté, tradičně ve čtvrtek **2. února 2006** od 13.00 na VŠE v Praze.

Na shledání na dalších akcích České statistické společnosti se těší

Výbor České statistické společnosti

SETKÁNÍ ZÁSTUPCŮ NÁRODNÍCH STATISTICKÝCH SPOLEČNOSTÍ, BUDAPEŠŤ, 23. ZÁŘÍ 2005

Gejza Dohnal

Z iniciativy našich maďarských kolegů se konalo 23. září 2005 v Budapešti setkání zástupců šesti národních statistických společností: maďarské, slovenské, české, rakouské, slovinské a rumunské [1]. Delegace byly ve složení:

- Za Rakouskou statistickou společnost: pánové Joachim Lamel (president společnosti), Daniel Dekic a slečna Pia Ebhart.
- Za Českou statistickou společnost: pan Gejza Dohnal (člen výboru společnosti).
- Za Maďarskou statistickou společnost: pánové Sándor Herman (president společnosti), Lrinc Soós a paní Éva Laczka.
- Za Rumunskou statistickou společnost: pánové Ilie Dumitrescu (člen výboru společnosti) a Ioan Goreac (tajemník společnosti).
- Za Slovenskou statistickou a demografickou společnost: pánové Peter Mach (president společnosti), Gejza Wimmer a Karol Pastor.
- Za Slovinskou statistickou společnost: Andrej Blejec (president společnosti).

Jednání probíhalo v jedné ze zasedacích místností Maďarského statistického úřadu v historické budově v centru Budapešti. Atmosféra byla od počátku velice přátelská. Zástupci jednotlivých společností v krátkých výstupech představili své společnosti, jejich cíle a problémy, se kterými se potýkají. Z těchto vystoupení bylo zřejmé, že hlavní cíle všech zúčastněných společností jsou velice podobné: vytvořit podmínky pro sbližování statistiků různého zaměření, především sblížení státní statistiky a různých směrů matematické statistiky, podpora mladých statistiků a osvětová činnost, zaměřená především na zvýšení kreditu statistiky v povědomí veřejnosti. Také problémy jsou ve všech zemích podobné: relativně malá komunikace mezi statistiky různého zaměření, „boj“ o zájem mladých statistiků a nízká informovanost spolu s nezájmem o problematiku prezentace a interpretace výsledků statistických šetření ze strany veřejnosti, médií a politiků. Tato vystoupení potvrdila hlavní

[1] Poněkud zde chyběli kolegové z Polska, důvody nebyly zcela zřejmé, vysvětlení bylo, že prý z časových důvodů (?).

myšlenku setkání, potřebu spolupráce a intenzivnější komunikace mezi statistickými společnostmi různých zemí. Ve druhé části programu vystoupil kolega Soós s příspěvkem věnovaným problematice statistické etiky. V roce 1985 zveřejnil International Statistical Institute (ISI) „Deklaraci profesionální etiky“ [2], která se stala základním dokumentem v této oblasti. Od té doby prochází řadou diskuzí a snah o doplnění či upřesnění. Zásadní problémy jsou dva. Za prvé, na koho se má tento kodex vztahovat - zda jen na „výrobce“, tedy na statistiky, kteří provádějí analýzu a zpracovávají data, nebo také na „uživatele“, tedy na ty, kdo výsledky statistických analýz a šetření interpretují široké veřejnosti a kteří velmi často nejsou profesionálními statistiky? Druhý problém je spojen s případnými sankcemi za nedodržení etického kodexu. Jak donutit statistiky i ty ostatní dodržovat etické zásady stanovené kodexem? Podobné otázky jsou zdrojem neustálých diskuzí na toto téma a diskutovalo se o nich i na tomto setkání. Třetí část jednání směřovala k návrhu společné deklarace o spolupráci. Hlavní diskutované body se týkaly společné publikace a případných společných akcí. Návrh na pravidelný čtvrtletní bulletin neprošel. Nakonec byla deklarována shoda v tom, že je třeba jakýsi bulletin vydávat, ale konkrétní podmínky (kdo, kde, jak často) nebyly dohodnuty. Podobně potřeba společných setkání zůstala na formulaci „nejméně jednou ročně“. Organizace příští schůzky v roce 2006 se ujala slovenská delegace a ústy předsedy Slovenské statistické a demografické společnosti pozvala všechny zúčastněné na příští rok do Bratislavy. V závěru dopoledního jednání byla zformulována Dohoda o spolupráci, kterou podepsali zástupci všech zúčastněných statistických společností.

Odpolední program pokračoval společným výletem do staroslavného Visegrádu. Tam byla připravena prohlídka zříceniny slavného hradu nad majestátním Dunajem, kdysi sídla maďarských králů, místa několika Visegrádských dohod. V dobách existence římské říše byla oblast dnešního Visegrádu severní hranicí impéria. Římané zde na skalní vyvýšenině postavili pevnost na obranu proti útokům různých germánských i jiných kmenů. V dobách středověkého maďarského království se Visegrád k roku 1009 připomíná jako župní sídlo a místo, kde bylo vybíráno clo na obchodní cestě podél Dunaje. Po mongolských nájezdech nechal v 50. letech 13. století král Béla IV. zbudovat mohutný hrad, na který se roku 1323 se dvorem přestěhoval král Karel Robert. V té době prožíval Visegrád období své největší slávy. Stal se na čas i hlavním městem říše s velkolepým královským palácem. V roce 1335 zde došlo k první dohodě: setkali se zde český král Jan Lucemburský, uherský král Karel Robert a polský král Kazimír Veliký a uzavřeli antihabsburské

[2] plný text lze nalézt na http://isi.cbs.nl/ethics.htm

spojenectví. Jan Lucemburský se zřekl nároků na polskou korunu za 20 tisíc kop pražských grošů a byla mu zde přiznána svrchovanost nad Slezskem. Další dohoda z roku 1339 ustanovila, že po smrti Kazimíra Velkého zasedne na polský trůn jeden ze synů Karla Roberta (Ludvík Uherský) pod podmínkou, že bude dodržovat privilegia Poláků a Kazimír Veliký nebude mít syna. Ve druhé polovině 15. století hrad významně přestavěl v renesančním stylu Matyáš Korvín. Královské sídlo bylo přestěhováno do nedaleké Budapešti a Visegrád sloužil jako venkovská rezidence uherských panovníků. To trvalo až do ovládnutí dolních Uher Turky. Po opakovaném dobývání v letech 1529 a 1543 se hrad proměnil v ruiny. Město v podhradí bylo znovu obydleno až na sklonku 17. století. Od roku 1878 až dodnes probíhají v areálu hradu archeologické výzkumy a rekonstrukční práce. Poslední významná dohoda zde byla učiněna roku 1991, kdy se zde zástupci ČSFR, Maďarska a Polska dohodli na spolupráci při vstupu do EU. Vznikla tak Visegrádská skupina, označovaná jako Visegrádská čtyřka (V4).

Naše jednání bylo završeno tak, jak se na toto slavné místo sluší - královskou hostinou. Ta přišla vhod, neboť při jednání v Budapešti nebyl čas ani na oběd. Před hostinou ve stylově (a trochu kýčovitě) zařízené restauraci rezervované jenom pro nás, jsme si měli ze svého středu vybrat krále. Ten dostal královský plášť a privilegium vybrat si k sobě královnu. Králem jsme zvolili rumunského kolegu Ilie Dumitrescu (snad kvůli jeho impozantní postavě) a on si vybral ke svému boku kolegyni z Rakouska, Piu Ebhart. Poté jsme zasedli k dlouhé tabuli a hostina mohla začít ...

Plný text uzavřené dohody čtenáři najdou na následujících stránkách (v originále anglicky i v českém překladu).

Addresa: Doc. RNDr. Gejza Dohnal, CSc.,
Ústav technické matematiky, Fakulta strojní ČVUT v Praze.
✉dohnal@nipax.cz

Dohoda o spolupráci

Z hlediska potřeby spolupráce a při vědomí toho, že posláním a cílem statistických společností je přinášet prospěch statistice, zástupci zúčastněných statistických společností vyjadřují svůj souhlas s tím, že jejich společnosti budou:

- podporovat vědecký pokrok a rozvoj jak teoretické, tak i aplikované statistiky v oblasti;
- poskytovat prostor pro diskusi o teoretických i aplikačních problémech ve statistice ve své oblasti a pokud bude třeba, zaujmou aktivní pozici při projednávání profesionálních záležitostí;
- udržovat kontakty s ostatními statistickými společnostmi v oblasti;
- pomáhat mladým statistikům v jejich profesionálním růstu, především usnadněním přístupu k profesním setkáním a konferencím;
- nabízet podporu nově vznikajícím statistickým společnostem;
- prosazovat dodržování Deklarace profesionální etiky;
- rozvíjet a podporovat spolupráci s ISI, zejména napomáhat rozšiřování zástupců členů ISI v regionu;
- ve svých společnostech propagovat základní principy oficiální statistiky, profesionální etiky a správné postupy při aplikaci statistiky v praxi;
- cíleně podporovat co nejlepší vztahy mezi akademickou a oficiální (státní) statistikou.

K dosažení výše uvedených cílů budou zúčastněné společnosti mimo jiné:

- nejméně jednou ročně organizovat společné setkání a
- poskytovat si vzájemně informace o svých aktivitách, včetně publikování bulletinu, v němž si budou vyměňovat názory a zkušenosti z oblastí, v nichž ve statistice působí.

Níže podepsané společnosti vyzývají ostatní statistické společnosti, aby se připojily k této dohodě. Podepsáno 23. září 2005 v Budapešti.

Rakouská statistická společnost: Joachim Lamel (prezident)
Česká statistická společnost: Gejza Dohnal (pověřený prezidentem)
Maďarská statistická společnost: Sándor Herman (prezident)
Rumunská statistická společnost: Ilie Dumitrescu (pověřený prezidentem)
Slovenská statistická a demografická společnost: Peter Mach (prezident)
Slovinská statistická společnost: Andrej Blejec (prezident)

Několik záběrů z jednání:

Nahoře zleva: Joachim Lamel, Daniel Dekic, Pia Ebhart, Gejza Wimmer a Karol Pastor

Gejza Dohnal a Peter Mach

Ilie Dumitrescu a Andrej Blejec

Nalevo zleva a shora: Eva Laczká, Peter Mach, maďarský tlumočník, Sándor Herman, Lörinc Sóos

Agreement on Cooperation

With a view to the need for cooperation, and also considering that the objectives and goals of statistical societies are to serve statistics, the representatives of signatory Statistical Societies agree to

- facilitate scientific progress and theoretical / practical development of statistics in the region;
- provide a regional forum for the discussion of theoretical and practical issues of statistics, and take up a position in professional matters, if required;
- maintain contact with the Statistical Societies of the region;
- assist young statisticians in their professional career by facilitating access to professional meetings and conferences;
- offer support to the nascent Statistical Societies of the region;
- promote the application of the Professional Ethics in the region;
- promote cooperation and relations with the ISI particularly to facilitate the enlargement of representatives of ISI members from the region;
- promote, within their respective Societies, the fundamental principles of official statistics, Professional Ethics and good practices in statistics, and
- support the objectives of best relations between academic and official statistics.

For achieving the above objectives the signatory Societies shall, among others,

- meet at least once a year, and
- provide mutual information on their activities, including the publication of a newsletter for the exchange of views and experience in their domains of statistics.

The signatory Societies invite other Statistical Societies to join this Agreement.

Signed on the 23rd of September, 2005, in Budapest

Austrian Statistical Society: President Joachim Lamel
Czech Statistical Society: For the president Gejza Dohnal
Hungarian Statistical Society: President Sándor Herman
Romanian Statistical Society: For the president Ilie Dumitrescu
Slovak Statistical and Demographical Society: President Peter Mach
Statistical Society of Slovenia: President Andrej Blejec

O INTUICI A PRAVDĚPODOBNOSTI

Josef Tvrdík

Podnětem k napsání této poznámky byla jedna hra mariáše na cyklistické dovolené v Heřmani u Písku. Pravidelné dovolené se stejnou partou jsou pro mne příležitostí jednou za rok si připomenout zákonitosti této tradiční hry a při tom se bavit tímto krásným jednoduchým stochastickým modelem mnoha životních příběhů. Hra prakticky neustále přináší okamžiky, kdy je vhodné užít pravděpodobnost k ocenění možných variant rozložení karet. Většinou stačí ohodnocení intuitivní a žádné velké počítání není potřeba, ale někdy se přece jen klasická pravděpodobnost může hodit.

Pokud mariáš nehrajete nebo vás vůbec nezajímá, tak neztrácejte čas dalším čtením. Pro ostatní připomínám, že v textu je užíván obecný jazyk mariášníků, tedy termíny jako plonk, mazat, šintovat a štych (nikoliv zdvih jako v bridži). Pokud někomu z vás taková slova v psaném českém textu vadí, tak už také dále nečtěte.

Podnětná situace nastala u hraní tzv. ostravické varianty mariáše. V této variantě jsou barvy různě drahé (zelené dvakrát, kule třikrát a červené čtyřikrát dražší než žaludy), výše výhry je úměrná dosaženému rozdílu, hlášená sedma stojí stejně jako rozdíl padesát (patrně krajovým vlivem taroků, v nichž pagát je také drahý), betl a durch se nehraje, stovky se nehlásí, protože hraje se na co největší rozdíl. Výhodou této ostravické varianty mariáše je, že nejsou hry složené bez sehrávky. Pokud se však nechcete ostravickou variantou zatěžovat, tak si představte, že popisovaná situace nastala při kulové stovce v běžném mariáši. Hráči sedí v pořadí J, A, M, hráč J je na výnosu a má následující karty:

kule (trumfy):	6 nejvyšších
zelené:	eso
červené:	eso, svršek, osma
talon:	dvě malé žaludy

Hráč J vynesl červenou osmu a podle svého přiměřeného očekávání uhrál 120. Po hře se proběhla diskuse s následujícím obsahem:
A: Proč jsi nenesl červené eso?
J: Proč bych to dělal?
A: Abys uhrál plonkovou desítku, cílem je přece uhrát co nejvíc.
M a J: Výnos osmou je lepší. Pravděpodobnost uhrát 130 je malá, srovnatelná s tím, že soupeři eso zabijí trumfem, což pak může vést k dalším obtížím. Vynést červené eso by byla riskantní hloupost.

Ač A je zdatný hráč mariáše i taroků, v tomto okamžiku intuitivně hodnotil situaci zcela odlišně od intuice obou ostatních a zatvrzele trval na svém názoru. Jak posoudit, čí intuice je správná? Pomůžeme si klasickou pravděpodobností.

Z pohledu hráče J zbývá 20 karet, počet možných rozdělení karet mezi dva hráče je $\binom{20}{10} = 184\ 756$. Pravděpodobnost, že z n ($n \leq 10$) zvolených karet má daný hráč právě k karet ($k \leq n$) z této n-tice je

$$P(x_n = k) = \frac{\binom{n}{k}\ \binom{20-n}{10-k}}{\binom{20}{10}}\,, \tag{1}$$

kde výraz $x_n = k$ značí, že počet karet z uvažované n-tice v ruce hráče A je roven k. Pravděpodobnost, že bude mít k-tici předem určených karet, je

$$P(x_{nk} = k) = \frac{\binom{20-n}{10-k}}{\binom{20}{10}}\,, \tag{2}$$

kde výraz $x_{nk} = k$ značí, že všech k určených karet z uvažované n-tice je v ruce hráče A. Např. pro $n = 5$, což je počet červených zbývajících v rukou A a M, a $k = 1$ to je pravděpodobnost, že karta v ruce hráče A bude plonková červená desítka, po dosazení do rovnice (2) je $P(x_{51} = 1) = 0.0271$.

Plonková červená desítka (jev D, $P(D) = 2 \times 0.0271 = 0.0542$) při výnosu červeným esem k uhrání sto třiceti nestačí. Ještě je nutné, aby hráč s plonkovou desítkou nemazal na zbývající dva štychy na červené. To znamená, že má oba trumfy (jev T) nebo nemá co mazat, tj. všechny tři zbývající ostré (zelená desítka, žaludské eso a desítka) jsou v ruce druhého hráče (jev Z). Je potřeba vyhodnotit pravděpodobnost

$$P[(D \cap T) \cup (D \cap Z)] = P(D)\ P[(T \cup Z) \mid D]\,.$$

Podmíněnou pravděpodobnost na pravé straně rovnice můžeme přepsat

$$P[(T \cup Z) \mid D)] = P(T \mid D) + P(Z \mid D) - P(T \cap Z) \mid D)$$

a pak pravděpodobnosti vpravo od rovnítka jednoduše spočítat (počet možných doplnění plonkové desítky je $\binom{15}{9}$, počet příznivých doplnění se snižuje pevně umístěnými kartami):

$$P(T \mid D) = \frac{\binom{13}{7}}{\binom{15}{9}} = 0.3429\,,$$

$$P(Z \mid D) = \frac{\binom{12}{9}}{\binom{15}{9}} = 0.0440\,,$$

$$P[(T \cap Z) \mid D] = \frac{\binom{10}{7}}{\binom{15}{9}} = 0.0240\,.$$

Pravděpodobnost uhrát 130 je pak

$$P(D)\ P[(T \cup Z) \mid D] = 0.0542 \times 0.3629 = 0.0197\,.$$

Pravděpodobnost, že při výnosu červeným esem ho protihráči zabijí trumfem - všech pět červených v jedné ruce (jev E), druhý hráč má aspoň jeden trumf (jev G) - je

$$P(E \cap G) = P(E)\ P(G \mid E)$$

Podle rovnice (1) je $P(x_5 = 0) = 0.0163$ a $P(E) = 2 \times 0.0163 = 0.0325$[1]. Podmíněnou pravděpodobnost $P(G \mid E)$ spočítáme takto: Počet možných doplnění karet hráče s pěti červenými je $\binom{15}{5}$, počet příznivých doplnění je $\binom{13}{5}$, tj. dva trumfy v ruce druhého hráče, plus 2 $\binom{13}{4}$, tj. jeden trumf v ruce druhého hráče, takže pak

$$P(E)\ P(G \mid E) = P(E)\ \frac{\binom{13}{5} + 2\binom{13}{4}}{\binom{15}{5}} = 0.0325 \times 0.9048 = 0.0294\,,$$

tedy pravděpodobnost zabití červeného esa je větší než pravděpodobnost uhrát sto třicet. Vynést červené eso je nevýhodné. Navíc výnos červeným esem může způsobit mariášnickou katastrofu. Při rozložení karet mezi hráče A a M

A: jeden trumf, všech zbývajících sedm zelených, eso a desítka žaludská
M: jeden trumf, všech pět zbývajících červených a zbývající čtyři žaludy

může mít hra následující průběh (čísla znamenají štychy):

1. J vynese červené eso, A zabije trumfem, M nenamaže desítku a prozřetelně vsadí na to, že v talonu nejsou červené.
2. A vynese zelenou desítku, M zabije trumfem, J dá plonkové zelené eso.
3. M vynese červenou desítku, J dá osmu, A namaže žaludskou desítku.
4. M vynese červeného krále, J dá svrška, A namaže žaludské eso.

[1] K výpočtům byl užit Matlab, hodnoty pravděpodobnosti jsou uváděny na čtyři desetinná místa, ale počítány s větší přesností. Zde nejde o chybu v násobení dvěma, ale o zaokrouhlování.

Zbylé štychy sice dohraje J na trumfy, ale dosáhne jen 70 proti stovce dosažené soupeři. Takový výsledek znamená nejen značnou finanční ztrátu, ale také dokonalou mariášnickou potupu. Pravděpodobnost takového rozložení karet je sice malá (2 / 184756), ale kladná. Při výnosu červenou osmou nic takového nastat nemůže, hráč J i při obzvláštní nepřízni osudu uhraje aspoň devadesát. Zcela bezpečný postup hry je vytrumfovat a pak pustit soupeřům dva štychy na malé červené. Tak hráč J uhraje sto zcela jistě, ale pravděpodobnost vyšší výhry je menší než při výnosu červenou osmou v prvním štychu a kromě toho tento postup postrádá sebemenší náznak zábavnosti.

Intuitivními argumenty se nám nepodařilo změnit bludný názor hráče A. Hrozil jsem mu, že jeho tmářství vyvrátím vědecky a za trest ho budu pranýřovat. Takže i druhou část hrozby teď splním. Vzhledem k dlouhodobému přátelství a zásluhám hráče A o úroveň naší hry tarokŭ i mariáše (popsaný úlet je výjimkou) se mi příčí ostouzet jeho jméno v tisku. Hrozbu tedy plním nepřímo:

- Označení hráčů A, M, J odpovídá počátečním písmenům jejich křestních jmen.
- Křestní jméno jednoho z hráčů je shodné s příjmením známého českého malíře.
- Příjmení hráče A se shoduje s jedním ve statistice velmi známým pseudonymem.
- Podobu i profesi hráče A můžete odhalit v cyklu ostravské televize o bydlení.

Abych nekončil hnusným prásknutím kamaráda, tak ještě dodatek. Intuice je někdy ošidná. Tak např. pravděpodobnost rozdělení dvou trumfů mezi dva hráče není 1/2, po dosazení do rovnice (1) totiž dostaneme hodnotu 0.5263. $P(G \mid E)$, tj. pravděpodobnost, že při pěti červených v jedné ruce bude mít druhý hráč aspoň jeden trumf, není 3/4, ale 0.9048, což je o dost více. Znalost rozdělení pravděpodobnosti je samozřejmě užitečná pro řešení rozhodovacích problémů, ale bohužel málokdy je nalezení rozdělení tak jednoduché jako v popsané hře. Intuice je často jediným použitelným nástrojem pro rozhodnutí, která zásadně ovlivňují náš život. Přeju dobrou intuici.

Poděkování: V tomto výzkumném projektu si každý hráč hradil své náklady sám. Děkuji.

Addresa: Doc. RNDr. Josef Tvrdík, CSc., Katedra informatiky a počítačů, Fakulta přírodovědecká, Ostravská univerzita.
✉tvrdik@osu.cz

ÚVOD DO PROBLEMATIKY DATA MININGU

Petr Klímek

Úvod

Mnoho organizací kumuluje ve svých databázích data, avšak co skutečně potřebují, jsou informace. Využití informace, uložené v datových souborech, kterou může organizace využít pro zvýšení konkurenceschopnosti, vyžaduje nejen nové nástroje, nové techniky, ale i nový způsob myšlení. Informace jsou potom podkladem pro získávání znalostí. Získávání znalostí dat v databázích nazýváme dolování dat (data mining). Vhodným zdrojem je datový sklad (data warehouse – DW). Data mining můžeme chápat jako součást procesu objevování znalostí v databázích (knowledge discovery in databases – KDD). KDD je chápán jako proces netriviálního objevování implicitních, dopředu neznámých a potenciálně použitelných vzorů z dat. Oproti tomu **dolování dat (DM) je pouze krokem v procesu KDD** založeným na aplikaci výpočetních technik, které na základě daných omezení (výpočetní efektivnost) poskytují enumeraci vzorů či modelů nad danými daty.

Existují i další, alternativní názvy, např. dolování znalostí z databází, extrakce znalostí, získávání informaci z dat, archeologie dat, bagrování znalostí, analýza dat apod. Objevování znalostí vede k extrakci zajímavých zákonitostí či informací vyšší úrovně, které mohou být studovány z dalších úhlů pohledu. Tento úkol je ve své podstatě interaktivní a iterativní.

Proces KDD zahrnuje podle [3] tyto fáze:

- *selekce* – data se vybírají nebo segmentují podle nějakého kritéria. Selekcí je například omezení všech osob na ty, které vlastní automatickou pračku. Pro některé algoritmy DM stačí selekcí vybrat pouze vzorky dat, není nutné zapojit do zpracování celý datový sklad;
- *předzpracování* – znamená čištění dat, kdy některá data jsou odstraňována, protože nejsou potřebná a bránila by efektivnímu vyhodnocení dotazu. Například při objevování znalostí o porodnosti je možné uvažovat z registru pacientů pouze ženy, a není tedy nutné přejímat atribut pohlaví. Součástí čištění je také úprava formátů dat, např. kód pohlaví se unifikuje na binární atribut s hodnotami 0, 1;

- *transformace* – nejsou přenášena pouze vyčištěná data, ale jsou rozšířena o další atributy např. z externích zdrojů (demografické atributy), které obohatí použitelnost dat;

- *dolování dat* – jde o stádium, které se zabývá extrakcí vzorů dat. Zde vybíráme vhodnou techniku DM (klasifikace, regrese, shlukování, neuronové sítě apod.). Dále v této fázi vybereme konkrétní algoritmus pro řešení DM úlohy. Nakonec tento krok obsahuje vlastní vyhledání zajímavých znalostí, jejichž forma závisí na zvolené metodě DM a může mít podobu klasifikačních pravidel nebo stromů, funkčních závislostí, logických pravidel atd.;

- *interpretace a vyhodnocení* – vzory identifikované systémem jsou vyhodnoceny jako znalosti, které mohou být použity k podpoře rozhodnutí managera. Rozhodování je vztaženo k úlohám týkajících se predikce, klasifikace apod. tak, že je sumarizován obsah databáze nebo jsou vysvětleny pozorované jevy.

Celý proces je interaktivní, tj. řízený uživatelem a využívající jeho schopnosti a znalosti. Apriorní, předem známá fakta hrají klíčovou roli především v přípravě dat. Každá databáze je připravována s určitým cílem. Uživatel má tedy alespoň přibližnou představu o tom, jaká data jsou v ní obsažena a jaký typ znalostí by pro něho mohl být užitečný. To ovšem neznamená, že by první fáze byla nepodstatná nebo jednoduchá. Na vhodné volbě cílů a přípravě dat často závisí úspěch celého KDD procesu. Proto je tento proces často iterativně opakován. Již získané znalosti pomáhají lépe specifikovat cíle a metody při opakovaném hledání. Situaci navíc často ztěžuje heterogenní prostředí. Různé druhy dat jsou uchovávány v různých typech databází – relačních, objektových, deduktivních, aktivních, hypertextových a multimediálních, časových, prostorových a jiných. Zajímavou výzvou je také hledání znalostí v distribuovaných databázích, například v prostředí Internetu.

1. Co je data mining?

Jak definovat pojem data mining? Obecná definice popisuje data mining jako proces výběru, prohledávání a modelování ve velkých objemech dat, sloužící k odhalení dříve neznámých vztahů mezi daty, za účelem získání konkurenční výhody.

1.1. Data mining v širším slova smyslu

DM je definován jako získávání dosud neznámých, ověřených a použitelných znalostí z rozsáhlých databází pro provádění klíčových manažerských rozhodnutí. Příklady obchodního zadání, které mohou vést k zavádění technik DM jsou:

- *klasifikace* (např. „Představuje toto hlášení o škodní události pojišťovací podvod?“);
- *odhad* (např. „Jaká je obchodní hodnota zákazníka?“);
- *předpovídání* (např. „Kteří zákazníci od nás pravděpodobně odejdou v průběhu nejbližších šesti měsíců?“);
- *analýza nákupního košíku* (např. „Které produkty se obvykle kupují společně?“);
- *seskupování podle podobností* (např. „Které skupiny zákazníků mají nějaké společné charakteristiky?“);
- *deskripce* (např. „Které atributy nejvíce charakterizují chování určité skupiny zákazníků?“).

1.2. Data mining v užším slova smyslu

Definice DM se někdy zužuje na využití nástrojů disponujících pokročilejšími analytickými technikami jako:

- *rozhodovací stromy*,
- *analýza asociací*,
- *vyhledávání shluků*,
- *umělé neuronové sítě a další.*

1.3. Další definice DM

- Analytický proces navržený k prozkoumání velkých objemů dat (většinou z oblasti výroby a obchodu) s cílem nalézt a ověřit konzistentní znaky a/nebo systematické vztahy mezi proměnnými. Proces se skládá ze tří etap: průzkumu, vytváření modelu či rozpoznávání obrazců a verifikace. [1]

- Proces nalezení smysluplných nových korelací, obrazců a trendů prozkoumáváním velkých objemů dat uložených úložištích s použitím metod rozpoznávání obrazců, statistických a matematických metod. [2]

2. Požadavky na dolování dat

V praxi je důležité mít možnost vyhodnotit různé produkty, které se pro data mining nabízí. Z hlediska technologií IS/IT je možné charakterizovat následující požadavky na data mining:

- *možnost práce s různými typy dat*,

  Metody data miningu zahrnují algoritmy pracující s různými datovými typy. Přestože nejčastějšími aplikacemi jsou systémy dolování nad relačními databázemi, s rozvojem možností nových datových typů se objevují jednak složité objekty, jednak data jako jsou texty, obrázky apod. Integrované metody dolování však zatím neexistují, k dispozici jsou spíše metody pro jednotlivé typy dat.

- *efektivnost a škálovatelnost dolovacích algoritmů*,

  Algoritmy dolování jsou většinou časově náročné. Měla by být známa jejich složitost, aby bylo možné odhadnout, jak se budou chovat na databázích různých objemů.

- *dolování z různých zdrojů dat*,

  Možnosti Internetu v dosažitelnosti různých zdrojů dat vedou k novým požadavkům na integraci těchto zdrojů a možnost aplikace globálních požadavků na dolování. To vede ke konstrukci distribuovaných algoritmů pro data mining.

- *ochrana soukromí a utajení dat*,

  Aplikace dolovacích algoritmů na chování zákazníka může vést k získávání a analýze osobních dat. Je důležité studovat, kdy může objevování znalostí narušit soukromí a jaká pravidla vůbec při zacházení s osobními daty aplikovat. Tato pravidla by se měla doplňovat s požadavky na utajení dat.

- *užitečnost výsledků.*

  Přímo s data miningem souvisí také požadavek na užitečnost výsledků získaných dolováním. Objevené znalosti by měly být užitečné pro rozhodování. Pro uživatele by měly být doprovázeny mírami nejistoty, ne-

přesnosti či spolehlivosti. Je žádoucí studovat kvalitu získaných znalostí, definovat užitečnost. K tomu mohou sloužit statistické, analytické, či simulační modely a nástroje.

3. Jakých modelů data mining používá?

Rozlišujeme dva typy modelů, a to **prediktivní**, jejichž cílem je předpovědět hodnoty nějakých atributů na základě již známých hodnot jiných atributů, a **deskriptivní**, které popisují vzory v existujících datech. Mezi disciplíny, na nichž je data mining založen, patří především intuitivní učení, strojové učení a statistika. Z nich také vycházejí modely, které se v data miningu používají. Pro řešení běžných problémů se v praxi velmi často uvažuje 7 typů modelů, jejichž přehled nabízí Tabulka 1.

Tabulka 1: Používané modely v DM a jejich použití

Model	Popis chování	Predikce
klasifikace		x
regrese		x
časové řady		x
neuronové sítě		x
shlukování	x	
exploratorní analýza	x	
asociační analýza	x	

Je nutné zdůraznit, že tento výčet není úplný a v literatuře a softwarových produktech je tento stručný výčet značně rozšířen.

Je zřejmé, že při použití technik DM je nutné použít vhodný software. Zde však musím poznamenat, že úspěšnost data miningu závisí nejen na dobré volbě modelu a statistické metody, ale v prvé řadě na dobré formulaci problému a použití správných dat. [2]

4. Jaké existují softwarové nástroje pro data mining? (nabídka největších firem na Internetu)

Data mining je prudce se rozvíjející oblastí, do které investuje v současné době mnoho softwarových společností. Stav na trhu v oblasti data miningu je tedy poměrně dynamický; pro účely tohoto článku je uvedeno v Tabulce 2 šest největších světových dodavatelů technologií pro data mining spolu s jejich internetovými adresami.

Tabulka 2: Přehled největších světových dodavatelů technologií pro data mining

Dodavatel	**Hlavní produkt**	**Kontakt pro ČR**
Angoss Software Corp.	Knowledge Studio	Speedware společnost s ručením omezeným http://www.speedware.cz
IBM Corporation	DB2 Intelligent Miner for Data	IBM ČR společnost s ručením omezeným http://www.ibm.cz
SAS Institute Inc.	Enterprise Miner	SAS Institute ČR společnost s ručením omezeným www.sas.com/offices/europe/czech
Silicon Graphics, Inc.	Mine Set	Silicon Graphics společnost s ručením omezeným http://www.sgi.cz
SPSS Inc.	Clementine	SPSS ČR společnost s ručením omezeným http://www.spss.cz
StatSoft Inc.	STATISTICA Data Miner	StatSoft CR společnost s ručením omezeným http://www.statsoft.cz

5. Jaké vlastnosti by měl mít software pro data mining?

- *Analytické schopnosti*,

  Všechny ze zmíněných nástrojů pro data mining mimo Mine Set (který neobsahuje např. neuronové sítě) obsahují dnes nejvíce využívané metody pro data mining: rozhodovací stromy, shlukování a modelování s využitím neuronových sítí i řadu dalších algoritmů. Šíře a možnosti parametrizace jsou pro různé produkty různé, produkty podporují i vytváření vlastních modelů.

- *snadnost ovládání a analytické práce*,

  V rámci data mining je vytváření určitého modelu často iterativní a obtížný proces. Při shlukové analýze je např. obvyklé vyzkoušení optimální metody (např. K-means, Kohonenova síť, pravděpodobnostní model) i testování optimálního počtu shluků. Při využití vícevrstvých neuronových sítí se chování může radikálně změnit na základě změny počtu neuronů, způsobu normalizace vstupních dat apod. Tyto a další důvody vedou k požadavku na intuitivní prostředí pro vytváření, správu, propojování a průběžné vyhodnocování modelů a zdrojových i modifikovaných datových sad.

- *konektivita na vstupní data, operativnost práce s datovými sadami*,

  Všechny zmíněné nástroje umožňují pružně importovat vstupní data s využitím různých metod vzorkování, vytvářet jejich podmnožiny a operativně s nimi manipulovat. Podporují dále různé funkce pro transformaci vstupních dat jako filtrování, normalizace, náhrada hodnot, změna distribučních vlastností atd.

- *vizualizace a statistické vyhodnocování dat a výsledků modelů*,

  Profilování, vizualizace a statistické zpracování vstupních dat i výsledků analýz je vždy součástí projektu data mining a podporují ho všechny uvedené softwarové nástroje bohatým způsobem.

- *architektura, platformy, výkon*,

  Uvedené softwarové produkty pracují jak na platformách MS Windows, tak jiných systémech typu UNIX i dalších. Tyto produkty pracují jak v jednopočítačovém režimu, tak v režimu s vyhrazenými mining servery a jejich výkon je tak dobře škálovatelný.

- *cena a licenční politika, lokální podpora.*

  Standardní startovací cenové relace uvedených softwarových produktů pro data mining jsou typicky v řádech začínajících desítkami tisíc USD. Knowledge studio je výjimkou a může být z hlediska ceny nejpřístupnější variantou. Všechny uvedené nástroje disponují v České republice lokální podporou.

Závěr

Data mining je zcela jistě novou, rychle se rozvíjející disciplínou, která nejen dává prostor ambiciózním výzkumným projektům, ale je již prakticky využívána v řadě oblastí. Ve světě obchodu je nejúspěšnější aplikací data miningu tzv. databázový marketing. Jedná se o způsob analýzy zákaznických databází, dovolující vyhledávat současné i budoucí preference zákazníků. Uvádí se, že s jejím použitím lze zvýšit prodej až o 20 %. Řada investičních společností využívá data mining metody k analýze finančních a akciových trhů. Data mining se uplatňuje i při detekci a prevenci pojišťovacích a daňových podvodů.

Z výše uvedeného je jasně patrný značný význam data miningu pro získání konkurenční výhody podniku. Použití metod (zejména statistických), které data mining využívá, by se zcela jistě mělo stát součástí tvorby znalostí v podniku. Metody data miningu se však neuplatňují pouze v ekonomické sféře. Tyto metody se používají rovněž i v řadě oborů a vědeckých aplikací jako jsou astronomie (automatická identifikace hvězd a galaxií), biologie (vyhledávání molekulových struktur), meteorologie (predikce a odhalování globálních klimatických změn) aj.

Literatura

[1] Adriaans, P., Zantige, D.: *Data Mining.* Addison-Wesley, Harlem 1996.

[2] Berry, M. J. A., Linoff, G. S.: *Data Mining Techniques.* Wiley, New York, 1997.

[3] Fayyad, U. M., Piatetsky-Shapiro, G. a kol. (eds.): *Advances in Knowledge Discovery and Data Mining.* MIT Press, Cambridge MA, 1996.

Addresa: Ing. Petr Klímek, Ph.D., Ústav informatiky a statistiky,
Fakulta managementu a ekonomie, Univerzita Tomáše Bati ve Zlíně.
✉klimek@fame.utb.cz

Všichni členové společnosti jsou srdečně zváni na

17. výroční schůzi České statistické společnosti

která se bude konat ve čtvrtek, 2. února 2006, od 13 hodin ve staré budově VŠE, nám. W. Churchilla, Praha 3, místnost č. 260. Na programu budou zprávy o činnosti a hospodaření společnosti, diskuze a odborná přednáška.

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. ISSN 1210 – 8022.
Předseda společnosti: Prof. RNDr. Jaromír Antoch, CSc., KPMS MFF UK Praha, Sokolovská 83, 186 75 Praha 8, e-mail: jaromir.antoch@mff.cuni.cz
Redakce: Doc. RNDr. Gejza Dohnal, CSc. a Mgr. Pavel Stříž;
e-mail: gejza.dohnal@fs.cvut.cz a striz@fame.utb.cz