Data mining: Hledání znalostí v databázích

Luboš Popelínský
Katedra teorie programování, Fakulta informatiky
Masarykova Universita, CZ - 602 00 Brno
Email: popel@fi.muni.cz

Abstract

We give a summary of the area of knowledge discovery in databases(KDD). After introductory part we specify the notion of knowledge in KDD and briefly explain reasons for KDD boom. We introduce the basic paradigm of inductive learning and discuss the role of visualization and statistics in KDD. After sumerizing the typical KDD tasks and after the list of the promising applications we introduce three systems, *C*4.5, *DBLearn*, and *CLEMENTINE*. *C*4.5 is a typical machine learning program for the efficient synthesis of decision trees. *DBLearn* extends relational DBMS by learning facilities. *CLEMENTINE* is an integrated tool which consists of data manipulation programs, visual programming, graphs as well as learning algorithms, neural networks and C4.5. After the conclusion we offer the list of electronic journals and archives relevant to KDD.

1 Data mining

Termín "hledání znalostí v databázích" [8, 9] se objevuje stále častěji jak v diskusích odborníků – databázových teoretiků i praktiků – tak i na stránkách neodborného tisku. Odhlédněme od často propagačních cílů těchto diskusí a podívejme se bez předsudků, ale též bez přehnaných očekávání, na tuto velmi zajímavou oblast databázového výzkumu i databázové praxe. V anglické literatuře se kromě výrazu "knowledge discovery in databases" používá nejčastěji data mining, ale též knowledge extraction, information harvesting nebo data archeology. Stejně často jako data mining můžeme číst nebo slyšet zkratku KDD, kterou budeme i my v dalším používat.

Znalosti

Znalosti byly v databázích hledány, a občas i nacházeny, od okamžiku instalace prvního databázového systému. Znalostí v tomto kontextu rozumíme takovou informaci, která není explicitně uložena v databázi, ale kterou lze na základě dat z databáze odvodit. Za znalost nepovažujeme většinou výsledek databázového dotazu, ale spíše přijatelné zobecnění tohoto dotazu. Pohybuje-li se například váha novorozených dětí v databázi mezi 2.9 a 5.1 kg, nemusí to jistě platit pro celou populaci nově narozených dětí. Z kvalitních dat však mohu získat znalosti o průměrné váze nebo o závislosti váhy na pohlaví.

Boom

Od začátku 90. let, kdy většina lidských činností má svůj stínový obraz v některé z databází, přestávají tradiční metody neustále se zvětšujícímu objemu dat v databázích stačit. Začínají se objevovat prototypy počítačových programů, které se specializují na podporu vyhledávání znalostí v databázích. Mnohdy jde jen o nový make-up nebo jen o nový vývěsní štít. Zákazník, není-li zrovna sběratelem kuriosit, nezíská nic. Mnohdy se však ukáže, že i "jen" zručným poskládáním běžně používaných statistických programů a úhledným zabalením do přátelského uživatelského rozhraní vznikne zajímavý a užitečný nástroj.

Induktivní učení

Kromě statistických metod se velice často v systémech KDD objevují neuronové sítě a/nebo metody induktivního učení (inductive learning) [1]. Metodami induktivního učení dokážeme shromáždit podobné objekty (učení bez učitele, unsupervised learning) nebo najít obecný popis třídy, tj. specifikované části databáze (učení s učitelem, supervised learning). Ten je zpravidla vyjádřen ve výrokové logice ve tvaru

if $Atribut_1 = hodnota_1$ **and** $Atribut_2 = hodnota_2$ **and** ...
then $Trida = jmeno\_tridy$

Metody induktivního učení můžemem použít i tehdy, když klasické statistické metody selhávají. Nepotřebují totiž žádné další informace o pravděpodobnostním rozložení dat. Stačí jen označit ty záznamy, které patří do třídy (pozitivní příklady) a ty, které do ní patřit nemají (negativní příklady). Základními induktivního učení operacemi jsou operace **generalizace** a **specializace**.

Hledáme-li popis dat, začínáme často od nejobecnějšího popisu (tj. předpokládáme, že všechny příklady patří do dané třídy), který poté specializujeme. Přidáváme vhodné dvojice $Atribut = hodnota$ tak dlouho, až je popis dostatečně charakteristický, tj. platí pro většinu pozitivních příkladů a naopak neplatí pro většinu negativních.

Vždy musíme rozlišovat mezi reálně existujícími závislostmi a takovými, které plynou z omezeného počtu příkladů v databázi. To je ovšem problém každého KDD systému. Programy induktivního učení jej řeší tak, že množinu příkladů rozdělí na dvě části: na učící množinu, která je zpravidla větší a na níž je nalezen charakteristický popis, a na testovací množinu, která slouží pro ověření přijatelnosti nalezeného popisu. Níže představíme program $C4.5$, který využívá dalších metod pro nalezeni nejlepšího popisu.

... a ti druzí

Ti druzí jsou mnohdy těmi prvními, a (možná) z komerčního pohledu důležitějšími. Standardní metody deskriptivní statistiky – výpočet minima a maxima, průměru či směrodatné odchylky, korelační a regresní analýza – by měly být součástí každého obecného KDD systému. Systém s obchodními ambicemi se neobejde bez kvalitního uživatelského rozhraní. A předem ztracený je takový program, který neposkytuje širokou škálu různých grafů a diagramů. Jako samozřejmost se jeví existence rozhraní KDD systému pro běžně používané databáze a schopnost zpracovávat reálné objemy dat. Bohužel ne vždy je to tak. Myslíte, že to autorům takových "KDD" systémů vadí?

SKICAT

Kromě obecných prostředků pro data mining vznikají též programy specializované na některou konkrétní databázi. *SKICAT* (Sky Image Analysis and Cataloging Tool) [9] vyvinutý v Jet Propulsion Laboratory a na kalifornské technice slouží k automatické katalogizaci objektů (hvězd, mlhovin, galaxii a artefaktů) z digitalizovaných snímků z palomarské observatoře. Dosahuje 94% přesnosti, je 10-100 rychlejší než jiné metody a dokáže najít a katalogizovat až 3x více objektů. Především však odstraňuje namáhavou a nezajímavou práci s gigabyty dat. A co je nejdůležitější, je opravdu používán. Takové programy právem patří do této třídy programů, i kdyby při jejich vývoji nikdo slova data mining nepoužil. Podle [4] sem patří i Meta-Dendral, který sloužil pro automatickou tvorbu báze znalostí expertního systému Dendral pro hmotnostní spektroskopii.

Typické úlohy KDD

Identifikace tříd, tj. homogenních podmnožin dané databáze je typickou úlohou KDD. V databázi zaměstnanců můžeme například najít takové, jejichž náklady na zdravotní péči jsou "příliš" vysoké. **Popis** takové **třídy**, další z klasických úloh KDD může vypadat takto:

if `věk>52 a nedobrovolný_kuřák` **then** `příliš_nemocen`

Analýza závislostí vyhledá takové závislosti mezi jednotlivými položkami databáze, které jsou zajímavé, například mezi vzděláním, pohlavím a platem. **Detekce odchylek** upozorní na významnou změnu, k níž došlo v obsahu databáze, ať je to zvýšený zájem o typ zubní pasty nebo volební preference v daném místě.

Aplikace

KDD systémy jsou dnes používány (nebo alespoň testovány) mnoha firmami i v méně poetických oblastech než je astronomie. Může to být analýza nákupů v supermarketech, hledání odchylek v nákladech na lékařskou péči (GTE), výběr akcií (Lockheed), detekce podvodů s kreditními kartami (Eurocard Nederland) nebo vyhledávání nadějných zákazníků, jimž jsou zasílány informace o novinkách (Microsoft). Patří sem i systémy pro detekci závad při výrobě elektronických součástek (Texas Instruments) nebo součástek amerického raketoplánu (NASA).

■

V druhé části popíšeme několik obecných prostředků pro KDD. Nebudeme se tedy zabývat těmi systémy, jež byly vyvinuty pro data mining v některé specifické databázi. Spíše si všimneme typických rysů a metod. Jedná se vesměs o komerční systémy.

2 Systémy pro vyhledávání znalostí

*C*4.5 [7]

je typickým představitelem algoritmů induktivního učení. Umí pracovat jak s nenumerickou informací, tak s přirozenými i reálnými čísly. Výsledkem je rozhodovací strom (decision tree), který může být převeden do množiny pravidel **if** *Podminka* **then** *Trida*. V podmínce se mohou vyskytovat i nerovnosti. *C*4.5 začíná pouze s výřezem (okénkem) dat, které postupně zvětšuje. Výpočet opakuje několikrát a z dosažených výsledků vybere nejcharakterističtější. Program je napsán v jazyku *C* a je dodáván ve zdrojovém kódu. Nevýhodou je neexistence databázového rozhraní. Neobsahuje žádné další statistické ani zobrazovací metody. Obsluha je velmi jednoduchá a vlastní výpočet je velmi rychlý. Zkušenější uživatel si může pomocí parametrů systém vyladit na míru dané aplikace. *C*4.5 byl úspěšně použit v řadě aplikací [6].

DBLearn [3]

je vyvíjen na několika kanadských universitách pro kanadskou vládu. Vychází od vlastního DBMS, který rozšiřuje o KDD prostředky. Má rozhraní k SyBase. Je impementován c *C* pro počítače SUN. Uživatel zapíše svůj požadavek ve tvaru podobném SQL. Tento požadavek je přeložen do SQL, vyhledaná data jsou zobecněna a zobrazena uživateli. *DBLearn* zpracovává dva typy KDD dotazů, charakteristické dotazy pro popis specifikované části databáze a porovnávací dotazy pro porovnání dvou částí databáze. Zápis dotazu na *"charakteristické pravidlo pro granty udělené počítačovým disciplínám na základě dat z relací* grant, organizace, typ_grantu *vzhledem k atributům* kolik_peněz, region" bude zapsáno

learn characteristic rule
for "Získané_granty"
from grant G, organizace O, typ_grantu T
where
O.kód_organizace = G.kód_organizace **and**
G.kód_grantu = T.kód_grantu **and** G.disciplina = počítače
in relevance to kolik_peněz, region

Výsledkem bude popis získaných grantů vyjádřený pomocí atributů kolik_peněz a region.

CLEMENTINE

Autor prohlašuje, že není nijak podporován firmou ISL. Věnuje-li produktu ISL velkou pozornost, je to jen díky dobrým zkušenostem s jejím programem. *CLEMENTINE* je integrovaný nástroj vyvinutý společností Integral Solutions Ltd. z Velké Británie. Firma má více než 400 zákazníků, z nichž téměř třetina je ze zahraničí. Kromě KDD firma produkuje software pro umělou inteligenci a pro řízení výroby elektronických obvodů.

CLEMENTINE se skládá z modulů pro vytřídění a předzpracování dat, přičemž interface na běžné relační databáze je k dispozici. Data mohou být graficky zobrazena několika způsoby. Součástí *CLEMENTINE* je neuronová síť a program *C*4.5. Cílem bylo vytvořit KDD systém se snadným přístupem k datům v databázi a exportem výsledků, s možností snadného a přehledného prohledávání databáze a s bezbolestným využitím neuronových sítí a induktivního učení. Přitom obsluha neměla být složitější než u spreadsheetu. Využitím vizuálního programování se podařilo splnit i tento cíl.

"Program" v systému *CLEMENTINE* je na předchozí stránce. Pomocí myši si vyberete ikonu. Začínáte s ikonou pro import dat (DRUG1). Chcete-li některé záznamy nebo některý atribut vynechat nebo sloupce pojmenovat, zvolíte ikonu ze zásobníku pro manipulaci s daty (Name Fields). Spojíte ji s předcházející ikonou a definujete vámi požadovanou akci. Po vykonání akce můžete zobrazit výsledek v textové nebo grafické formě opět výběrem ikony a jejím připojením do grafu. Můžete též vytvořit nové atributy z existujících (Na_to_K) zadáním aritmetického výrazu. Před použitím učících algoritmů jste se rozhodli některé sloupce zrušit. Použijete tedy opět ikonu pro manipulaci s daty (Discard Fields). Před vlastním učením musíte jen definovat typy jednotlivých sloupců (Define Types). Pak ještě připojíte pětiúhelníkové ikony pro neuronovou síť (Drug Net) a pro syntézu pravidel pomocí *C*4.5 (Drug Rule) (vpravo od Define types). Po chvíli se vám objeví ikony s výsledkem učení. Chcete-li porovnat výsledky obou algoritmů, připojíte tyto ikony k (Define Types) a přidáte ikonu pro analýzu (analysis).

CLEMENTINE je implementována v jazyku *POP* a *Prolog* a dodává se pro počítače SUN. Implementační jazyky na první pohled exotické dovolují velice snadno systém rozšiřovat o moduly napsané v obou implementačních jazycích a v jazyku *C*.

3 Závěr

Cílem KDD systémů je to, co by každý zručný statistik obeznámený s obsahem databáze a způsobem získávání dat zvládl, jen kdyby měl dost času. Dá se předpokládat, že se další vývoj bude ubírat dvěma směry. První, vycházející z induktivního učení, bude hledat cestu k databázím, a to především zefektivněním algoritmů a jejich přizpůsobení velkému množství dat. Druhý směr bude rozšiřovat existující DBMS o učící algoritmy. KDD jako oblast společného zájmu teoretiků i praktiků v současnosti zažívá boom. Nechme se překvapit, jak dlouho bude trvat, a těšme se, co zajímavého KDD ještě přinese.

4 Kde hledat další informace o data mining

Základní četbou jsou sborníky z KDD workshopů, které se od roku 1991 až do loňska konaly jako součást amerických konferencí o umělé inteligenci. Letos koncem srpna se v Montrealu konala první konference o KDD. Součástí letošní evropské konference o strojovém učení ECML'95 v Heraklionu byl i workshop věnovaný úloze statistických metod a učení v KDD. V prosinci v Singapuru navazuje na konferenci o deduktivních a objektově orientovaných databázích DOOD'95 workshop o KDD. Napřesrok se ve Vídni jako součast evropského informatického symposia bude konat konference o KDD. Tento rok vychází též dvě monografie věnované KDD [5], [2]. Kromě toho existuje několik elektronických časopisů a elektronických informačních serverů, jejichž přehled uvádíme.

Elektronické časopisy

KDD Nuggets

vydává Grigory Piatetski-Shapiro, GTE Labs. USA jednou za 2-3 týdny. Je základním fórem pro výměnu nejnovějších informací o KDD. Objednat lze na adrese `kdd-request@gte.com`. Stará čísla najdete na anonymním ftp `ftp.gte.com` v adresáři `pub/kdd/nuggets/`.

Machine Learning List

moderuje Michael Pazzani, Univ. California, Irvine. Vychází jednou za 1-2 týdny a obsahuje informace o konferencích a nových publikacích z celé oblasti strojového učení. Objednat lze na adrese `ml-request@ics.uci.edu`. Stará čísla najdete na anonymním ftp `ics.uci.edu` v adresáři `pub/ml-list/`.

AI and Statistics

moderuje Dave Fischer, Univ. of Waterloo, Canada. Vychází každé 3 týdny. Objednat lze na adrese `ai-stats-request@watstat.uwaterloo.ca`.

Elektronické archívy

Knowledge Discovery Mine, GTE Labs, USA

`http://info.gte.com/˜kdd/` obsahuje neustále aktualizovaný katalog KDD systémů, archiv časopisu KDD Nuggets, seznam relevantních publikací a ukazatele na další informační zdroje.

The Machine Learning Database Repository at Univ. California Irvine

anonymní ftp na `ics.uci.edu`. V adresáři `pub/machine-learning-databases/` najdete data pro testování učících systémů a v adresáři `pub/machine-learning-programs/` vlastní programy.

References

[1] Brůha I.: Machine Learning. In: Sborník SOFSEM'91, Jasná pod Chopkom, 1991.

[2] Fayyad U.M., Piatetski-Shapiro G., Smyth P., Uthurusamy R.(eds.): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press 1995.

[3] Han J., Cai Y., Cercone N.: Knowledge Discovery in Databases: An Attribute-Oriented Approach. In: Sborník konference VLDB'92, str. 547-559, Vancouver 1992, Kanada.

[4] Holsheimer M., Siebes A.: Data mining: the search of knowledge in databases. Technická zpráva CS-R9406, CWI Amsterodam, 1994.

[5] Holsheimer M., Kloesgen W., Mannila H., Siebes A.: A data mining architecture. Vyjde 1995.

[6] Michie D., Spiegelhalter D.J., Taylor C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood Publ. 1994.

[7] Quinlan J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publ. 1992.

[8] Piatetski-Shapiro G., Frawley W.J (eds.): Knowledge discovery in databases. AAAI/MIT Press 1991.

[9] Piatetski-Shapiro G..: Knowledge discovery in databases: from research to development. In: MLnet Workshop on Industrial Applications of Machine Learning, str. 15-32, Dourdan, 1994, France.