

TECHNIKY DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ

METHODS OF KNOWLEDGE DISCOVERY IN DATABASES

Pavel Procházka

Anotace:

Základní informace o procesu dobývání znalostí z databází; Popis metod – neuronové sítě, fuzzy systémy; Výhled do budoucna

Klíčová slova:

informace, znalost, databáze

Abstract:

Basic informations about process of knowledge discovery; Description of used methods – neural networks, fuzzy systems; An outlook for the future

Keywords:

information, knowledge, database

ÚVOD

V současnosti existuje celkem široká řada producentů softwaru kolekce nástrojů, které se často označují souhrnným názvem *systémy pro podporu rozhodování (DSS – decision support systems)*. Poptávka po takových systémech je motivována především zejména snahou managementu obchodních společností opřít svá strategická ale i operativní rozhodnutí o argumenty skryté v často obrovském množství dat shromažďovaných po dlouhá léta v podnikových informačních systémech.

V datech uchovávaných v podnikových databázích, které byly instalovány původně z důvodů vedení účetnictví, řízení skladového hospodářství a automatizace dalších ekonomických či technologických agend, jsou skutečně zaznamenávány projevy chování daného ekonomického, technologického nebo sociálního systému. Jejich analýza pak může napomoci přijetí důležitých strategických rozhodnutí. Zvláště nachází uplatnění při formulování optimální obchodní a cenové strategie obchodní společnosti.

Právě tento článek chce ukázat na používané metody při získávání znalostí z databází s použitím metod umělé inteligence.

ZÁKLADNÍ INFORMACE O PROCESU DOBÝVÁNÍ ZNALOSTÍ Z DATABÁZÍ

Do vědeckého povědomí se pojem „Dobývání znalostí z databází“ dostal na přelomu 80. a 90. let minulého století. První konference na toto téma se konaly v Americe, které se věnovaly umělé inteligenci obecně.

Technologický základ pro zrození „Dobývání znalostí z databází“ byly právě databázové technologie, které představují osvědčený prostředek k uchovávání a vyhledávání informací. Nejenom databázové technologie přispěly k rozvoji toho vědního oboru, další důležitou disciplínou, která ho ovlivnila byla statistika představující osvědčený prostředek, jak modelovat a analyzovat závislosti v datech.

Podle Fayyada(1996) lze dobývání znalostí definovat jako: netriviální získávání implicitních, dříve neznámých a potenciálně užitečných informací z dat. Na rozdíl od standardního použití statistických metod a metod umělé inteligence(také strojového učení) se

v procesu dobývání znalostí z databází již klade důraz i na přípravu dat pro analýzu a na interpretaci výsledných znalostí. Při přípravě dat se obvykle z dat uložených ve složité struktuře, např. datového skladu, vytváří jedna tabulka obsahující relevantní údaje (atributy) o sledovaných objektech.

Cílem procesu dobývání znalostí je získat co nejvíce relevantních informací vhodných k řešení daného problému. V případě dobývání znalostí z databází můžeme mluvit o různých typech úloh. Jde především o

- klasifikaci nebo predikci
- deskripci
- hledání zrnků tzv. nuggetů

Při predikci nebo klasifikaci se snažíme nalézt znalosti použitelné pro klasifikaci nových případů. V tomto případě požadujeme, aby získané znalosti co nejlépe odpovídaly danému konceptu (připouštíme větší množství méně srozumitelných dílčích znalostí tak).

Při deskripci je cílem nalézt dominantní strukturu nebo vazby, které jsou skryté v daných datech. Požadujeme srozumitelné znalosti pokrývající daný koncept (dáváme tedy přednost menšímu množství méně přesných znalostí)

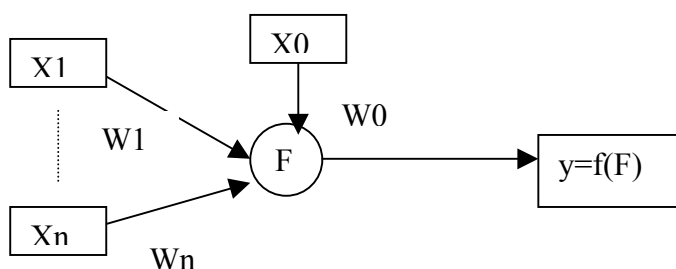
V případě, že hledáme zrnka nebo-li nuggety, požadujeme vyhledání zajímavých (nových, překvapivých) znalostí, které nemusí plně pokrývat daný koncept.

Úloh dobývání znalostí je celá řada např. se může jednat o segmentaci a klasifikaci klientů banky (rozpoznávání problémových a rizikových klientů), predikci vývoje různých veličin, analýzu příčin poruch v telekomunikačních sítích, segmentaci a klasifikaci klientů pojišťovny, určení příčin poruch automobilů, rozbor databáze pacientů, analýzu nákupního košíku, aj..

POPIS METOD – NEURONOVÉ SÍTĚ, FUZZY SYSTÉMY

Neuronové sítě

Umělá neuronová síť je systém sestávající z výpočetních jednotek – neuronů, které jsou mezi sebou vzájemně propojeny spoji ohodnocenými váhami a tímto propojením. Schopností adaptovat tyto váhy – učit se – na základě trénovacích vzorů umožňují realizovat kvalitativně novou funkci implicitně obsaženou v trénovacích datech. Důležitou vlastností neuronových sítí je kromě schopnosti učit se, tedy nacházet závislosti v trénovacích datech a ty reprezentovat pomocí vah, také schopnost zevšeobecňování (generalizace) získaných poznatků, tedy schopnost správně reagovat i na neznámé vstupy, na které nebyla neuronová síť naučena. Umělé neuronové sítě jsou inspirovány poznatky z oblasti neurofyzologie, tedy funkcí nervových buněk a mozku živých organismů



Obr.1 – formální neuron

V současné době existuje řada prací, které se zabývají neuronovými sítěmi, jejich modifikacemi, učícími pravidly a rychlejšími učícími algoritmy včetně řady aplikací. Nejvíce zájmu se soustřeďuje kolem vícevrstvé perceptronové sítě, která je nejznámější a nejrozšířenější. Existují desítky algoritmů učení této neuronové sítě. Tento článek si neklade za cíl podrobně popisovat tyto algoritmy, čtenář se může dozvědět více z použité literatury.

Fuzzy systémy

Anglické slovo fuzzy znamená „roztřepený“; vlastnost je fuzzy, když ji některé objekty mají více a některé méně. Například *tlustý* není žádná přesná hranice mezi tlustými a netlustými lidmi, tenhle je tlustý, ale onen je tlustší. Mohou mít fuzzy pojmy logiku? Co je to tedy je ta fuzzy logika? Tento termín má minimálně dva významy: V širokém smyslu se vžil termín fuzzy logika pro téměř jakékoli metody práce s fuzzy pojmy, množinami apod., např. ve fuzzy řízení. V úzkém smyslu jde o formální systémy podobné klasickému predikátovému počtu, tj. jde o studium vztahu důsledku mezi fuzzy výroky.

Adaptivní, popř. samonastavující se, fuzzy systémy umožňují při návrhu modelu statického či dynamického systému využít zdroje informace systematickým způsobem. Fuzzy modely se mohou stát součástí algoritmu řízení, který pracuje v reálném čase a může sloužit pro analýzu statických a dynamických vlastností modelovaného procesu pro potřeby zdokonalení operativního řízení, monitorování a diagnostiky.

VÝHLED DO BUDOUCNA

V drtivé většině se databázemi, ze kterých se dobývají znalosti, myslí relační databáze. U těchto databází se předpokládá vzájemná nezávislost záznamů z hlediska pořadí v databázi. Existují samozřejmě složitější data: časová (např. časové řady), prostorová (např. data z geografických informačních systémů) nebo strukturální (např. data o chemických sloučeninách). Na druhé straně stojí data nestrukturovaná (např. texty). Spolu s tím se objevují nové oblasti aplikací dobývání znalostí, které řeší své specifické problémy. Někdy stačí mírně adaptovat existující postupy, někdy je třeba zásadně změnit kroky předzpracování a transformace dat, a někdy přicházejí na řadu zcela nové metody.

Novým rozvíjejícím se oborem v této oblasti je např. dobývání znalostí z textů, tzv. text mining, který můžeme chápat jako speciální typ úlohy dobývání znalostí z databází. Zatímco u databází pracujeme s údaji uloženými v pevné struktuře, zde máme co do činění s nestrukturovatelným textem. Hlavním problémem tedy je jak vhodně reprezentovat textový dokument, aby bylo možné použít některý z algoritmů.

ZÁVĚR

Tento článek se snažil čtenáři přiblížit nový vědní obor – dobývání znalostí z databází a chtěl v krátkosti představit některé používané metody z oblasti umělé inteligence.

Tyto metody mají velký potenciál stát se široce používanými v informačních systémech obchodních i neobchodních společností, protože dokáží pomoci při řízení a rozhodování v těchto společnostech.

Literatura:

Mařík M., Štěpánková O., Lažanský J.; *Umělá inteligence 1*; Academia Praha; ISBN 80-200-0496-3
Mařík M., Štěpánková O., Lažanský J.; *Umělá inteligence 2*; Academia Praha; ISBN 80-200-0504-8
Mařík M., Štěpánková O., Lažanský J.; *Umělá inteligence 3*; Academia Praha; ISBN 80-200-0472-6
Mařík M., Štěpánková O., Lažanský J.; *Umělá inteligence 4*; Academia Praha; ISBN 80-200-1044-0
Berka P.; *Dobývání znalostí z databází*; Academia Praha; ISBN 80-200-1062-9

Kontaktní adresa:

Ing. Pavel Procházka, pracoviště: KII, PEF ČZU Praha 6 – Suchbátka, 6. p. rektorát
email: prochazkap@tiscali.cz, tel: +420 605 821 072