MOŽNOSTI VYUŽITIA GENETICKÝCH ALGORITMOV PRI ZÍSKAVANÍ ÚDAJOV Z ROZSIAHLYCH DATABÁZ

THE UTILIZATION POSSIBILITY OF GENETIC ALGORITHMS AT THE KNOWLEDGE DISCOVERY IN LARGE DATABASES

Michal KEBÍSEK, Pavol TANUŠKA

Autori: ***Ing. Michal Kebísek, Ing. Pavol Tanuška, PhD.***
Pracovisko: ***Katedra aplikovanej informatiky a automatizácie, Materiálovotechnologická fakulta STU***
Adresa: ***Paulínska 16, 917 24 Trnava***
Tel.: ***00421 33 544 77 34*** *Email:* ***kebisek@mtf.stuba.sk***

Abstract

This contribution deals with problem of utilization of genetic algorithms at the knowledge discovery in large databases. The process of data mining and its requirements are described too. Genetic algorithms and Neuronal networks are included as techniques and methods.

Tento článok pojednáva o problematike využitia genetických algoritmov pri získavaní údajov z rozsiahlych databáz. V článku je popísaný proces dolovania dát a požiadavky na dolovanie dát. Z používaných metód a techník sú podrobnejšie popísané genetické algoritmy a neurónové siete.

Key words

genetics algorithms, neuronal networks, data warehouse

algoritmy genetické, siete neurónové, sklad dátový

Úvod

Pre koniec dvadsiateho storočia je charakteristický explozívny nárast schopnosti generovať a zhromažďovať dáta. V rôznych oblastiach, ako je veda, priemysel, administratíva, štátna správa, sú dáta zhromažďované a uchovávané vo veľkých objemoch. Každý náš telefonát, nákup, prístup na Internet alebo len návšteva lekára, znamená veľké množstvo vyprodukovaných dát, ktoré vstupujú do transakčných systémov rôznych organizácií. Týmto organizácie získavajú potenciálne bohatstvo, ktoré tvorí základ ich tzv. pamäte. To vytvára naliehavú potrebu vzniku novej generácie metód a nástrojov pre podporu získavania použiteľných informácií a znalostí zo stále rýchlejšie sa zväčšujúceho objemu elektronických dát.

Dolovanie dát (DM)

„Dolovanie dát je špecifický proces získavania dát pre rozhodovanie z veľmi rozsiahlych dátových skladov, a to extrakciou relevantných, vopred neznámych informácií.“ Rovnako ako sa osobné počítače stali účinnými a užívateľsky priateľskými, nové nástroje na dolovanie v dátach boli vyvinuté na získanie výhod z rastúcej sily výpočtovej techniky. Postupy v dolovaní dát vznikajú zákonite ako odpoveď na nové a stále častejšie vyjadrované potreby manažérskych rozhodnutí. Niektoré definície DM sú zakotvené v špecifických analytických postupoch, ako sú neurónové siete, genetické algoritmy a iné. Iné definície DM sú zamieňané s definíciami, ktoré sa týkajú ukladania dát do dátových skladov (data warehousing). Vytváranie dátových skladov a DM sú komplementy. Dátový sklad slúži na uloženie dát, ale neslúži na premenu dát na informácie. Data mining mení dáta na informácie a informáciu na znalosť.

Relatívne nedávne udalosti značne zmenili pohľad na danú problematiku a prelomili mnohé bariéry. V dôsledku toho sa rozšírilo používanie pokročilých analytických postupov, ktoré tu boli mnoho rokov, a samozrejme, tiež rady najnovších postupov. Medzi kľúčové faktory, ktoré teraz umožňujú prevádzať data mining v ďaleko väčšom rozsahu než predtým, patrí:

- lepší prístup k dátam a prístup k oveľa väčšiemu množstvu dát,
- dramatický rast sily výpočtovej techniky, zvlášť v prípade desktopov pracovníkov, ktorí pracujú s informáciami,
- väčší význam vzdelania – väčšina profesionálov má dnes minimálne aspoň základné štatistické vzdelanie,
- dramatické zmeny v použiteľnosti softvéru, vrátane grafických užívateľských rozhraní, sprievod-cov, učebných programov a schopnosť vytvoriť obvyklé rozhrania - všetky tieto nástroje uľahčujú učenie,
- vznik profesionálnych spoločností špecializujúcich sa na metodológiu data miningu a na realizácie projektov tohto typu.

Prostriedky na dolovanie dát používajú dáta na vytváranie modelov reálneho sveta. Výsledkom tohto modelovania je popis vzorov a vzťahov v dátach. Tieto modely sa dajú použiť dvojakým spôsobom. Vzory a vzťahy v dátach poskytujú znalosti, ktoré môžu ovplyvniť následné akcie v aplikačnej sfére (napr. dolovanie dát zo záznamov o predajoch v supermarkete môže viesť k prerovnaniu tovaru v regáloch). Pri druhom type použitia vzorov je ich možné využiť k predpovediam. Napr. z analýzy typov zákazníkov vo vzťahu k ponukám sa dá určiť taká podmnožina zákazníkov, ktorí s veľkou pravdepodobnosťou budú reagovať na písomnú ponuku a týmto rozposlať ponukové letáky.

Požiadavky na dolovanie dát

Z hľadiska technológií IS/IT je možné charakterizovať nasledujúce požiadavky na DM:

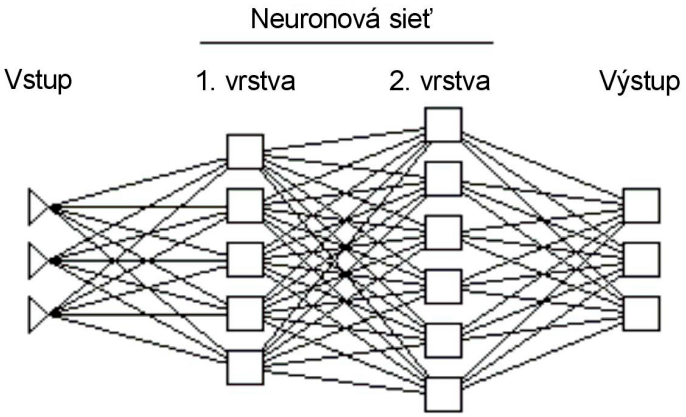
1. možnosť práce s rôznymi typmi dát: metódy DM zahrňujú algoritmy pracujúce s rôznymi dátovými typmi. Napriek tomu, že najčastejšími aplikáciami sú systémy dolovania nad relačnými databázami, s rozvojom možností nových dátových typov v štandarde SQL92 a v štandarde SQL3 sa objavujú jednak zložité objekty, jednak dáta, ako sú texty, obrázky a pod.. Integrované metódy dolovania však zatiaľ neexistujú, k dispozícii sú skôr metódy pre jednotlivé typy dát.
2. efektívnosť a škálovateľnosť dolovacích algoritmov: algoritmy dolovania sú väčšinou časovo náročné. Mala by byť známa ich zložitosť, aby bolo možné odhadnúť, ako sa budú správať na

databázach rôznych objemov. Pretože samotné databázy a DW sú rozsiahle, ťažko sa uplatnia algoritmy majúce v závislosti na veľkosti vstupu exponenciálnu zložitosť.

3. vyjadrovanie požiadaviek na dolovanie a prezentácia výsledkov: pretože uplatnenie dolovaných znalostí z DW je predovšetkým v riadení, požiadavky na znalosti by malo byť možné formulovať v jazykoch vyššej úrovne. Rovnako tak aj prezentácia výsledkov by mala byť prevedená tak, aby bola zrozumiteľná a priamo použiteľná užívateľom. S dolovaním dát je teda neoddeliteľne spätá problematika vizualizácie získaných dát.
4. interaktívnosť dolovania na rôznych úrovniach abstrakcie: keďže nie je dopredu známe, aké znalosti budú získané, mali by mať dotazovacie nástroje možnosť zachovať niektoré zaujímavé cesty v dolovaní na prípadné ďalšie využitie. Do interaktívneho spracovania by mali byť zahrnuté možnosti dynamickej zmeny vzoriek dát, hĺbka dolovania, zmien úrovní abstrakcie, voľby rôznych uhlov pohľadov a podobne.
5. dolovanie z rôznych zdrojov dát: možnosti Internetu v dosiahnuteľnosti rôznych zdrojov dát vedie k novým požiadavkám na integráciu týchto zdrojov a možnosť aplikácie globálnym požiadavkám na dolovanie. To vedie ku konštrukcii distribuovaných algoritmov pre DM.
6. ochrana súkromia a utajenia dát: aplikácia dolovacích algoritmov na správanie zákazníka môže viesť k získavaniu a analýze osobných dát. Je dôležité stanoviť, kedy môže objavovanie znalostí narušiť súkromie a aké pravidlá pri zaobchádzaní s osobnými dátami aplikovať. Tieto pravidlá by sa mali doplňovať s požiadavkami na utajenie dát.

Neurónové siete a genetické algoritmy

Neurónové siete sú údajové modely, ktoré simulujú štruktúru ľudského mozgu. Ako mozog, tak i neurónové siete sa učia z množiny vstupov a dolaďujú svoje parametre modelu vzhľadom na tieto nové znalosti, aby našli schémy v údajoch. Predstavujú novú modernú techniku predikatívneho modelovania s vynikajúcou veľkou variabilitou možných modelov a jednoduchosťou modifikácie ich návrhu. O možnosti vytvorenia umelých neurónových sietí informoval poprvýkrát americký fyziológ McCulloch už v roku 1921. Model umelého neurónu navrhli McCulloch a jeho žiak Pits v roku 1943. Najdokonalejší systém na spracúvanie informácií totiž predstavuje ľudský mozog a umelé neurónové siete sa snažia napodobniť jeho stavbu a funkciu. Vlastnosťami, ktorými sa neurónové siete odlišujú od počítačov, sú schopnosť učiť sa a schopnosť vykonávať svoju funkciu aj vtedy, ak veľká časť neurónov tejto siete je vyradená z funkcie. Táto sieť je tak isto ako mozog zložená z jednotlivých neurónov, ktoré sú prepojené sieťou spojení.



Obr. 1 Neurónová sieť

Základom je koncept umelého neurónu, ktorý napodobňuje činnosť neurónu v ľudskom mozgu. V algoritmoch DM sa používa model viacvrstvého perceptrónu so vstupnou a výstupnou vrstvou. V neurónových sieťach mnoho vstupov generuje výstup, ktorý je nelineárnou funkciou váženého súčtu týchto vstupov. Ak sieť rozpozná príslušnosť vstupného objektu k určitej triede, aktivuje výstup zodpovedajúci tejto triede a ostatné výstupy zostanú pasívne. Váhy priradené každému zo vstupov sú získavané na základe procesu učenia, kedy sú generované výstupy porovnávané s tzv. cieľovými výstupmi (známymi hodnotami). Získané odchýlky medzi známymi hodnotami a získanými výstupmi slúžia ako spätná väzba pre úpravu váh. Tento algoritmus spätnej úpravy váh a prahov sa nazýva spätné šírenie (back propagation).

Používajú sa tam, kde je potrebné pochopiť zložité vzťahy medzi jednotlivými premennými. Typicky je to predikcia stavu určitého systému. Neurónové siete sú nelineárne svojim dizajnom a nepotrebujú explicitne špecifikovať funkcionálny tvar, ako to potrebuje nelineárna regresia. Výhodou je, že nie je potrebné mať na mysli nejaký špeciálny model, keď sa spúšťa analýza. Neurónové siete môžu tiež nájsť interakčné efekty (ako sú efekty z kombinácie veku a pohlavia), ktoré musia byť explicitne vyjadrené v regresii. Nevýhodou je ťažšia interpretácia výsledného modelu s jeho váhovými vrstvami a transformáciami.

Neurónové siete sú užitočné pre predpovedanie cieľovej premennej, keď sú údaje značne nelineárne, ale nie sú veľmi užitočné, keď tieto vzťahy údajov je potrebné vysvetliť. Neurónové siete majú oproti klasickým technikám lepšie výsledky na dátach postihnutým šumom (viď. problémy nasadenia DM). Ich ostatné vlastnosti sú však horšie. Predovšetkým neposkytujú zrozumiteľné vyjadrenie výsledku. Získaná znalosť je implicitne uschovaná vo vektore nastavenia siete, ktorá sa potom správa ako dobre fungujúca čierna skrinka. Proces KDD by však mal poskytnúť človeku užitočnú novú znalosť vo forme zrozumiteľného popisu. Práve prevedenie znalosti uloženej v nastavení neučenej neurónovej siete do čitateľného tvaru je jedným zo súčasných problémov DM metód využívajúcich techniku neurónových sietí. Okrem toho sú neurónové siete výrazne pomalšie pri učení, ako sú napr. rozhodovacie stromy (systém založený na rozhodovacom strome bol pri testoch 500 - 100 000 krát rýchlejší ako neurónová sieť). V súčasnosti neurónové siete, ktoré sme schopní zostaviť, ani zďaleka nemôžu súťažiť s mozgom akéhokoľvek zvieraťa, no napriek tomu majú oblasti nasadenia, kde sú veľmi úspešné. Budúcnosť umelých neurónových sietí bude závisieť od toho, či sa preukáže možnosť vytvárania oveľa komplexnejších sietí, ktoré by boli schopné simulovať ešte zložitejšie myslenie.

GA je metóda kombinatorickej optimalizácie založenej na podobnosti s procesmi v biologickom vývoji, ktorá využíva princípy odvodené z procesu vývoja človeka. Základná myšlienka spočíva v tom, že v evolučnom vývoji prežívajú iba najodolnejšie druhy. Úvodná množina (populácia) popisov (organizmov) postupne vylepšuje svoju kvalitu algoritmom, ktorý vytvára nové popisy zostavené z častí najlepších popisov populácie. Nepoužiteľné popisy zanikajú. Vzniká druhá generácia popisov a algoritmus opakuje svoju činnosť na nej. Tak vznikajú ďalšie generácie, pokiaľ nie je dosiahnutá dostatočná kvalita, alebo pokiaľ sa proces nezastaví na nemožnosti zlepšiť kvalitu populácie. Nevýhodou uvedenej techniky je fakt, že na získanie rozumných výsledkov sa požaduje veľké množstvo učiacich príkladov, typicky 10 generácií po 50-100 organizmoch. To samozrejme predlžuje dobu učenia.

Záver

Rôzne algoritmy a aplikácie sa na prvý pohľad javia ako veľmi rozdielne, ale často zistíme, že spolu zdieľajú množstvo spoločných komponentov. Porozumenie DM a odvodzovanie modelov na tejto úrovni objasňuje správanie iných DM algoritmov a pre užívateľa umožňuje

jednoduchšie pochopenie ich celkového prínosu a použiteľnosti v celom procese získavania údajov.

Neexistuje nijaká univerzálna DM metóda a výber odpovedajúceho najlepšieho algoritmu pre konkrétne aplikácie je tiež veľmi zložitý. Prakticky veľká časť aplikačného úsilia spočíva v dôkladnej formulácii problému, v optimalizácii detailov algoritmu vybranej DM metódy.

Literatúra:

[1] ADRIANS,P., ZANTIGE,D. *Data Mining.* Addison-Wesley, 1998.

[2] FAYYAD,U. Data Mining and Knowledge Discovery in Databases: Implication for Sicentific Databases. Proc. In *Ninth Int. Conf. on Scientific and Statistical Database Management.* Washington, Olympia, Computer Society, 1997.