Oponentský posudek na bakalářskou práci

Dolejšek, J.: Získávání znalostí z databází

Katedra logiky FF UK, září 2013

Cílem práce bylo podat celkový přehled problematiky dobývání znalostí z databází a na praktické implementaci porovnat dvě z používaných metod, tzv. rozhodovací stromy a neuronové sítě. **Zadaný cíl práce splnila.**

Práce je rozdělena do 5 kapitol. V úvodní kapitole jsou uvedeny cíle práce a použité technologie. Druhá kapitola obsahuje stručný úvod do problematiky, rozdělení základních úloh řešených v dané oblasti a náčrt standardizované metodiky CRISP-DM. Třetí kapitola ilustruje na konkrétním příkladu dat z UNIFITTEST testů provedených na FTVS problém čištění dat a předběžné analýzy.

Čtvrtá kapitola je jádrem celé práce. Nejprve zavádí základní pojmy tzv. dataminingu a uvádí stručný přehled některých metod a modelů (rozhodovací stromy, associační pravidla, evoluční algoritmy, bayesovské metody). Další část je věnována rozhodovacím stromům a účícímu algoritmu ID3. Je podrobně popsána jeho praktická implementace a diskutováno jeho použití na konkrétním vzorku dat. V další části jsou pak představeny neuronové sítě, a učící algoritmus backpropagation. Jsou popsány dvě konkrétní implementace (autorova a cizí) a opět je diskutováno jejich použití na konkrétních datech.

Závěrečná kapitola podává shrnutí práce, diskuzi vhodnosti jednotlivých metod a podněty k případné další práci.

Z formálního hlediska práce neobsahuje mnoho překlepů ani gramatických chyb, některé formulace jsou však trochu neobratné (pozn. 12, druhá věta pozn. 15, zakončení pozn 14.: „... definice funkce signum, která pro $x < 0 = -1$“, věta ze závěru: „Neuronové sítě mají relativně náročnější koncept.“). Sazba je pěkná a cenné je i zařazení obrázků. Některým grafům nicméně chybí popisy os (obr. 10, 11) což je činí v podstatě bezcennými. Poněkud zbytečný se zdá jednostranný tisk a tisk příloh, které by bylo vhodnější přiložit na nějakém paměťovém médiu.

Po obsahové stránce je nutno zmínit některé nedostatky. Čtenáři nemusí být úplně jasné, proč byl zařazen oddíl věnující se metodice CRISP-DM. Obrázek, uvedený v tomto oddíle, je pěkný, ale mnoho toho neříká. Název oddílu 4.2 „Stručný přehled metod“ je poněkud zavádějící, v oddíle jsou představeny zejména *modely*, učící metodou jsou pouze genetické algoritmy. Zdá se, že autor práce sám si není rozdílu úplně vědom. Popis algoritmu Apriori je nic neříkající stejně jako odstavec o Neuronových sítích. Oddíle 4.4 o neuronových sítích by velmi prospěl schematický obrázek neuronové sítě. Tyto nedostatky by se daly shrnout pod hlavičku „stylistické neobratnosti“.

Dále práce obsahuje i některé další nedostatky:

- v algoritmu id3 (výpis na str. 18) by měla řádka 31 být přesunuta ven z cyklu; (takto to není chybné, „jen“ velmi neefektivní)
- Zvláštní je poslední věta před oddílem 4.4.1. Je pravda, že použití neuronové sítě je typicky relativně rychlé, nicméně i tak se nedá srovnávat např. s rychlostí rozhodovacích stromů.

- na str. 21 by možná bylo vhodné alespoň zmínit, že $\overrightarrow{x} \cdot \overrightarrow{y}$ je skalárním součinem dvou vektorů.
- na str. 22 by mělo být „Lze dokázat, že algoritmus *skončí*“, nikoliv „Lze dokázat, že algoritmus *je finitní*“.
- na str. 23, poslední rovnost by měla být „$\delta_k = o_k(1 - o_k)(t_k - o_k)$“ místo „$\delta_k = o_k(1 - k)(t_k - o_k)$“.
- na str. 24 by se autor rozhodně neměl přiznávat k tomu, že neví, zda není v jeho implementaci chyba. Nicméně po zběžném přehlédnutí se zdá, že algoritmus je opravdu správně implementovaný.

Navzdory výše uvedeným nedostatkům bych chtěl vyzdvihnout fakt, že autor prací prokázal, že se v této široké problematice orientuje. Navíc považuji praktickou stránku práce za její těžiště. Z tohoto hlediska se dá vytknout snad jen volba dat (jsou dostupná mnohem vhodnější data, viz např. `http://archive.ics.uci.edu/ml/`). Z tohoto důvodu navrhuji práci hodnotit stupněm **velmi dobře** nebo **výborně**.

V Praze dne 27. srpna 2013 *Mgr. Jonathan Verner, PhD.*