Posudek diplomové práce **Šimona Rajčana: Magazine sales prediction**

Předkládaná diplomová práce se zabývá aktuální problematikou dobývání znalostí a jejím využitím v praxi, a to zejména při odhadu prodejnosti vydávaných časopisů. V současné době tyto odhady vycházejí především ze zkušeností odpovědných zaměstnanců jednotlivých nakladatelství a jsou založené na jejich vlastním více či méně subjektivním úsudku. Nesprávné odhady (ať už nadhodnocené anebo podhodnocené) jsou ovšem spojeny s neúměrně vysokými finančními náklady, které se každé nakladatelství z pochopitelných důvodů snaží udržet v rozumných mezích. Vhodným prostředkem by přitom mohl být systém poskytující adekvátní odhad prodejnosti daného časopisu automaticky na základě znalosti dřívějších dat o jeho prodeji.

Cílem předkládané diplomové práce proto bylo prostudovat existující regresní metody a na základě jejich vzájemného porovnání některé z nich vybrat, implementovat a experimentálně otestovat na reálných datech poskytnutých Ruským nakladatelstvím magazínů Burda. Důležitým kritériem při volbě vhodného přístupu přitom měly být jeho paměťové a výpočetní nároky. Nejefektivnější algoritmus, resp. kombinace algoritmů se pak měly stát základem aplikace použitelné pro automatickou predikci prodejnosti vydávaných časopisů. Navrhované řešení mělo podporovat práci s daty uloženými v relační databázi a jeho součástí měla být i odpovídající dokumentace.

Předkládaná práce, bohužel, svůj původní cíl nesplňuje. Diplomant sice ve své práci uvádí základní myšlenky tří modelů (rozhodovacích stromů, vrstevnatých neuronových sítí a k-NN-klasifikátoru) vybraných pro řešení zadané úlohy, práce však postrádá jak jejich přesný popis, tak také podrobnější analýzu jejich funkce a rigorózní porovnání jejich vlastností. Vzhledem k množství a charakteru zpracovávaných dat by navíc bylo vhodnější uvažovat spíše o rekurentních modelech, které při predikčních úlohách využívají zpětné vazby, lépe zobecňují a bývají i robustnější vzhledem k případnému šumu na vstupních datech.

Značné rezervy má předkládaná práce rovněž v experimentální analýze a implementaci zvolených metod. Z celkového množství 303341 posloupností, z nichž každá obsahuje 200 záznamů o prodeji jednotlivých magazínů příslušnými prodejci, použil autor pro testování vybraných modelů pouze 3 takové posloupnosti. Při vyhodnocování testovaných metod navíc diplomant zvolil pokaždé jinou metodiku. Získané výsledky tedy nejsou ani reprezentativní, ani vzájemně porovnatelné. Přiložený software implementovaný pod systémem Matlab počítá jen s předem naučenými neuronovými sítěmi. Ty ovšem běžný zaměstnanec nakladatelství nejspíš nebude schopný jakkoliv modifikovat, např. pro později získaná data či jiný magazín. Matlab přitom podporuje automatické vytváření (učení) nových modelů na základě předkládaných trénovacích dat.

Problematiky obecnějšího předzpracování vstupních dat (např. klastrování zpracovávaných vzorků, korelační analýzy pro uvažované atributy, doplnění chybějících hodnot atributů aj.) se předkládaná práce nedotýká téměř vůbec. Podobně se diplomant nevěnoval ani alternativním (a pro uživatele dostupnějším) prostředkům pro dobývání znalostí a práci s neuronovými sítěmi, např. Weka, Encog. Práce je navíc psaná špatnou angličtinou s velkým počtem překlepů a gramatických chyb, např. "Begging" na str. 19, překlep ve vztahu na str. 9, neodpovídající počet příznaků na str. 22 a 23, a mnoho dalších. Obrázky nejsou opatřené popisky a jejich označení nekoresponduje s odkazem použitým v textu, např. obr. 3.9.1 a 3.9.2 zmiňované na str. 15. Reference uvedené v seznamu použité literatury nejsou úplné a často se týkají neověřených zdrojů.

Vzhledem k výše uvedeným důvodům práce Šimona Rajčana nesplňuje požadavky na diplomovou práci, a proto nedoporučuji uznat ji jako práci diplomovou.

V Praze, 2. 9. 2013

Doc. RNDr. Iveta Mrázová, CSc.
KTIML MFF UK