

Posudek diplomové práce **Vojtěcha Hlaveše**
“Metody srovnávání pravidel získaných z dat”

Předkládaná diplomová práce se zabývá problematikou vzájemného porovnávání kvality pro sady pravidel vygenerované pomocí různých nástrojů pro dobývání znalostí. Oblast dobývání znalostí patří v současné době k nejrychleji se rozvíjejícím oborům moderní informatiky. Důvodem je především pokrok ve vývoji výpočetní techniky umožňující relativně snadné pořizování, ukládání a zpracování velkých objemů dat. Dalším důležitým faktorem je pak dostupnost technologií usnadňujících návrh a vývoj systémů pro dobývání znalostí. Perspektivní aplikační úlohy pocházejí mimo jiné z oblasti marketingu a medicíny.

Algoritmy používané pro generování souborů asociačních pravidel (resp. pravidel klasifikačních a rozhodovacích) většinou nevedou k jednoznačným výsledkům. Standardní míry kvality (např. podpora, spolehlivost a pokrytí) jsou typicky definované pouze pro jednotlivá pravidla a fundované porovnání celých sad pravidel navzájem mezi sebou, bohužel, nepodporují.

Cílem této práce proto bylo zmapovat stávající metody pro měření a vzájemné porovnávání kvality extrahovaných pravidel. Diplomant měl přitom analyzovat přednosti a nedostatky jednotlivých existujících metod, a to jednak teoreticky na základě vlastností těchto metod, tak také rozбором výsledků získaných při použití vybraných metod na reálná data. Práce pak měla být završena návrhem a praktickým ověřením vhodných modifikací jedné či vícero z těchto metod, které by měly posilovat přednosti, resp. oslabovat nedostatky uvažovaných technik.

Ve své práci diplomant navrhl původní metodu pro měření kvality souborů pravidel generovaných různými metodami (4ft-Miner, AQ21, fuzzy-neuronové sítě a klasifikační stromy) a pro vzájemné porovnání těchto souborů. Vycházel přitom z rozšíření měř známých pro práci s (čtyřpolní) kontingenční tabulkou a z tzv. charakteristiky souboru pravidel, která zohledňuje počet objektů, pro které uvažovaná pravidla platí, resp. neplatí, anebo pro ně nedokáže rozhodnout. Rozšíření tzv. ROC prostorů na celé soubory pravidel umožňuje lepší grafické znázornění získaných výsledků.

Navržené techniky uchazeč aplikoval při měření kvality souborů pravidel vygenerovaných výše uvedenými metodami pro čtyři různé sady reálných dat. Vygenerovaná pravidla však autor ve své práci nikde neuvádí a s výjimkou pravidel pro 4ft-Miner je ani není možné jednoduchým způsobem získat. Z textu práce dále není zřejmá nutnost generování nikdy nesplněných pravidel uvažovaná na str. 34 (zřejmě by bylo postačující doplnit pouze kontingenční tabulku). Poněkud nelogicky působí i značení zavedené pro charakteristiku souboru pravidel a (opačné) pořadí os zvolené pro ROC prostory. Práce sama je však až na poměrně řídké výjimky - např. překlepy (index na předposl. ř. ve vztahu (3.5), na 5. ř. na str. 41, na 3. ř. zdola na str. 80), stylistické (na str. 35, resp. 7.-4. ř. zdola na str. 46) a gramatické chyby (9. ř. zd. na str. 41, 4. ř. zdola na str. 69) napsána kultivovaným jazykem a má pěknou grafickou úpravu. Výše uvedené nedostatky by však diplomant byl schopen snadno odstranit.

Závěrem tedy bezesporu lze konstatovat, že práce splňuje svůj původní cíl. Diplomant pronikl dostatečně hluboko do problematiky dobývání znalostí z databází (zejména v oblasti asociačních pravidel, jejich generování a vzájemného porovnávání jejich kvality). Prokázal schopnost samostatně řešit konkrétní úlohy z oblasti návrhu a implementace systémů pro dobývání znalostí. Vlastní výsledky a zkušenosti dokázal objektivně vyhodnotit. Práce Vojtěcha Hlaveše tedy jednoznačně splňuje požadavky kladené na diplomovou práci, a proto doporučuji uznat ji jako práci diplomovou.

V Praze, 14. 5. 2007


RNDr. Iveta Mrázová, CSc.
KSI MFF UK