Oponentský posudek disertační práce
Finding Optimal Decision Trees

Autor disertační práce: **Mgr. Ing. Petr Máša**
Rok dokončení práce: **2006**
Místo obhajoby: **Fakulta informatiky a statistiky VŠE**
Obor: **Informatika**

Předložená disertační práce se zabývá problémem konstrukce optimálních rozhodovacích stromů, tedy problémem, o kterém je všeobecně známo, že je obecně NP těžký (nechci zde zacházet do detailů v jakém smyslu, či diskutovat, v jakém smyslu je NP úplný; podstatné pro nás je, že nelze počítat s tím, že by mohl existovat v reálné praxi použitelný algoritmus, který by konstruoval optimální rozhodovací stromy) a který je, možná právě vzhledem ke své obtížnosti, stále aktuální. Potěšil mě autorův přístup k rozhodovacím stromům, na které se dívá nejen jako na zápis rozhodovacího algoritmu (či na zápis znalostí získaných z dat), jak je v současné literatuře obvyklé, ale jako na nástroj pro reprezentaci a realizaci výpočtů s pravděpodobnostními distribucemi. Tento přístup sice není nový, ale v současné době v odborné komunitě zabývající se dolováním znalostí zcela zapomenutý. Neexistenci jakýchkoliv odkazů na původní literaturu plně omlouvám, neboť se jedná o práce z šedesátých a sedmdesátých let minulého století, které vycházely převážně ve francouzštině (autoři C.F. Picard, B. Bouchon, M. Terrenoire, D. Tounissoux a další) a navíc tyto práce nenazývaly studované struktury rozhodovacími stromy, ale dotazníky (questionnaire).

Musím pochválit i celkovou strukturu práce. První kapitola uvádí rámec použití rozhodovacích stromů jako jednu z technologií pro analýzu dat a získávání znalostí z dat. Druhá kapitola je standardním přehledem toho, co se obvykle o rozhodovacích stromech v současné literatuře uvádí. Již jsem řekl, že mi nevadí neexistující návaznost na práce francouzských matematiků z teorie dotazníků. Mrzí mne však, a to nejen na autora předložené disertační práce, ale na celou komunitu zabývající se tvorbou rozhodovacích stromů pro získávání znalostí, že zcela opomíjejí skutečnost, že výběr proměnné pro přiřazení uzlu dle kritéria maximálního snížení entropie (či maximálního zvýšení informace) popisovaného v předložené disertaci na straně 36 použili již koncem 60. let D. Knuth i S. Guiasu (D.E. Knuth: *Optimal Binary Search Trees*, Acta Informatica, 1971, 14-25; S. Guiasu: *On the Most Rational Algorithm of Recognition*, Kybernetik, 1968, *stránky neznám*). Přestože v současné době prakticky všichni autoři odkazují na Quinlana, dá se odkaz na práce z šedesátých let získat nahlédnutím do „programátorské bible“ D. Knutha *The Art of Computer Programming*, či na osobních stránkách D. Knutha.

Kapitola čtvrtá obsahuje výsledky autorovy snahy formalizovat pojmy týkající se rozhodovacích stromů. Formalizace je navržena takovým způsobem, aby bylo možno studovat vlastnosti jím navrhovaných algoritmů matematickými metodami, tedy tak, aby autor nebyl při posuzování uváděných algoritmů odkázán pouze na hodnocení experimentálních výsledků. Tento aparát mu vhodným způsobem umožňuje zavést na struktuře všech rozhodovacích stromů dvě relace ekvivalence, s jejichž pomocí pak snadno zavádí potřebný pojem optimality, který poukazuje na nejmenší rozhodovací strom v dané třídě ekvivalence. Kdyby znal autor výše zmíněné starší práce o dotaznících, mohl svůj pojem s použitím zavedených pravděpodobnostních distribucí generalizovat tak, že by nehovořil o velikosti stromu, ale například o jeho střední délce (taková generalizace má však smysl jen pro některé aplikace, například pro použití stromů při rozhodování, a proto možná není pro uvedenou práci relevantní).

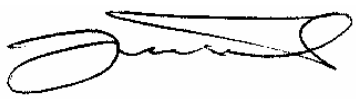
Hlavním autorovým přínosem je návrh operací umožňujících efektivní prořezávání (pruning) neoptimálních rozhodovacích stromů. V práci uvedené teoretické studium vlastností rozhodovacích stromů a především autorem navržených algoritmů ukazuje, že autor je schopný používat přesný formální aparát. Nalezl předpoklady, za kterých jím navržený algoritmus vždy nalezne optimální rozhodovací strom. Samotné teoretické výsledky však příliš nevypovídají o skutečných praktických aplikacích. Jestliže autor ukazuje, že jeho pojem věrnosti či přesnosti (faithfulness – musím přiznat, že příliš nerozumím, proč autor zvolil právě toto slovo) je z teoretického hlediska velice málo omezující (autor ukazuje, že míra všech distribucí, které tuto vlastnost nemají, je rovna 0), pak z praktického hlediska je tomu právě naopak. Uvažujeme-li distribuce, které jsou výsledkem statistických odhadů

získaných z dat, pak naopak budou „věrné“ distribuce vzácné. To však nijak nesnižuje význam zařazení teoretické části do disertace. Pouze podtrhuje důležitost páté kapitoly, ve které je činnost navrhovaných algoritmů ověřována na reálných (bohužel málodimenzionálních) datech. V této souvislosti bych chtěl zmínit, že naopak to, že autor uvažuje pouze distribuce, které nazývá „plné“ (full), nikterak dopad jeho výsledků nelimituje. Sám říká, že se jedná pouze o technický předpoklad, který mu zjednodušuje další vyjadřování. Z mého hlediska se jedná o téměř zbytečné omezení.

Původní výsledky předložené v disertační práci tedy zahrnují originální teoretický aparát umožňující jak popis čistě teoretických vlastností (třídy ekvivalence rozhodovacích stromů), tak popis operátorů umožňujících ve výrazně větší míře prořezávat stromy zkonstruované přímým algoritmem. V tomto aparátu odvozená tvrzení jsou správná a zdůvodňují korektnost autorem navrženého algoritmu pro „věrné“ distribuce. Pro distribuce „nevěrné“ je použití popsaného algoritmu heuristické a nemusí vést k optimálnímu řešení. Dle mého odhadu skutečně ve většině případů pro mnohodimensionální situace také k optimálnímu řešení nepovede. Odpovědím na otázky takovéhoto druhu se autor v disertaci důsledně vyhýbá, nicméně by v průběhu obhajoby mohl odpovědět na otázku, zda si myslí (či dokonce má již v současné době prakticky vyzkoušeno), že jím navržený algoritmus umožní nalézat skutečně lepší řešení, než algoritmy dosavadní. Jiná otázka se může týkat i autorových tvrzení o algoritmické složitosti popisovaných postupů. Jeho tvrzení, přestože si myslím, že jsou správná, nejsou v práci podložena detailnější analýzou. I když to není nikdy explicitně vyjádřeno, předpokládám, že má autor vždy na mysli algoritmickou složitost *nejhoršího případu* (worst case complexity). Z tohoto pohledu se mi zdá zajímavou otázka: je vůbec reálné u problémů takovéto složitosti studovat *střední algoritmickou složitost*? Má již nějaké zkušenosti s použitím algoritmu na skutečně mnohodimensionální data (alspoň 100 veličin)?

Při čtení předložené disertační práce se mi stále vracela jedna myšlenka, kterou zde nemohu nezmínit. Vždycky jsem totiž potěšen, když dostanu k oponování disertační práci napsanou v angličtině. Jedná se o závěrečnou práci dokazující, že se příprava k vědecké činnosti neminula svým cílem. Do vědecké přípravy také zcela samozřejmě patří vybavit uchazeče schopnostmi prezentovat výsledky odborné činnosti na mezinárodním fóru, a to si v současné době nelze představit bez dostatečného zvládnutí odborné angličtiny v mluvené i písemné formě. Proto je také na naprosté většině vysokých škol výuka cizích jazyků a odpovídající zakončující zkouška povinnou součástí studijního plánu. Bohužel tomu tak není na VŠE. A proto si myslím, že když už škola nemá žádné náklady spojené s výukou cizích jazyků v doktorských studijních programech, mohla by alespoň zajistit těm uchazečům, kteří jsou ochotni psát disertaci v angličtině, jazykovou korekturu. Ta totiž předložené disertační práci skutečně chybí.Špatná úroveň angličtiny je největší slabinou předložené práce.

Závěrem bych tedy chtěl shrnout, že cílem předložené disertační práce bylo vybudovat teoretický aparát pro práci s rozhodovacími stromy a navrhnout algoritmus pro konstrukci optimálního (nejmenšího) rozhodovacího stromu pomocí co největšího možného počtu použití operátoru prořezávání. Dále pak bylo cílem ukázat některé teoretické vlastnosti navrženého algoritmu. Těchto cílů bylo jednoznačně dosaženo a, dle mého názoru, tyto výsledky jsou pro disertační práci zcela dostatečné. Jsem přesvědčen, že předložená disertační práce splňuje požadavky na disertační práce kladené zákonem č. 111/1998 Sb. i předpisy VŠE, a proto ji doporučuji k obhajobě před komisí Fakulty informatiky a statistiky VŠE pro obor informatika.



V Nučicích 10. srpna 2006 Radim Jiroušek

Příloha 1

Tabulka hodnocení doktorské práce

(vyplňuje školitel a recenzenti - příloha recenzního posudku)

Jméno doktoranda: …Petr Máša
Název práce: …FINDING OPTIMAL DECISION TREES

Jméno hodnotitele: …Radim Jiroušek

Kritéria hodnocení	*Nepřijatelné*				*Průměrné/dobré*					*Výborné*		
	1	2	3		4	5	6	7		8	9	10
1. Vědecký význam - přínos práce, novost myšlenek a metod							X					
2. Aktuálnost a smysluplnost cíle práce											X	
3. Zmapování stavu zkoumané oblasti na základě světové i tuzemské literatury, analýza stavu (silné a slabé stránky vzhledem k současným požadavkům na řešenou problematiku)								X				
4. Praktický význam (aplikovatelnost výsledků práce)							X					
5. Vymezení zkoumané oblasti							X					
6. Vymezení cíle práce											X	
7. Metodika dosažení stanoveného cíle								X				
8. Definice používaných pojmů										X		
9. Struktura textu a jasnost vyjadřování								X				
10.Formální úroveň práce										X		

Přemýšlel jsem, ve kterém z uvažovaných kritérií se má odrazit špatná úroveň angličtiny. Nakonec jsem ji nezohlednil v žádném.

Datum: 10.8.2006

podpis .