

Oponentní posudek disertační práce Ing. Petra Máši

Disertační práce Ing. Petra Máši nazvaná *Finding Optimal Decision Trees* svým obsahem směřuje do oblasti strojového učení a dobývání znalostí z databází. Přesněji řečeno, je věnována problematice konstrukce rozhodovacích stromů z dat.

Práce je členěna do šesti kapitol a tří příloh a má celkem 141 stran. První kapitola obsahuje stručný úvod do analýzy dat a dobývání znalostí. Kapitola dvě představuje stručný úvod do problematiky rozhodovacích stromů, kapitola tři formuluje problém, na jehož řešení se doktorand zaměřil – jsou zde tedy popsány cíle disertační práce. Jádrem disertace jsou kapitoly 4 a 5. Čtvrtá kapitola obsahuje formální definice, věty včetně důkazů a samotný algoritmus, kapitola pátá experimentálně srovnává navržený algoritmus s algoritmem standardním. Kapitola 6 shrnuje přínosy práce a naznačuje směry dalšího výzkumu. Jednotlivé přílohy pak doplňují práci o terminologické informace i o detailní výsledky testování autorova algoritmu. Je třeba ocenit, že doktorand zvolil obtížnější cestu sepsání disertace v anglickém jazyce. To ale možná bylo příčinou ne vždy dostatečně precizního způsobu vyjadřování.

Hlavním vědeckým výsledkem prezentovaným v práci je návrh algoritmu pro prořezávání rozhodovacích stromů tak, aby výsledný strom byl optimální (z hlediska počtu listů) aniž by došlo ke zhoršení přesnosti klasifikace (pro distribuci). Jako dvě základní operace prořezávání autor navrhuje náhradu podstromu listem a záměnu rodiče s potomkem. Autor se věnuje teoretickým vlastnostem algoritmu (otázce složitosti i schopnosti nalézt optimální strom), navržený algoritmus je rovněž testován na několika datových souborech. Algoritmus je přitom chápán jako „doplněk“ k nějakému existujícímu algoritmu pro tvorbu stromů (v experimentech se používá CART).

K předložené práci mám následující připomínky:

- Vzhledem k tomu, že tématem práce je algoritmus prořezávání, měly být zmíněny jiné přístupy – např. prořezávání v C4.5 rovněž umožňuje měnit „vnitřní strukturu“ stromu
- Opakovaně se uvádí, že navržený algoritmus prořezávání může být propojen s libovolným algoritmem pro tvorbu stromů – toto tvrzení je třeba zdůvodnit neboť spojení navrženého algoritmu s algoritmem CART se zdá být zásadní
- V práci není zmíněn způsob implementace algoritmu

Jisté výhrady mám k experimentální části práce, kde je navržený algoritmus testován na datech. Jak uměle generovaná data, tak reálná data o pojistných smlouvách jsou velmi málo atributů. Výsledný neprořezaný strom nebude tedy příliš košatý a navržený algoritmus prořezávání tak plně neprokáže své možnosti. Při testování se navíc navzájem porovnává CART s původním prořezáváním a CART s autorovým prořezáváním, ale chybí informace o výsledcích CART bez prořezávání. Na základě tabulek se totiž zdá, že za některé chyby (v tom smyslu, že nebyl rekonstruován původní strom) může CART jako takový: vzhledem k tomu, že počet chyb ve většině testů klesá s rostoucí velikostí trénovací množiny, CART asi pro málo příkladů nepovažoval některá větvení za dostatečně signifikantní a tudíž vyprodukoval menší strom než bylo potřeba.

V rámci obhajoby mám na doktoranda následující otázky:

- Navržený algoritmus předpokládá práci s binárními vstupními atributy (a je ukázáno jak binarizovat atributy jiné) – co se ale bude dělat s daty, která jsou rozdělena do více tříd?
- Do jaké míry je implementace přenositelná na jiné algoritmy pro tvorbu stromů

Přes výše uvedené připomínky mohu konstatovat, že doktorand prokázal schopnost samostatné vědecké práce a že jeho disertace obsahuje původní vědecké výsledky. Předloženou disertační práci tedy **doporučuji** k obhajobě před příslušnou komisí.

V Praze dne 25.7.2006

Prof. Ing. Petr Berka, CSc.

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: Prof. Ing. Petr Berka, CSc.....

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	

Datum:27.7.2006.....

.....

podpis .