Posudek oponenta na disertační práci
Computational Intelligence Methods in Metalearning
od Mgr. Jakuba Šmída

Práce se zabývá aktuálním tématem metaučení, které doporučí vhodný ML algoritmus pro konkrétní datovou sadu.

Práce analyzuje současné metody metaučení. Identifikuje potenciál ke zlepšení ve zpracování informace o jednotlivých atributech, jejichž využití komplikuje nestejný počet atributů jednotlivých databází. Algoritmy převzaté ze zpracování jazyka apod. nezajišťují některé očekávané vlastnosti.
Autor formalizuje pojem metriky a detailně analyzuje přechod metriky definované na atributech na metriku v prostoru databází. Navrhuje algoritmus Attribute Assignment, který metriku z atributů převede na metriku na databázích. Analyzuje podmínky, za kterých vlastnosti metriky a semi-metriky zůstanou zachovány, uvádí protipříklady kdy zachovány být nemusí.

Pro provedení experimentů bylo třeba analyzovat a předzpracovat velké množství dat ML repository OpenML, odkud také přebírá globální informace o datasetech. V experimentech se autor úzce zaměřuje na ověření jím navrženého aogoritmu a učení parametrů metriky. Na vyšší úrovni volí pro doporučení algoritmus k-NN nejbližších sousedů ($k=17$) bez experimentů s různým k nebo jiným modelem.
V první sadě experimentů porovnává algoritmus využívající jen globálních dat, autorem navržených metod a kombinace obojího při různých metrikách. Uvítala bych porovnání s některým z algoritmů prezentovaných v sekci 4.1., které rozšiřují globální informace podobným směrem.
Druhá sada experimentů hledá pomocí evolučního programování vhodnou funkci vzdálenosti na atributech. Toto učení prozatím nevedlo k lepším výsledkům než nejlepší z původních expertních návrhů, ukazuje ale cestu pro další vývoj.
Z mého pohledu se experimenty příliš rychle zaměřily na velmi obecnou funkci náchylnou k přeučení a pominuly zjevné parametry k ladění: váhu kategoriálního/spojitého selektoru, nastavení vzdálenosti k dummy proměnným.
I přes tyto výhrady experimenty demonstrují užitečnost zpracování ne-globálních atributů a použitelnost navrženého algoritmu.

Stylisticky je práce značně různorodá, oceňuji matematickou přesnost ve větách o metrikách a semimetrikách a přehledné tabulky srovnání výsledků jednotlivých algoritmů. Místy je těžké se v práci orientovat.

Hlavní význam práce vidím v korektní a srozumitelné agregaci informačních vektorů různé délky algoritmem, který bere v potaz všechny atributy a výsledek nezávisí na jejich uspořádání. Dostupnost systému na internetu umožňuje přímé využití vědeckou komunitou.

Práce prokazuje předpoklady autora k samostatní tvořivé práci, obsahuje původní výsledky, které byly publikovány na mezinárodní úrovni. Doporučuji aby předložená práce byla uznána jako práce disertační.

V Praze dne 27.7.2016

Mgr. Marta Vomlelová, Ph.D.
KTIML MFF UK

Detailní připomínky k textu:

- Proč se při normalizaci (sekce 5.4.3.) dělí rozdílem max-min a ne výběrovým rozptylem? Tím by normalizovaná veličina měla příhodnější rozptyl.

- Není dostatečně popsán výpočet při pouze částečném uspořádání (např. uváděná Definice 2.6. Spearman's rank koeficientu předpokládá lineární uspořádání, podobně řádek 13 Algoritmu 5 předpokládá znalost rank pro každý algoritmus).
- V Corollary 4 chybí předpoklad, že selektory pokryjí všechny atributy (jinak nemusí platit coincidence).
- Použité symboly nejsou v práci jednotné a často mají více významů: n (v Algoritmu 6 řádky 6 a 7), σ (v Definici 28 zobrazení atributů do R, v Algoritmu 17 vector distance measure), λ (v Algoritmu 7 funkce, v Definici 22 konstanta).
- Citace ARFF by měla směřovat na Waikato universitu resp. nějaký zdroj.