Jakub Šmíd: Computational Intelligence Methods in Metalearning

Vyjádření školitele

Práce Jakuba Šmída se zabývá problematikou výběru algoritmu strojového učení vzhledem k nové úloze reprezentované datovou množinou. Jde o problém, který spadá do aktuální oblasti strojového učení, tzv. meta-učení zaměřené na dobývání znalostí z dat. Princip řešení spočívá v myšlence, že algoritmy strojového učení se na podobných datech chovají podobně. Stěžejní část předkládané práce se zabývá způsoby jak definovat podobnost datových množin a jak ji využít pro doporučení vhodných algoritmů strojového učení.

Vlastní práce je rozdělena do devíti kapitol. V úvodní kapitole autor představuje motivaci a kontext problému a představuje cíle a strukturu práce. Druhá kapitola je věnována vysvětlení základních pojmů a přístupů z oblasti meta-učení. Obsahuje přehled relevantní literatury včetně původních prací autora. Ve třetí kapitole se autor věnuje aplikaci teoretických vlastností metrik pro definici vzdálenosti datových množin založené na pevně zvoleném počtu globálních atributů. Kapitola 4 obsahuje vlastní výsledky autora, které se týkají definice vzdálenosti datových množin založené na vlastnostech atributů dat. Důležitou částí tohoto přístupu je návrh algoritmu přiřazování atributů dvou datových množin. Pátá kapitola je úvodem k experimentální části a popisuje několik databází experimentů strojového učení, které byly použity pro získání historických dat a extrakci atributů za účelem experimentálního ověření navržených algoritmů. Kapitola 6 se zabývá experimenty, které na praktických příkladech ukazují vývoj metriky mezi datovými množinami pomocí genetického algoritmu. Sedmá kapitola se zabývá rozšířením algoritmů pro vzdálenost datových množin, kde je relaxována podmínka trojúhelníkové nerovnosti. Autor zavádí optimalizační algoritmus založený na genetickém programování a ukazuje jeho výsledky na několika experimentech. Osmá kapitola obsahuje popis několika řešení, které posilují generalizaci navržených algoritmů. Jsou zde uvedeny přístupy koevoluce a vícekriteriální optimalizace a pokročilé techniky genetického programování jako bootstraping a bloat control. Devátá kapitola shrnuje a komentuje výsledky práce včetně návrhu dalších směrů výzkumu.

Práce Jakuba Šmída obsahuje několik původních výsledků, které rozšiřují oblast meta-učení o návrhy dosud neuvažovaných způsobů doporučení vhodného modelu k předkládaným datům.

První oblastí vlastního výzkumu je určování vzdálenosti datových množin. Na rozdíl od běžně užívaných metod, pracujících s omezenou množinou globálních meta-dat, se Jakub zaměřil na metriky založené na jednotlivých atributech. K tomu bylo nutno vyřešit problém zarovnání a vzájemného optimálního přiřazení atributů (kapitola 4). Autor aplikoval obecné vlastnosti metrik k odvození podmínek pro několik návrhů metrik a semi-metrik na datových

množinách. Vzniklé algoritmy konstrukce měr vzdálenosti datových množin jsou podrobně analyzovány i z hlediska časové složitosti. Výsledky shrnují Věta 12 a Algoritmy 9 (speciální případ) a 12 (zobecnění algoritmu 9). Korektnost konstrukce metriky na základě atributů umožnila pracovat odlišně s různými typy atributů (např. kategoriální a číselné atributy), a také kombinaci lokálních atributů s globálními atributy, které byly užívány doposud (Věty 14 a 16 ukazují korektnost algoritmu 12). Přístup k vytváření metrik je poměrně obecný a lze ho aplikovat na širší třídu problémů.

Jedním z teoretických výsledků jsou i podmínky, kdy lze parametry metrik dále optimalizovat, aniž by došlo k porušení jejich vlastností. Pro optimalizaci parametrů metrik a semi-metrik použil Jakub metody z oblasti evolučních algoritmů, konkrétně genetické algoritmy (kapitola 6) a genetické programování (kapitola 7). Využitím pokročilých metod evolučních algoritmů, jako je bootstrap, koevoluce a vícekriteriální optimalizace, se podařilo překonat problémy s přeučením a bloatem stromů genetického programování, a dosáhnout dobré generalizace výsledných metrik. Experimenty ukázaly, že optimalizované metriky jsou úspěšnější při doporučování modelů strojového učení. Stejně tak se ukázal velmi úspěšným přístup kombinující lokální a globální atributy.

Za třetí důležitý přínos práce považuji vytvoření celého postupu doporučování modelů strojového učení pro nová data na základě předchozích výsledků. Na základe teoretických výsledků Jakub implementoval nevržené algoritmy v rámci systému, který je otevřený, dokáže kombinovat různé metriky a optimalizovat jejich parametry. Vzhledem k časové náročnosti daných výpočtů je implementace distribuovaná a schopná pracovat v heterogenním prostředí. Pro dobré fungování doporučovacích metod je důležitá rozsáhlá databáze výsledků předchozích experimentů. Jakub se nejprve podílel na vzniku lokální databáze experimentů modelů systému Weka na vybraných datech z databáze UCI machine learning repository, většinu experimentů v práci pak realizoval na datech z nově vzniklé evropské iniciativy OpenML, ke které také aktivně přispívá.

Autor se v práci věnuje velmi aktuálním problémům současné umělé inteligence a svými výsledky přispívá k porozumění problémů v oblastech strojového učení a dobývání znalostí z dat. I na základě jeho publikační historie lze doložit, že dosažené výsledky jsou relevantní a na mezinárodní úrovni. *Domnívám se, že předkládanou prací i průběhem svého studia Jakub Šmíd jasně prokázal svou schopnost samostatné vědecké práce a doporučuji uznání této práce jako doktorské disertace*.

V Praze dne 15. srpna 2016

Roman Neruda
Ústav informatiky AV ČR, v.v.i