Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Matěj Sochor		
Název práce	Semi-supervised Learning from Unfavorably Distributed Data		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Umělá inteligence

Autor posudku	Mgr. Martin Pilát, Ph.D.	**Role**	vedoucí
Pracoviště	KTIML MFF UK		

Text posudku:

Cílem práce bylo zkoumat metody semi-supervised učení, které by byly schopny se vypořádat s neoznačkovanými daty, která mají jinou distribuci než data označkovaná. Speciálně potom s takovými neoznačkovanými daty, která patří do tříd, které se v označkovaných datech vůbec nevyskytují. Ukazuje se, že za určitých podmínek mohou tato neoznačkovaná data procesu učení uškodit a vést k horším výsledkům než v případě, kdy taková data vůbec použita nejsou (a model se trénuje pouze na označkovaných datech). Studentovi se v předložené práci podařilo navrhnout metodu založenou na filtrování nepříznivých dat, která tento problém zmenšuje nebo odstraňuje. Navíc práce lépe charakterizuje situace (množství dat), za kterých k problému dochází.

Celá práce je rozdělena do čtyř kapitol. V prvních dvou autor popisuje napřed základní principy strojového učení včetně všech technik použitých v práci a následně se věnuje přímo samotnému problému řešenému v práci, tedy semi-supervised učení z nepříznivě distribuovaných dat. Obě tyto kapitoly jsou precizně napsány s rozumným množstvím detailů tak, aby zbytek práce byl pochopitelný i pro čtenáře, který není seznámen s oblastí semi-supervised učení. Autor se zbytečně nevěnuje detailům a oblastem, které nejsou pro práci důležité a vhodně odkazuje na jinou existující literaturu.

Samotné jádro práce je soustředěno v kapitole třetí a čtvrté. Ve třetí kapitole student popisuje navrženou metodu filtrování nepříznivě distribuovaných dat, která je založena na použití autoenkodérů pro snížení dimenze dat a následném využití shlukování pro filtrování dat. Celá metoda je založena čistě na učení bez učitele a je obecná – nepředpokládá použití žádného konkrétního algoritmu pro semi-supervised učení. Popis metody je dostatečně podrobný a jasný. Autor práci doplnil o vhodná schémata, která metodu dobře vysvětlují.

Ve čtvrté kapitole student provádí experimenty a vyhodnocuje výsledky metody. Podařilo se ukázat, že v některých nastaveních představená metoda zlepšuje výsledky semi-supervised učení. Navíc se podařilo během těchto experimentů ukázat, že překvapivě nezáleží na poměru příznivě

a nepříznivě distribuovaných neoznačkovaných dat, jak se původně myslelo, ale především na samotném množství příznivě distribuovaných dat. Experimenty v této kapitole jsou provedeny velmi pečlivě a ukazují zajímavé vlastnosti jak navržené metody, tak samotného problému. Student při srovnání používá vhodné baseline, které ukazují limity jak navržené metody, tak libovolné jiné obecné metody založené na myšlence filtrování.

Obecně je práce velmi pěkně napsána, obsahuje všechny potřebné informace, ale nezachází do zbytečných detailů (na ty vhodně odkazuje). Experimenty jsou provedeny pečlivě a tam, kde je to vhodné a účelné, i statisticky vyhodnoceny. V některých experimentech by bylo lepší provést více opakování, to ale nebylo možné vzhledem k velké časové náročnosti především algoritmu semi-supervised učení (samotná studentem navržená metoda běží řádově rychleji, než použitý algoritmus semi-supervised učení). Student v práci jasně toto omezení počtu opakování experimentů zmiňuje a je si ho vědom při komentování výsledků a vytváření závěrů.

Na základě výše uvedeného práci doporučuji k obhajobě.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 26. června 2020

Podpis: