FAKULTA JADERNÁ A FYZIKÁLNĚ INŽENÝRSKÁ
KATEDRA MATEMATIKY
TROJANOVA 13
120 00 PRAHA 2
ČESKÁ REPUBLIKA

Posudek oponenta na diplomovou práci

Bc. Viktor Bílek: Vliv ztrátové funkce na detekci anomálií

Detekce anomálií je téma, které se objevuje téměř ve všech oblastech a odvětvích, kde je potřeba analyzovat data. Často je spojeno s pojmy jako je detekce odlehlých pozorování, robustnost, kontrola kvality a klasifikace. Student ve své diplomové práci se věnoval tomuto tématu z pohledu modelů pro odhad pravděpodobnostního rozdělení, kde využíval metod strojového učení. Tím došlo k propojení klasického statistického přístupu a v dnešní době moderních metod, vyžadujících velký výpočetní počítačový výkon.

Práce je strukturována do pěti kapitol, přičemž první dvě jsou věnovány teoretickému úvodu do detekce anomálií a popisu modelů pro odhad hustoty pravděpodobností, přičemž je kladen důraz na modely využívající neuronové sítě. Jak metodě FFJORD, tak metodě součtových produktových sítí, se student věnoval podrobněji ve svém výzkumném úkolu tudíž v této práci je uvedena jejich rešerše. Nutno však zmínit, že tyto úvodní kapitoly jsou dobře a srozumitelně sepsané. Třetí kapitola popisuje detekci anomálií jakožto klasifikační problém a za tímto účelem zmiňuje i stručný úvod do binární klasifikace spolu s popisem potřebných pojmů a metrik pro vyhodnocování kvality jednotlivých modelů. V posledních dvou kapitolách je hlavní nový přínos této práce, a to modifikace ztrátové funkce, její zavedení na kvantilovém intervalu a otestování těchto modifikací spolu s dříve definovanými modely jak na simulovaných tak reálných datech.

Celkově je práce velmi čtivá, s přípustným množstvím překlepů. Co bych částečně vytkl, je na jednu stranu snaha o matematický zápis pomocí definic a vět, ale poté chybějí odkazy na důkazy, přestože jsou to věty všeobecně známé. To je pak v kontrastu s tím, že v místech, kde je samotný přínos studenta, tento formální zápis naopak chybí. Protože se práce věnuje metodám odhadů hustot pravděpodobností, tak mi tam chybí porovnání s klasickými přístupy jako jsou jádrové odhady, modely Gaussovských směsí a poté Bayesovský přístup v klasifikační úloze. Dále je z provedené simulační studie znát, že student se musel vypořádat s řadou problémů spojených se zpracováním mnoha datových sad, velkým množstvím iterací a některé metody sám implementoval v jazyce Julia. V práci to ale není bohužel zmíněno a podle počtu commitů na studentově githubu se to těžko hodnotí. Taktéž by stálo za zmínku uvést náročnost prováděných výpočtů a možnosti reproduktibility provedených studií.

Ke zvoleným metodám a jejich implementaci nemám co vytknout, ale měl bych několik dotazů k samotnému zvolenému přístupu pro detekci anomálií, a zvláště pak ke studii na reálných datech.

Otázky a poznámky k obhajobě:

- V závěru zmiňujete možný problém s přetrénováním a důsledným odhadem pravděpodobnostního rozdělení normálních dat. Zkuste stručně vysvětlit hlavní výhodu vámi zvoleného řešení pomocí metod FFJORD a SPTN proti odhadům hustot pomocí smíšeným modelům GMM a s maximálně věrohodným odhadem pomocí EM algoritmu a oproti neparametrickým jádrovým odhadům.
- Jak jsou zvolené přístupy citlivé na tzv. „prokleti dimenzionality". Pro jaká data byte zkoumané metody doporučil a naopak, kde je jejich omezení.
- Na straně 56 je zavedena pořádková statistika a zmíněno, že gradient se počítá pro daný výběr dat a iterativně se najde další parametr, pomocí kterého dostaneme jiný výběr. Co zaručuje konvergenci tohoto přístupu a proč není možné, že algoritmus bude jen přepínat mezi dvěma podvýběry?
- Panuje podle vás vztah mezi očekávaným procentuálním množstvím anomálií a doporučeným kvantilovým intervalem? Pokud ano, tak jaký?
- Výběr reálných dat z UCI pro simulační studii, spolu se způsobem porovnávání kvality zvolených metod, je dle mého názoru nešťastný. Zvolený soubor obsahuje jak datasety, na kterých metody dosahují AUC 1, nebo téměř jedna a liší se o setiny, tak datasety, kde je AUC dokonce pod 0,5. Průměrné pořadí na takovýchto datech poté ztrácí vypovídající hodnotu. Dále některé datasety obsahují pozorování od několika jedinců a nejsou v tom případě nezávislá. Tudíž jejich náhodné dělení na trénovací, validační a testovací část by tím mělo být ovlivněno. Okomentujte prosím, proč jste zvolil právě tento výběr a proč jste volil průměrné pořadí pro porovnávání mezi metodami.
- Ve výběru reálných dat, je mnoho velmi známých souborů, používaných jako benchmarky pro binární klasifikaci. Přičemž klasický přístup pomocí ML binární klasifikace dosahuje lepších výsledků než uvedené metody (při porovnání AUC). V čem je zvolený přístup přínosnější.
- V práci se píše o použití křížové validace, ale poté je provedeno 5 různých náhodných rozdělení na trénovací, testovací a validační soubor. Toto náhodné dělení, ale nezajistí, že každá anomálie bude právě jednou v testovací části. U souborů, kde je jen několik anomálií to mže hrát zásadní vliv, kor pokud nepoužijeme stejné dělení jak pro různé hyperparametry, tak pro různé metody.

Vzhledem k tomu, že bylo splněno zadání a podařilo se implementovat a otestovat navrhnuté ztrátové funkce, tak i přes zmíněné nedostatky doporučuji práci k obhajobě a v případě zodpovězení uvedených otázek navrhuji udělit známku **A (výborně).**

V Praze dne 21. května 2021 .....

Ing. Jiří Franc, Ph.D.