Posudek na diplomovou práci

Metody strojového učení v částicové fyzice

Autor: Bc. Miroslav Kubů, 5. ročník, KM FJFI, akademický rok 2019/2020

Školitel:
Ing. Petr Bouř, KM FJFI ČVUT Praha

Předložená diplomová práce se zabývá použitím metod strojového učení na vyhodnocování dat z experimentu NOvA v laboratoři FNAL, USA. Tímto navazuje na předchozí bakalářskou práci a výzkumný úkol, jejichž závěry rozšiřuje. Předložená práce tematicky plně odpovídá zadání diplomové práce. K jednotlivým vybraným aspektům práce se vyjádřím v následujících bodech.

1. V úvodní kapitole autor přibližuje fyzikální cíle experimentu NOvA. Následující kapitola 2 vysvětluje základní parametry pro měření kvality klasifikace úloh. V následujících dvou kapitolách (3,4) autor sumarizuje základní klasifikační metody z oblasti umělých neuronových sítí a metody klasifikace založené na rozhodovacích stromech a lesech. V těchto kapitolách popisuje i některé heuristiky, které se používají jednak pro urychlení numerických algoritmů učení, jednak pro eliminaci jevu přeučení klasifikátorů, kdy se snižuje generalizace klasifikátoru za současného trvalého poklesu optimalizované penalizační funkce. Zmíněné kapitoly mají víceméně popisný rešeršní charakter, ale jejich zpracování je přehledné a poskytuje dobrý vhled do problematiky klasifikačních úloh strojového učení aplikovaného v oblasti fyziky vysokých energií.

2. Následující pátá kapitola je věnována klasifikaci neutrinových interakcí z experimentu NOvA pomocí metod uvedených v předešlých kapitolách. Autor inovativně navrhuje různé kombinace klasifikátorů (ResNet, ML, Random Forest, AdaBoost a Gradient Boosting) a za referenční klasifikátor bere CVN s inception moduly. Následuje obsáhlá část ve které jsou diskutovány klady a zápory jednotlivých metod dokladované histogramy výstupů klasifikátorů a maticemi záměn.

3. Na základě výše uvedených experimentů bylo dosaženo zlepšení klasifikace pro třídy ν_μ a ν_e (v procentech 84.84 $\rightarrow$ 86.11, resp. 77.42 $\rightarrow$ 79.88) což je při klasifikaci tohoto typu dat relativně dobré zlepšení.

4. K textu práce mám následující připomínky, které jsou pouze formálního charakteru:

5. - str. 26, ř. 2 shora: je zde první zmínka o lineárním klasifikátoru. Je myšleno prosté dělení prostoru nadrovinou, nebo se jedná o kompletnější strukturu využívající jako diskriminační plochy nadroviny?
   - str. 30, Definice 3.1.2.: pojmem sigmoidální aktivační funkce se v oboru ML obecně rozumí jakákoli spojitě diferencovatelná omezená monotónní funkce z R do R.
   - výraz 3.9.: v souladu s předešlou definicí by mělo být L_{CCE} namísto L. V řádku pod a výrazu 3.12. taktéž.

- str. 34 nahoře: empirická riziková funkce J je hned následně zmíněna jako cílová funkce. Postrádám explicitní informaci, že funkci L nadále považujeme za součást klasifikujících algoritmů.
- str. 38, ř. 3 odspodu: vysvětlit (upřesnit) pojem "jednokanálový".

6. V rámci obhajoby bych si dovolil autorovi práce položit následující dotazy:
   - na str. 57 autor uvádí rozdělení datových množin na trénovací, validační a testovací v poměru 80%, 10% a 10%. S ohledem na poměrně velký počet dat vznáším otázku proč nebylo pro konstrukci testovací množiny použito více dat, třeba na úkor dat trénovacích.
   - je-li odpovědí na předešlou otázku nízký poměr dat jedné třídy vůči celkovému počtu dat, vznáším otázku zda autor nezvažoval využití nějakých augmentačních technik na zvýšení počtu dat reprezentujících danou třídu, eventuálně proč nebylo simulováno více dat dané třídy pro fázi učení (při učení s učitelem (supervised learning) pracujeme přece se simulovanými daty).
   - za důležité v rámci široké kooperace na experimentech ve FNAL považuji možnost oponování výsledků ostatními týmy kooperace experimentu NOvA. Vznáším dotaz, zda data použitá v této práci jsou ostatním týmům k dispozici, případně za jakých podmínek, a to zejména členům české části kooperace na experimentu NOvA.

Celkově v diplomové práci autor prokázal schopnost komplexního zpracování zadaného tématu, počínaje počátečním netriviálním procesem získání relevantních dat z experimentu, dále sofistikovaným původním návrhem klasifikačního algoritmu a finálně masivními výpočty a vyhodnocováním.

Vzhledem k výše uvedeným skutečnostem doporučuji práci uznat jako práci diplomovou a navrhuji hodnocení této práce stupněm

A - výborně.

Praha, 25. ledna 2021

Ing. František Hakl, CSc.
Oddělení strojového učení
Ústav informatiky AV ČR