



Posudek oponenta závěrečné práce

Oponent práce: Ing. Daniel Vašata, Ph.D.
Student: Bc. Michaela Kučerová
Název práce: Metody pro balancování binárních dat
Obor / specializace: Znalostní inženýrství
Vytvořeno dne: 5. února 2024

Hodnotící kritéria

1. Splnění zadání

- [1] zadání splněno
- [2] zadání splněno s menšími výhradami
- [3] zadání splněno s většími výhradami
- [4] zadání nesplněno

Zadání bylo splněno bez výhrad.

2. Písemná část práce

85 /100 (B)

Práce je logicky strukturovaná a po jazykové stránce výborná. Zdroje jsou relevantní a správně citované. Po obsahové a typografické stránce mám k práci několik výhrad.

Nesouhlasím s komentářem v části 1.3, že když jsou dvě třídy separovatelné a mají jasnou rozhodovací hranici, není třeba provádět balancování. To možná platí, pokud člověk zná rozhodovací hranici dopředu, pak není ani potřeba trénovat žádný model. Jinak to ale neplatí a i pro teoreticky dokonale separovatelné třídy mohou mít modely problém se tu hranici správně naučit, pokud budou data nevyvážená.

Občas se objevuje odsazení odpovídající novému odstavci uprostřed věty, která pokračuje po vzorečku nebo rovnici, jako např. na str. 9.

Studentka se nachytila (stejně jako i mnoho výzkumníků) na klasickou GAN past, kterou představuje vztah (2.2). Ve skutečnosti tato rovnost neplatí, protože na levé straně je konstanta (hodnota min max) a na pravé straně je funkce od G a D.

Ve vztahu (2.4) je chyba v tom, že zde nejsou uvedeny střední hodnoty, bez kterých vztah nedává smysl, i kdyby tam byly, tak je tento vztah nekonzistentní s větou, která ho uvozuje - totiž, že je to účelová funkce. Účelová funkce to může být pouze bez min max členu na začátku.

Trochu mi také vadilo, že se některé metody vyskytují v práci dvakrát. Například TVAE, ale i další, jsou uvedeny v části 2.2 i v kapitole 3.

Studentka zavádí ROC křivku v části 4.2 naprosto neodpovídajícím komentářem, ze kterého není nic zřejmé. Přitom se jedná o téma, které měla v rámci minimálně jednoho

magisterského kurzu do podrobně vysvětlené.

Na konci části 4.3 je velmi zvláštní (řekl bych spíše nesprávná) věta: "The distance is the measure of how well the generated samples approximate the minority class distribution."

V generátoru GAN části navrhovaného modelu LIT-GAN se neprovádí žádné generování, tutíž by se asi neměla nazývat generátorem. Generátor a diskriminátor (který by bylo správnější nazývat critic - protože to je ve Wassersteinově smyslu) jsou dvě sítě trénované adversariálním způsobem, ale generativní nejsou (to je pouze celek díky té VAE části).

Vzhledem k tomu, jak je model kombinován (kdy generátor GAN části vlastně není žádný generátor) podle mě nedává smysl to takhle kombinovat a stačila by pouze jedna síť se třemi částmi - encoder VAE, decoder VAE, který by byl totožný s "generátorem" GANu a pak diskriminátor GANu. To už je ale spekulativní poznámka a muselo by se to vyzkoušet.

Kromě těchto technických detailů je ovšem text práce srozumitelný, čtivý a metodologicky korektní, takže předchozí výtky pouze mírně snižují celkové hodnocení v této části posudku.

3. Nepísemná část, přílohy 99 /100 (A)

Nepísemnou částí práce byla rozsáhlá sada experimentů, která je opravdu hodnotná představuje poměrně unikátní srovnání. Přílohou práce jsou přehledné zdrojové kódy které umožňují experimenty replikovat.

4. Hodnocení výsledků, jejich využitelnost 95 /100 (A)

Z pohledu závěrečné práce jsou výsledky velmi přínosné. Jedná se zejména o unikátní porovnání velkého množství metod, které je možné využít v případě nevybalancovaných dat. Navržená metoda by asi chtěla ještě důkladnější výzkum, než by bylo možné ji publikovat, ale i tak to není špatný směr.

Celkové hodnocení 90 /100 (A)

Celkově se mi práce líbila. Především experimentální část vyžadovala opravdu velké úsilí a provedené a prezentované experimenty jsou přínosné. V písemné části jsem měl nějaké poznámky, ale i tak si myslím, že se jedná o velmi zdařilou práci, která je na velmi vysoké úrovni. Proto navrhuji hodnocení stupněm A.

Otázky k obhajobě

Bylo při prováděných experimentech patrné, že metody, které mají samy o sobě hodně parametrů (jako například různé generativní modely neuronových sítí), jsou v případě malých dat horší než klasické neparаметrické metody (jako např. SMOTE)?

Jinými slovy - dá se vypořádat nějaká souvislost mezi úspěchem metody v závislosti na její komplexitě a velikosti datasetu?

Instrukce

Splnění zadání

Posudte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posudte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Písemná část práce

Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posudte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti.

Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posudte správnost používání formálních zápisů obsažených v práci. Posudte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posudte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Nepísemná část, přílohy

Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce – opakovatelnost experimentů.

Hodnocení výsledků, jejich využitelnost

Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Celkové hodnocení

Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.