FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Matouš Kozák
Student:	Bc. Lukáš Děd
Název práce:	Výběr reprezentativních vzorků z datových sad pro detekci malwaru
Obor / specializace:	Počítačová bezpečnost
Vytvořeno dne:	17. března 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)

Obsah diplomové práce odpovídá zadání a práce neobsahuje věcné chyby.

Kapitola 1 popisuje možnosti předzpracování dat a obsahuje známé metody, které se běžně pro předzpracování využívají. Kapitola 2 detailně popisuje instance selection algoritmy, které jsou použity v práci. Oceňuji přiložení pseudokódu pro jednotlivé algoritmy. Kapitola 3 obsahuje krátký úvod do klasifikačních algoritmů a běžně používaných metrik. Kapitola 4 obsahuje velmi detailní popis PE file formátu. Nicméně, z mého pohledu se jedná o kapitolu, která by v práci vůbec být nemusela a pokud ano tak ve velmi zredukované formě. Kapitola 5 obsahuje detailní popis jak byly datasety EMBER a SOREL-20M předzpracovány. Kapitola 6 obsahuje popis navrhnutých vylepšení metody Parallel Instance Filtering. Kapitola 7 porovnává metody z Kapitoly 2 a Kapitoly 6 na datasetech EMBER a SOREL-20M. Použitá literatura je aktuální a relevantní k tématu.

Několik drobností co bych práci vytknul:
- Edited Nearest Neighbors (2.2.1), je v některých pozdějších částech práce nazývána jako Wilson editing. Bylo by vhodné tento název zmínit i v kapitole 2.2.1.
- V části 5.3 zabývající se laděním hyperparametrů pro předzpracování datasetů, je uvedeno, že kompletní výsledky jsou obsaženy v příloze. Nicméně, není zde uveden odkaz pro rychlou orientaci a výsledky se mi ani manuálně najít nepodařilo.
- V části 7.2 jsou výsledky v textu rozděleny do skupin, pro přehlednost, by bylo vhodné

skupiny označit i v přiložených tabulkách.
- Práce obsahuje malý počet chyb a překlepů, které ale nebrání pochopení textu.
- Doporučil bych konzistentní formát nadpisů, konkrétně občas chybějící použití velkých písmen na začátku slov.
- Formátování některých grafů v části "Appendix A" by mohlo být vylepšeno odstraněním záporných hodnot na ose y u metrik, které mohou nabývat pouze nezáporných hodnot.
- U záznamů 14. a 34. v seznamu použité literatury chybí unikátní identifikátor použité literatury.

3. Nepísemná část, přílohy 95 /100 (A)

Většina přiložených zdrojových kódů je v jazyce Python, jazyk běžně používány v oblasti strojového učení. Oceňuji, že některé algoritmy byly také přepsány do jazyka C, pro efektivnější použití v praxi. Příloha obsahuje také zdrojové kódy nutné pro zreprodukování experimentů a spolu s faktem, že oba použité datasety jsou veřejně přístupné je možné výsledky práce zopakovat/ověřit.

4. Hodnocení výsledků, jejich využitelnost 90 /100 (A)

Práce obsahuje velmi detailní porovnání instance selection algoritmů při použití s klasifikátorem kNN. Ve výsledcích se autor zaměřuje hlavně na porovnání přesnosti a redukce velikosti. Z výsledků vyplívá, že se jedná o kompromis a vysoká přesnost je vykoupena nižší redukovatelností. Pro lepší aplikovatelnost výsledků by bylo vhodné zjistit jak redukce sady ovlivňuje i jiné klasifikační algoritmy. Očekával bych také porovnání i z hlediska falešné pozitivity, která je oblasti detekce malwaru velmi důležitá, zvláště pro vyhodnocení praktického využití. Při porovnání časové náročnosti jednotlivých instance selection algoritmů, metody PIF a RPIF, které využívají paralelismus, dosahují velmi dobrých výsledků. Nicméně, pro férové porovnání časové náročnosti by bylo vhodné výsledky očistit o použitý paralelismus.

I přes malé nedostatky, práce dává čtenáři dobrý přehled aktuálních metod pro výběr reprezentativní sady vzorků a obsahuje doporučení pro uživatele, které metody je vhodné využít s algoritmem kNN.

Celkové hodnocení 91 /100 (A)

Diplomová práce je zdařilá a splňuje zadání. Oceňuji velmi rozsáhlé porovnání instance selection algoritmů spolu s detailním popisem použitých metod.

Práci hodnotím známkou A a doporučuji k obhajobě.

Otázky k obhajobě

Při porovnání časové náročnosti použitých instance selection algoritmů, metody PIF a Repeated PIF dosahují velmi podobných výsledků. Jak si vysvětlujete tyto výsledky? (Z popisu algoritmu RPIF bych očekával, že bude RPIF časově náročnější)

Instrukce

Splnění zadání

Posuďte, 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, posuďte 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 posuďte, 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. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 52/2021, článek 3.

Posuďte, 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.