FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Tomáš Čejka, Ph.D.
Student:	Bc. Petr Skružný
Název práce:	Automatická optimalizace datových sad síťového provozu
Obor / specializace:	Počítačová bezpečnost
Vytvořeno dne:	29. května 2022

Hodnotící kritéria

1. Splnění zadání

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

Práce se zabývá analýzou a optimalizací datových sad síťového provozu, které slouží k trénování modulů pro klasifikaci a detekci bezpečnostních událostí. Práce obsahuje rešerši existujících použitelných přístupů a z nich pak množinu implementuje. Implementace je využitelná jako modul do systému Active Learning Framework (ALF), který automaticky trénuje detektory bezpečnostních hrozeb.

2. Písemná část práce 58 / 100 (E)

Velkou část práce tvoří popisy existujících algoritmů a zdá se, že je tento rozsah na úkor podrobného vyhodnocení implementovaných metod nad datovými sadami. Úvodní kapitola nedostatečně popisuje motivaci, řešený problém a cíle práce, kterých následně autor dosáhl. Sice autor píše v úvodu o "aktivním učení", ale kontext je vysvětlen až na straně 31.

Zdá se, že byly možnosti optimalizace zkoumány pouze na datové sadě/sadách jednoho problému a to detekce DNS over HTTPS. Jedná se sice o důležitou datovou sadu pro aktuálně probíhající výzkum, avšak problematika optimalizace datových sad by měla být zkoumána a vyhodnocována nad větším počtem různých datových sad.

Textová část celé diplomové práce je celkově velice slabá a to jak po jazykové tak typografické stránce. Práce obsahuje značné množství překlepů a gramatických chyb. Odkazy na sekce na několika místech v textu nejsou správně uvedeny. Výpisy algoritmů není vhodné označovat jako Obrázky.

3. Nepísemná část, přílohy 80 /100 (B)

Výsledkem práce je funkční zdrojový kód v jazyce Python, který je použitelný jako rozšíření systému ALF. Tato implementace obsahuje 6 přístupů pro optimalizaci datové sady, které byly otestovány v práci. Dále při řešení diplomové práce vznikla sada python notebooků, ve kterých probíhaly experimenty.

Diplomovou práci by bylo vhodné doplnit o popis instalace a použití vyvinutého nástroje a dokumentaci zdrojového kódu.

4. Hodnocení výsledků, jejich využitelnost 70 /100 (C)

Problematika optimalizace a vyhodnocování datových sad pro síťovou bezpečnost je velice důležitá, což ukazuje i úspěch současných publikací v celosvětové vědecké komunitě. Odevzdaná diplomová práce tuto oblast adresuje, ale k publikaci výsledků a využití v praxi by bylo potřeba mnohem důkladnější vyhodnocení a interpretace výsledků testů.

Celkové hodnocení 60 /100 (D)

Zadání práce bylo splněno, ale kvalita textové části práce je velice nízká. Výsledkem je implementace několika použitelných metod optimalizace datové sady síťového provozu, ale jejich vyhodnocení v práci nebylo příliš důkladné (v textu práce je popsáno testování nad jednou datovou sadou). Z textu se zdá, že zhodnocení výsledků je založeno spíše na domněnkách místo obecně měřitelných faktů (př. str. 50: "Vzhledem k objemu nastavitelných parametrů si dovolím tvrdit, že důladná analýza použití autoenkodérů nad zadaným problémem by vydala na stejné, ne-li větší, množství práce jako práce stávající a mohla by být zajímavým tématem pro budoucí výzkum."). Z popsaných důvodů navrhuji celkové hodnocení D.

Otázky k obhajobě

Tabulka 8.1 srovnává několik testovaných metod.
Je možné interpretovat výsledky a najít důvody významně rozdílných hodnot F1 skóre mezi metodami K Means a Autoenkodér při srovnatelné míře redukce? Tzn. čím se vlastně výsledné redukované datové sady liší (mezi sebou a od originální datové sady)?

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.