FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Petr Pulc
Student:	Bc. Jan Rudolf
Název práce:	Posouzení důležitosti zařízení podle chování na síti
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	6. června 2023

Hodnotící kritéria

1. Splnění zadání

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

Problematické je částečné splnění bodu jedna zadání, kdy se student pouze povrchně seznámil s daty jako takovými, ale neposkytl hlubší vhled do zkoumané domény. Například chybí poměrně zásadní informace o roli registrovaných a efemérních portů v komunikaci. S ohledem na tento nedostatek bylo pak pro studenta komplikované posoudit kvalitu dodaných dat, ke kterým z pochopitelných důvodů vedoucí nemohl dodat žádné garance (protože ani vedoucí tyto garance často nemá).

Větším problémem je pak faktické nesplnění bodu dva zadání, kdy student popsal jen zdánlivě náhodně vybranou skupinu algoritmů a jejich výběr (až na Page Rank) nijak neodůvodnil s ohledem na studovanou doménu. Zcela chybí rešerše existujících přístupů ke klasifikaci zařízení podle síťového provozu, přičemž minimálně pro IoT zařízení takové přístupy existují a jejich popis je veřejně dostupný (BAI, Lei, et al. Automatic device classification from network traffic streams of internet of things, a další). Neveřejně pak student mohl konzultovat existující přístupy klasifikace zařízení vyvíjené kolegy vedoucího práce, což ale také neprovedl.

2. Písemná část práce 40 / 100 (F)

Rozsah práce je včetně prázdných stran na spodním limitu doporučovaného rozsahu. Přitom úplně zbytečně. Kapitoly nula a jedna jsou informačně velmi chudé, nerozvádí do dostatečné hloubky ani to málo ze splněných bodů jedna a dva zadání.

Kapitoly nula a jedna jsou navíc problematicky hodnotitelné, protože obsahují omezené množství vlastních myšlenek autora – obsahují značné množství nepřímých citací a autocitací. Často tak chybí komentář, který by provázal citované informace se zkoumanou

doménou v teoretické rovině. Některé vložené obrázky jsou pak evidentně zkopírované z bakalářské práce autora, což dále snižuje přínos práce. Zároveň vložené vzorce jsou často špatně opsané – například sumy mají špatně definované meze.

Obecně pak text práce obsahuje netriviální množství nepřesností, které by byly určitě snadno odhalitelné při včasných konzultacích s vedoucím práce. Příkladem je už jen nepochopení některých informací v dodaném datasetu. Například pole „SHA" obsahuje hash spustitelného souboru a „device id" je spolehlivý identifikátor zařízení (jen není garantované, že udaná směrovost komunikace je vždy správná v relaci s „device id"). Nebo v předchozí části hodnocení zmíněná role portů v komunikaci, které by pomohly vyčistit zdrojový dataset od potenciálně problematických vzorků. Jenže popis datasetu se v textu omezil jen na metainformace, neobsahuje popis vlastního průzkumu dat jako takových. Zároveň některé myšlenky jsou vysloveně nešťastné, jako například uvedení „prokletí dimenze" u konfigurace hyperparametrů, i když je tento pojem spojený s počtem dimenzí dat jako takových.

Stěžejním problémem je ale gramatická a typografická úroveň práce: od špatných shod podmětu s přísudkem, přes archaismy, nevhodné slovní obraty, špatné slovosledy, nedokončené věty, nevhodně vsazené citace do toku textu, legendy pod tabulkami, špatně vsazené interpunkce u vzorců, nebo rastrové (JPEG!) obrázky místo vektorových. Smutnou zajímavostí je samotná existence kapitoly nula, která poukazuje na to, že student zřejmě práci sázel ve spěchu. Odkaz na neznámou sekci v prvním odstavci strany 36 (navíc ve špatně formulované větě, kdy se student dvakrát odkazuje na vrcholy grafu) už považuji za detail.

Problematické je pak i posouzení rozdílů mezi obrázky 3 a 4, které na první pohled působí duplicitně, i proto, že jsou umístěné ob stránku. Nebo tabulka 2.3, která obsahuje v principu netabulková data (neexistuje závislost v rámci jednoho řádku). A podobných typografických prohřešků (výrazně snižujících srozumitelnost) obsahuje práce v prvních třech kapitolách velké množství.

Kvalita textové části práce se pak od kapitoly číslo čtyři výrazně zlepšuje. I když obsahuje stále netriviální množství gramatických a drobných typografických chyb. Navíc podle přiložených zdrojů ani zdaleka nepokrývá všechny provedené experimenty, jedno jak byly úspěšné.

3. Nepísemná část, přílohy 70/100 (C)

I s ohledem na netriviální datovou sadu, která je navíc komplikovaná anonymizací dat, jsou výstupy studenta mimo textovou část poměrně dobré. A to jak po stránce kódu, i výstupů experimentů.

V práci jsou podrobně prozkoumané jak přístupy využívající heuristiku založenou jen na počtu komunikací, tak metodu řazení uzlů pomocí algoritmu odvozeného z Page Rank nebo XGBoost klasifikátoru; a to včetně mnoha provedených experimentů nad různými nastaveními algoritmů.

V práci ale chybí hlubší prozkoumání předzpracování a filtrace dat (které vlastně chybí i v textu práce), což velmi pravděpodobně způsobuje i nestabilní výsledky některých přístupů.
Stejně tak chybí zhodnocení algoritmu pro odhad dopadu výpadku zařízení, který je v textu práce popsán pouze teoreticky, ale další experimenty chybí; I kdyby měly pouze

konstatovat úroveň zranitelnosti sítě jednotlivých zákazníků nějakou jednoduchou metrikou.

4. Hodnocení výsledků, jejich využitelnost 60 /100 (D)

Od této práce jsme si původně slibovali především rešerši existujících přístupů v doméně síťového provozu a jejich aplikaci na nám dostupná (omezená) data. Jen částečným splněním prvních dvou bodů zadání se ale těžiště práce přesunulo k aplikaci obecně používaných algoritmů na námi dodaná data.

Tato změna není nezbytně nutně špatná, protože přináší nový pohled na data jako taková, bez zatížení předpoklady doménové znalosti. Je tak například zcela v pořádku, že experimenty inspirované algoritmem Page Rank nedosahovaly očekávaných výsledků ani při zásadních úpravách. Výsledky algoritmu XGBoost jsou pak příjemným překvapením a velmi pravděpodobně budeme přístup validovat na větším vzorku dat.

Nedostatečné prozkoumání domény, datasetu a existujících přístupů má ale značně negativní vliv na možnost přímého využití výsledků. Problémem je, že i když algoritmus XGBoost dosahoval vysoké schopnosti generalizovat nad dodanými daty, udržení si vysoké přesnosti natrénovaného modelu v čase (nad nepročištěnými daty) je z praxe velmi složité a studentem pravděpodobně neprozkoumané.

Navíc textová část diskutující výsledky je místy zmatečná a ne zcela dostatečná pro plné posouzení přínosů práce.

Celkové hodnocení 55 /100 (E)

Práce se obecně pohybuje na hraně doporučení připuštění k obhajobě.

Student předkládá poměrně rozsáhlé výsledky. A to včetně netriviálního množství vlastního kódu, ač se typicky jedná o spuštění učení dostupného algoritmu se sadou vybraných hyperparametrů a polo-automatickým vyhodnocením výsledků. Zároveň podle rozsahů přiložených notebooků věnoval značné úsilí samotnému výzkumu, který z podstaty nemusí vždy končit skvělým výsledkem.

Velmi problematická (až nedostatečná) je ale textová část práce, která je informačně značně chudá (ač student podle příloh měl značné množství informací z širokého spektra experimentů). Navíc působí dojmem, že si jí student před odevzdáním ani nestihl zkontrolovat.

Otázky k obhajobě

1. Pokuste se interpretovat následující záznam z datasetu (část, neanonymizovaný): {source_ip: 192.168.0.2, source_port: 17848, destination_ip: 142.251.39.110, destination_port: 443, sha: 795bc80bcbe9d079b4050b6a35ef45afc892c4389198b982b62375295c5b1e4d}. Využijte jakékoliv veřejně dostupné zdroje.

2. Jaká další informace o síťovém provozu nedodaná v datasetu (můžete se inspirovat popisem NetFlow) by, podle Vás, pomohla zpřesnit klasifikaci zařízení na síti?

3. Na základě jakých extrahovaných příznaků se algoritmus XGBoost nejčastěji rozhodoval?

4. Jak se měnily významné příznaky s ohledem na rozdělení dat (podle dní / podle zákazníků / nedělená data / …)? Je možné vypozorovat korelaci s ad-hoc navrženým dělením na „malé“ a „velké“ zákazníky?

5. Jak stabilní v čase jsou predikce XGBoost (t.j. jak dobře predikuje model naučený např. v pondělí na datech z dalších dnů týdne)?

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.