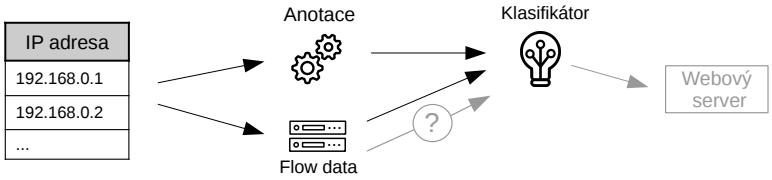
BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2021
http://excel.fit.vutbr.cz

Automatizovaná anotace síťových zařízení a následná klasifikace podle statistického chování

Pavel Eis*



Abstrakt
Automatická klasifikace zařízení v počítačové síti lze využít pro detekci anomálií v síti a také umožňuje aplikaci bezpečnostních politik dle typu zařízení. Pro vytvoření klasifikátoru zařízení je stěžejní kvalitní datová sada, jejichž veřejná dostupnost je nízká a tvorba nové datové sady je složitá. Cílem práce je vytvořit nástroj, který umožní automatizovanou anotaci datové sady síťových zařízení a vytvoření klasifikátoru síťových zařízení, který využívá pouze základní údaje o síťových tocích. Výsledkem této práce je modulární nástroj poskytující automatizovanou anotaci síťových zařízení využívající systém ADiCT sdružení Cesnet, vyhledávače Shodan a Censys, informace ze služeb PassiveDNS, TOR, WhoIs, geolokační databáze a informace z blacklistů. Na základě anotované datové sady je vytvořeno několik klasifikátorů klasifikujících síťová zařízení podle používaných služeb. Výsledky práce nejen výrazně zjednodušují proces vytváření nových datových sad síťových zařízení, ale zároveň ukazují neinvazivní přístup ke klasifikaci síťových zařízení.

Klíčová slova: Automatizovaná anotace síťových zařízení — Klasifikace síťových zařízení — Statistická klasifikace

Přiložené materiály: Veřejný kód

*aisik004@gmail.com, *Fakulta informačních technologií, Vysoké učení technické*

1. Úvod

Pro efektivní správu počítačové sítě je důležité mít přehled o zařízeních připojených do sítě, na základě kterých lze detekovat různé anomálie v chování zařízení a je také jednodušší aplikovat jednotné bezpečnostní politiky na stejná zařízení. Jedním ze způsobů získávání informací o zařízeních je sledování jejich dlouhodobé a krátkodobé aktivity.

Zařízení pod správou síťového administrátora bývají stovky i tisíc, manuální identifikace tedy není možná. Proto je nutné zkoumat a vyvíjet automatizované způsoby klasifikace síťových zařízení. Pro automatizovanou klasifikaci se nejčastěji využívají modely strojového učení, které jsou závislé na dostupnosti datových sad, z kterých se učí. Veřejně dostupných datových sad je ale málo a tvorba vlastní datové sady je velmi pracná. Jednou z nejtěžších částí tvoření nové datové sady je její anotace, tedy správné označení komunikace síťového zařízení (kterou využívá model ke svému učení) štítkem, který reprezentuje dané zařízení.

Jakmile je datová sada vytvořena, existuje několik různých přístupů ke klasifikaci provozu a síťových zařízení. Souhrnný článek o profilování provozu síťových zařízení [1] definuje tři základní přístupy ke klasifikaci síťových zařízení. Prvním z nich je přístup zkoumající jednotlivé toky, u kterého se většinou vytváří grafy závislostí v komunikaci mezi hosty. Tato technika dále zkoumá vlastnosti těchto grafů, především

hledá různé vzory a závislosti. Další technikou je statistický přístup, který zkoumá souhrnné statistiky o provozu daného zařízení (např. objem provozu v určitém směru). Poslední technikou profilování síťových hostů je inspekce obsahu paketů. Výhodou prvních dvou technik je zachování soukromí (nepotřebují zkoumat obsahy paketů a uživatelských dat). To ale vede k jejich o něco menší přesnosti oproti přímé inspekci obsahu paketů. Statistické metody zároveň mají i velkou škálovatelnost a flexibilitu a proto byl tento přístup zvolen pro klasifikaci síťových zařízení na vytvořené datové sadě. Existuje několik prací [2, 3, 4] využívajících statistickou analýzu provozu pro profilování a klasifikaci síťových zařízení. Tato práce využívá vybraných statistik z předchozích prací jako základ pro modely strojového učení klasifikující síťová zařízení.

Tato práce adresuje problematiku nedostatku datových sad síťových zařízení vytvořením modulárního nástroje, který dokáže automatizovaně anotovat síťová zařízení. Pro každou IP adresu získává nástroj informace ze systémů ADiCT od sdružení Cesnet, globálních vyhledáváčů síťových zařízení Shodan a Censys, ze systémů PassiveDNS, WhoIs, TOR, různých blacklistů a geolokačních databází. Zároveň je nástroj navržen tak, aby bylo dostatečně jednoduché přidat další datové zdroje nebo upravit výslednou strukturu anotace.

Využitím automatizovaného nástroje vznikla anotovaná datová sada obsahující bezmála 2 300 síťových zařízení. Datová sada byla dále využita pro vytvoření modelů strojového učení, které klasifikují zařízení podle druhů běžících služeb. Průměrná přesnost modelů se pohybuje okolo 93 % (přesnosti se liší podle typu klasifikovaného zařízení).

V kapitole 2 jsou popsány datové zdroje, které využívá automatizovaný anotátor k anotaci a kapitola 3 popisuje konkrétní způsoby využití datových zdrojů k anotaci. V kapitole 4 je popsána struktura anotátoru a vytvořená datová sada síťových zařízení. Poslední částí článku (kapitola 5) je zhodnocení klasifikátorů síťových zařízení natrénovaných na vytvořené datové sadě.

2. Zdroje dat využitelné pro klasifikaci síťových zařízení

Projekt ADiCT (Asset Discovery, Classification and Tagging) sdružení Cesnet slouží pro sběr informací o zařízeních v síti především pomocí pasivního monitorování provozu[1] a následnou analýzu těchto sbíraných dat pro odvození dodatečných vysokoúrovňových informací. Systém je stále ve vývoji ale již nyní obsahuje několik modulů pro sběr dat, které získávají informace o entitách IP adres a MAC adres nacházejících se v monitorované síti sdružení Cesnet. Systém ADiCT v aktuální podobě obsahuje několik použitelných modulů obsahující informace o zařízeních, např. modul *Service labels*, který sbírá informace o otevřených portech zařízení z probíhající komunikace.

Ke zjištění informací o síťových zařízeních lze použít kromě systému ADiCT i několik externích zdrojů, které shromažďují a získávají informace o velkém počtu zařízení z globálního prostoru IP adres. Takovými zdroji jsou vyhledávače Shodan[2] [5] a Censys[3] [6], které pravidelně skenují globální adresový prostor a sbírají informace o zařízeních připojených do internetu. Oba vyhledávače postupně skenují porty zařízení[4] a pokud je některý port otevřen, snaží se získat i baner běžící služby na daném portu. Kromě seznamu všech otevřených portů vyhledávače udržují informace, které se povedlo získat analýzou textu baneru běžící služby (např. název a verzi konkrétní aplikace běžící na daném portu nebo název operačního systému nainstalovaném na daném zařízení).

Další užitečnou informací je prezence zařízení blacklistu. Většina veřejných blacklistů sdružuje identifikátory (nejčastěji IP adresy) zařízení, které jsou blokovány z určitého důvodu (např. seznam IP adres zařízení generujících DDoS útoky). Pokud se nějaké zařízení nachází na daném blacklistu, je to užitečná informace popisující zařízení. Podobně přínosná je informace o tom, zda je zařízení používano sítí TOR k anonymizaci síťové komunikace. U uzlů sítě TOR je možné dále rozlišit, zda se jedná o běžný uzel sítě TOR nebo o výstupní uzel (tzv. `exit node`).

Systém DNS bývá často zneužíván útočníky, protože využívají otevřené struktury systému DNS, díky které mohou udržovat a ovládat škodlivé zařízení a domény skrze internet (útoky phishing, spamové kampaně, ...). Škodlivé domény a IP adresy je možné detekovat pozorováním a analýzou provozu služby DNS, ale tento přístup je problematický kvůli obtížnosti analýzy provozu v reálném čase. Proto vznikl systém PassiveDNS [7], který vytváří částečné replikace zón systému DNS záchytem živého provozu (především odpovědí systému DNS). Na základě informací ze souhrnného článku o analýze provozu systému PassiveDNS [8] lze použít informace ze systému Pas-

[1] Několik málo sledovaných informací pochází i z externích zdrojů

[2] https://www.shodan.io/

[3] https://censys.io/

[4] Skenován je pouze určitý výčet nejčastěji používaných portů

siveDNS k získání informací o škodlivých aktivitách daného zařízení, např. odhalení používání generovaných doménových jmen (tzv. `DGA` - hodí se pro skrývání škodlivé aktivity za náhodná doménová jména) nebo k odhalení tzv. `fast flux` domén (zneužívají časté aktualizace DNS záznamů pro zakrytí stop škodlivých aktivit).

Základní informace o zařízení lze získat ze systému WhoIs a geolokačních databází, které dohromady poskytnou informace o fyzické lokaci zařízení (především stát a do určité míry přesnosti i město) a číslo autonomního systému daného zařízení.

V této kapitole bylo představeno několik datových zdrojů, které poskytují informace o zařízeních v IPv4 adresovém prostoru. Z pohledu IPv6 jsou zatím některé z představených zdrojů nepoužitelné, především se jedná o zdroje spoléhající na skenování globálního adresového prostoru (vyhledávače Censys a Shodan), který je v případě IPv6 příliš rozsáhlý. Lze ale využít různých heuristik[5], které redukují potřebný čas na proskenování adresového prostoru. Vzhledem k tomu, že práce se zabývá klasifikací zařízení na úrovni ISP (Internet Service Provider), který si musí udržovat situační povědomí a zná svůj adresový prostor, nemělo by navíc být problém definovat přesný seznam adres, o kterých chce správce zjistit nějaké informace. Tento seznam už by měl pokrývat dostatečně malý adresový prostor a samotný ISP má možnost si daný rozsah oskenovat. Poslední překážkou může být překlad adres. Se zvyšujícím se nasazením IPv6 je ale předpoklad, že problém překladu adres se bude postupně redukovat.

3. Návrh anotace datové sady

Tato kapitola shrnuje postup přiřazovaní štítků ke všem síťovým zařízením, které uživatel zadá na vstupu automatizovaného nástroje, který má za úkol provést anotaci. Vzhledem k většímu počtu datových zdrojů je vhodné jejich oddělení. Proto je jako výsledný formát výstupu anotace zvolen formát JSON, který je dobře strojově zpracovatelný a má také dobrou lidskou čitelnost.

3.1 Typ zařízení a podporované služby

Typ zařízení je definován službami, které jsou spuštěny a provozovány na daném zařízení. Tuto informaci lze často odvodit na základě síťových portů, protože většina služeb používá ke své komunikaci předem definovaná čísla portů.

Zdroje dat ADiCT, Shodan i Censys poskytují pro každé dostupné zařízení výčet otevřených portů, který lze dále převést na názvy služeb, které jsou registrované pod danými čísly portů pomocí veřejného registru[6]. Konkrétní název služby je užitečný, ale při trénování klasifikátoru je často kladen důraz na určitou množinu služeb, např. je úkolem natrénovat klasifikátor klasifikující webové služby. Proto lze služby rozdělit do skupin podle jejich účelu a zařízení tak přiřadit obecnější štítek, který reprezentuje typ zařízení. Např. služby běžících na portech 80, 443, 8080, 8443 (protokoly HTTP a HTTPS) lze celkově označit štítkem `web server`. Několik příkladů překladu čísla portu (služby) na typ zařízení je zobrazeno v tabulce 1. Tyto štítky jsou aktivně využívány systémem ADiCT a základ převodní tabulky byl navržen v bakalářské práci Josefa Koumara [9]. Některé štítky ale byly při analýze této převodní tabulky doplněny (např. štítek ICS - Indrustrial Control System, který je často odhalen vyhledávači Shodan a Censys). Celkově je možné na základě převodové tabulky přiřadit zařízení 40 různých štítků definující typ zařízení.

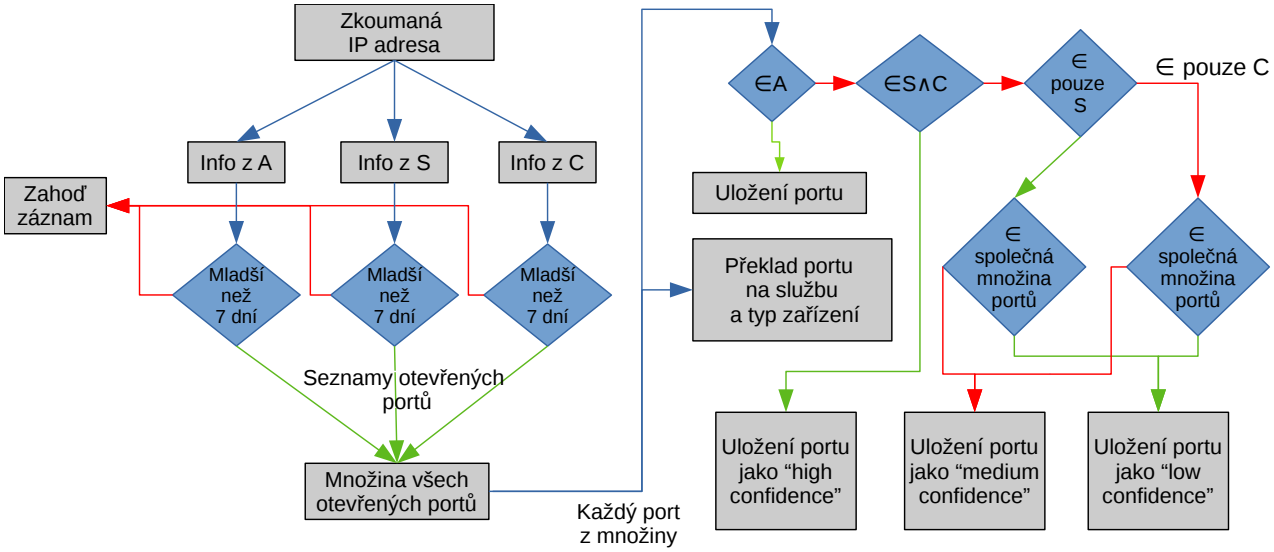
Typ zařízení (štítek)	**Služby (čísla portů)**
`file server`	FTP (20, 21), TFTP (69), NFS (111), ...
`database server`	MySQL (3306), PostgreSQL (5432), ...
`mail server`	SMTP (25), IMAP (143), POP3 (110), ...

Tabulka 1. Část seznamu z převodové tabulky převádějící služby na typ zařízení. Typ zařízení je přiřazený štítek a zahrnuté služby ukazují čísla portů služeb, které spadají pod daný štítek.

Informace o otevřených portech získané pomocí nástrojů ADiCT, Shodan a Censys jsou často velmi podobné, někdy se ale liší. Je tedy nutné navrhnout postup, podle kterého bude docházet ke sjednocení informací v případě, kdy nástroje dojdou k rozdílným závěrům. Celý proces sjednocení informací o otevřených portech je znázorněn na obrázku 1. Nejprve je nutné zkontrolovat, zda informace nejsou příliš staré (maximálně 7 dní) a poté se sjednotí všechny otevřené porty. Pokud je číslo portu dostupné v systému ADiCT, je tato informace dostatečně důvěrná, protože se port

[5] https://www.internetsociety.org/blog/2015/02/ipv6-security-myth-4-ipv6-networks-are-too-big-to-scan/

[6] https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml



Obrázek 1. Způsob přidělení štítků o dostupných službách na zařízení a typu zařízení. Znak **A** reprezentuje službu **A**DiCT, znak **S** reprezentuje službu **S**hodan a znak **C** reprezentuje službu **C**ensys. Pokud je určitá podmínka (modré kosočtverce) splněna, proces pokračuje po zelené šipce. Pokud podmínka splněna není, pokračuje se po šipce červené. Nejprve dojde ke kontrole, zda jsou informace ze systémů ADiCT, Censys a Shodan aktuální. Pokud ano, seznamy portů se sjednotí a zkoumá se, v jakých službách se daný port vyskytuje. Společná množina portů je rovna sjednocení portů, které skenují na zařízeních zároveň obě služby Shodan i Censys. Pokud je port přítomen pouze u vyhledávačů Shodan nebo Censys, je potřeba přidat informaci o důvěryhodnosti.

objevil přímo v síťovém provozu a je uložena ve výsledné struktuře anotace (viz výpis 1). Pokud se ale informace o otevřeném portu objeví pouze u vyhledávačů Shodan a Censys, je nutné rozlišit míru důvěry. Je-li port detekován oboumi vyhledávači, je důvěra v informaci vysoká, míra důvěry se ale snižuje na střední, pokud je port detekován pouze jedním vyhledávačem a na nízkou, pokud je port detekován pouze jedním vyhledávačem a jedná se o port, který pravidelně skenují oba vyhledávače (protože je podezřelé, že porty, které by měly odhalit oba vyhledávače, odhalil pouze jeden).

Výpis 1. Ukázka struktury výsledné anotace podle informací získaných ze všech interních i externích služeb.

```
'device_type': ["web server",
                "file server"],
'open_ports': [80, 443, 22],
'open_ports_scanners':
    {"21": "medium_confidence"},
'services': ["HTTP", "HTTPS", "SSH",
             "FTP"],
'applications': [
    {'name': "OpenSSH",
     'version': "7.9"},
    {'name': "Apache httpd",
     'version': None}],
'os': "Ubuntu",
'passiveDNS': {'domains_count': 124,
    'dga_domains_count': 1, ...}
'blacklists': ["scanner/dshield",
    "bruteforce/bruteforceblocker"],
'geoip': {'country': "CZ", 'city': "Brno",
        'longitude': 24, 'latitude': 42},
'whois': {'asn': 197451, ...}
```

Pokud běžící služba na určitém portu poskytuje uvítací baner, jsou často schopny vyhledávače Shodan a Censys vyextrahovat název běžící aplikace na daném portu, včetně její verze a také název nainstalovaného operačního systému (všechny tyto informace jsou podmíněny jejich přítomností v baneru běžící služby). Tyto informace jsou velmi užitečné a jsou také uloženy do výsledné anotace zařízení.

3.2 Dodatečné informace o zařízení

Pro každé zařízení lze dohledat záznamy v systému PassiveDNS (pro účely této práce je použit systém PassiveDNS sdružení Cesnet), z kterých lze získat výčet domén, které se na daném zařízení vyskytovaly. Dané domény lze zkontrolovat klasifikátorem DGA domén, zda mezi nimi není náhodně vygenerované doménové jméno, které může indikovat škodlivou aktivitu. Z celkového výčtu domén lze získat jejich celkový počet a také zkontrolovat velikost TTL záznámů doménových jmen. Vysoký počet doménových jmen může indikovat škodlivou aktivitu `fast_flux_domains` a velmi nízké TTL je také typické pro podezřelou aktivitu (jedná se pouze o velmi základní analýzu, ne vždy se jedná o škodlivou aktivitu).

U každého zařízení je dále zkontrolována přítomnost

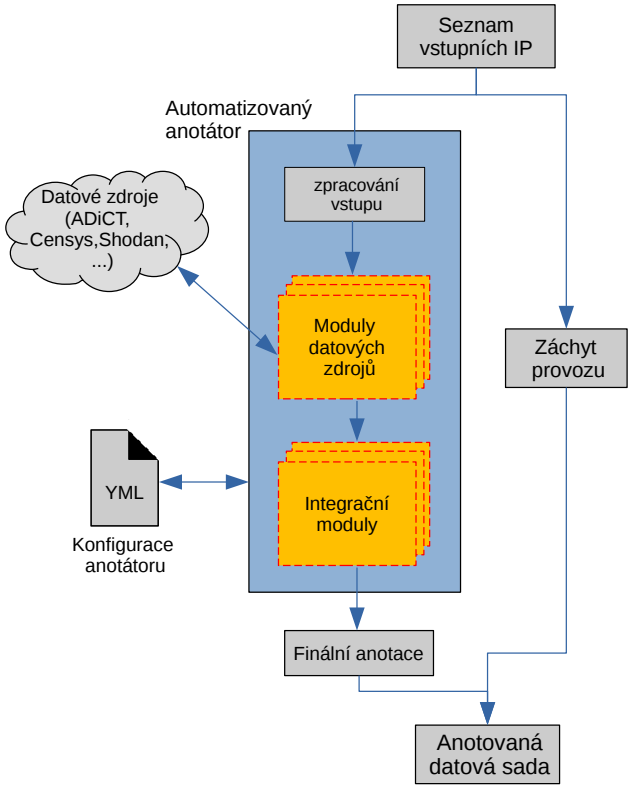
mnost na předem definovaných blacklistech (např. seznam botnetů, . . .) a také přítomnost na seznamu uzlů sítě TOR. Jako dodatečné vlastnosti zařízení jsou uloženy informace o fyzické lokaci zařízení (stát, město a souřadnice) a informace ze systému WhoIs.

4. Automatizovaná anotace síťových zařízení

Při návrhu architektury automatizovaného anotátoru bylo důležité myslet na jeho univerzálnost a také jednoduchost úprav (pro zaručení znovupoužitelnosti nástroje). Jednoduché musí být především přidání nového datového zdroje informací použitelného pro rozšíření anotace, jelikož odlišné experimenty spojené s klasifikací síťových zařízení mohou mít odlišné požadavky na anotaci. Z těchto důvodů má uživatel možnost zvolit si v konfiguraci nástroje jednotlivé datové zdroje, které si přeje použít pro anotaci a zároveň může zvolit i jejich pořadí, v kterém jsou spuštěny. Následně nástroj kromě vstupního seznamu IP adres (popř. celých síťových prefixů) umí také extrahovat seznam IP adres určených k anotaci přímo ze zachyceného provozu (datové sady) ve formátu CSV nebo PCAP.

Celý proces anotace je rozdělen do fáze získání informací ze zvolených zdrojů a následné integrace informací do výsledné podoby anotace. Průběh je lépe znázorněn na obrázku 2. Datové zdroje získávají informace o IP adrese v pořadí, v kterém jsou uvedeny v konfiguraci. Toto pořadí je spíše důležité u integrátorů, které postupně integrují podobné informace z více zdrojů do jedné finální podoby a díky striktnímu pořadí mohou používat výstupy předchozích integrátorů. Pod integrací si lze mimo jiné představit proces integrování portů znázorněn na obrázku 1. Dále jsou integrovány informace z vyhledávačů Shodan a Censys, např. názvy a verze aplikací. Z důvodu optimalizace byla do konfigurace přidána možnost nastavit počet vláken, v kterých má anotátor běžet. Vzhledem k síťovým požadavků datových zdrojů tak při více vláknech dojde k výraznému urychlení.

Přidání nového datového zdroje je přímočaré. Stačí vytvořit nový modul, který dědí z abstraktní třídy bázového modulu definující základní funkcionalitu každého datového zdroje a poté přidat tento modul do seznamu povolených modulů v konfiguraci anotátoru. Anotátor si při spuštění vždy načte seznam povolených modulů a automaticky se postará o jejich import a přidání do řetězce datových zdrojů, nic z toho už programátor nemusí řešit. Úplně stejný postup lze aplikovat i na přidání nového integrátoru.



Obrázek 2. Celý postup anotace datové sady. Na začátku je nutné specifikovat seznam vstupních IP adres. Pro každou IP adresu jsou získány základní informace (např. hostname) a následně je spuštěn každý modul komunikující s určitým datovým zdrojem (podle nastavené konfigurace). Informace z jednotlivých datových zdrojů jsou následně integrovány do finální podoby. Provoz stejných vstupních IP adres je ukládán a s hotovou finální anotací vzniká anotovaná datová sada.

4.1 Výsledná anotovaná datová sada síťových zařízení

Anotátor byl odzkoušen přímo v praxi a byl použit pro anotaci datové sady zachycující týdenní provoz 2 300 unikátních zařízeních ze sítě univerzitního charakteru. Přímo v datové sadě se podle výsledné anotace nachází přes 1000 zařízení provozujících operační systém Windows, 600 zařízení provozující webové služby (štítek `web server`), 250 zařízení provozujících databázové služby a další. Výsledná anotovaná datová sada je použita pro vytvoření několika klasifikátorů odlišujících jednotlivé typy zařízení podle statistického chování, který je popsán v následující kapitole.

Míra důvěry ve výslednou anotaci závisí na použitých datových zdrojích a také může záviset na způsobu využití datové sady. V případě otevřených portů záleží, kolik zdrojů daný port detekovalo, navíc míra důvěry v detekci systémem ADiCT je vysoká vzhledem k zisku informace přímo ze síťového provozu (ve vytvořené datové sadě ze všech otevřených portů - 45 644 - jich

Kategorie statistik	Jednotlivé rysy
Agregační statistiky	Celkový a průměrný počet toků, přenesených paketů a bytů, průměrná velikost paketu, průměrný počet toků na hosta, průměrný počet paketů za tok, průměrná doba toku, ...
Unikátní hodnoty	Počet unikátních komunikujících hostů a cílových portů, autonomních systémů a států, počet unikátních protokolů.
Top N statistiky	Nejvíce komunikující porty (včetně počtu toků), nejvíce komunikující protokoly (včetně počtu toků).
Poměry	Počet unikátních zdrojových portů ku počtu unikátních zdrojových adres.

Tabulka 2. Přehled jednotlivých kategorií statistických údajů včetně příkladů konkrétních použitých statistik pro rozhodování modelu. Uvedené statistiky jsou z pohledu příchozího provozu, v případě odchozího provozu se počítají unikátní hodnoty zdrojových portů, protože nás zajímají především porty, které využívá sledované zařízení.

detekoval systém ADiCT 99,5 %). Spolehlivost informací o aplikacích je poměrně vysoká, protože pokud je aplikace detekována přímo v baneru služby, je vysoká pravděpodobnost běhu této služby, stejná míra spolehlivosti platí i pro operační systém. Informace ze systému PassiveDNS jsou spolehlivé, míra spolehlivosti informací z blacklistů záleží na organizaci provozující daný blacklist. Geolokační informace a informace ze systému WhoIs bývají spíše orientační.

5. Klasifikace síťových zařízení podle statistického chování

Pro klasifikaci síťových zařízení byl použit přístup využívající statistického chování zařízení. K výpočtu statistik jsou použity pouze základní Flow data[7]. Pro každé zařízení jsou vypočteny souhrnné statistiky (profil) za každou hodinu jejich provozu, tedy celkem 24 profilů zařízení za 1 den. Jednotlivé statistiky se dají rozdělit do 5 různých kategorií, jejichž základní přehled je uveden v tabulce 2. V tabulce jsou uvedeny 4 základní kategorie statistických údajů, které jsou vypočteny pro každou IP adresu pro každý směr provozu, tedy jednou jsou vypočteny pro všechny příchozí toky komunikace a poté jsou vypočteny pro všechny odchozí toky komunikace. Tento přístup umožňuje přidat pátou kategorii statistických údajů, kterou jsou poměry mezi příchozí a odchozí komunikací, které jsou vypočteny pro většinu statistik ze 4 základních kategorií jednosměrného provozu.

Všechny tyto údaje jsou uloženy jako profil zařízení po jednotlivých hodinách, které jsou poté použity ke klasifikaci zařízení. Jako algoritmus pro klasifikaci byly zvoleny náhodné lesy[8], protože u statistického chování produkují dobré výsledky, nejsou příliš náchylné na přetrénování (pokud je použit dostatečný počet stromů) a také je možné u každého modelu zjistit váhu jednotlivých rysů, kterou jim náhodný les přiřazuje.

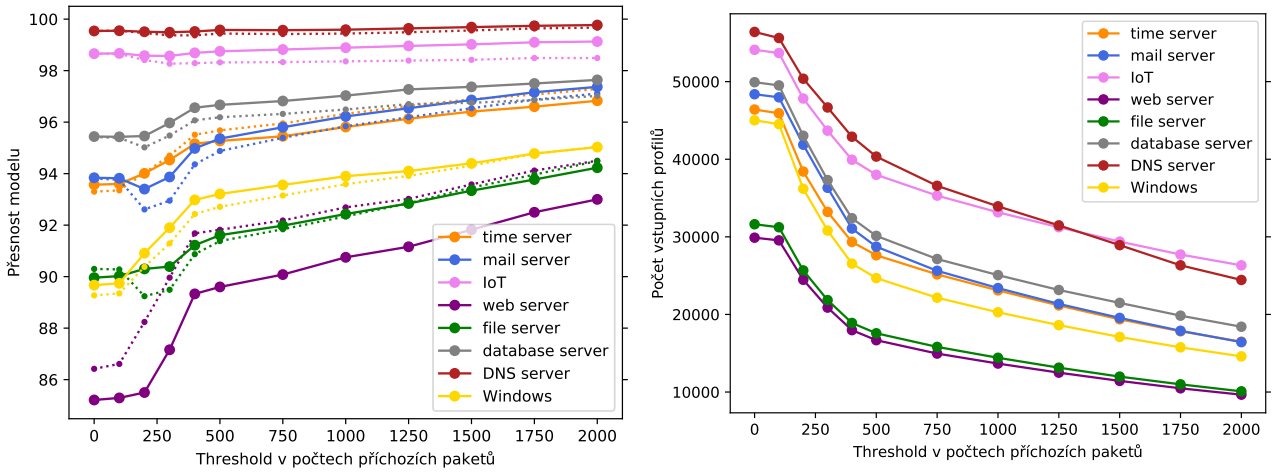
Jelikož každé zařízení v datové sadě může provozovat více služeb, byl zvolen přístup použití více klasifikátorů, kde každý z nich se zaměřuje na klasifikaci jednoho konkrétního typu zařízení. Každému profilu zařízení byl u konkrétního modelu přiřazen štítek dané služby tehdy, pokud je daný štítek přítomen v anotaci zařízení (i v přítomnosti více dalších štítků) a v případě nepřítomnosti je přiřazen štítek `other`. Prvotní výsledky přesnosti klasifikace jsou vidět v tabulce 3 (sloupec *bez ADASYN*). Lze vidět, že především metrika F1-score[9] je kvůli nevyváženosti některých štítků v datové sadě nízká. Z tohoto důvodu byla na minoritní třídy použita metoda generování syntetických vzorků ADASYN [10], která dává větší důraz na generování takových vzorků, v jejichž okolí se nachází méně ostatních vzorků a jsou tak obtížněji predikovatelné klasifikátorem. Výsledky po aplikaci metody ADASYN jsou zobrazeny v tabulce 3 (sloupec *s ADASYN*).

Po analýze klasifikačních výsledků po aplikaci metody ADASYN bylo odhaleno, že klasifikátor špatně klasifikuje převážně profily, které jsou význačné nízkou komunikací (málo provedených toků, přenesených paketů). Proto byly provedeny experimenty, které ukazují vliv nízké komunikace profilu na přesnost klasifikátoru. Experimenty znázorňující přesnosti izo-

[7] https://gitlab.com/Aisik/dataset_annotator/-/blob/master/additional_notes.md

[8] https://en.wikipedia.org/wiki/Random_forest

[9] https://en.wikipedia.org/wiki/F-score



Obrázek 3. Přehled přesnosti jednotlivých klasifikátorů klasifikujících určitý typ zařízení na základě velikosti thresholdu filtrace vstupních profilů určených ke klasifikaci (obrázek vlevo). Plná čára představuje celkovou přesnost klasifikátoru, přerušovaná čára představuje metriku F1-score. Obrázek vpravo ukazuje vliv určitého thresholdu na počet odfiltrovaných vstupních profilů.

Predikovaný štítek	**Přesnost; F1-score** bez ADASYN	**Přesnost; F1-score** s ADASYN
database server	92,7 %; 69,4 %	95,2 %; 95,3 %
DNS server	99,1 %; 82,9 %	99,5 %; 99,5 %
file server	89,9 %; 88,4 %	89,9 %; 89,7 %
IoT	97,6 %; 81,7 %	98,6 %; 98,6 %
mail server	91,2 %; 65,5 %	93,5 %; 93,4 %
time server	90,6 %; 69,8 %	93,0 %; 92,5 %
web server	84,4 %; 86,3 %	84,6 %; 85,9 %
Windows	85,4 %; 91,3 %	89,7 %; 90,0 %

Tabulka 3. Přehled přesností a metriky *F1-score* jednotlivých klasifikátorů před použitím metody ADASYN pro generování syntetických vzorků dat a po použití metody ADASYN.

lovaných klasifikátorů (klasifikujících pouze jeden daný typ zařízení) vzhledem k filtraci minimálního počtu příchozích paketů, které musí profil obsahovat, aby byl puštěn ke klasifikaci, jsou znázorněny na obrázku 3. Dále byla provedena analýza ideálního nastavení algoritmu náhodných lesů pomocí vyzkoušení mnoha různých kombinací (tzv. hyper parameter tuning), která vedla ke zlepšení klasifikace v nízkých jednotkách desetin procenta.

Z výsledků experimentů je vidět, že v případě nulové filtrace se přesnost jednotlivých modelů pohybuje v rozmezí 85 - 99 % a jejich přesnost se zvyšuje aplikací striktnějšího vstupního filtru (thresholdu) do klasifikátoru. Tento threshold ale nelze zvyšovat na příliš vysoké hodnoty, protože by jinak nedošlo ke klasifikaci výrazné části provozu. Proto je důležité najít vhodnou rovnováhu podle požadavků na přesnost dané klasifikace a požadavků na množství vstupních profilů síťových zařízení určených ke klasifikaci.

6. Závěr

Cílem práce bylo vytvořit nástroj, který usnadní anotaci datových sad síťových zařízení. Dalším cílem bylo demonstrovat využitelnost datové sady vytvořením klasifikátoru síťových zařízení.

Výsledný nástroj pro automatizovanou anotaci uživateli zjednodušuje anotaci datové sady na pouhou definici seznamu zařízení určených k anotaci a díky modulární struktuře umožňuje jednoduše přidat další datové zdroje potřebné k anotaci. Následně bylo vytvořeno několik klasifikátorů síťových zařízení využívajících statistického chování zařízení ke klasifikaci o průměrné přesnosti 93 %.

Výzkum různých využití metod strojového učení je v některých doménách značně omezen kvůli nedostatku trénovacích dat. Nové experimenty s automatizací procesu vytváření nových datových sad mohou toto omezení výrazně redukovat a časem určitě i odstranit úplně.

Automatizovaný anotátor může být v budoucnu rozšířen o další datové zdroje, které mohou obohatit a zpřesnit výstupy anotace. U klasifikace zařízení po-

dle statistického chování je možné se dále zaměřit na jejich dlouhodobé chování, např. tedy místo klasifikace zařízení podle chování za poslední hodinu vypočítat souhrnné statistiky za poslední týden.

Poděkování

Rád bych poděkoval svému vedoucímu práce Ing. Martinu Žádníkovi, PhD. za odborné vedení práce a informačně bohaté konzultace, které mi věnoval. Poděkování patří také Ing. Václavu Bartošovi, PhD. ze sdružení Cesnet, který mi také poskytl mnoho užitečných rad v průběhu práce.

Literatura

[1] Mingda Wang, Hangyu Hu, and Guangmin Hu. A survey on traffic-behavioral profiling of network end-target. In *Proceedings of the ACM Turing Celebration Conference-China*, pages 1–7, 2019.

[2] Pavel Novák. Identifikace podobných zařízení vzhledem k chování v počítačové síti. Bakalářská práce, Masarykova univerzita, Fakulta informačních technologií, 2020.

[3] Ahmad Jakalan, Jian Gong, and Shangdong Liu. Profiling ip hosts based on traffic behavior. In *2015 IEEE International Conference on Communication Software and Networks (ICCSN)*, pages 105–111. IEEE, 2015.

[4] Bingdong Li, Mehmet Hadi Gunes, George Bebis, and Jeff Springer. A supervised machine learning approach to classify host roles on line using sflow. In *Proceedings of the first edition workshop on High performance and programmable networking*, pages 53–60, 2013.

[5] John Matherly. Complete guide to shodan. *Shodan, LLC (2016-02-25)*, 1, 2015.

[6] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. A search engine backed by internet-wide scanning. CCS '15, page 542–553, New York, NY, USA, 2015. Association for Computing Machinery.

[7] Florian Weimer. Passive dns replication. In *FIRST conference on computer security incident*, page 98, 2005.

[8] S. Torabi, A. Boukhtouta, C. Assi, and M. Debbabi. Detecting internet abuse by analyzing passive dns traffic: A survey of implemented systems. *IEEE Communications Surveys Tutorials*, 20(4):3389–3415, 2018.

[9] Josef Koumar. Automatické rozpoznávání síťových zařízení a jejich závislostí. Bakalářská práce, České vysoké učení technické v Praze, Fakulta informačních technologií, 2020.

[10] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.