ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Learn and Predict Metasploit Exploit Rank from Available Vulnerability Information**
Jméno autora:	Karel GAVENČIAK
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	KK
Oponent práce:	Gustav Šourek
Pracoviště oponenta práce:	KP

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **náročnější**

Hodnocení náročnosti zadání závěrečné práce.

Zadani je relativne primocare avsak student musel kombinovat znalosti strojoveho uceni a sitove bezpecnosti.

Splnění zadání **splněno**

Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Zadani je splneno beze zbytku.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Zvoleny postup odpovida literature a autor pouzil odpovídající datove zdroje a klasifikacni metody.

Odborná úroveň **B - velmi dobře**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Práce je po odborne strance relativne dobra, autor pouziva standardni techniky a nastroje, nektere odborne úvahy, především v oblasti klasifikatoru a ROC krivek, mi prijdou trochu zavadejici, nicmene se však autor musel, nad ramec standardniho studia, seznámit tez s tematikou ohodnocovani exploitu v pocitacove bezpecnosti, coz ocenuji.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Práce je dobře strukturovana, prehledna, a psana velmi peknou a ctivou angličtinou.

Výběr zdrojů, korektnost citací **A - výborně**

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Vse v poradku.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Viz zaverecne hodnoceni.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Predlozena práce se zabyva predikovanim urovne zavaznosti exploitu, tj. chyb v pocitacovych systémech, které je možno zneuzit k prevzeti kontroly. Informace potrebne k predikci autor ziskava propojenim verejne dostupných databazi s historickými záznamy o znamych zranitelnostech (repozitar NVD) s udaji, které zranitelnosti slo skutecne zneuzit spolecne se stupnem spolehlivosti utoku (databaze Metasploit). Zdrojova data pak prevadi na feature vektory zpracovatelne standardnimi metodami strojového uceni, s jejichz pomoci predikcni ulohu stupne spolehlivosti resi. Autor diskutuje několik ruznych zpusobu extrakce a normalizace features z textového popisu zranitelnosti, provádí experimenty, interpretuje a zhodnocuje výsledky.

Text práce se mi celkove velmi libi, chvalim peknou angličtinu a obecnou typografii. Samotny postup mi prijde prehledny a velmi standardni, s pouzitim bezne dostupných technik a triviálních baseline pristupu. Ocenil bych o trochu vice inovace, nicmene chvalim experimenty se zavedenim features pro negace sloves, zpetnou intepretaci modelu a zhodnocuji ze na bakalářskou praci mi to prijde velmi pekne.

Jedinou faktickou vytku mam k obsazenym argumentacim temat strojového uceni. Například z tvaru vynesnych ROC krivek mam dojem, ze všechny obsahuji pouze jediny bod a tedy ze byl uvazovan pouze jediny klasifikator. To odpovida i faktu, ze přes podrobny popis zakladni funkce modelu SVM a Random Forest není popsano, jakym zpusobem lze získat vysledne skore kazde zranitelnosti, tj. nejaky odhad posteriorni pravdepodobnosti prislusnosti do dane tridy (napr. logisticka regrese vzdalenosti od rozdělující nadplochy SVM, nebo celkove pomery trid ve vybranych listech stromu u Random Forest). Bez tohoto skore nevidim ve vyhodnocovani pomoci ROC krivek prilis velky smysl a tez prislusny prechod z multi-class klasifikace na binarni verze, s agregovanim jejich ROC score, mi prijde trochu zavadejici, vhodnejsi by bylo spise uvest napr. matice zamen. Podobne nektere argumentace kolem samotnych klasifikatoru, napr. ze diky velkému množství dat jste byli schopni vytvořit Random Forest se 100 stromy, jsou trochu unahlene (RF pridavanim stromu neoverfituji, pokud Vas to napadlo). V nasledne modifikaci ulohy na binarni verzi predikce samotne existence exploitu je však vse v poradku a autor tez dosahuje velmi vysoke presnosti.

Otazky na autora jsou tedy, v kontextu předchozího odstavce, nasledujici:

1) Vyjasnete vypocet TPR a FPR (str. 22), predevsim “Predicted positive/negative” a spravnost vypoctu ROC.
2) Vyjasnete prosim vase pouziti ROC krivek a postup vypoctu celkových vyhodnocovacich metrik (AUC a accuracy) pro vami predstaveny multi-class klasifikacni problem.
3) Nalezi tedy vsechny ROC krivky pres ruzne tridy v jednom grafu stejnemu multi-class klasifikatoru?
4) Nasledne mi pak většina ROC krivek prijde velmi blizko diagonale, zhodnotte prosim co to znamena a vztah k celkove accuracy.

Az na tyto drobne nejasnosti mi práce prijde velmi pekna (pripadne pri jejich vyjasneni i vyborna), a tez prilozeny kod hodnotim jako prehledny.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **B - velmi dobře.**

Datum: 5.6.2016 Podpis: Šourek