Problematika předzpracování dat v úlohách dobývání znalostí

Doktorand: Ing. Jiří Zettel

Školitel: prof. Ing. Petr Berka, CSc.

Téma práce

- Proč je nutné data předzpracovat?
  - Specifické požadavky na reprezentaci dat pro úlohy KDD
  - Problémy s kvalitou dat
- Činnosti které budu řešit
  - Formát
  - Chybějící a odlehlé hodnoty
  - Příliš mnoho objektů, atributů
  - Numerické atributy (diskretizace)
  - Kategoriální atributy
  - Nevyvážené rozdělení dat
  - Vytváření odvozených atributů

Téma práce

- Převážně první 3 kroky z metodiky CRISP-DM
  1. Porozumění problematice
  2. Porozumění datům
  3. Příprava dat (60 – 80% času)
  4. Modelování
  5. Vyhodnocení výsledků
  6. Využití výsledků

Téma práce

- Zvolená aplikační oblast = bezpečnost IS
  - Síťově orientované systémy detekce průniku, zaměřené na anomálie
    (anomaly-based network intrusion detection system)
  - Používají nástroje KDD pro detekci anomálií
  - Zkoumaná data = data ze síťové komunikace (vybraná podmnožina)
  - Převážně znamená detekci odlehlých hodnot
    - Ve stylu učení s učitelem (trénovací množina obsahuje data s označením třídy)
    - Nebo semi-supervised learning (trénovací množina obsahuje pouze data bez útoků)
    - Nebo bez učitele (data bez označení třídy)

Aktuální stav výzkumu

Projekty s předzpracováním dat pro KDD

1. Mining Mart
   - vznikl na univerzitě v Dortmundu, cca 2000 - 2006
   - zaměřili se na nastavení osvědčených postupů pro předzpracování dat a jejich znovu použití
2. SumatraTT 2 (Transformation Tool)
   - ČVUT FEL, centrum aplikované kybernetiky, r. 2000 – 2008
   - předpracování dat různých zdrojů (CSV, SQL, DBF, XML, Weka, Lisp, apod.)
3. DataPreparator
   - Autorka pí. Božena Stewart, University of Western Sydney
   - Zabývala se různými typy operátorů pro předzpracování dat (čištění, diskretizace, chybějící a odlehlé hodnoty, odvozené proměnné, numerace nominálních atributů, škálovatelnost numerických atributů, atd.)

Předzpracování dat pro systémy detekce průniku

1. směr výzkumu: Zpracování pouze základních rysů hlaviček paketů
   - PŘÍKLAD
     - projekt PHAD (Packet Header Anomaly Detector, Mahoney a Chan, 2001)
     - Ve fázi učení prochází všechny pole hlaviček paketů a počítá rozmezí hodnot na úrovni spojové vrstvy (Ethernet), síťové (IP) a transportní (TCP, UDP, ICMP)

2. směr výzkumu: Rysy odvozené z jednotlivých síťových spojení (SCD)
   - datové instance jednosměrné síťové spojení, ne pouze jednotlivé pakety
   - nejpoužívanější rysy patří časová statistická měření (např. počty paketů, střední délka paketů, apod.)
   - PŘÍKLAD
     - Yamada a kol. (2007) používá pro detekci útoků na webové servery v případě šifrovaných spojení (SSL nebo TLS)

Předzpracování dat pro systémy detekce průniku

3. směr výzkumu: Rysy odvozené z vícenásobných spojení (MCD)
   - Základní rysy v rámci několika toků dat
   - PŘÍKLAD
     - systém používající NetFlow záznamy (logy), vyvinul Lakhina a kol. (2005)
     - počítá entropii v délce pěti minut pro vybrané rysy
     - pakety, které skenují porty počítače, budou mít nízkou entropii pro cílovou IP adresu, ale vysokou entropii pro cílový port

4. směr výzkumu: Detekce anomálií podle obsahu uživatelských dat paketu
   - Detekce časově náročnější
   - Několik článků se zabývá N-gram analýzou
   - PŘÍKLAD
     - PAYL (Payload-based Anomaly Detector) - vytvořen autory Stolfo a Wang (2004)
     - Každý datový obsah produkuje frekvenční distribuci bajtů, ten je pak centroid (model)
     - Při detekci se počítá Mahalanobisova vzdálenost

Disertační práce

Hypotézy

1. Normální data (bez útoků) za síťové komunikace určité služby (např. www) mají společné charakteristiky
2. Data s útokem se od normálních dat odlišují (anomálie)
   - Anomálie je možné detekovat pomocí metod KDD
3. Je možné navrhnout společné prvky pro předzpracování dat z určité služby
   - Je možné definovat také typový postup (metodika)
   - Je možné definovat nové procedury

Vědecké metody

1. fáze - Analýza / syntéza
   - Výzkumné otázky převážně z činnosti „Porozumění problematice a datům“
     - Jak detailně fungují zkoumané síťové služby (www)?
     - Jak vypadají normální data?
     - Jak vypadají útoky, anomálie?
     - Jaké jsou problémy kvality dat?
     - Jaké jsou potřeby pro předzpracování dat?
     - Jaké je současné vědecké poznání? (rešerše)

Vědecké metody

2. fáze - Návrh artefaktů a validace

- Činnost „Příprava dat“
  - Transformace dat na záznamy v datové matici
  - Vytvoření datové množiny (trénovací a testovací)
  - Návrh nových rysů a dalších transformací
    - např. shlukování, feature extraction, redukce (irrelevantní rysy, zmenšení dimenzionality), apod.
    - Implementované v nástroji RapidMiner nebo vytvořené v jazyce Python (možno v rámci RapidMiner) např. pomocí open-source knihovny Pandas

PROTOTYP

Vědecké metody

2. fáze - Návrh artefaktů (validace)
   - Aplikování metod KDD na předzpracovaná data
   - Ověření úspěšné detekce anomálií
   - Výzkumné otázky
     - Je předzpracování úspěšné pro detekci útoků?
   - Re-implementace algoritmů

   - Metodika předzpracování dat pro systémy detekce průniku
   - Případně popis datových transformací dle standardu PMML (Predictive Model Markup Language)

4. fáze - Případová studie
   - Výzkumné otázky k ověření znovu použitelnosti
     - Jsou navržené artefakty opětovně použitelné (pro nové případy)?

   - (Analýza) možnosti implementace jako plugin do open-source systémů IDS (např. Snort)

Předpokládaný přínos a cíle

- Vědecký přínos
  - Navázat na rozpracovaný výzkum
  - Vyřešit některé problémy s předzpracováním dat nastíněné v jiných výzkumech, specificky pro systémy detekce průniku
  - Navrhnout vhodné metody předzpracování dat, tak aby zohledňovaly aktuální data současných webových aplikací a aktuální útoky
- Přínos pro praxi
  - Možná implementace algoritmů nebo procedur do komerčních systémů (open-source)

Harmonogram

Rok	-	Výzkumná činnost	Publikační činnost
2	Zima	• Povinné studium, Rešerše • Navázání kontaktu s významnými pracovišti • **Obstarávání a prozkoumávání dat** (veřejně dostupných, případně vytvořených na vlastním web serveru nebo jiné)	
2	Léto	• Povinné studium, Rešerše • **Prozkoumávání dat** ze síťové komunikace • **Transformace dat** na záznamy v datové matici • Hledání podobných probíhajících projektů na významných pracovištích	
3	Zima	• **Transformace dat** na záznamy v datové matici • Volba vhodné metody dobývání znalostí • Prozkoumání možnosti podílet se na projektu probíhajícím na významném pracovišti	• Vystoupení na dni doktorandů
3	Léto	• **Návrh nových rysů** a dalších transformací • **Implementace a testování** algoritmů	• Možná publikační činnost jako spoluautor

Harmonogram pokračování

Rok	-	Výzkumná činnost	Publikační činnost
4	Zima	• **Návrh nových rysů** • **Implementace a testování** algoritmů • Detekce anomálií (útoku) pomocí zvolené metody dobývání znalostí	• Konferenční příspěvek pro "Data a znalosti"
4	Léto	• Soupis **metodiky** • **Detekce anomálií** (útoku) pomocí zvolené metody dobývání znalostí • Prozkoumání další možné implementace jako součást komerčních systémů (open-source)	• Publikační činnost ve zvoleném časopise
5		• **Případová studie** na datech ze zvolené oblasti • Porovnání výsledků předzpracování dat s jinými autory	• Vystoupení na konferenci "Interdisciplinary Information Management Talks" (Cyber Security topic) nebo na vybrané mezinárodní konferenci • Publikační činnost ve zvoleném časopise