Problematika předzpracování dat v úlohách dobývání znalostí

Disertační projekt

Doktorand: Ing. Jiří Zettel
Školitel: prof. Ing. Petr Berka, CSc.
Katedra informačního a znalostního inženýrství, VŠE Praha

Rozšířený abstrakt

Předzpracování dat je důležitým krokem v procesu dobývání znalostí. Podle praktických zkušeností zabere cca 60-80% z celého procesu dobývání znalostí. [1] Na jedné straně se data potýkají s problémy s kvalitou dat a na druhé straně existují specifické požadavky konkrétních modelů na reprezentaci dat.

Navrhovaný projekt se pohybuje v prvních třech fázích metodiky pro dobývání znalostí CRISP-DM. Konkrétně se tedy budu věnovat porozumění problematice, porozumění datům a samotnému předzpracování dat. Pro předzpracování dat jsem vybral konkrétní aplikační oblast, jedná se o **systémy detekce průniku založené na rozpoznání anomálií** (A-NIDS, anomaly-based network intrusion detection system). Tyto systémy v reálném čase analyzují síťovou komunikaci a používají nástroje dobývání znalostí k detekci anomálií. V současné době se jako nevíce perspektivní oblast pro výzkum předzpracování dat jeví systémy, které posuzují anomálie podle obsahu uživatelských dat paketů.

Předzpracováním dat pro metody dobývání znalostí se zabývalo několik výzkumných projektů. Na univerzitě v Dortmundu v letech 2000-2006 vznikl **Mining Mart**. V něm se zaměřovali na nastavení osvědčených postupů pro předzpracování dat a jejich znovu použití. Rozpracovali typové úlohy převážně pro marketing. Na ČVUT FEL, v centru aplikované kybernetiky vznikl v letech 2000-2008 projekt **Sumatra TT** (Transformation Tool). Zabývá se předzpracováním dat z různých zdrojů. Další významný projekt je **DataPreparator** vyvinutý pí. Boženou Stewart na University of Western Sydney. Autorka se zabývala různými typy operátorů pro předzpracování dat (čištění, diskretizace, chybějící hodnoty, odlehlé hodnoty, odvozené hodnoty, numerace, apod.).

V provedené rešerši metod předzpracování dat pro systémy detekce průniku jsem se setkal se čtyřmi hlavními směry výzkumu. První se zabývá zpracováním pouze **základních rysů hlaviček paketů**. Hlavičky paketů se mohou analyzovat na úrovni spojové, síťové a transportní vrstvy modelu OSI. Druhý směr výzkumu se zabývá **rysy odvozenými z jednotlivých síťových spojení** (single connection derived features). Zde jako datové instance používají jednosměrné spojení, ne pouze jednotlivé pakety. Mezi nejčastější rysy zařazují časová statistická měření (počty paketů, střední délka paketů, apod.). Třetím směrem by pak byly **rysy odvozené z vícenásobných spojení**.

[1] D. Pyle, *Data preparation for data mining*, vol. 1. morgan kaufmann, 1999.

Tam jsou rysy vytvářeny v rámci několika toků dat. Např. počítaná entropie v délce pěti minut pro vybrané rysy (IP adresa, port). Posledním směrem výzkumu je detekce anomálií podle **obsahu uživatelských dat paketu**. Tam jsou detekce časově nejnáročnější. Několik článků se zabývá N-gram analýzou.

Pro svůj projekt jsem stanovil několik **hypotéz**. Předpokládám, že normální data (bez útoků) ze síťové komunikace určité služby (např. www) mají společné charakteristiky. Tyto charakteristiky pak umožňují vytvořit model normálních dat pomocí metod KDD. Data s útokem se pak od normálních dat odlišují, obsahují tzv. anomálie. V posledním kroku předpokládám, že je možné také navrhnout společné prvky pro předzpracování dat z určité služby. Stanovené hypotézy bych rád ověřil prostřednictvím návrhu SW artefaktů, tak abych navázal na předchozí výzkumy a přinesl další nová poznání. Rád bych také vyřešil některé nastíněné problémy s předzpracováním dat.

V první fázi tedy detailně zkoumám chování zvolené síťové služby a data generovaná při používání služby (převážně forma textových logů). Následně budu data převádět na matici, řešit problémy s kvalitou, zkoumat v datech útoky a v neposlední řadě také specifické potřeby pro zvolenou metodu KDD. Budu pracovat na vytvoření **nových procedur** pro předzpracování dat (např. shlukování, feature extraction, redukce, apod.). Tyto algoritmy plánuji implementovat v nástroji RapidMiner s použitím jazyka Python (např. open-source knihoven Pandas). Algoritmy budu opakovaně testovat, musí být vhodné k vytvoření modelu KDD. Vedle konkrétních algoritmů bych rád uvedl **ucelenou metodiku** popisující kroky předzpracování dat, případně také popis datových transformací dle standardu **PMML**. **Validace** artefaktů pak bude provedena testem úspěšné detekce útoků podle anomálií. To znamená, že vytvořený model aplikuji na množinu testovacích dat. Na závěr bych rád ověřil znovu-použitelnost artefaktů formou **případové studie** pro nová data. Vzhledem k tomu, že celý výstup výzkumu bude na úrovni prototypu softwaru, rád bych nakonec vyhodnotil budoucí možnost implementace jako plugin do open-source systémů IDS (např. Snort).

Aktuální status projektu je práce na rešerši metod předzpracování dat pro systémy detekce průniku. Vedle toho si současně také obstarávám data z různých zdrojů (např. publikovaná data pro KDD Cup 1999) a tato data převádím na záznamy v datové matici vhodné pro zpracování metodami KDD. Datový zdroj zavádím do nástroje RapidMiner a budu vytvářet modely pro detekci. Nyní se tedy věnuji přípravě prostředí a analýze některých zveřejněných výzkumů / projektů. V následujících obdobích (2. – 5. ročník) budu pracovat na následujících aktivitách: obstarávání a prozkoumávání dat, transformace dat na záznamy v datové matici, volba vhodné metody KDD, návrh nových rysů a dalších transformací, implementace a testování algoritmů, soupis metodiky, detekce anomálií, případová studie.

Plánuji spolupracovat s kolegy z VŠE, kde máme průnik v tématech (kontaktní osoba doc. Svátek). Rád bych také navázal kontakt a případnou spolupráci s jiným významným pracovištěm. Co se týče konferenční činnosti, plánuji vystoupení na konferenci „Data a znalosti“ a později na konferenci "Interdisciplinary Information Management Talks" (Cyber Security topic). U publikací chci začít s lokálními.