Analýza prúdu prichádzajúcich udalostí použitím rôznych metód pre analýzu údajov

Výskumný zámer, Informačné systémy, FIIT STU

Matúš Cimerman

Vedúci práce: Ing. Jakub Ševcech

V súčasnosti pozorujeme vysoký nárast záujmu v oblasti analýzy údajov. Vhodné použitie metód a techník na analýzu prináša hodnotné výstupy pre používateľa, ktoré môžu môžu byť použité pre strategické rozhodnutia v podnikoch. Najčastejší postup je aplikovaním metód ako napríklad lieviková analýza alebo rozhodovacie stromy nad statickou kolekciu dát. Avšak tento prístup neposkytuje okamžitý výsledok pre používateľa. Nedostatkom takýchto prístupov je nutnosť dáta najskôr zozbierať a uložiť, čo je dnes, kedy vznikajú milióny záznamov za deň, veľký problém.

1 Analýza stavu problematiky

Pri dolovaní v prúde dát čelíme niekoľkým výzvam: objem, rýchlosť (frekvencia) a rozmanitosť. Veľký objem dát, ktoré vznikajú veľmi rýchlo je potrebné spracovať v ohraničenom časovom intervale, často v takmer reálnom čase (závisí od kontextu problému). Objem dát sa neustále zvyšuje, potenciálne až do nekonečna. Identifikujeme niekoľko najviac zasiahnutých oblastí, ktoré sú zdrojmi týchto dát, a to senzory, počítačové siete, sociálne siete a Internet Vecí (angl. Internet of Things). Na informácie generované z takýchto zdrojov sa často pozeráme ako na neohraničené a potenciálne nekonečné prúdy údajov.

Spracovanie a nasledujúca analýza týchto prúdov je komplexná úloha. Pre aplikácie je kritické spracovať údaje s nízkou odozvou, pričom riešenie musú byť presné, škálovateľné a odolné voči chybám. Nakoľko sú prúdy neohraničené vo veľkosti a potenciálne nekonečné, môžeme spracovať len ohraničený inter-

val prúdu. Potom hovoríme, že dáta musia byť spracované tak ako vznikajú. Tradičné metódy a princípy pre spracovanie statickej kolekcii údajov nie sú postačujúce na takéto úlohy.

Dolovanie a extrakcia informácií z prúdiacih dát dnes predstavuje niekoľko výziev. Jednou z oblastí je bezpečnosť a dôvernosť v dolovaní dát. Návrh modelu môže byť navrhovaný na pôvodných dátach, ale nasadenie by malo byť realizované na anonymizovaných dátach. Identifikujeme dve hlavné výzvy pre uchovanie bezpečnosti pri dolovaní v prúdoch. Prvou je nekompletnosť informácií, informácie často prichádzajú nekompletné alebo neaktuálne. Druhá výzva nadväzuje na prvú, a to že dáta sa môžu v čase meniť a vyvýjať. Môže sa meniť štruktúra, komplexita a ich presnosť. Preto, vopred definované bezpečnostné pravidlá a politiky nemusia byť časom aktuálne a pokrývať stanovenú oblasť. Medzi známe techniky dolovania v prúdoch údajov považujeme nasledujúce:

- Klastrovanie, existuje niekoľko výskumov, ktoré sa venovali špeciálne klastrovaniu implementovaním napríklad k-mediánu a inkrementálnych algoritmov.
- Klasifikácia, opäť existuje niekoľko známych výskumov, ktoré s venujú problému klasifikácie. Napríklad použitím dát z reálneho sveta a umelých dát, pričom implementujú algoritmy, ktoré triedia dáta na základe porovnaní medzi týmito dvoma vzorkami.
- Počítanie frekvencie a opakovaní, použitím posuvných okien a inkrementálnych algoritmov na detekciu vzorov v prúde.
- Analýza časových radov použitím symbolickej reprezentácie šasových radov v prúde dát. Takáto reprezentácia nám umožňuje redukciu veľkosti prenášaných dát. Táto technika pozostáva z dvoch hlavných

krokov, aproximácia po častiach a následná transformácia výsledku do diskrétnych veličín.

Predspracovanie prúdu je dôležitým krokom v aplikáciach reálneho sveta. Nakoľko dáta prichádzajú z nehomogénneho sveta, môžu byť zašumené, nekompletné, duplicitné alebo často obsahovať hodnoty, ktoré sa značne líšia od ostatných. Predspracovanie prúdiach údajov je potrebné čo najviac automatizovať. Existuje potreba pre implementovanie metód strojového učenia, ktoré sa adaptujú v čase s meniacimi sa dátami. Avšak, tieto modely a metódy by mali byť synchronizované s prediktívnymi modelmi. Pri aplikovaní strojového učenia a prediktívnych modelov je nutnosť uvažovať historické dáta. Čím podstatne narastá zložitosť celého prístupu. Hlavné výskumné problémy rozdeľujeme do dvoch hlavých kategórií, *dátovo orientované* a *orientované na úlohu*. Dátovo orientované techniky používané pri spracovaní prúdiacich údajov:

- Vzorkovanie je proces výberu dátovej vzorky na spracovanie podľa pravdepodobnostného modelu. Problém pri vzorkovaní je v kontexte analýzy prúdiach dát je potenciálne nekonečný dataset, resp. nevedomosť jeho skutočnej veľkosti.
- Kontrolovanie zaťaženia je proces zahadzovania niektorých, potenciálne nepotrebných vzoriek dát. Opäť tu nástava problém v kontexte prúdu. Môže nastať moment, kedy práve zahodená vzorka dát je pre aktuálnu analýzu príznačná a dôležitá.
- Agregácia je proces vypočtu štatistických údajov nad prúdmi. Problém je pri veľkých tokoch v prúde, pričom je tiež potreba pohľadu do minulosti.
- Aproximačné algoritmy použitie aproximačných algoritmov má za následok podstatné zrýchlenie spracovania a analýzy prúdov za predpokladu

istej chybovosti. Chybovosť je zvačsa ohraničená.

- Posuvné okno, tento prístup vznikol s potrebou analýzy definovaného časového okna z prúdiacih údajov.

2 Rozpracovanie problému

Analýza prúdu dát a vo všeobecnosti analýza dát predstavuje najväčšiu výzvu v oblasti informácií. Pre správne aplikovanie analytického modelu a teda vykonanie analýza je často potrebné mať vynikajúce znalosti z oblasti štatistiky, či strojového učenia. Zároveň je nutnosť mať vynikajúcu znalosť domény odkiaľ pochádzajú dáta. Analytici, ktorí sú experti v danej doméne často nemajú detailné znalosti z oblasti štatistiky, či strojového učenia sa. Existuje preto potreba navrhnúť nástroje, ktoré pomôžu v analýze prúdiach údajov bez toho aby museli byť špecialistami ďalších doménach, ktoré niesú až tak potrebné pre analytika. Je pochopiteľné, že analytik musí poznať model, ale nieje nutné aby poznal aj jeho detailné vnutorné fungovanie.

Takéto prístupy a metódy vznikli postupom času pre analýzu statických kolekcií údajov, napríklad lieviková analýza (angl. funnel analysis) alebo metóda vnárania sa (angl. drill down). Tieto metódy sú jednoduché na použitie a výsledky sú pochopiteľne interpretovateľné aj pre menej znalých používateľov. Avšak, tieto metódy niesú tak, ako boli navrhnuté, použiteľné pri analýze prúdu prichádzajúcich udalostí.

Zatiaľ čo väčšina algoritmitkých metód pre prúdy údajov je dostupná a preskúmaná, výzvu predstavuje ich nasadenie a aplikovanie v reálnom svete. Jednou z takýchto výziev je vytváranie čo najjednoduchších a pritom špecializovaných modelov, ktoré sú reaktívne. Jeden z aspektov jednoduchého modelu je správna kombinácia novovznikajúcich údajov (online) a offline údajov.

Použitie online a offline modelov sa vzájomne vylučuje, avšak práve ich kombinácia môže najviac obohatiť celý analytický model o poznanie z histórie (offline údaje). Spracovanie offline resp. statických kolekcií údajov je veľmi dobre preskúmaná oblasť kde spracovanie prebieha vo väčšine prípadov v dávkach (angl. batch processing) s použitím MapReduce modelu. Spracovanie online údajov spolu so strojovým učením môže prispieť pri rozhodovaní sa v reálnom čase pre dosiahnutie okamžitých výsledkov alebo úpravu parametrov modelu.

Adaptívne modely a systémy sú často nastavené rôznymi parametrami. Je preto snaha minimalizovať závislosti medzi týmito parametrami. Je tiež dôležite minimalizovať počet parametrov, ktoré môže meniť a nastavovať používateľ z pohľadu použiteľnosti systému.

3 Tézy

V našej práci sa zameriavame na výber a aplikovanie správneho analytického modelu v oblasti spracovania prúdiach údajov. Čiastočne chceme skúmať tiež závilosti parametrov zvoleného modelu a tiež dopad alebo efektívnosť pridania offline modelu. Medzi hlavné výskumné problémy v tejto oblasti zaraďujeme:

- Predspracovanie, dolovanie a spracovanie prúdiach údajov v takmer reálnom čase.
- Aplikovanie správneho analytického modelu a jeho udržateľnosť v čase.
- Minimalizácia závislostí a možná adaptácia vnútorných parametrov modelu.
- Kombinácia online a offline modelu v situáciach kde to má zmysel.

[1] KREMPL, Georg, et al. Open challenges for data stream mining research. ACM SIGKDD Explorations Newsletter, 2014, 16.1: 1-10.
Diskusia o ôsmich otvorených problémoch pri dolovaní a analýze v prúdoch dát.
[2] GABER, Mohamed Medhat; ZASLAVSKY, Arkady; KRISHNASWAMY, Shonali. Mining data streams: a review. ACM Sigmod Record, 2005, 34.2: 18-26.
Prehľadový článok o teoretických základoch dolovania v prúdoch dát. Známe techniky pre dolovanie a systémy. Aktuálne výskumné vázvy v tejto oblasti.