CLASSIFICATION IN DATA STREAMS USING ENSEMBLE METHODS

Martin Jarosch
Master Degree Programme (2), FIT VUT
E-mail: xjaros34@stud.fit.vutbr.cz

Supervised by: Martin Hlosta
E-mail: ihlosta@fit.vutbr.cz

Abstract: In today's world we produce data faster and in bigger scale than ever before. This raises need for online analysis of tremendous amounts of data, which are constantly produced. Data stream mining is one of the means that allow us to do so. This work gives brief introduction into data stream classification and presents three fundamentally different state-of-the-art algorithms based on ensemble of classifiers. Finally this work provides plan of implementation for ensemble classification system.

Keywords: data stream, classification, ensemble, CSHT, DoS, DWAA, MAS, system

1 ÚVOD

Požadavky na rychlé a přesné zpracování obrovského, teoreticky až nekonečného množství dat, vytváří nové prostředí proudu dat pro již známé úlohy dolování z dat. Jednou z těchto úloh je klasifikace v proudu dat, které se v minulých letech dostalo zvýšené pozornosti a v budoucnosti bude nabývat na důležitosti. Cílem tohoto příspěvku je seznámit čtenáře s problematikou a aktuálními trendy klasifikace v datových proudech. V první části textu je uveden stručný úvod do klasifikace v proudu dat. Následně jsou představeny tři algoritmy, které budou implementovány a nakonec je popsán návrh jejich implementace.

2 KLASIFIKACE

Klasifikace je formou analýzy dat, umožňující extrahovat skryté informace, které je možné použít pro tvorbu inteligentních rozhodnutí. Dokáže rozdělit, klasifikovat data do jednotlivých tříd, určených diskrétním neseřazeným návěštím. Jsme pak schopni odhadnout hodnoty diskrétních atributů na základě předchozích dat.

Klasifikátor je nutné nejprve natrénovat na již označených datech, kdy při dostatečném množství trénovacích dat jsme schopní vytvořit modely, které odpovídají s velkou přesností současnému trendu dat. Na základě naučeného modelu se pak algoritmus rozhodne, jakou kategorii prvku přidělí.

3 DATOVÉ PROUDY

Datové proudy se vyznačují několika specifickými vlastnostmi, které dělají jejich dolování obtížnější. Jsou potenciálně nekonečné, nevíme kdy a jestli vůbec někdy skončí. Data tudíž nelze uložit a analyzovat v celku jako při běžném dolování, musíme vycházet ze získaných modelů a sumárních charakteristik. Algoritmy musí být jednoprůchodové, náhodný přístup k datům je příliš drahý. Data se můžou rychle a dynamicky měnit a to obsahem i rychlostí přijímaných dat. Proto bývá z pohledu uživatele i zpracování dat vyžadována odezva v reálném čase.

Dalším podstatným problémem je *Concept drift*, neboli změna konceptu dat. Nastává z důvodu časové proměnlivosti obsahu datových proudů. Problémem je, že mnoho skutečných aplikací závisí na skrytých souvislostech, ovládaných pro nás skrytými nebo neznámými změnami. Změnu konceptu rozlišujeme na změnu *náhlou* nebo změnu *pozvolnou*. Klasifikační algoritmus proto musí být dostatečně citlivý, aby zachytil i pozvolnou změnu konceptu, musí rychle reagovat na náhlé změny, ale zároveň musí být robustní proti šumu.

4 KLASIFIKACE POMOCÍ SOUBORU KLASIFIKÁTORŮ

Klasifikace pomocí souboru klasifikátorů se liší od dříve vyvinutých metod snahou použít několik klasifikačních modelů zároveň. Nepoužíváme jeden komplexní model, ale řídíme množinu jednoduchých klasifikačních algoritmů (např. Naive Bayes). Tento přístup je podporován jak teoretickými, tak i praktickými důvody. Podstatou je získat co nejvíce expertních názorů a ty následně sloučit do jednoho výsledku, který zaručuje vyšší správnost. Pokud natrénujeme několik rozdílných klasifikátorů a při klasifikaci budeme převádět kombinaci jejích výstupů na jeden, můžeme, ale nemusíme překonat přesnost nejlepšího klasifikátoru. Určitě však dosáhneme výrazného snížení rizika produkce zvláště špatného výsledku.

Vzhledem k rapidnímu vývoji nových algoritmů, využívajících soubor klasifikátorů, byly vybrány tři, principiálně rozdílné algoritmy pro implementaci.

4.1 ALGORITMUS CSHT

Algoritmus detekuje změnu konceptu pomocí testu hypotézy (*hypothesis test*). Algoritmus přijímá jednotlivé trénovací bloky dat a předpokládá, že nové vzorky pro klasifikaci mají stejnou distribuci dat, jako poslední přijatý blok trénovacích dat. Proto je nutné po přijetí nového trénovacího bloku dat, aktualizovat celý systém, aby odpovídal nové distribuci dat.

Soubor klasifikátorů se aktualizuje natrénováním nového klasifikátoru na nejnovějším bloku dat a přezkoumáním, zda se stávající klasifikátory shodují s aktuálním konceptem. Porovnání konceptů dat se provede detekcí změny distribuce dat již natrénovaných modelů a aktuálního bloku dat za pomoci testu hypotézy. Pro podrobnější pochopení doporučuji specifikaci algoritmu [1]. Došlo-li ke změně, je klasifikátor vyloučen. Použitelným klasifikátorům je následně přiřazena váha dle jejich úspěšnosti a výsledek je získán váženým hlasováním.

4.2 ALGORITMUS PRO DETEKCI DOS

Tento algoritmus uvedený v [2] byl navržen pro detekci datových proudů způsobujících DoS útoky, ale po provedení několika úprav může být použit pro obecnou klasifikaci. Odlišností algoritmu oproti ostatním je, že trénuje více klasifikačních algoritmů na jednom bloku dat.

Na každém novém bloku dat jsou natrénovány různé klasifikační algoritmy. Následně se porovná jejich úspěšnost a vybere se množina klasifikátorů s nejlepšími výsledky na novém bloku dat. Algoritmus pak dosahuje lepších výsledků díky větší rozmanitosti klasifikačních algoritmů.

4.3 ALGORITMUS DWAA

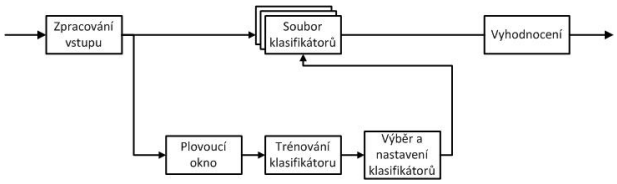
Poslední metoda je založena na odměňovací strategii. V mnoha metodách je použito odměňování nebo penalizace klasifikátorů dle jejich výsledků, ve většině jsou např. přidávány pevné hodnoty. Algoritmus DWAA z [3] velikost odměny klasifikátoru zvyšuje v poměru k počtu správných odpovědí, ostatních klasifikátorů. Pokud správně klasifikuje pouze jeden klasifikátor, je mu výrazně, ale ne příliš zvýšena váha, aby mohl snadněji ovlivnit celkové rozhodnutí. Opatrnost je na místě, protože se může jednat o šum.

Klasifikátory se trénují na předposledním bloku dat a vyhodnocují se na nejnovějším. Vyhodnocení vynuluje váhy klasifikátorů a dle správných odpovědí přičte odpovídající odměny. Nejhorší klasifikátor je pak nahrazen novým. Nakonec je možné provést úpravu vah, aby byl rozdíl mezi lepšími a horšími klasifikátory větší.

5 NÁVRH IMPLEMENTACE

Pro implementaci zmíněných algoritmů byl vytvořen jednotný systém dle obrázku 1. Systém se skládá z několika modulů, které můžou být dále přizpůsobeny. Zpracování vstupu obstarává příjem dat, ta jsou následně rozdělena na trénovací, již označená data a data pro klasifikaci. Trénovací data se hromadí v plovoucím okně, dokud není jejich počet dostatečný pro natrénování nového klasifikátoru. Následně se dle implementace algoritmu aktualizuje soubor klasifikátorů a plovoucí okno se vyprázdní. Neoznačená data jsou co nejdříve klasifikována jednotlivými klasifikátory a dle vybraného algoritmu je zvolena výstupní hodnota. Výsledek je pak odeslán na výstup. Pro klasifikaci byl implementován klasifikačních algoritmus Naive Bayes a budou implementovány i další např. rozhodovací stromy.

Vzhledem k možné rychlosti datového proudu, budou algoritmy i celý systém naprogramovány v jazyce C# a .NET Frameworku 4, který poskytuje funkce pro efektivní paralelismus a rozložení zátěže.



Obrázek 1: Zpracování datového proudu.

6 ZÁVĚR

Cílem práce je porovnat přínosy a výkon jednotlivých přístupů. Na základě získaných výsledků můžeme vypozorovat vodítka pro možný návrh modifikace těchto algoritmů. Implementované algoritmy budou využity v MAS (*Malware analysis system*), analytickém systému vyvíjeném v rámci projektu Systém pro zvýšení bezpečnosti v prostředí Internetu analýzou šíření škodlivého kódu, který je implementován na FIT VUT. Právě z tohoto systému budou pocházet data, pro vyhodnocení algoritmů.

PODĚKOVÁNÍ

Tento příspěvek vznikl za podpory výzkumného záměru MSM0021630528, grantů TAČR TA01010858 a FIT-S-11-2.

REFERENCE

[1] CHEN, H., MA, S., JIANG, K.: Detecting and adapting to drifting concepts, In *Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2012, s. 775-779, ISBN 978-1-4673-0025-4.

[2] YAN, J., YUN, X., ZHANG, P., aj.: A New Weighted Ensemble Model for Detecting DoS Attack Streams, In *Proceedings of the 3th IEEE/WIC/ACM International Conference on Web Intelligence and Inteligent Agent Technology (WI-IAT)*, 2010, s. 227-230, ISBN 978-1-4244-8482-9.

[3] WU, D., WANG, K., HE, T., aj.: A Dynamic Weighted Ensemble to Cope with Concept Drifting Classification, In *Proceedings of the 9th International Conference for Young Computer Scientists (ICYCS)*, 2008, s. 1854-1859, ISBN 978-0-7695-3398-8.