ARTIFICIAL IMMUNE SYSTEMS FOR SPAM DETECTION

Michal Hohn

Master Degree Programme (2), FIT BUT

E-mail: xhohnm00@stud.fit.vutbr.cz

Supervised by: Josef Schwarz

E-mail: schwarz@fit.vutbr.cz

Abstract: This work deals with creating a hybrid system based on the aggregation of artificial immune system with appropriate heuristics to make the most effective spam detection. This work describes the main principals of biological and artificial immune system and conventional techniques to detect spam including several classifiers are given.

Keywords: Bayes, email, spam, artificial immune system, lymphocytes

1. ÚVOD

V dnešní době existuje nepřeberné množství komunikačních kanálů, ať se jedná o *VoIP*, *Skype*, *Instant Messaging*, sociální sítě a v neposlední řadě elektronická pošta (dále jen email). Jakákoliv varianta komunikace má svá úskalí a v případě emailů je to nevyžádaná pošta (dále jen spam).

Tato práce se zabývá umělými imunitnímy systémy (dále jen UIS). Cílem není vytvořit systém, který by tvořil/přijímal/odesílal poštu a následně je zpracovával v integrovaném filteru, ale vytvořit program na bázi UIS, který by efektivně analyzoval email a na základě vhodných heuristik by rozhodl, zda se jedná o nevyžádanou poštu, či nikoliv.

2. BIOLOGICKÝ IMUNITNÍ SYSTÉM

UIS využívá principy fungování lidského imunitního systému, který se skládá ze 4 základních úrovní: **Fyzická** - kůže, **Fyziologická** – kyselina mléčná a mastné kyseliny, **Vrozený imunitní systém**, **Adaptivní imunitní systém**.

2.1. VROZENÝ IMUNITNÍ SYSTÉM

Je to ta část imunitního systému, kterou máme geneticky zakódovanou a zdědíme ji po našich rodičích. Důležitou vlastností vrozené imunity je ta, že **není specifická**. To znamená, že funguje na nejnižší biologicko-chemické úrovni a její chování by se dalo popsat jako „předprogramované“. Zde se vyskytují *Makrofágy, Neutrofily, Dendrické buňky, Langerhansovy buňky a Imunoglobuliny*.

2.2. ADAPTIVNÍ IMUNITNÍ SYSTÉM

Adaptivní imunitní systém je **specifický**, to znamená, že útočí na předem neznámý druh patogenu a je aktivován pouze, je-li stimulován vrozenou imunitou. Tento systém je tvořen dvěma druhy lymfocytů. **Lymfocyt B (B-cells)** se tvoří v kostní dřeni. **Lymfocyt T (T-cells)** je také vytvořen v kostní dřeni, ale dozrává v brzlíku.

Každý *patogen* je specifikován svým *antigenem*, díky tomu je každá cizorodá látka v biologickém systému unikátně popsatelná. Každý lymfocyt má na svém povrchu *receptory*, kterými jednoznačně rozpoznává *patogeny*.

3. UMĚLÝ IMUNITNÍ SYSTÉM

V UIS se používají umělé lymfocyty s detektory (analogie receptorů) pro detekci jednotlivých slov emailu, následně jsou detektory modifikovány a filtrovány podle účinnosti detekce. Dále v UIS existují tři základní algoritmy. **Pozitivní selekce** ruší neproduktivní lymfocyty. **Negativní selekce** je procesem selekce lymfocytů na základě schopnosti/úspěšnosti rozpoznáva spam zprávy. **Klonální selekční algoritmus** popisuje proces reprodukce a expanze úspěšných lymfocytů.

3.1. LYMFOCYT

Lymfocyt s detektory tvoří hlavní část UIS. Detektory mohou být tvořeny regulárním výrazem, konkrétním slovem, IP adresou apod. Dále obsahuje 2 váhové proměnné: **spam_matched** – počet emailů označených za spam, **msg_matched** – počet emailů, na které se lymfocyt navázal. Dále je vhodné časové razítko vytvoření lymfocytu.

3.2. UČENÍ

Strojové učení kandidátních detektorů v UIS probíhá ve třech fázích. **První fáze** je učení se ze souborů, které jsou zvolené uživatelem a jsou označeny příznakem spam/ham (ham – označení pro vyžádaný email). Soubor je rozdělen na malé části, obvykle slova a tyto části jsou ohodnocovány +-5 body k celkovému ohodnocení. Plus je pro ham, mínus pro spam. Pokud je celkové ohodnocení kladné, víme, že slovo se většinou vyskytuje v hamu a naopak.

Layer	User (from files)	AIS	User feedback (false negative / false positive)
Ham/Spam	+5 / -5	+1 / -1	+10 / -10

Tabulka 1: Ohodnocení slov/částí na jednotlivých úrovních.

Druhá fáze probíhá při běhu UIS, kdy se aktualizuje množina částí/slov a jejich ohodnocení +-1 bodem. **Třetí fáze** je zásah uživatele, který má možnost změnit ohodnocení +-10 body. Jedná se o případy *false positive / false negative* klasifikování. Ale jelikož není vytvářen program na práci s emaily viz. Úvod, tak třetí fázi můžeme ignorovat.

3.3. BAYESŮV FILTR

Bayesův filtr je považován v UIS za jeden z nejlepších klasifikátorů emailů. Výsledkem je hodnota v intervalu < 0 , 1 >, tedy pokud výsledek je 0,88 tak víme, že email je na 88% spam.

$$BayesScore = \frac{\prod_{matching\_lymphocytes} \frac{spam\_matched}{msg\_matched}}{\prod_{matching\_lymphocytes} \frac{spam\_matched}{msg\_matched} + \prod_{matching\_lymphocytes} 1 - \frac{spam\_matched}{msg\_matched}} \quad (1)$$

Pomocí rovnice (1) se počítá Bayesovo skóre, pomocí lymfocytů, které se úspěšně navázaly na email. Aktualizování lymfocytů probíhá po klasifikování, kdy se inkrementují proměnné *spam_matched* a *msg_matched*. Dále je nutné stanovit práh, který když se překročí, tak je email klasifikován za spam.

4. VLASTNÍ PŘÍNOS

Místo Bayesova skóre jsem použil vlastní klasifikační rovnici (2). Využívám lymfocyty se slovy/detektory, které mají pozitivní i negativní ohodnocení. Při vytváření lymfocytů mají všechny lymfocyty nastavenou proměnnou *msg_matched=1*. Lymfocyty vybrané ze slov s kladným ohodnocením mají nastaveno *spam_matched = 0* a ty se záporným ohodnocením *spam_matched = 1*. Poznamenejme, že ohodnocení lymfocytů se po provedení klasifikace mění (viz. tabulka 1, AIS).

$$MyScore = \frac{\sum_{matching\_lymphocytes} \frac{spam\_matched}{msg\_matched}}{\sum_{matching\_lymphocytes} \frac{spam\_matched}{msg\_matched} + \sum_{matching\_lymphocytes} 1 - \frac{spam\_matched}{msg\_matched}} \quad (2)$$

Toto je výhodné zejména v situaci, že po příchodu velkého množství hamů přijde jeden spam. To má za následek, že výsledné skóre je výrazně pod prahovou hodnotou: *Spam_matched* << *msg_matched*, takže spam je klasifikován jako ham, respektive *false positive*. Ve většině UIS je detektor lymfocytu reprezentován regulárním výrazem, který se provádí nad emailem. V mé implementaci jsou detektory/slova uloženy v nebinárním uspořádaném stromu, což snižuje složitost porovnávání z kvadratické složitosti na lineární.

Tělo emailu je rozděleno na slova, případně podřetězce. Tyto slova jsou vyhledávána v uvedeném stromu, kdy nad každým uzlem se provádí dodatečné heuristiky. Například máme spamové slovo *viagra*, tak aby systém detekoval obměnu *v14gr@* provádí se reverzivní převedení i = *{ i, 1, !, | }*, a = *{ a, 4, @ }* atd.

Emaily, které mají skóre v intervalu (prahová hodnota +-0,08) mohou projít dodatečnými heuristickými testy ve druhé klasifikační úrovni, jako je například detekce HTML tagů, jestli email obsahuje větší množství odkazů, obrázků, nebo je psán velkými písmeny.

4.1. VÝSLEDKY

Za zdroj emailů jsem zvolil korpusy z *http://spamassassin.apache.org/publiccorpus/* . Trénovací množina obsahovala 1000 hamů a 1000 spamů. Z těchto emailů bylo natrénováno 23688 lymfocytů. K otestování byla připravena množina obsahující 9349 emailů (6951ham, 2398 spam). Emaily přicházely v rovnoměrném rozložení. Po zpracování všech emailů byl počet lymfocytů 51535. Práh pro použité skóre byl nastaven na 0.367, vypnuta druhá úroveň klasifikace.

HAM	SPAM	False positive	False negative	Úspěšnost	Čas/email
6743	2206	192	208	95,7108%	9,8 ms

Tabulka 2: Dosažené výsledky.

Dalším experimentem jsem zjistil, že pokud bude použita druhá klasifikační úroveň, může celková úspěšnost klasifikace dosáhnout hodnoty až 98% (přičemž false positive 3x klesne, false negative klesne 2x).

5. ZÁVĚR

V této práci jsem popsal biologický a umělý imunitní systém a principy použitelné pro detekci spamu. Byl popsán životní cyklus lymfocytu a proces jeho učení. Navržený klasifikátor dle vztahu (2) se jeví jako velmi perspektivní. Naměřenou úspěšnost 95,71% lze považovat za velmi slibnou (v článcích [3, 4] je dosažená úspěšnost 94%).

REFERENCE

[1] Tschudin, CH., Meyer, T., Yamamoto, L.: Artificial Immune Systems, University of Basel, 2009

[2] Neuwirth, D.: Umělé imunitní výpočetní systémy, Brno, 2007

[3] Oda, T., White, T.: Immunity for spam: an analysis of an artificial immune system for junk email detection, Carleton University, Ottawa ON, Canada

[4] Oda, T., White, T.: Developing an Immunity to Spam, Carleton University, Ottawa ON, Canada