

Adaptive Spam Detection Inspired by the Immune System

Alaa Abi-Haidar* and Luis M. Rocha

Department of Informatics, Indiana University, Bloomington IN 47401, USA
and
Instituto Gulbenkian de Ciência, Oeiras, Portugal
*aabihaid@indiana.edu

Abstract

This paper proposes a novel solution to spam detection inspired by a model of the adaptive immune system known as the cross-regulation model. We report on the testing of a preliminary algorithm on six e-mail corpora. We also compare our results with those obtained by the Naive Bayes classifier and another binary classification method we developed previously for biomedical text-mining applications. We obtained very encouraging results which can be further improved with development of this bio-inspired model. We show that the cross-regulation model is promising as a bio-inspired algorithm for spam detection in particular, and binary classification in general. Finally, we also present evidence that our bio-inspired model is relevant for understanding immune regulation itself.

Introduction

Spam detection is a binary classification problem in which e-mail is classified as either ham (legitimate e-mail) or spam (illegitimate or fraudulent e-mail). Spam is very dynamic in terms of advertising new products and finding new ways to defeat anti-spam filters. The challenge in spam detection is to find the appropriate threshold between ham and spam leading to the smallest number of misclassifications, especially of legitimate e-mail (false negatives). To avoid confusions, ham and spam will be labeled as positives and negatives respectively.

The vertebrate adaptive immune system, which is one of the most complex and intelligent biological systems, learns to distinguish harmless from harmful substances (known as pathogens) such as viruses and bacteria that intrude the body. These pathogens often evolve new mechanisms to attack the body and its immune system, which in turn adapts and evolves to deal with changes in the repertoire of pathogen attacks. A weakly responsive immune system is vulnerable to attacks while an aggressive one can be harmful to the organism itself, causing autoimmunity. Given the conceptual similarity between the problems of spam and immunity, we investigate the applicability of the cross-regulation model of T-cell dynamics (Carneiro et al., 2007) to spam detection.

Below we offer a short review of related work in spam detection, a brief introduction to the adaptive immune system, and the cross-regulation model (Carneiro et al., 2007). In the following section, the bio-inspired cross-regulation algorithm and its application to spam are discussed. In the Results section, the experiments and implementation of the model vis a vis the other binary classification models are discussed.

Spam Detection

Spam detection has recently become an important problem with the ubiquity of e-mail and the rewards of no-cost advertisement that can reach the largest audience possible. Spam detection can target e-mail headers (e.g. sender, receiver, relay servers...) and/or content (e.g. subject, body). Machine learning techniques such as support vector machines (Carreras and Marquez, 2001; Kolcz and Alspecter, 2001), Naive Bayes classifiers (Sahami et al., 1998; Metsis et al., 2006) and other classification rules such as Case-Based Reasoning (Fdez-Riverola et al., 2007) have been very successful in detecting spam in the past. However, they generally lack the ability to detect spam drift since they rely on training on fixed corpora, features and rules. Research in this area is now focusing on concept drift in spam, with very promising results (Delany et al., 2006a; Méndez et al., 2006; Tsymbal, 2004; Kolter and Maloof, 2003). In addition, social-based spam detection models (Boykin and Roychowdhury, 2005; Chirita et al., 2005) have recently become relevant and competitive. Artificial Immune System (AIS) based algorithms (Oda, 2005; Bezerra and Barra, 2006; Yue et al., 2007) are another area of exciting development. The AIS models are inspired by diverse responses and theories of the natural immune system (Hofmeyr, 2001) such as negative selection, clonal selection, danger theory and the immune network theory. Our bio-inspired spam detection algorithm is based instead on the cross-regulation model (Carneiro et al., 2007), which is a novel development in AIS approaches to spam detection.

The Adaptive Immune System

The immune system, and more specifically, the vertebrate adaptive immune system, is a complex network of cells that distinguish between harmless and harmful substances or antigens—usually proteins or fragments of proteins and certain types of carbohydrate polymers that can be recognized by the immune system. When harmful antigens are discovered, an immune response to eliminate them is set in motion. Recognizing harmless self antigens, which obviously should not lead to an immune response to eliminate them, is resolved by a process known as positive and negative selection of T-cells which takes place in the thymus. It is in the thymus that T-cells develop and mature; only T-cells that have failed to bind to self antigens are released, while the rest of the T-cells is culled. The mature T-cells are allowed out of the thymus to detect harmful nonself antigens. They do this by binding to antigen presenting cells (typically B-cells, macrophages and dendritic cells) that collect and present antigens through MHC complexes after breaking them by lysosome. The specific T-cells that are able to bind to the presented antigens then stimulate B-cells that start a cascade of events leading to antibody production and the destruction of the pathogens or tumors linked to the antigens. However, it is possible that T-cells and B-cells, which are also trained in the thymus, could mature before being exposed to all self antigens. Even more problematic is the somatic hypermutation that ensues in lymph nodes after the activation of B-cells. At this stage, it is possible to generate many mutated B-Cell clones that could bind to harmless self antigens. Either situation can cause auto-immunity by generating T-cells capable of attacking self antigens. One way around this is by a process called costimulation which involves the co-verification of self antigens by both T-cells and B-cells before the antigen is identified as harmful pathogen and attacked. To further insure that the T-cells do not attack self, another type of T-cells known as T regulatory cells, are formed in the thymus where they mature to avoid recognizing self antigens. These regulatory T-cells have the responsibility of preventing autoimmunity by suppressing other T-cells that might bind and kill self antigens.

The Cross-regulation Model

The cross-regulation model, proposed by Carneiro et al. (2007), aims to model the process of discriminating between harmless and harmful antigen—typically harmless self/nonself and harmful nonself. The model consists of only three cell types: Effector T-Cells (E), Regulatory T-Cells (R) and Antigen Presenting Cells (A) whose populations interact dynamically, ultimately to detect harmful antigens. E and R are constantly produced, while A are capable of presenting a collection of antigens to the E and R. T-cell proliferation depends on the co-localization of E and R as they form conjugates (bind) with the antigens presented by A cells (this model assumes that A can form conjugates with a maximum

of two E or R). The population dynamics rules of this model are defined by three differential equations, which can be, for every antigen being presented by an A, summarized by the following three laws of interaction:

1. When E bind to the A, they proliferate with a fixed rate.
2. When R bind to the A, they remain in the population.
3. if an R binds together with an E to the same A, the R proliferates with a certain rate and the E remains in the population but does not proliferate.

The E and R proliferation rates in this model are fixed to 200%, which is the exactly the process of duplication or production of one extra copy. Finally, the E and R die at a fixed death rate. Carneiro et al. (2007) showed that the dynamics of this system leads to a bistable system of two possible stable population concentration attractors: (i) the co-existence of both E and R types identifying harmless self antigens, or (ii) the progressive disappearance of R, identifying harmful antigens. An illustration of the three rules is shown in figure 1 and more details on the model are available in the original paper (Carneiro et al., 2007).

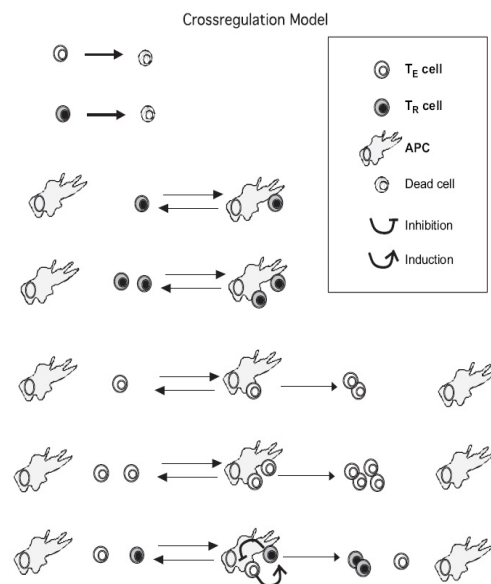


Figure 1: Figure is courtesy of Carneiro et al. (2007). The Cross-regulation Model. The diagram illustrates the interactions underlying the dynamics of A, E and R as assumed in the model in which A can only form conjugates with a maximum of two T cells.

The Cross-regulation Spam Algorithm

In order to adopt the cross-regulation algorithm for spam detection, which we named the Immune Cross-Regulation Model (ICRM), one has to think of e-mails as analogous to

the organic substances that upon entering the body are broken into constituent pieces by lysosome in A. In biology, these pieces are antigens (typically protein fragments) and in our analogous algorithm they are words extracted from e-mail messages and processed to become features¹. Thus, in this model, antigens are words or potentially other features. For every antigen there exists a number of virtual E and R that interact with A which present a sample of the features of a given e-mail message. In other words, the A correspond to the e-mail. The general ICRM algorithm is designed to be first trained on N e-mails of “self” (a user’s outbox) and harmless “nonself” (a user’s inbox). However, in the results described here, it was not possible to directly obtain outbox data; we are currently working on collecting outbox data for future work. In addition, the ICRM is also trained on “harmful nonself” (spam arriving to a given user). Training on or exposure to ham e-mails, in analogy with Carneiro’s et al model (Carneiro et al., 2007), is supposed to lead to a “healthy” dynamics denoted by the co-existence of both E and R with more of the latter. In contrast, training on or exposure to spam e-mails is supposed to result in much higher numbers of E than R. When e-mail features occur for the first time, a fixed initial number of E and R, for every feature, are generated. These initial values of E and R are different in the training and testing stages; more weight to R for ham features, and more weight to E for spam features is given in the labeled training stage. While we specify different values for initializing the proportions of E and R associated with e-mail features, depending on whether the algorithm is in the training or the testing stage, the ICRM is based on the exact same algorithm in both stages. An illustration comparing the artificial model to the biological one is shown in figure 2. The ICRM algorithm begins when an e-mail is received and cycles through three phases for every received e-mail:

In the **pre-processing phase**, HTML tags are not stripped off and are treated as other words, as often done in spam-detection (Metsis et al., 2006) . All words constituting the e-mail subject and body are lowercased and stemmed using Porter’s algorithm (Porter, 1980) after filtering out common English stop words and words of length less than 3 characters. A maximum of n processed unique features (words, in this case) are randomly sampled and presented by the virtual A which corresponds to the e-mail. These virtual antigen presenting cells have n_A binding slots per feature, i.e. $n \times n_A$ slots per e-mail message. The breaking up of the e-mail message into constituent portions (features) is inspired by the natural process in Biology, but is further enhanced in this model to select the first and last $\frac{n}{2}$ features in the e-mail. The assumption is that the most indicative information is in the beginning (e.g. subject) and the end of the e-mail (e.g. signature), especially

¹Naturally, features other than words are possible (e.g. bigrams, e-mail titles)

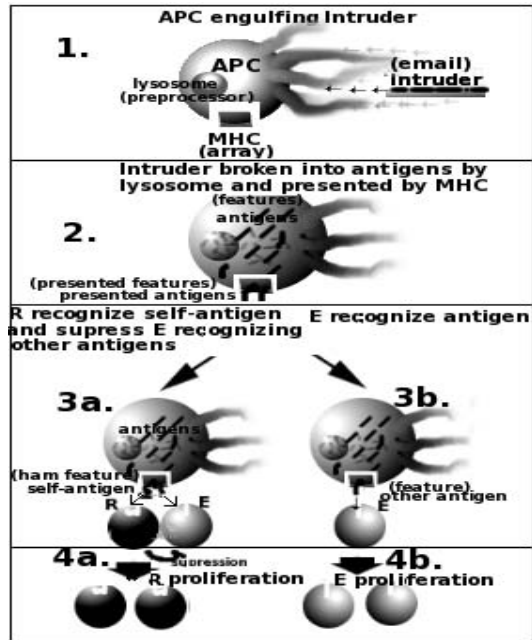


Figure 2: An illustration of the cross-regulation model (and its mapping to spam detection). In step 1, the intruder (received e-mail) is engulfed by an A (e-mail representer array). In step 2, the intruder is broken down by lysosome (preprocessor which strips html tags, filters out stop words and short words and porter stems a selection words) into antigens (features) which are then sampled and presented through MHC (an array residing in the memory) so that in step 3 specific E or R T-cells (virtual E and R residing in memory) can recognize it and bind to it. In step 3a, an R recognizing what probably is a self-antigen (ham feature) shares the A with an E recognizing a probably nonself-antigen (new or spam feature). In step 4a, the R suppresses the E which then excites the R to make it proliferate with a higher rate giving the antigen recognized by E more tolerance (making the novel feature more ham since it co-occurred with a ham feature). In step 3b on the other hand, the E is not suppressed by any R and thus it proliferates in step 4b making the system more immune to the antigen recognized by E (making the feature E recognize one more spam feature). After step 4, the whole intruder (e-mail) is judged based on its antigens (features) on whether it is bad or good (spam or ham) as explained in the decision phase of the algorithm.

concerning ham e-mails. Nevertheless, the feature selection problem will be studied in more detail in future work.

In the **interaction phase**, feature-specific R_g and E_f are allowed to bind to the corresponding antigens presented by A, which are arbitrarily located on its array of feature slots. Every adjacent pair of A slots is dealt with separately: the E_f for a given feature f proliferate only if they do not find themselves sharing the same adjacent pair of A binding slots with R_g , in which case only the R_g , associated with feature g , proliferate. The model assumes that novel ham features k tend to have their E_k suppressed by R_g of other pre-occurring ham features g because they tend to co-occur in the same message. As for the algo-

rithm’s parameters, let n_A be the number of A slots per feature. Let $(E_{0_{ham}}, R_{0_{ham}})$ and $(E_{0_{spam}}, R_{0_{spam}})$ be the initial values of E and R for features occurring for the first time in the training stage for spam and ham respectively. For the testing stage we have $(E_{0_{test}}, R_{0_{test}})$. Moreover, $E_{0_{ham}} \ll R_{0_{ham}}$, $E_{0_{spam}} > R_{0_{spam}}$ and $E_{0_{test}} > R_{0_{test}}$. Therefore, a feature f initially occurring in a ham e-mail would have $R_f \gg E_f$ and vice versa for spam. In the ICRM implementation hereby presented, a major difference from Carneiro’s et al model (Carneiro et al., 2007) was tried: the elimination of cell death. This is a rough attempt to provide the system with long term memory. Cell death can lead to the forgetfulness of spam or ham features if these features do not reoccur in a certain period of time as shown later on.

In the **decision phase**, the arriving e-mail is assessed based on the relative proportions of R and E for its n sampled features. Features with more R are assumed to correspond to ham while features with more E are more likely to correspond to spam. The proportions are then normalized to avoid decisions based on a few highly frequent features that could occur in both ham and spam classes. For every feature f , the feature score is computed as follows:

$$score_f = \frac{R_f - E_f}{\sqrt{R_f^2 + E_f^2}}, \quad (1)$$

indicating an unhealthy (spam) feature when $score_f \leq 0$ and a healthy (ham) one otherwise. $score_f$ varies between -1 and 1. For every e-mail message e , the e-mail immunity score is simply:

$$score_e = \sum_{\forall f \in e} score_f. \quad (2)$$

Note that a spam e-mail with no text such as the cases of messages containing exclusively image and pdf files, which surpass many spam filters, would be classified as spam in this scheme—e-mail e is considered spam if $score_e = 0$. Similarly, e-mails with only a few features occurring for the first time, would share the same destiny, since the initial E is greater than R in the testing stage $E_{0_{test}} > R_{0_{test}}$ which would result in $score_e < 0$.

Results

E-mail Data

Given the assumption that personal e-mails (i.e. e-mails sent or received by one specific user) are more representative of a writing style, signature and themes, it would be preferable to test the ICRM on e-mails from a personal mailbox. Unfortunately, this is not offered by the most common spam corpus

of *spamassassin*² and similarly for *ling-spam*³. In addition, the ICRM algorithm requires timestamped e-mails, since order of arrival affects final E/R populations. Timestamped data is also important for analyzing concept drifts over time, thus we cannot use the *PUI*⁴ data described by Androutsopoulos et al. (2000b). Delany’s spam drift dataset⁵, introduced by Delany et al. (2005), meets the requirements in terms of timestamped and personal ham and spam however its features are hashed and therefore it is not easy to make tangible conclusions based on their semantics. The *enron-spam*⁶ preprocessed data perfectly meets the requirements as it has six personal mailboxes made public after the enron scandal. The ham mailboxes belong to the employees *farmer-d*, *kaminski-v*, *kitchen-l*, *williams-w3*, *beck-s* and *lokay-m*. Combinations of five spam datasets were added to the ham data from *spamassassin* (s), *HoneyProject* (h), *Bruce Guenter* (b) and *Georgios Paliouras*’ (g) spam corpora and then all six datasets were tokenized (Metsis et al., 2006). In practice, some spam e-mails are personalized, which unfortunately cannot be captured in this dataset since the spam data comes from different sources. Only the first 1000 ham and 1000 spam e-mails of each of the corpora are used, as shown in table 1.

Table 1: Enron datasets

Dataset	ham + spam	ham:spam	[ham, spam] time range
Enron1	farmer-d + gp	1000:1000	[12/99, 06/00], [12/03, 01/05]
Enron2	kaminski-v + sh	1000:1000	[12/99, 05/00], [05/01, 07/05]
Enron3	kitchen-l + bg	1000:1000	[2/01, 06/01], [08/04, 03/05]
Enron4	williams-w3 + gp	1000:1000	[4/01, 01/02], [12/03, 06/04]
Enron5	beck-s + sh	1000:1000	[1/00, 11/00], [05/01, 03/05]
Enron6	lokay-m + bg	1000:1000	[6/00, 7/01], [08/04, 10/04]

ICRM Settings and Parameters

For each of the six enron sets, we ran each algorithm 10 times. Each run consisted of 200 training (50% spam) and 200 testing or validation (50% spam) e-mails that follow in timestamp order. From the 10 runs we computed variation statistics for the F-score⁷, and Accuracy performance.

In the e-mail pre-processing phase, we used $n = 50$, $n_A = 10$, $E_{0_{ham}} = 6$, $R_{0_{ham}} = 12$, $E_{0_{spam}} = 6$, $R_{0_{spam}} = 5$, $E_{0_{test}} = 6$ and $R_{0_{test}} = 5$. These initial E

²<http://spamassassin.apache.org/publiccorpus/>

³<http://www.aueb.gr/users/ion/publications.html>

⁴<http://www.iit.demokritos.gr/skel/i-config/downloads/enron-spam/>

⁵<http://www.comp.dit.ie/sjdelany/Dataset.htm>

⁶<http://www.iit.demokritos.gr/ionandr/publications/>

⁷The F1-measure (or *F-Score*) is defined as $F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$, where $Precision = \frac{TP}{(TP+FP)}$ and $Recall = \frac{TP}{(TP+FN)}$ and $Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$ measures of the classification of each test set, where TP, TN, FP and FN denote true positives, true negatives, false positive and false negatives respectively (Feldman and Sanger, 2006)

and R populations for features occurring for the first time are chosen based on the initial ratios chosen by Carneiro et al. (2007) and were then empirically adjusted to achieve the best F-score and Accuracy results for the six enron datasets. Finally, the randomization seed was fixed in order to compare results to other algorithms and search for better parameters. The ICRM was compared with two other algorithms that are explained in the following two subsections. The ICRM was also tested on shuffled (not in order of date received) validation sets to study the importance of e-mail reception order. The results are shown in table 2. The mean and variance of the results are also plotted on the F-score vs Accuracy axes as shown in figure 3.

Naive Bayes

We have chosen to compare our results with the multinomial Naive Bayes (NB) with boolean attributes (Jensen et al., 1996) which has shown great success in a previous research (Metsis et al., 2006). In order to fairly compare NB with ICRM, we selected the first and last unique $n = 50$ features. The Naive Bayes classifies an e-mail as spam in the testing stage if it satisfies the following condition:

$$\frac{p(c_{spam}) \cdot \prod_{f \in e-mail} p(f|c_{spam})}{p(c_{spam}) \cdot \sum_{c \in \{c_{spam}, c_{ham}\}} \prod_{f \in e-mail} p(f|c)} > 0.5, \quad (3)$$

where f is the feature sampled from an e-mail, and $p(f|c_{spam})$ and $p(f|c_{ham})$ are the probabilities that this feature f is sampled from a spam and ham e-mail respectively, while c is the union of spam and ham emails. The results are shown in table 2 and plotted in figure 3.

Variable Trigonometric Threshold (VTT)

We developed the VTT as a binary classification algorithm and implemented it as a protein-protein abstract classification tool⁸ using bioliterature mining (Abi-Haidar et al., 2007, 2008). VTT is itself inspired by another case-based spam detection algorithm (Fdez-Riverola et al., 2007). Briefly, VTT’s strategy is to make a selection of most significant preprocessed words ranked by a score $S(w) = |p_{ham}(w) - p_{spam}(w)|$ where $p_{ham}(w)$ and $p_{spam}(w)$ are the probabilities of a word w of occurring in the ham and spam training datasets which in our case are batches of 200 e-mails each. Naturally, a selection of 650 words would be fairly sufficient. The e-mails are then reduced to vectors of these 650 words. Then, the probabilities of co-occurring pairs of words (w_i, w_j) in these vectors are computed using $p_{ham}(w_i, w_j)$ and $p_{spam}(w_i, w_j)$. Then the trigonometric measures of the angle α , of this vector with the p_{ham} axis: $\cos(\alpha)$ is a measure of how strongly terms are exclusively associated with training ham e-mails, and similarly $\sin(\alpha)$

with training spam ones. Then, for every e-mail e , we compute the sum of all pairs’ measures to study the e-mail e ’s likelihood of being ham or positive $P(e)$ and spam or negative $N(e)$:

$$P(e) = \sum_{(w_i, w_j) \in e} \cos(\alpha(w_i, w_j)), \quad (4)$$

$$N(e) = \sum_{(w_i, w_j) \in e} \sin(\alpha(w_i, w_j)) \quad (5)$$

and finally the decision of whether an e-mail is ham or spam is made using the VTT equation:

$$\begin{cases} e \in ham, & \text{if } \frac{P(e)}{N(e)} \geq \lambda_0 + \frac{\beta - np(a)}{\beta} \\ e \in spam, & \text{otherwise} \end{cases} \quad (6)$$

where λ_0 is a constant threshold for deciding whether an e-mail is positive (spam) or negative (ham) obtained through exhaustive parameter search. For this experiment $\lambda_0 = 1.3$ produces the best results. Another parameter is β which was used in the abstract classification experiment to regulate $np(a)$ which counts the number of tagged protein in an abstract a but will be ignored in spam detection for the sake of simplicity. Therefore, equation 6 can be reduced to classify e as ham if $\frac{P(e)}{N(e)} \geq 1.3$ or as spam otherwise. The results are shown in table 2 and plotted in figure 3 then discussed in the discussion section.

Table 2: F-score and Accuracy mean +/- sdev of 10 runs for 50% spam enron data sets with the first two columns using ICRM (the first one applied on ordered e-mail, the second one on shuffled timestamps of testing data, and the last two using Naive Bayes and VTT.

Dataset		ICRM		Other Algorithms	
		Ordered	Shuffled	Naive Bayes	VTT
Enron1	F-score	0.9 ± 0.03	0.9 ± 0.03	0.89 ± 0.04	0.91 ± 0.04
	Accuracy	0.9 ± 0.03	0.9 ± 0.03	0.87 ± 0.05	0.9 ± 0.04
Enron2	F-score	0.86 ± 0.06	0.85 ± 0.06	0.92 ± 0.07	0.82 ± 0.23
	Accuracy	0.85 ± 0.06	0.83 ± 0.07	0.93 ± 0.05	0.86 ± 0.13
Enron3	F-score	0.88 ± 0.04	0.88 ± 0.04	0.93 ± 0.03	0.86 ± 0.08
	Accuracy	0.87 ± 0.05	0.87 ± 0.05	0.92 ± 0.04	0.85 ± 0.07
Enron4	F-score	0.92 ± 0.05	0.92 ± 0.04	0.92 ± 0.05	0.95 ± 0.03
	Accuracy	0.92 ± 0.05	0.92 ± 0.05	0.91 ± 0.06	0.95 ± 0.03
Enron5	F-score	0.92 ± 0.03	0.87 ± 0.06	0.94 ± 0.04	0.84 ± 0.13
	Accuracy	0.91 ± 0.03	0.87 ± 0.05	0.95 ± 0.03	0.87 ± 0.09
Enron6	F-score	0.89 ± 0.04	0.9 ± 0.04	0.91 ± 0.02	0.88 ± 0.05
	Accuracy	0.88 ± 0.05	0.89 ± 0.05	0.9 ± 0.03	0.87 ± 0.07
Total	F-score	0.9 ± 0.05	0.89 ± 0.05	0.92 ± 0.04	0.88 ± 0.12
	Accuracy	0.89 ± 0.05	0.88 ± 0.06	0.91 ± 0.05	0.88 ± 0.08

Discussion

As clearly shown in table 2 and figure 3, ICRM, NB and VTT are very competitive for most enron datasets, indeed the performance of ICRM is statistically indistinguishable from VTT (F-score and Accuracy p-values 0.15 and 0.63 for the paired t-test validating the null hypothesis of variation equivalence), though its slightly lower performance

⁸The Protein Interaction Abstract Relevance Evaluator (VTT) tool is available at <http://casci.informatics.indiana.edu/VTT/>

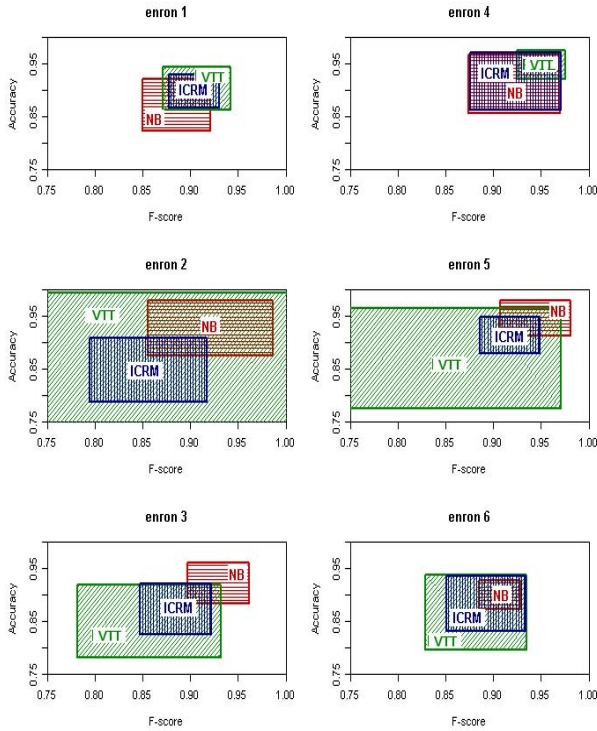


Figure 3: F-score vs Accuracy (mean and standard deviation) plot comparison between ICRM (vertical blue), NB (horizontal red) and VTT (diagonal green) for each of the six enron datasets. A visualization of table 2.

against NB is statistically significant (F-score and Accuracy p-values 0.01 and 0.02 for the paired t-test, rejecting the null hypothesis of variation equivalence with 0.05 level of significance).

More particularly, we investigate VTT’s performance deviations between enron 2 and enron 4 and notice that the average number of top 650 features that are ham features is only 10.22 for enron 2 (having many spam and very few ham indicative features) while it is 75.02 for enron 4 (having relatively more ham and less spam indicative features) this giving us the maximum deviations off 43.40, which is the mean of ham features’ constituency of the top 650 features for all enron sets. Enron 4’s Inbox (williams-w3), contained 619 automatically generated notification e-mails of the exact same contents with a subtle variation in the filename id, as shown via *Enron Explorer*⁹, an online visualization tool of the publicly available enron data. The peculiarity of enron 4 is also manifested in Metsis’ Naive Bayes results (Metsis et al., 2006). We think that the huge proportion of spam indicative features for enron 2 (similarly but less so for enron 5) is due to the huge spam drift and diversity of spamassassin and HoneyProject spanning four years mostly in 2001, 2002

⁹<http://enron.trampolinesystems.com/focus/338815>

and 2005 which is not available in the barely six months lifespan of ham. This diversity gives VTT many highly indicative spam features that only occur in spam and much less, if at all, in ham. This leads to many ham misclassifications for the few indicative features (out of 650) that are selected for the training. A fix to this could be by either by increasing the threshold beyond 650 features or balancing the number of top 650 indicative ham and spam features as clearly is the case for enron 4, or by finding a synchronous spam and ham data. VTT’s disadvantage of the features selection is paid off by its advantage of using feature co-occurrence of the top 650 features which is not the case in any of ICRM and NB. This might not be a fair comparison yet a modification to VTT would result in a modified VTT for another project and similarly, the use of co-occurrences with ICRM and NB will be pursued for a more advanced ICRM. From here onwards, we proceed with the comparison between ICRM and NB only.

Table 3: ICRM vs NB F-score and Accuracy mean +/- sdev for spam to ham ratio variations for mean of the six enron datasets.

		50% spam	30% spam	70% spam
ICRM	F-score	0.9 ± 0.05	0.91 ± 0.03	0.79 ± 0.12
	Accuracy	0.89 ± 0.05	0.86 ± 0.05	0.83 ± 0.08
NB	F-score	0.92 ± 0.04	0.86 ± 0.07	0.79 ± 0.07
	Accuracy	0.91 ± 0.05	0.84 ± 0.07	0.74 ± 0.01

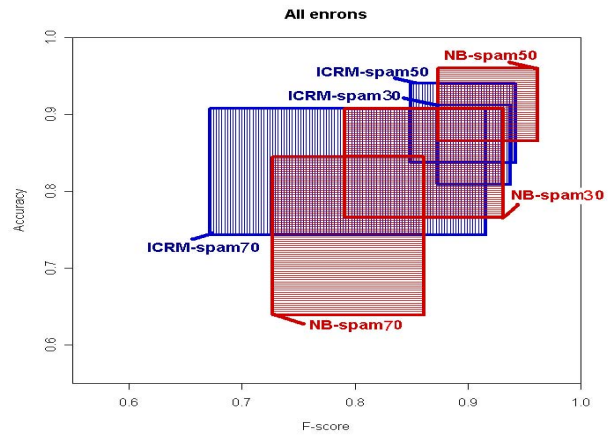


Figure 4: F-score vs Accuracy plot comparison between ICRM (vertical blue) and NB (horizontal red) with different spam to ham ratio variations 30:70 (spam30), 70:30 (spam70) and 50:50 (spam50) for the mean of the six enron datasets.

As shown in table 3 and figure 4, the ICRM can be more resilient to ham ratio variations¹⁰. While the performance of both algorithms was comparable for 50% spam (though significantly better for NB), the performance of NB drops for

¹⁰The 30% and 70% spam results were balanced for the evaluation by randomly sampling from the 70% class, reducing it to 30%

30% spam ratio (5% lower F-score than ICRM) and 70% spam ratio (9% less accurate than ICRM) while ICRM relatively maintains a good performance. The difference in performance is statistically significant, except for F-Score of the 70% spam experiment, as the p-values obtained for our performance measures clearly reject the null hypothesis of variation equivalence: F-Score and Accuracy p-values are 0 and 0.01 for 30% spam, and Accuracy p-value is 0.01 for 70% spam (p-value for F-Score is 0.5 for this case). While one could argue that NB's performance could well be increased, in the unbalanced spam/ham ratio experiments, by changing the right hand side of equation 3 to 0.3 or 0.7, this act would imply that, in real situations, one could know a priori the spam to ham ratio of a given user. The ICRM model, on the other hand, does not need to adjust any parameter for different spam ratios—it is automatically more reactive to whatever ratio it encounters. It has been shown that spam to ham ratios indeed vary widely Meyer and Whateley (2004); Delany et al. (2005), hence we conclude that the ICRM's ability to better handle unknown spam to ham ratio variations is more preferable for dynamic data classification in general and spam detection in particular.

In most Enron sets, the shuffled e-mails in the test set did very slightly worse than the ordered-by-reception-date ones. This observation was however statistically insignificant according to a t-test with p-value greater than 0.05 and thus it accepts null hypothesis of similarity between the two performances showing no importance of order for the ICRM dynamics. To further study the resilience of ICRM and its adaptive ability to catch concept drifts, we trained both ICRM and NB on the first 200 emails and then tested them on sequential overlapping slices of 200 emails. Our results showed very little decay in performance for both methods in most data sets (Abi-Haidar and Rocha, 2008). Therefore, we conclude that the data sets are not appropriate to study the effects of concept drift. In future work, we plan to test the ICRM on more appropriate data sets for the study of concept drift in spam (Delany et al., 2005, 2006b).

The three modifications to the original cross-regulation model, namely training on both ham and spam classes, feature selection and cell death elimination have quite improved the performance of the algorithm to make it rival with traditional binary classifier. The first modification's improvement was mostly manifested in enron 4 which cannot only rely on positive training for the majority of exact uninformative e-mails it has. Nonetheless, it is debatable whether the automatically generated messages in enron 4 should be classified as ham or not. The selection of the first and last features boosted the performance of both ICRM and NB about 2% in terms of F-score and Accuracy yet we are still working on making a better selection without totally disregarding the message body. The elimination of cell death also improved the overall performance of ICRM about 1%, especially in terms of long term memory. We are currently

experimenting with a carrying capacity for the E and R concentrations that could be promising for future work.

Conclusion

The observations made based on the artificial immune system can help us guide or further deepen our understanding of the natural immune system. For instance, ICRM's resilience to spam to ham ratio show us how dynamic is our immune system and functional independently of the amount of pathogens attacking it. In addition, the three modifications made to the original model can be very insightful: The improvements made by training on both spam and ham (rather than only ham or self) reinforce the theories of both self and nonself antigen recognition by T-cells outside the thymus. The feature selection makes us wonder whether the actual T-cell to antigen binding is absolutely arbitrary. Finally, the elimination of cell death may reinforce the theories behind long lived cells as far as long term memory is concerned.

In this paper we have introduced a novel spam detection algorithm inspired by the cross-regulation model of the adaptive immune system. We have compared it with Naive Bayes and another binary classification tool called VTT. Our model has proved itself competitive with state of art spam binary classifiers in general and resilient to spam to ham ratio variations in particular through interestingly unique results that can be further improved by integration, hopefully in the near future. The overall results, even though not stellar, seem quite promising especially in the area of tracking concept drifts in spam detection. This original work should be regarded not only as a promising bio-inspired method that can be further developed and even integrated with other methods but also as a model that could help us better understand the behavior of the T-cell cross-regulation systems in particular, and the natural vertebrate immune system in general.

Acknowledgements

We thank Jorge Carneiro for his insights about applying ICRM on spam detection and his generous support and contribution for making this work possible. We also thank Florentino Fdez-Riverola for the very useful indications about spam datasets and work in the area of spam detection. We would also like to thank the FLAD Computational Biology Collaboratorium at the Gulbenkian Institute in Oeiras, Portugal, for hosting and providing facilities used to conduct part of this research.

References

- Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Retschsteiner, A., Verspoor, K., Wang, Z., and Rocha, L. (2007). Uncovering protein-protein interactions in the bibliome. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume ISBN 84-933255-6-2, pages 247–255.

- Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Retchsteiner, A., Verspoor, K., Wang, Z., and Rocha, L. (2008). Uncovering protein-protein interactions in abstracts and text using linear models and word proximity networks. *Genome Biology*. inpress.
- Abi-Haidar, A. and Rocha, L. (2008). Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics. In *Proceedings of the 7th International Conference on Artificial Immune Systems (ICARIS 2008)*. Lecture Notes on Computer Science, Springer-Verlag . inpress.
- Androutsopoulos, I., Koutsias, J., Chandrinou, K., Paliouras, G., and Spyropoulos, C. (2000a). An evaluation of Naive Bayesian anti-spam filtering. *Proceedings of the workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning, 2000*, pages 9–17.
- Androutsopoulos, I., Koutsias, J., Chandrinou, K., and Spyropoulos, C. (2000b). *An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages*. ACM Press New York, NY, USA.
- Bezerra, G. and Barra, T. (2006). An Immunological Filter for Spam. *International Conference on Artificial Immune Systems (ICARIS 2006)*, LNCS, pages 446–458.
- Boykin, P. and Roychowdhury, V. (2005). Leveraging social networks to fight spam. *Computer*, 38(4):61–68.
- Carneiro, J., Leon, K., Caramalho, Í., van den Dool, C., Gardner, R., Oliveira, V., Bergman, M., Sepúlveda, N., Paixão, T., Faro, J., et al. (2007). When three is not a crowd: a Cross-regulation Model of the dynamics and repertoire selection of regulatory CD4 T cells. *Immunological Reviews*, 216(1):48–68.
- Carreras, X. and Marquez, L. (2001). Boosting Trees for Anti-Spam Email Filtering. *Proceedings of RANLP-2001, 2001* pages 58–64.
- Chirita, P., Diederich, J., and Nejdil, W. (2005). MailRank: using ranking for spam detection. *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 373–380.
- Delany, S. J., Cunningham, P., and Smyth, B. (2006a). Ecue: A spam filter that uses machine learning to track concept drift. In Brewka, G., Coradeschi, S., Perini, A., and Traverso, P., editors, *ECAI 2006, 17th European Conference on Artificial Intelligence, August 29 - September 1, 2006, Riva del Garda, Italy, Including Prestigious Applications of Intelligent Systems (PAIS 2006)*, *Proceedings*, pages 627–631. IOS Press.
- Delany, S. J., Cunningham, P., and Tsybmal, A. (2006b). A comparison of ensemble and case-base maintenance techniques for handling concept drift in spam filtering. In Sutcliffe, G. and Goebel, R., editors, *Proceedings of the 19th International Conference on Artificial Intelligence (FLAIRS 2006)*, pages 340–345. AAAI Press.
- Delany, S. J., Cunningham, P., Tsybmal, A., and Coyle, L. (2005). A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*, 18(4–5):187–195.
- Fdez-Riverola, F., Iglesias, E., Díaz, F., Méndez, J., and Corchado, J. (2007). SpamHunting: An instance-based reasoning system for spam labelling and filtering. *Decision Support Systems*, 43(3):722–736.
- Feldman, R. and Sanger, J. (2006). *The Text Mining Handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press.
- Hofmeyr, S. (2001). An Interpretative Introduction to the Immune System. *Design Principles for the Immune System and Other Distributed Autonomous Systems*.
- Jensen, F., Jensen, F., and Jensen, F. (1996). *Introduction to Bayesian Networks*. Springer-Verlag New York, Inc. Secaucus, NJ, USA.
- Kolcz, A. and Alspector, J. (2001). SVM-based filtering of e-mail spam with content-specific misclassification costs. *Proceedings of the TextDM*, pages 1–14.
- Kolter, J. and Maloof, M. (2003). Dynamic weighted majority: a new ensemble method for tracking concept drift. *ICDM 2003. Third IEEE International Conference on data mining 2003*, pages 123–130.
- Mason, J. (2002). SpamAssassin corpus, 2002. *U RL* <http://spamassassin.apache.org/publiccorpus>.
- Méndez, J., Fdez-Riverola, F., Iglesias, E., Díaz, F., and Corchado, J. (2006). Tracking Concept Drift at Feature Selection Stage in SpamHunting: an Anti-Spam Instance-Based Reasoning System. *Proceedings of the 8th European Conference on Case-Based Reasoning, ECCBR-06*, pages 504–518.
- Metsis, V., Androutsopoulos, I., and Paliouras, G. (2006). Spam Filtering with Naive Bayes—Which Naive Bayes? *Third Conference on Email and Anti-Spam (CEAS)*, pages 125–134.
- Meyer, T.A. and Whateley, B. (2004) SpamBayes: Effective open-source, Bayesian based, email classification system *Proceedings of the First Conference on Email and Anti-Spam (CEAS)* <http://ceas.cc/papers-2004/136.pdf> .
- Oda, T. (2005). *A Spam-Detecting Artificial Immune System*. Masters thesis, Carleton University.
- Porter, M. (1980). An algorithm for suffix stripping (1980). *Program*, 14:130–137.
- Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. *Learning for Text Categorization: Papers from the 1998 Workshop*, pages 55–62.
- Tsybmal, A. (2004). The problem of concept drift: definitions and related work. *Informe técnico: TCD-CS-2004-15, Department of Computer Science Trinity College, Dublin*, <https://www.cs.tcd.ie/publications/techreports/reports>, 4:15.
- Yue, X., Abraham, A., Chi, Z., Hao, Y., and Mo, H. (2007). Artificial immune system inspired behavior-based anti-spam filter. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 11(8):729–740.