Available online at www.sciencedirect.com
ScienceDirect
BioSystems 92 (2008) 215–225
Bio Systems
www.elsevier.com/locate/biosystems

Identification of SPAM messages using an approach inspired on the immune system

T.S. Guzella [a,*], T.A. Mota-Santos [b], J.Q. Uchôa [c], W.M. Caminhas [a]

[a] Department of Electrical Engineering, Federal University of Minas Gerais, Belo Horizonte (MG) 31270-010, Brazil
[b] Department of Biochemistry and Immunology, Federal University of Minas Gerais, Brazil
[c] Department of Computer Science, Federal University of Lavras, Lavras (MG) 37200-000, Brazil

Received 8 July 2007; received in revised form 23 February 2008; accepted 23 February 2008

Abstract

In this paper, an immune-inspired model, named innate and adaptive artificial immune system (*IA-AIS*) is proposed and applied to the problem of identification of unsolicited bulk e-mail messages (SPAM). It integrates entities analogous to macrophages, B and T lymphocytes, modeling both the innate and the adaptive immune systems. An implementation of the algorithm was capable of identifying more than 99% of legitimate or SPAM messages in particular parameter configurations. It was compared to an optimized version of the naïve Bayes classifier, which has been attained extremely high correct classification rates. It has been concluded that *IA-AIS* has a greater ability to identify SPAM messages, although the identification of legitimate messages is not as high as that of the implemented naïve Bayes classifier.

Keywords: Artificial immune system; SPAM identification; Continuous learning; Innate and adaptive immunity; Regulatory T cells

1. Introduction

The problems caused by the high volume of unsolicited bulk e-mail messages, commonly referred to as SPAM, circulating throughout the Internet are familiar to practically every e-mail user, whose mailboxes are flooded by these messages every day. Besides the precious time spent for their removal, network bandwidth is wasted for its delivery. Solutions to this problem can be described as either legal or technical. Although some countries have recently adopted legislation in this area (e.g. Carpinter and Hunt, 2006), it turns out that technical approaches remain indispensable, because SPAM can be sent from virtually anywhere in the world, and tracking the actual sender of messages may be difficult.

Recently, artificial immune systems (AISs), have emerged as a novel soft computing paradigm (de Castro and Timmis, 2002), exploring the cognitive capabilities of the immune system (Varela et al., 1988; Cohen, 1992). Due to the fact that AISs are a relatively new area, it is interesting to pursue a greater understanding of the biological models and mechanisms, which can serve as inspiration to the development of new algorithms. However, it is important to emphasize that simply using biologically inspired ideas and models is not, *per se*, enough to ensure that the resultant system will attain a good performance. In this context, an interesting work by Freitas and Timmis (2007) discusses the need of considering the target problem and its characteristics when designing an AIS.

The inspiration for the application of an immune-inspired algorithm to the problem of SPAM identification arises from similarities between pathogenic microorganisms and such messages, some of them briefly summarized below:

- Just like any living organism, SPAM messages are constantly "evolving", through changes in message features, such as alternative word spelling (e.g. "fr33" instead of "free"), on an attempt to evade SPAM filters.
- SPAM messages can be identified by their contents, using a mechanism similar to the one used by the natural immune system: pattern matching.

* Corresponding author at: Department of Electrical Engineering, Federal University of Minas Gerais, Avenue Antonio Carlos 6627, Belo Horizonte (MG) 31270-901, Brazil. Tel.: +55 31 3499 2625; fax: +55 31 3499 4188.
E-mail addresses: tguzella@cpdee.ufmg.br (T.S. Guzella), tomaz@icb.ufmg.br (T.A. Mota-Santos), joukim@ginux.ufla.br (J.Q. Uchôa), caminhas@cpdee.ufmg.br (W.M. Caminhas).

0303-2647/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2008.02.006

Therefore, based on the metaphor of an e-mail message as a microorganism to be identified as either pathogenic (a SPAM message) or non-pathogenic (a legitimate message), this work investigates the use of an AIS to identify SPAM messages. The model presented is a major improvement of an initial proposal by Guzella et al. (2005), with refinements on several points, such as the algorithm (modeling an additional class of T cells, referred to as regulatory T cells), and on the application to the target problem. The latter includes considering the importance of user feedback, when the correct classifications of previously analyzed messages are fed back into the classifier.

This paper is organized in the following way: a brief description of the immune system is presented in Section 2, while Section 3 comments on the problem of SPAM identification. Section 4 presents the proposed model and its application to the problem in detail. Finally, Section 5 presents the obtained results, when comparing the proposed model with a very popular algorithm, the naïve Bayes classifier, followed by the final conclusions in Section 6.

2. Immunological Background

This section outlines some basic concepts and components of the immune system. It is intended as a brief, self-contained description. For more details, the reader is invited to consult a specialized textbook such as Janeway et al. (2001).

In the human immune system, molecular patterns in pathogenic microorganisms are identified by innate immunity cells such as macrophages and dendritic cells. This identification occurs upon interaction between the pathogen associated molecular pattern (PAMP) and receptors (pattern matching structures) for PAMP (r-PAMP), expressed on the membrane of these cells. The innate system is composed of cells immediately available to respond to a limited variety of pathogens, which can be identified by patterns that do not occur in body cells. Hence, innate immune system receptors are said to be gene-encoded, to emphasize that they are germline-selected, because recognition is based on molecular patterns conserved in pathogens.

The adaptive immunity, which includes B and T lymphocytes, has a much greater recognition capability. Lymphocytes are capable of identifying antigens, a different kind of molecule that can be recognized by specific lymphocyte receptors and antibodies. These receptors are encoded by a random rearrangement of gene segments, which allows the generation of a very large number of receptors with unique antigen specificity. Although originated from the same precursors, B lymphocytes develop in the bone marrow, while T cells acquire immuno-competency in the thymus. The two types of lymphocytes have distinctive roles: B cells are responsible for secreting antibodies, while helper T lymphocytes can activate stimulated B cells, cytotoxic T cells eliminate host cells infected by intracellular pathogens and regulatory T cells directly regulate the activation of other B and T cells. Therefore, two requirements must be fulfilled for a B cell to be activated. First, its receptor must be occupied, reflecting the recognition of an antigen. Second, it must receive co-stimulation signals from a helper T cell, after the presentation of peptides coupled to major histocompatibility complex (MHC) molecules by the B cell. An immune response initiated in this way is referred to as thymus-dependent, due to the need of participation of T cells. This model, usually referred to as two-signal model, and several others taking it as starting point, identify that activation requires some other signal, besides specific recognition of the antigen.

From an evolutionary point of view, an adaptive immune system is important to cope with the appearance of new pathogens and their ability to evade recognition (Cohn, 2005). The advantage of having both an innate and adaptive system is the interaction between them, which allows the selection of an appropriate response. The innate immunity produces signalling proteins, called cytokines, that lead to inflammation and take part in the activation of cells from the adaptive immunity. On the other hand, the adaptive system designates cells of the innate system to eliminate pathogens.

The clonal selection theory (CST) (Burnet, 1959) provides a model to the establishment and maintenance of the "immunological memory". The theory postulates that an antigen induces the production of antibodies specifically reactive to it, by selecting a B lymphocyte capable of secreting such antibodies. This lymphocyte secretes antibodies and reproduces, initiating the immune response. After the elimination of the invading pathogen, some of the generated clones become memory cells, creating the structure for a more avid response to an identical or similar agent encountered later.

A key idea in the CST is the self/nonself discrimination, which can be described as the system's ability to react against external, harmful agents (nonself or pathogens), while remaining unresponsive to internal and harmless components (self). Due to the random generation of receptors for B and T lymphocytes, it is necessary to ensure that these cells will not react destructively to self antigens. As originally proposed in the CST, a process denominated negative selection would delete (eliminate) all B and T cells that reacted against self antigens. Because of this, negative selection is usually referred to as a recessive mechanism, because it explains self-tolerance based on the absence of self-reactive cells (Coutinho, 2005). However, it is known that some self-reactive cells eventually escape from deletion (Schwartz, 2005), and it is usually agreed that the theory does not account for self-tolerance (Cohen, 1992). Hence, it is necessary that additional, dominant, mechanisms prevent the activation of the cells not eliminated by negative selection (Coutinho, 2005). One such mechanism is mediated by regulatory T cells, capable of regulating the activation of self-reactive cells (Sakaguchi, 2004). As reviewed by von Boehmer (2005), these cells exert their functions through secretion of cytokines and direct cell contact.

3. Formulation of the Problem and Related Work

At first glance, identification of SPAM could be considered as a straightforward text classification problem. However, an important aspect of SPAM filtering is the constant change in message features, on an attempt to prevent recognition by content-based filters. Therefore, it is important that a classifier is able to receive feedback, supplied by the user, which can help

the system capture these new features. In contrast to this feature of SPAM and the importance of user feedback, algorithms proposed in the literature are usually evaluated in the classical batch-training scenario. An exception is a recent work by Delany et al. (2005), which also analyzes a classifier assuming that, from time to time, the misclassified messages will be used for training, with the correct classification indicated by the user. Wang et al. (2006) have also considered the dynamic characteristic of SPAM, compared to an adaptive filtering problem, but have not experimentally analyzed the performance in such scenario.

A message $\vec{M}$ can be represented as a vector of words $\vec{w}_i$, $i = 1, 2, \ldots, N$, according to Eq. (1), where N is the total number of words in the message. Each word, in turn, can be represented as a vector of characters c_j from an alphabet S, as described by Eq. (2):

$$\vec{M} = [\vec{w}_1 \quad \vec{w}_2 \quad \cdots \quad \vec{w}_i \quad \cdots \quad \vec{w}_N], \qquad i = 1, 2, \ldots, N \tag{1}$$

$$\vec{w}_i = [c_1 \quad c_2 \quad \cdots \quad c_j \quad \cdots \quad c_{n_i}], \qquad j = 1, 2, \ldots, n_i, \quad c_j \in S \tag{2}$$

In this view, alternative spelling of words can be seen as simply changing the value of some character c_j in a vector representing a word in Eq. (2). This technique is based on the fact that a word or sentence in a message is recognized by filters only if it is a perfect match, while humans can, in most cases, easily identify the original word given the new spelling and the message context. In contrast, techniques that are robust to small changes in some characters may face problems, due to the recognition of legitimate words that differ in few characters, such as *free* and *tree*, *low* and *law*. In this case, it is necessary to introduce a mechanism capable of taking the message context into consideration.

A classifier is defined as an entity that maps an instance $\vec{M}$ to one element of the set $\{l, s\}$, where l and s are labels for legitimate (negative class) and SPAM messages (positive class), respectively. It is important to recognize that, in most cases, there's an asymmetry regarding the classification mistakes. This occurs because classifying a SPAM message as legitimate (a false negative) is generally not as troublesome as misclassifying a legitimate message as SPAM (false positive). In the former, the user will just have to remove the message, while, in the latter case, this mistake can cause several problems, because missing a legitimate message is usually unacceptable. Based on this asymmetry, some authors have suggested using performance indices that assign to false positives a greater cost than false negatives. One approach is to consider a false positive as being γ times more costly than false negatives, so that it will be accounted as γ mistakes (Androutsopoulos et al., 2000). However, as commented by Carpinter and Hunt (2006), an ideal value for γ is difficult to be determined, depending on how likely is the user noticing that a message has been misclassified, and also on the importance of the message. Due to this reason, this paper adopts the conventional indices of correct classification rates, relating the relative number of SPAM and legitimate messages correctly classified. However, it should be emphasized that false positives should be minimized, sometimes even in exchange of more false negatives.

A recent review on the various methods used for SPAM filtering has been conducted by Carpinter and Hunt (2006). For this reason, only immune-inspired approaches, along with a concept-drift-based model proposed by Delany et al. (2005), are reviewed here. Currently, the most popular approach is the naïve Bayes classifier, presented by Sahami et al. (1998), with the suggestions and optimizations proposed by Graham (2002). These optimizations are known to produce much better results than the original proposal, and Graham reports correctly classifying legitimate messages with rates exceeding 99% (Graham, 2002). This classifier has become the standard for comparison with new developments. However, must of the papers that use this classifier have not included the optimizations, which, in a certain way, makes the comparisons unfair.

In the Bayesian framework, the probability that a given representation of a message, denoted as $\vec{x} = [x_1 \ x_2 \ \cdots \ x_n]$, belongs to a class $c \in \{s, l\}$ is given by Eq. (3), where $P(\vec{x}|c)$ and $P(c)$ are the probabilities that a message classified as c is represented by $\vec{x}$ and the message belongs to class c, respectively, and $P(\vec{x})$ is the *a priori* probability of a random message represented by $\vec{x}$ (e.g. Lai, 2007):

$$P(c|\vec{x}) = \frac{P(\vec{x}|c)P(c)}{P(\vec{x})} = \frac{P(\vec{x}|c)P(c)}{P(\vec{x}|s)P(s) + P(\vec{x}|l)P(l)} \tag{3}$$

The naïve classifier is obtained by assuming that the components x_i, $i = 1, 2, \ldots, n$ are conditionally independent, so that $P(\vec{x}|c)$ is given by (4), and Eq. (3) is reduced to (5):

$$P(\vec{x}|c) = \prod_{i=1}^{n} P(x_i|c) \tag{4}$$

$$P(c|\vec{x}) = \frac{\prod_{i=1}^{n} P(x_i|c)P(c)}{\prod_{i=1}^{n} P(x_i|s)P(s) + \prod_{i=1}^{n} P(x_i|l)P(l)} \tag{5}$$

Graham (2002) has suggested to determine the message's probability using the probabilities of occurrences of each word, i.e. by Eq. (6), where n_i^{s} and n_i^{l} are the number of occurrences in SPAM and legitimate messages, respectively, of the ith word:

$$p_i = \frac{n_i^{\mathrm{s}}}{n_i^{\mathrm{s}} + n_i^{\mathrm{l}}} \tag{6}$$

Therefore, the probability that the message belongs to class s (i.e. a SPAM message) is given by Eq. (7) where p_i is the

probability of the *i*th word in the message:

$$p^{s}_{\text{message}} = P(s|\vec{x}) = \frac{\prod_{i=1}^{n} p_i}{\prod_{i=1}^{n} p_i + \prod_{i=1}^{n}(1 - p_i)} \quad (7)$$

Hence, the classifier has just to use the number of occurrences of all (or some) words in SPAM and legitimate messages to obtain the message's probability and, using a threshold, determine the classification. It has been suggested that only the words with the most significant probabilities should be taken into consideration. This significance is given by the absolute difference between each probability and 0.5 (a "neutral" value). In addition, if a word unlisted in the database is found in the message being analyzed, after conducting several tests, it has been suggested that a probability of 0.4 should be used.

Up to this date, there are three immune-inspired models that have been designed aimed at this problem. A system using regular expressions for pattern matching has been proposed by Oda and White (2003a), and later extended by Oda and White (2003b). It assigns a weight to each detector (lymphocyte), which is incremented when it binds (i.e. recognizes an expression) to a SPAM message, and decremented if it binds to a legitimate message. After presenting a message to the population of detectors, the message weight is given by the sum of the positive and negative weights. If this sum is greater than a certain threshold, the message is flagged as SPAM. When the system misclassifies a message, it can be corrected by updating all the lymphocytes that match that message, by either incrementing or decrementing the weights, depending on the mistake made. It was used to classify a dataset composed of 1200 SPAM and legitimate messages, correctly identifying 90% of the former, and 99% of the latter, after being trained using 1600 and 1000 SPAM and legitimate messages, respectively. Another approach, based on a competitive antibody network named Supervised Real-Valued Antibody Network (SRABNET) has been presented by Bezerra et al. (2006). This system represents messages as binary feature vectors (representing the occurrence or not of a given set of words), and uses a procedure for dimensionality reduction, removing words that appear less than 5% and more than 95% in all the documents used for training. When applied to a corpus composed of 481 SPAM and 618 legitimate messages, correct classification rates between 97% and 98% were obtained, using 10-fold cross-validation. In contrast, the naïve classifier implemented has obtained very low classification rates, according to the authors, due to feature selection method used. Recently, Yue et al. (2007) have proposed an incremental clustering model, for grouping similar SPAM messages. This is done using a score calculated from some features of each message (such as the sender's IP address, and links in the body). However, no results in terms of correct classification rates of SPAM and legitimate messages have been reported by the authors.

Finally, an interesting work is a case-based reasoning classifier, proposed by Delany et al. (2005), which can track concept drift (a change in a target concept, which, in that work, are features of SPAM messages) in messages being analyzed. In the problem of identification of SPAM messages, the target concept remains constant, but the data distribution changes, a situation denominated virtual concept drift (Delany et al., 2005), which indicates the need to rebuild the classifier. Using misclassified messages to train the classifier at the end of each day, it was concluded that the proposed classifier can handle the concept drift in SPAM messages, performing well in comparison with the naïve Bayes classifier. In fact, this is the only work up to date that has analyzed the performance of the classifier considering feedback information.

Most of the considered approaches, with the exception of Oda and White (2003a) are based on using feature vectors with a constant structure and organization representing the message contents (for a detailed discussion of some methods, see Leopold and Kindermann, 2002). These approaches usually attain a high compression rate, because the classification model is based on only a subset of the words, selected based on feature selection methods applied during training. However, they may not scale well when considering real-life performance in this specific problem, due to the fact that the structure of these feature vectors is constant, and it is not possible to consider a pattern that was not initially selected when building the model, thereby compromising the ability to adapt to newly received messages. In this case, if one wants to consider a new word that was not initially presented, the only way is to retrain the classifier, which will be extremely inefficient if performed constantly. The naïve Bayes classifier with the optimizations by Graham (2002), does not suffer from this problem, because it performs feature selection on-line, by selecting the words in the message with the most "relevant" probabilities. In a way to consider this problem, Delany et al. (2005) analyzes the possibility of periodically re-applying the feature selection process, allowing new patterns to be considered by the classifier. It was concluded that this process reduces the average classification error, but it was not discussed if this operation can be conducted efficiently.

4. Proposed Model

The model presented in this paper, named innate and adaptive artificial immune system (*IA-AIS*), combines aspects of the negative selection (reviewed by Ji and Dasgupta, 2007) and clonal selection (CLONALG, de Castro and Von Zuben, 2002) algorithms. It is composed of macrophages, along with B and T cells, modeling both the adaptive and innate systems and the interaction between B and T cells (helper and regulatory). The motivation for incorporating these components and interactions is the possibility of designing a more biologically plausible model, an aspect advocated by Stepney et al. (2005).

The innate system has recently become of great interest in the artificial immune systems community. Inspired by the danger model, Greensmith et al. (2005) have proposed the dendritic cell algorithm (DCA), while Tedesco et al. (2006) have described an approach combining dendritic cells and T cells. In addition, the two models proposed by Sarafijanović and Le Boudec (2005) and Dasgupta et al. (2005), to be described next, also incorporate the innate system. For more details, see Twycross and

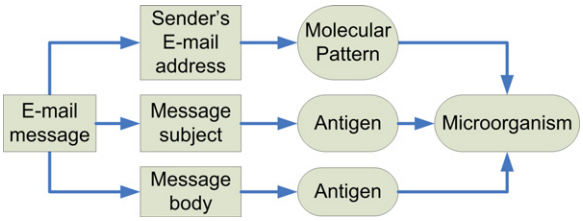

Fig. 1. Generation of a microorganism out of an e-mail message.

Aickelin (2005), who proposed a conceptual framework for the incorporation of the innate system and Twycross and Aickelin (2006).

The two-signal model has also been considered in some algorithms, as reviewed by Freitas and Timmis (2007). Kim and Bentley (2002) have proposed the dynamic clonal selection algorithm (DynamicCS), where the second signal is derived from human expert confirmation that a mature detector has, indeed, being activated by anomalous traffic. In artificial immune system for e-mail classification (*AISEC*), an artificial immune system for e-mail sorting proposed by Secker et al. (2003), user feedback constitutes the second signal leading to the proliferation of an activated cell. Ayara et al. (2005) presented an adaptive error detection (AED) system for automated teller machines, based on *AISEC*, where immature detectors need feedback supplied by a human expert to become competent. A different approach has been used by Sarafijanović and Le Boudec (2005), which proposed a model for anomaly detection in mobile *ad hoc* networks, where the second signal is an indication that a network node is conducting suspicious activities or malfunctioning. Finally, Dasgupta et al. (2005) have described multilevel immune learning algorithm (MILA), where the second signal is delivered to an stimulated B cell only if helper T cells are activated and suppressor T cells are not. In contrast, in *IA-AIS*, there's a true interaction between B and T cells, which is more accurate from a biological point of view.

An e-mail message is represented as a microorganism, where the message subject and body, which usually contain text, are analogous to antigens, while the sender's e-mail address is analogous to molecular patterns, as presented in Fig. 1. This definition is application dependent, and should be made so that features of an entry that directly define it as nonself are selected as molecular patterns, while other features which depend on each other and require a more elaborate analysis are used to generate antigens. This is due to the fact that molecular patterns are recognized by macrophages, while B and T cells recognize antigens. In the case of SPAM messages, this definition of a microorganism arises from the observation that all messages received after training the system, sent by someone whose e-mail address was present in SPAM messages used for training, will be SPAM. It is assumed that the user indicates that certain messages should not be used in this way, such as when a friend sends a SPAM message intended to be used for training.

For the generation of a microorganism, characters are encoded using a custom 6-bit encoding, specifically designed and tailored for this problem, as argued by Freitas and Timmis (2007). In this encoding, visually similar characters (for example, *O* and *0*, *E* and *3*) are represented by values that differ in only 1 or 2 bits (similar to the gray code, e.g. Salomon, 2004). Hence, a text sequence, such as the message contents, is represented by an antigen, with each word being analogous to a peptide in the antigen. As affinity measure (e.g. de Castro and Timmis, 2002), the relative number of matching bits in two sequences is used. Using this measure, in conjunction with the developed codification, it is possible that a cell with a receptor "test" binds to an antigen "t3st", if the activation threshold is properly defined. In this sense, the alternative spelling of words does not, in most cases, prevent the system from recognizing a pattern.

The system is trained by storing antigens and molecular patterns obtained from the data used for training, as presented in Fig. 2, with the contents extracted from SPAM and legitimate messages being represented as nonself and self data, respectively. Macrophages are generated using molecular patterns obtained from nonself entries that do not occur in the set of self molecular patterns. Regulatory T cells, on the other hand, are generated using randomly selected self antigens, which are processed and used to encode their receptors. The generation of B and helper T cells follows a similar procedure, using nonself antigens. An additional procedure is negative selection of these two types of cells, with the elimination of candidates that recognize self antigens. In the end, the surviving lymphocytes are added to the respective populations.

For the classification of a microorganism, depicted in Fig. 3, it is initially presented to the macrophage population, which analyzes its molecular patterns. If at least one macrophage is activated, the microorganism is classified as nonself, and the adaptive system is stimulated, through the generation of B and helper T cells specific for the antigens in the microorganism.

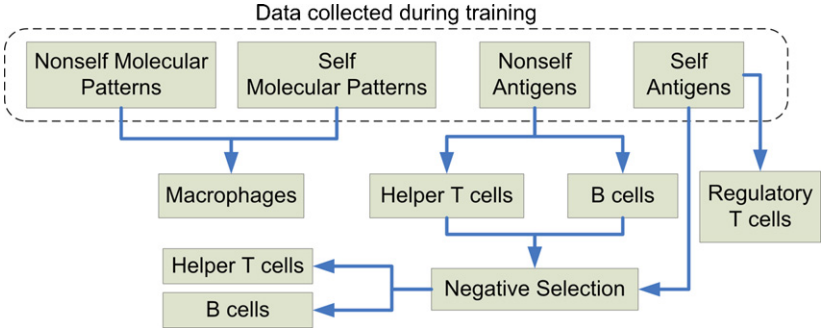

Fig. 2. Training procedure.

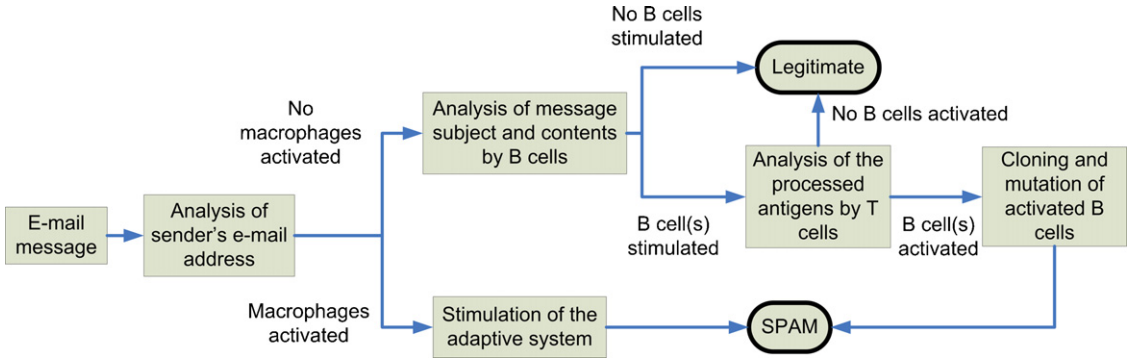

Fig. 3. Procedure for the classification of an e-mail message using the proposed model.

Otherwise, the analysis continues, with the presentation of the antigens to the B cell population. If no B cells bind to these antigens, the microorganism is classified as self, and the analysis is over. If, on the other hand, at least one B cell is stimulated, the antigen is processed and presented by the stimulated B lymphocytes to the T cell population. Depending on the reaction of T cells, the microorganism will be classified as self or nonself.

For an stimulated B cell to be activated, it must interact with helper and regulatory T cells. Before this, it processes the antigen, as illustrated in Fig. 4, by producing a sequence containing the peptides (words) in the region and in its vicinity that has lead to its stimulation. This procedure accounts, in the proposed model, for the presentation of peptides, derived from the antigen, attached to MHC molecules. In order to determine how close peptides have to be to the stimulation region, it is used a parameter, referred to as *antigen processing window*, as shown in Fig. 4. During the interaction, the stimulated B cell receives co-stimulation and suppression signals, emitted by helper and regulatory T cells, respectively. Co-stimulation signals can be thought of as the recognition of features as nonself by helper T cells, which will attempt to stimulate the initiation of a response. Suppression signals, on the other hand, are related to the recognition of self (normal) features, leading to the suppression of stimulated cells. The final outcome will be given by the combination of these two effects, depending on the magnitudes of the emitted signals, leading the stimulated B cell to be either activated or suppressed. In the latter case, it remains unresponsive, and does not influence the classification of the antigen. The B cell will be activated only if the sum of the number of times that all activated helper T cells have been activated exceeds that of regulatory T cells. The importance of this procedure is that it will attempt to capture the correct context of the expression that has stimulated each B cell. Considering the existence of degeneracy in the recognition, which can have undesired effects as discussed in Section 3, this is an important procedure. In addition, this interaction is a mechanism aimed at controlling the activation of self-reactive B cells, capable of recognizing expressions that are typical of legitimate messages, which may have escaped from negative selection.

When a cell is created, a *time to live*, an integer value representing a countdown to its death, is defined. This value is decremented after each antigen presentation, with the elimination of cells with a time to live equal to zero. In order to simulate the competition for antigen recognition, a cell-specific *bonus* is used: when activated, a cell will have its time to live incremented by this bonus, so that highly stimulated cells are kept in the population, and unstimulated cells are replaced by new ones. A related approach has been used in the aging mechanism of Cutello et al. (2007), where each cell maintains a counter that will lead to its elimination when it reaches a given value. In Cutello et al. (2007), when a cell is cloned, its age is reset, allowing each new clone to explore the search space. Therefore, the time to live of all cells that have been activated is defined in the same way, independently of the number of activations. In contrast, in *IA-AIS*, if a cell is repeatedly activated, it would take a long time to die. This is based on the principle that a given pattern

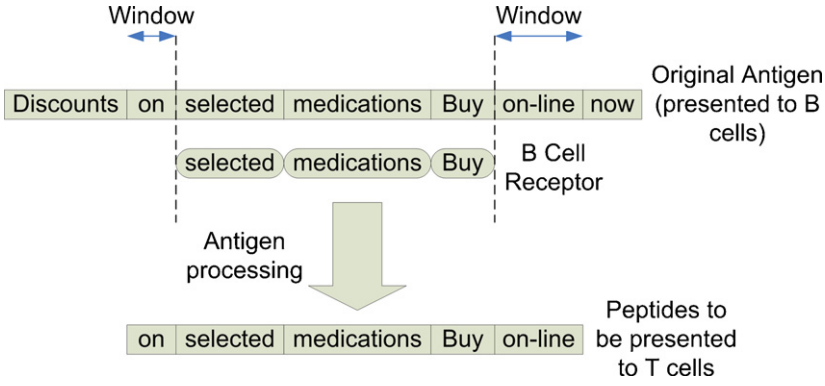

Fig. 4. Antigen processing by an stimulated B cell, for a processing window equal to 1.

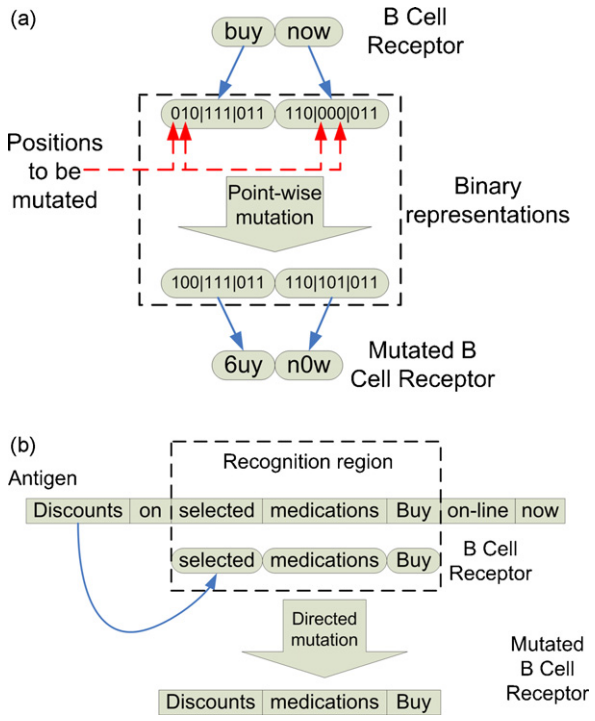

Fig. 5. Mutation operators: (a) point-wise mutation; (b) directed mutation.

typical of SPAM messages does not abruptly stop being used in all messages at once, so that a cell does not become suddenly useless, thereby rewarding cells that have been activated several times.

In this model, only B cells are allowed to reproduce, as a simplification. Given a B cell which recognized an antigen with affinity α, and has been successfully activated after interacting with T cells, the number n of clones to be generated is given by Eq. (8), where N_c is the maximum number of clones to be generated per activated B cell:

$$n = N_c \exp(-(1-\alpha)) \tag{8}$$

Each clone is then subjected to mutation, using the two operators depicted in Fig. 5. Point-wise mutation of the receptors, which are represented by binary strings, is applied with a probability β of mutating each bit, given by Eq. (9), where α is the affinity between the original B cell and the presented antigen, and ζ is a positive value, referred to as mutation factor:

$$\beta = 1 - \exp(-\zeta(1-\alpha)) \tag{9}$$

Therefore, cloning is proportional to the affinity, while the mutation probability is inversely proportional. Another mutation operator employed is named directed mutation, and replaces one word in the cell receptor with a pattern present in the recognized antigen. The only requirement for the selection of this pattern is that it must not be located in the recognition region, that is, the set of determinants recognized by the activated cell. The two mutation operators modify the clones according to two principles. Point-wise mutation generates clones that have a greater ability to recognize small variations in the epitopes previously encountered. These variations are typical, as discussed in Section 3, of alternative spelling of words. On the other hand, directed mutation will generate clones that are capable of recognizing variations in the structure of an antigen, characteristic of when, for example, an specific word in a message is replaced by a synonymous or related word. Finally, after generating the mutated clones, some of them are added to the B cell population. Currently, this selection process is random, i.e. the clones to be added are randomly selected out of the set of produced ones, unlike as proposed by de Castro and Von Zuben (2002), where only the clones with the greatest affinities are added. This is done because, in the specific application of SPAM filtering, it is considered that a low affinity does not necessarily imply that a clone is not important. Consider, for example, the case of directed mutation, where a clone has a receptor mutated by replacing one of the patterns used to encode its receptor. This clone may not have a high affinity at the time of its production, but only at a future time, when it is stimulated by a previously unseen antigen. If this is not the case, this clone will be eliminated, once its time to live reaches zero.

After classifying a microorganism, the system is updated, by decrementing the time to live of cells, eliminating those that reach zero and adding new cells, generated using randomly selected patterns. In order to ensure that the population size remains constant, it may be necessary to add more new cells, or remove cells that have not reached a time to live equal to zero. In the latter case, the cells to be removed are those with the lowest time to live. Using a constant population size is a simplification, and allows the system to better distribute candidate cells around the search space initially, when it is not storing much knowledge.

The procedures for correcting the classifier after a false positive or a false negative are shown in Fig. 6. In both cases, it is necessary to remove the stored patterns, eliminate cells that recognize these patterns and generate cells of appropriate types that will take their appearance into consideration. The latter is an initial stimulation for the classifier, so that similar messages encountered in the future will be correctly identified. The role of user intervention should be clear. It is an optional procedure, that can reinforce the classifier (in the case that it has correctly classified an entry) or correct it (if it has incorrectly classified the entry). In the former case, it merely involves adding the molecular patterns and antigens originated from the microorganism to the respective sets of patterns, depending on whether the microorganism is pathogenic or not. In fact, *IA-AIS* works even if no feedback is supplied at all, although probably achieving a low performance in this case.

5. Experimental Results

The proposed model was evaluated using SPAM messages available at Dornbos (2002), and legitimate messages from one of the authors. This dataset was selected because it includes a large number of SPAM messages (approximately 20,000 messages). The corpus selected was composed of a total of 2555 SPAM and 2513 legitimate messages, where 50% of the messages of each class were used for training. The classifiers were initially trained and then used to iteratively classify messages and received feedback, similar to the procedure used by Delany et al. (2005). The available messages were randomly shuffled and

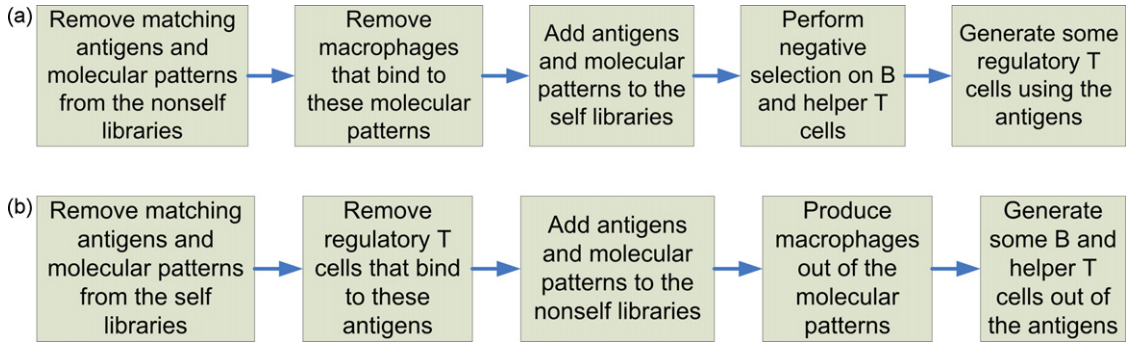

Fig. 6. Procedures for correcting the classifier: (a) correction of a false positive; (b) correction of a false negative.

then partitioned, generating the training set and the testing set. It should be noted that this process removes the time ordering of the messages, and the reason for using it can be understood through an example. Although a sequence of variations in a pattern such as "*free*" to "*fre3*", and then to "*fr33*", could be present in the dataset, by randomizing the messages it is possible to evaluate the classifiers considering multiple orderings of these variations (e.g. by presenting a message containing the pattern "*fr33*", then another containing "*free*", and so on). This is important because SPAM messages with similar modified patterns can be sent by multiple users, in the sense that a person may receive, at the same time, two SPAM messages, one containing a pattern "*fre3*", and another containing "*fr33*". Hence, the process used allows the performance assessment under several hypothetical conditions, in terms of modified patterns present in the dataset.

The testing sets were separated in 10 parts, so that the classifiers would only receive feedback after classifying each part. Therefore, the feedback is delayed, accounting for the possibility that the user will only verify the analyzed messages from time to time. The classifiers have been evaluated in five different partitions of the messages into training and testing sets, with the repetition of the classification of each partition five times, to allow a comparison of the standard deviation of the obtained results. Although this is a small sample size, from which it is, in general, difficult to extract statistical information, it was verified that the obtained results tend to be contained in a well defined interval of values. This is due to the introduction of feedback, which, in a certain way, counters the stochastic characteristic of the classifier. The naïve Bayes classifier was found to produce very low standard deviations in the obtained results, also due to the introduction of classifier feedback.

In *IA-AIS*, the effects of the number of patterns used to encode B cell receptors (BCRs) and the antigen processing window length have been evaluated. The cases considered were when the former was randomly, uniformly distributed in the intervals [1, 5], [2, 5], [3, 5] and [4, 5]. The equivalent parameters for helper and regulatory T cells were kept constant and equal to 1, because of the large number of possible combinations. In addition, as the signal computation phase can be time consuming, it is desirable to keep these two parameters as low as possible. On the other hand, varying the antigen processing length (W) affects, indirectly, the contribution of T cells, because more or less patterns in the neighborhood of the region that has stimulated each B cell will be taken into consideration. Some of the remaining parameters used are presented in Table 1. Finally, the activation threshold for both B and T cells was kept constant for all cells. This choice was made because of the specific feature of the encoding used. By keeping the threshold constant, little variations in small and large words are treated differently, as in the former fewer absolute variations are allowed in order to stimulate or activate the cell. To understand this, consider two hypothetical words, where one letter in each one is randomly replaced by another letter in the alphabet. In this scenario, there's a greater chance that the larger word will still be recognized by a person, because the absolute variation is lower that in the smaller word.

Table 1
Values for some of the parameters in *IA-AIS*

Parameter	B cells	T cells	Macrophages
Time to live	400	300	5000
Time to live bonus	300	200	5000
Binding threshold	0.98	0.98	0.99
Number of cells added per iteration	15	10	3
Maximum population size	10,000	10,000	2000
Number of patterns used to encode receptors	Various	1	–
Mutation factor	0.4	–	–

In order to evaluate cloning and selection, two configuration sets were tested, with the relevant parameters shown in Table 2. In the first configuration, a set of values for the number of clones per activated B cell, the number of B cells to be cloned, the number of clones to be added to the population, and the time to live for the added clones were chosen empirically. In the second configuration, the parameters were adjusted so that more clones would

Table 2
The two configuration sets used for evaluation

Parameter	Configuration 1	Configuration 2
Time to live for B cell clones	300	100
Maximum number of clones per activated cell	2	3
Maximum number of cells to be cloned	25	50
Maximum number of clones to be added	20	75

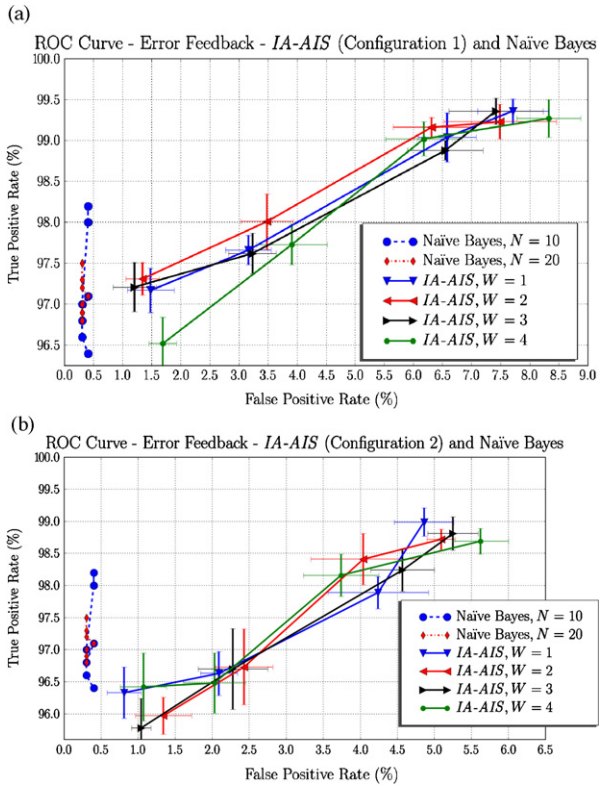

Fig. 7. ROC curves for both classifiers, considering error feedback: (a) configuration 1; (b) configuration 2.

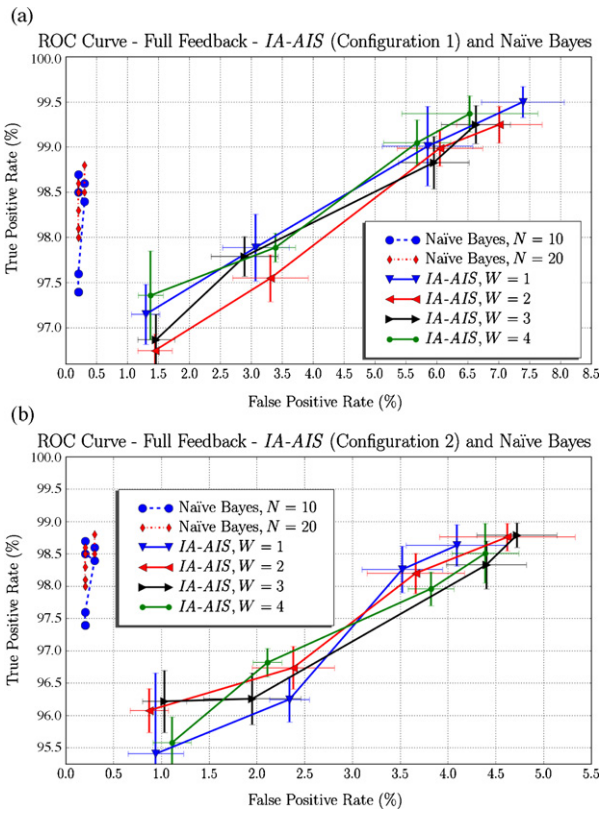

Fig. 8. ROC curves for both classifiers, considering full feedback: (a) configuration 1; (b) configuration 2.

be produced and added, but with a lower time to live. Therefore, the second configuration produces more clones on activation, counter-balancing this with a higher selective pressure on these clones, which will die more quickly unless stimulated. Finally, half of all the clones were submitted to point-wise mutation, and the other half to directed mutation.

The naïve Bayes classifier was evaluated by varying the number of patterns effectively used to determine the message's probability (N) and the classification threshold (β). When supplying feedback, there are several possibilities. One can increment the number of occurrences of all the patterns in the message, or only the ones used to compute the message's probability. In addition, the increment can be constant or variable, depending on the probability of the patterns (given by Eq. (6)). During the simulations, the number of occurrences for all patterns was incremented by 1. No further experiments have been conducted in this step, because the obtained results have been found to be excellent.

The obtained receiver operating characteristic (ROC) curves are shown in Figs. 7 and 8, considering the cases where only the incorrectly classified and all messages were used for feedback, respectively, and where the error bars represent the standard deviations. The results are reported in terms of the true positive (TPR) and false positive rates (FPR), the relative numbers of SPAM messages correctly identified as SPAM, and the number of legitimate messages incorrectly classified as SPAM, respectively. Due to the fact that the standard deviation in the results obtained using the naïve Bayes classifier is relatively low (0.1% at most), and to make the figures clearer, these curves show the data obtained using only $N = 10$ and $N = 20$, and omitting the error bars.

According to the error feedback ROC curves, it can be verified that the naïve Bayes classifier attains lower false positives in its entire range of true positive rates, in comparison with *IA-AIS*. The latter poses as a better alternative to the former in the case of a demand for high TPRs, as the naïve Bayes does not obtain TPRs exceeding 98.2%. Considering the full feedback curves, a similar situation is observed for the case of configuration 1. The second configuration, on the other hand, does not pose as an alternative to the naïve Bayes classifier, as the TPRs attained by these two are relatively close.

For *IA-AIS*, considering the first configuration and error feedback, true positive and false positive rates vary from 99.3% to 96.5% and from 8.4% to 1.2%, respectively. In the second configuration, these rates vary from 99% to almost 95.8%, and from 5.6% to 0.7%, respectively. It can be observed that the first configuration attains a better ability to identify SPAM messages, while making more mistakes when classifying legitimate messages, in comparison with the second configuration. This is due to the fact that the latter configuration tends to be less exploratory, and clones are less prone to be self-reactive than newly generated cells. When full feedback is used, in the first configuration the TPR varies between 99.5% and 96.8% as the interval for the number of patterns in BCRs is varied from [1, 5] to [4, 5], respectively. The FPR, on the other hand, varies between 7.5% and 1.3%. The window length appears to be inversely proportional to the standard deviation of TPRs, espe-

cially as the number of number of patterns in receptors increases. In the second configuration, the TPR decreases from 98.8% to 95.5% as the number of patterns increases, while the FPR increases from almost 4.7% to 0.8%. On the other hand, no clear conclusion can be drawn from the effects of the window length. Once again, the second configuration reaches lower FPRs, at the expense of lower TPRs. When comparing the two types of feedback analyzed, it can be concluded that, in the first configuration tested, the type of feedback has little effect on the performance, indicating that reinforcing the correctly classified messages is not really necessary. In the second configuration, the differences are noticeable, especially when the number of patterns used to encode receptors is large.

The naïve Bayes classifier has attained excellent results in both feedback situations, especially in terms of true negative rates. Considering error feedback (Fig. 7), it can be seen that it is capable of correctly identifying between 96.4% and 98.2% of SPAM messages, and 99.5% and 99.8% of legitimate messages. The TPR decreases as the classification threshold β is increased, for each value of N, while the FPR remains approximately constant. In this situation, increasing the number of patterns N to be considered also decreases the TPRs, although in a smaller amount in comparison with the classification threshold, with no observable effect on the FPRs. When analyzing the results obtained considering full feedback, shown in Fig. 8, it can be observed that between 97.4% and 98.7% of SPAM messages are correctly identified, in conjunction with between 99.7% and 99.9% of legitimate messages. The consequences of increasing the threshold β are the same, leading to smaller TPRs. On the other hand, increasing the number of patterns to be considered tends to increase the TPRs. This occurs because more information is used to determine the message's probability. When comparing the results obtained in both situations, it can be verified that including feedback on all the messages has increased true negative and true positive rates by an average value of 0.1% and 0.5%, respectively. Therefore, incorporating full feedback in this implementation of the naïve Bayes classifier can, in fact, lead to better results in comparison with error feedback.

Finally, during the simulations executed, it has been verified that the innate system would be activated $(15.7 \pm 0.9)\%$ of the times, with no false positives. In addition, some tests conducted in the absence of feedback indicate that the TPRs decreased by between 3% and 4.5%, while the FPRs are increased by between 5% and 6.5%, respectively, indicating that *IA-AIS* appears to be very dependent on feedback, especially in regard to the identification of legitimate messages. In the case of the naïve Bayes classifier, the lack of feedback results in a small effect on the false positive rates, while the true positive rates drops to values between 94% and 94.5%.

6. Conclusions

In this paper, an artificial immune system for the identification of SPAM messages, named *IA-AIS* was presented. The model includes macrophages, representing the innate immune system, and B and T cells. Its application to the problem of identification of SPAM is based on the recognition of the sender's e-mail address by a macrophage population, and the message's subject and contents by B and T cell populations. An important aspect is the consideration of user feedback, which requires the classifier to be able to acquire knowledge not presented during the initial training. A particularly interesting feature of the proposed model is that it emphasizes the importance of interactions between cells in the immune system, and not simply pattern matching.

It has been verified that *IA-AIS* can attain high classification rates, correctly classifying more than 99% of SPAM or legitimate messages, with different balances between false positives and false negatives depending on some parameters. In comparison with the naïve Bayes classifier with the optimizations proposed by Graham (2002) (which are not considered by most of the work in the literature) it is concluded that *IA-AIS* poses as an interesting alternative when high true positive values are preferable, at the expense of more false positives, although such situation is rare in practical scenarios. This conclusion does not necessarily imply the inappropriateness of the model proposed, as only some parameters have been analyzed to obtain the results. An important advantage of *IA-AIS* is that it guarantees that, in the case of a legitimate message misclassified as SPAM, it will not be incorrectly classified in the future, because the message patterns will be used to prevent the production of cells capable of recognizing this message. This is an important characteristic because incorrectly classifying a message that has been previously corrected is highly frustrating for the user, indicating that the classifier is unable to acquire knowledge obtained after training.

Future investigations to be conducted are the impact of the number of patterns used to encode T cell receptors, and methods for automatic setting some of the classifier parameters, based on user preferences. The good results obtained have motivated the development of a prototype (available at http://www.cpdee.ufmg.br/~tguzella), incorporating implementations of both *IA-AIS* and a naïve Bayes classifier with the optimizations proposed by Graham (2002).

Acknowledgments

This work was supported by UOL, through its Bolsa Pesquisa program (process number 20060519110414a), PQI/CAPES and CNPq (process number 307178/2004-8). The authors would like to thank an anonymous reviewer and the editor-in-chief for their insightful comments, which helped improve the quality of the paper.

References

Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Spyropoulos, C.D., 2000. An experimental comparison of naïve bayesian and keyword-based anti-SPAM filtering with personal e-mail messages. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Inf. Ret, pp. 160–167.

Ayara, M., Timmis, J., de Lemos, R., Forrest, S., 2005. Immunizing automated teller machines. Lect. Notes Comput. Sci. 3627, 404–417.

Bezerra, G.B., Barra, T.V., Ferreira, H.M., Knidel, H., de Castro, L.N., Zuben, F.J.V., 2006. An immunological filter for SPAM. Lect. Notes Comput. Sci. 4163, 446–458.

Burnet, F.M., 1959. The Clonal Selection Theory of Acquired Immunity. Cambridge University Press.

Carpinter, J., Hunt, R., 2006. Tightening the net: a review of current and next generation SPAM filtering tools. Comput. Secur. 25 (8), 566–578.

Cohen, I., 1992. The cognitive principle challenges clonal selection. Immunol. Today 13 (11), 441–444.

Cohn, M., 2005. The common sense of self–nonself discrimination. Springer Semin. Immunol. 27 (1), 3–17.

Coutinho, A., 2005. The Le Douarin phenomenon: a shift in the paradigm of developmental self-tolerance. Int. J. Dev. Biol. 49, 131–136.

Cutello, V., Nicosia, G., Pavone, M., Timmis, J., 2007. An immune algorithm for protein structure prediction on lattice models. IEEE Trans. Evol. Comput. 11 (1), 101–117.

Dasgupta, D., Yu, S., Majumdar, N.S., 2005. MILA—multilevel immune learning algorithm and its application to anomaly detection. Soft Comput. 9, 172–184.

de Castro, L.N., Timmis, J., 2002. Artificial Immune Systems: A New Computational Intelligence Approach, 1st ed. Springer.

de Castro, L.N., Von Zuben, F.J., 2002. Learning and optimization using the clonal selection principle. IEEE Trans. Evol. Comput. 6 (3), 239–251.

Delany, S.J., Cunningham, P., Tsymbal, A., Coyle, L., 2005. A case-based technique for tracking concept drift in SPAM filtering. Knowl.-Based Syst. 18, 187–195.

Dornbos, J., 2002. SPAM: What Can You Do About It? Visited on August 2006. URL: www.dornbos.com/spam01.shtml.

Freitas, A.A., Timmis, J., 2007. Revisiting the foundations of artificial immune systems for data mining. IEEE Trans. Evol. Comput. 11 (4), 521–540.

Graham, P., 2002. A Plan For SPAM. Visited on August 2006. URL: www.paulgraham.com/spam.html.

Greensmith, J., Aickelin, U., Cayzer, S., 2005. Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. Lect. Notes Comput. Sci. 3627, 153–167.

Guzella, T.S., Uchôa, J.Q., Mota-Santos, T.A., Caminhas, W.M., (in Portuguese) 2005. Proposal of a pattern recognition model inspired on the immune system: an application to the identification of SPAM. In: Proceedings of the VII Brazilian Congress on Neural Networks (CBRN 2005), vol. 1.

Janeway, C.A., Travers, P., Walport, M., Shlonmchik, M., 2001. The Immune System in Health and Disease, 5th ed. Garland Publishing, Oxford.

Ji, Z., Dasgupta, D., 2007. Revisiting negative selection algorithms. Evol. Comput. 15 (2), 223–251.

Kim, J., Bentley, P.J., 2002. Towards an artificial immune system for network intrusion detection: an investigation of dynamic clonal selection. In: Proceedings of the IEEE CEC, pp. 1015–1020.

Lai, C.-C., 2007. An empirical study of three machine learning methods for SPAM filtering. Knowl.-Based Syst. 20 (3), 249–254.

Leopold, E., Kindermann, J., 2002. Text categorization with support vector machines. How to represent texts in input space? Mach. Learn. 46, 423–444.

Oda, T., White, T., 2003a. Developing an immunity to SPAM. Lect. Notes Comput. Sci. 2723, 231–242.

Oda, T., White, T., 2003b. Increasing the accuracy of a SPAM-detecting artificial immune system. In: Proceedings of the IEEE CEC, vol. 1, pp. 390–396.

Sahami, M., Dumais, S., Heckerman, D., Horvitz, E., 1998. A Bayesian Approach to Filtering Junk e-mail. Tech. Rep. WS-98-05, AAI Press.

Sakaguchi, S., 2004. Naturally arising CD4+ regulatory T cells for immunologic self-tolerance and negative control of immune responses. Annu. Rev. Immunol. 22, 531–562.

Salomon, D., 2004. Data Compression, 3rd ed. Springer-Verlag.

Sarafijanović, S., Le Boudec, J.-Y., 2005. An artificial immune system approach with secondary response for misbehavior detection in mobile ad hoc networks. IEEE Trans. Neural Networks 16 (5), 1076–1087.

Schwartz, R.H., 2005. Natural regulatory T cells and self-tolerance. Nat. Immunol. 6 (4), 327–330.

Secker, A., Freitas, A.A., Timmis, J., 2003. *AISEC*: an artificial immune system for e-mail classification. In: Proceedings of the IEEE CEC, vol. 1, pp. 131–138.

Stepney, S., Smith, R.E., Timmis, J., Tyrrell, A.M., Neal, M.J., Hone, A.N.W., 2005. Conceptual frameworks for artificial immune systems. Int. J. Unconv. Comp. 1 (3), 315–338.

Tedesco, G., Twycross, J., Aickelin, U., 2006. Integrating innate and adaptive immunity for intrusion detection. Lect. Notes Comput. Sci. 4163, 193–202.

Twycross, J., Aickelin, U., 2005. Towards a conceptual framework for innate immunity. Lect. Notes Comput. Sci. 3627, 112–125.

Twycross, J., Aickelin, U., 2006. libtissue—implementing innate immunity. In: Proceedings of the IEEE CEC, pp. 499–506.

Varela, F.J., Coutinho, A., Dupire, B., Vaz, N.M., 1988. Cognitive networks: immune, neural and otherwise. In: Perelson, A.S. (Ed.), Theoretical Immunology. Part 2. SFI Studies in the Sciences of Complexity. Addison-Wesley, Boston, pp. 359–375.

von Boehmer, H., 2005. Mechanisms of suppression by suppressor T cells. Nat. Immunol. 6 (4), 338–344.

Wang, B., Jones, G.J.F., Pan, W., 2006. Using online linear classifiers to filter SPAM e-mails. Pattern Anal. Appl. 9, 339–351.

Yue, X., Abraham, A., Chi, Z.-X., Hao, Y.-Y., Mo, H., 2007. Artificial immune system inspired behavior-based anti-SPAM filter. Soft Comput. 11, 729–740.