On using an ensemble approach of AIS and SVM for text classification

Catarina Silva*†, Mário Antunes*‡, Bernardete Ribeiro†, Manuel Correia‡

*School of Technology and Management, Computer Science and Communications Research Centre (CIIC)
Polytechnic Institute of Leiria, Portugal
{catarina,mario.antunes}@ipleiria.pt

†Department of Informatics Engineering, Center for Informatics and Systems (CISUC)
University of Coimbra, Portugal
bribeiro@dei.uc.pt

‡Center for Research in Advanced Computing Systems (CRACS)
Faculty of Science, University of Porto, Portugal
mcc@dcc.fc.up.pt

***Abstract*—Artificial Immune Systems (AIS) and Support Vector Machines (SVM) are grounded on two radically different conceptual paradigms, each one having intrinsic distinctive features suitable to be successfully applied in dynamic real world applications. One of such applications is the classification of textual documents where each approach individually has proved to obtain promising results.**

In this paper we aim to present an hybrid system for text classification based on the ensemble of both AIS and SVM approaches. In AIS we explore a binary classification methodology derived from an immunological model which stats that for activation thresholds for T-cells activation is based on the recent history of their iterations with the environment. Regarding the SVM we take advantage of a non-evolutionary implementation that produced remarkable results with text classification. We report some preliminary results on the well-known Reuters-21578 benchmark, showing promising classification performance gains, resulting in a classification that improves upon all baseline contributors of the ensemble committee.

Keywords: Support Vector Machine, Artificial Immune System, Text Classification, Tunable Activation Thresholds, Ensembles

I. Introduction

The vertebrate Immune System (IS) has proved to be a very interesting and emergent source of inspiration for the development of innovative solutions applied to computer science and engineering fields, like anomaly detection and classification. Artificial Immune Systems (AIS) have incorporated the immune models and processes in a conceptual framework for the purpose of outperforming other detection and classification methods in such dynamic behaviors.

In the last decades the production of textual documents in digital form has increased exponentially, due to the increased availability of hardware and software [1]. As a consequence, there is an ever-increasing need for automated solutions to organize the huge amount of digital texts produced, in applications such as document processing and visualization, Web mining, digital information search and patent analysis. The task in text classification is often defined as assigning previously defined classes to documents (natural language texts) by analysing their content. While many techniques have successfully been used in tackling the problem of text classification, current research is focused on kernel-based algorithms mainly due to their performance accuracy and sparsity of the final solution. Examples are Vapnik's Support Vector Machine (SVM) [2] which implement the principle of structural minimization and different solutions based on committees of kernel-based machines, such as boosting.

This paper aims to evaluate the appropriateness of using an evolutionary Tunable Activation Threshold (TAT)-based approach to improve the overall performance of text classification. We propose an ensemble committee approach composed by an AIS and an SVM implementation. The results were obtained with the Reuters-21578 data set and show that a conjugated decision of the immune-based approach with SVM improve the final classification decision.

The rest of the paper is organized as follows. We start by presenting in Section II the fundamentals of the baseline AIS and SVM learning systems. We then proceed in Section III by describing the proposed approach to deal with text classification. Then, we show and discuss in Section IV the results obtained on processing the Reuters-21578 data set. Finally, in Section V we discuss the conclusions of our work and terminate by delineating some future research directions.

II. Background

Here we describe the fundamentals of the immunological model adopted in the proposed AIS for anomaly detection, as well as SVM and committee-based learning systems.

A. Artificial Immune Systems

The IS evolved to become a highly complex defense mechanism that has the ability to recognize foreign substances (pathogens) and to distinguish between those that correspond to the harmless (self) from those that are related to some form of intrusion (non-self) [3]. It is composed by two main layers of defense: *innate* and *adaptive*. The former only recognizes specific known substances and its behavior is similar in all individuals of the same species.The latter is apparently unique to each individual and is able to "learn" and to recognize new forms of abnormal pathogens that gradually change throughout time, thus providing a highly sophisticated adaptive form of pathogen recognition. The IS is also supported by a complex set of cellular structures. Among them, the Antigen Presenting Cell (APC) digests and converts pathogens into small *peptides* which are then presented to the lymphocytes (*T-cells*), through a molecular structure denominated "MCH/Peptide Complex". T-cells have a specific set of *receptors* that *binds* with a certain degree of affinity with the peptides that are being presented by APCs.

Artificial Immune Systems (AIS) are defined as adaptive systems inspired by theoretical immunology and observed immune processes and models, which can be applied to a problem solving [4]. Nowadays, there exists already a full body of theoretical work involving models and algorithms devised by theoretical immunologists that describe and successfully predict certain aspects of the IS behaviour. This constitute the basis and source of inspiration behind the developments of AIS [4] for several domain applications, like anomaly detection [5] and classification [6]–[8].

B. A model for tuning T-cells activation thresholds

Autoimmunity is still an open issue among the immunologists research community and was partially explained by Burnet's Negative Selection (NS) theory [9] and Matzinger's Danger Theory (DT) [10]. Besides their contributions on explaining the IS self/non-self discrimination, both models failed to clarify the presence of mature auto-reactive lymphocytes in normal healthy individuals that may cause autoimmune diseases. Interestingly, those auto-reactive lymphocytes circulate in healthy individuals and are prevented from mounting an harmful immune response against its own body tissues. Moreover, this observed IS function may be related to another one, that is how the immune system adjusts its response to the environmental context in which antigens are recognised [11].

In this direction, Grossman postulated that lymphocytes tune up and update their responsiveness according to the stimuli received from the environment. Grossman's TAT model [11] and Carneiro's simplified model [12] were behind the TAT model we adopted to build an AIS for anomaly detection [13]. The underpinning idea is that lymphocytes tune up and update their responsiveness according to the stimuli received from the environment. According to the model, those auto-reactive lymphocytes that are continuously stimulated by self antigen end up by adapt its level of responsiveness and thus prevent from mounting an immune response. Those that are sporadically stimulated with a strong stimulus become activated and are thus able to start an immune response [11], [13]. The TAT functioning can be described as follow:

1) T-cell activation process is controlled by the activity of two specific enzymes that respond to antigenic signals (S) delivered by the APC: Kinase (K) and Phosphatase (P).
2) T-cell activation is a switch-type response that requires that K supersedes P, at least transiently.
3) At each given moment in time, T-cells interacts with the peptide presented by APCs and receive a stimulus that depends on the *affinity* between its receptor (T-cell Receptor (TCR)) and the peptide ligand, causing the cell to adapt by increasing or decreasing its activation threshold.
4) During the T-cell conjugation with an APC, TCR stimulus result in a faster increase of both kinase ans phosphatase enzymatic activities.
5) If, after the conjugation of T-cell and APC, the kinase level supersedes that of phosphatase, then the T-cell is activated. Otherwise, the T-cell remains in a quiescent state.
6) T-cells react differently to the signals they receive from APC. (through pairs "MHC/peptide"), *adjusting* its threshold of activation proportionally to the signals received from the APC

Thus, within this model, each T-cell has its own responsiveness and tuning updated according to the history of intracellular interactions between T-cell and APC, which means that different cells with different antigen-specificity end up having different activation thresholds as they are exposed to different stimuli.

The activation threshold increases gradually if the signals received are recurrent and decreases in the absence of signals. T-cells should be *activated* if, in a given period of time, the signals received from the APC are higher than the current threshold. Notice that this can happen if a T-cell does not receive signals from an APC for some time and ends up with a substantially decreased threshold, thus becoming much easier to activate in the presence of higher signals.

We also made the following simplifications:

- both K and P are exposed to the same stimulus S;
- P's basal value (P_0) is higher than K's (K_0);
- S_0 is the initial value for S;
- K's turnover rate (τK) is higher than P's (τP);
- K's slope (ϕK) is higher than ϕP's;

In order to reduce the number of simulation parameters and therefore to simplify their run time optimization and consequently the overall processing time, we have chosen to derive K_0, P_0, τK and ϕP according to the following:

- we assume $K_0 = S_0 \cdot \tau\text{K}$ and $P_0 = S_0 \cdot \tau\text{P}$;
- we derived $\tau K = \tau \cdot \tau P$, with $\tau = \frac{\tau K}{\tau P}$;
- we also derived $\phi P = \phi \cdot \phi K$, with $\phi = \frac{\phi P}{\phi K}$;
- the IS's speed of response is given by a constant value (t);

Assuming $\{P_0, K_0\}$ as the basal values, in each iteration i the values for K and P are given by the following equations (1 and 2):

$$K_i = \begin{cases} \min((S+S_0)\cdot\tau K, K_{i-1} += \phi K \cdot t); & \text{if } (S+S_0)\cdot\tau K) > K_{i-1} \\ \max((S+S_0)\cdot\tau K, K_{i-1} -= \phi K \cdot t); & otherwise \end{cases} \quad (1)$$

$$P_i = \begin{cases} \min((S+S_0)\cdot\tau P, P_{i-1} += \phi P \cdot t); & \text{if } (S+S_0)\cdot\tau P) > P_{i-1} \\ \max((S+S_0)\cdot\tau P, P_{i-1} -= \phi P \cdot t); & otherwise \end{cases} \quad (2)$$

In this simple set-up, the activation threshold corresponds to the difference between P and K activities. The higher the value of the former relative to the latter, the greater the difficulty to activate the T-cell. Under these conditions, those T-cells that receive continuous or sufficiently frequent antigenic signals from APC become unresponsive and those that rarely see their antigen remain sensitive to a further activation [12].

C. Immunological metaphor

The native features of TAT immunological model, namely its adaptive and dynamic self tolerance and non-self discrimination processes described previously, highlights an intrinsic self/non-self discrimination function that can be intuitively transposed to non-biological and contextual application domains. One such appealing application is the text filtering, in which a set of texts should be filtered out before reaching a user or application. The best known example are the email spam filtering applications. Generally speaking, a spam filtering comprises the classification of incoming email messages between legitimate messages (ham) and unsolicited ones (spam). The ham messages corresponds to the "self" profile and is composed by the mostly used words in each one's email conversations. Otherwise, the spam (or non-self) messages has a subset of words whose relationship may indicate an unsolicited kind of information. Also in this example, the meaning of ham and spam changes over time. From one hand, one frequently initiate conversations about new subjects and with previously unused email addresses. On the other hand, the meaning of what is deemed by spam also changes continuously. Two examples about the dynamic change of spam are the innovative ways by which some assumed spam words are written in order to evade the spam filters (like, "v1agra" and "vIaGr@"), as well as the inclusion of new unseen and previously undeclared spam words in the unsolicited email messages.

Another appealing example is the text classification like the processing of Reuters-21578 Text Categorization Test Collection. Although it is a multi-class, multi-label and heterogeneous corpus of news articles, the underpinning idea is that it can be seen in a binary classification perspective and thus suitable to fall into an anomaly detection problem. Considering the whole data set, those documents belonging to a particular class we are looking for represents the abnormal behavior. Otherwise, those related to the remaining classes should define the normal behavior. The Reuters-21578 documents data set is the most popular benchmark test data set and has been used to test several classifier algorithms, being SVM the one that has achieved the best performance [14].

In the TAT-AIS, as with other AIS [4], there is a direct mapping of artificial system components with its relevant biological IS counterparts. To better contextualise the TAT immunological metaphor, Table I depicts the main immunological components and its corresponding artificial counterparts when applied specifically to text filtering.

Immune System	**AIS counterpart**
Antigen	A text document, representing a contextual behaviour (class).
APC	A text document, representing a contextual bulk of information carried on by the artificial antigens.
Phagocytosis	Data preprocessing, which consists on producing the chunks of information to be presented by the APC.
Peptide	The words in the document, representing the chunks of information contained in the APC.
MHC/Peptide ligand	A word representative of artificial peptides.
T-Cell	Artificial detector.
T-Cell Receptor (TCR)	A word representative of the artificial detectors.
T-Cell Repertoire	List of available detectors.
T-cell Activation	According to TAT dynamics, a detector with $K > P$
Affinity	Distance measure between the words representative of peptides and T-cells
Autoimmune response	Corresponds to a false text classification raised by the system.
Non-self	Documents set from the class we intend to classify.
Self	A set of documents from the remaining classes.

Table I: Immunological TAT metaphor for text classification.

D. Support Vector Machines

SVM are a learning method introduced by Vapnik [2] based on his Statistical Learning Theory and Structural Minimization Principle. When using SVM for classification, the basic idea is to find the optimal separating hyperplane between the positive and negative examples. The optimal hyperplane is defined as the one giving the maximum margin between the training examples that are closest to it. Support vectors are the examples that lie closest to the separating hyperplane. Once this hyperplane is found, new examples can be classified simply by determining on which side of the hyperplane they are. Figure 1 shows a simple two-dimensional example, the optimal separating hyperplane and four support vectors.

a) Foundations: SVM start from a simple linear maximum margin classifier. Given an independent and identically distributed sample $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_l, y_l)$, where $\mathbf{x}_i$ for $i = 1, ..., l$ is a feature vector of length l and $y_i = \{+1, -1\}$ is the class label for $\mathbf{x_i}$, find a classifier with the decision function $f(x)$, such that $y = f(x)$, where y is the class label for x.

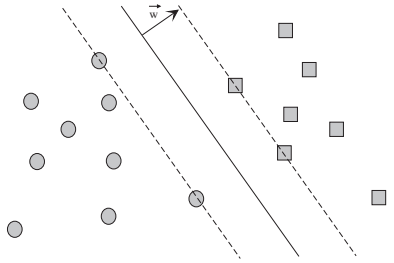


Figure 1: Optimal Separating Hyperplane.

The performance of the classifier is measured in terms of classification error, as defined in (3).

$$E(y, f(x)) = \begin{cases} 0 & if \quad y = f(x), \\ 1 & otherwise. \end{cases} \tag{3}$$

Learning machines have a set of adjustable parameters, λ. Given the above classification task, the machine will tune its parameters λ to learn the mapping $\mathbf{x} \rightarrow y$. This will result in a mapping $\mathbf{x} \rightarrow f(\mathbf{x}, \lambda)$, which defines the particular learning machine. The performance of this machine can be measured by the expectation of the test error, as shown in (4).

$$R(\lambda) = \int E(y, f(\mathbf{x}, \lambda)) \quad dP(\mathbf{x}, y) \tag{4}$$

This is called the expected risk and requires that at least an estimate of $P(\mathbf{x}, y)$ is known, which is not available for most classification tasks. Thus the empirical risk measure, defined in (5), has to be used.

$$R_{emp} = \frac{1}{l} \sum_{i=1}^{l} E(y, f(\mathbf{x}, \lambda)). \tag{5}$$

The empirical risk provides a measure of the mean error over the available training data and most training algorithms implement its minimization (Empirical Risk Minimization), i.e., minimize the empirical error using maximum likelihood estimation for parameters λ. These conventional training algorithms do not consider the capacity of the learning machine and this can result in over fitting, i.e., using a learning machine with too much capacity for a particular problem.

In contrast with ERM, the goal of SRM [2] is to find the learning machine that yields a good trade-off between low empirical risk and small capacity. There are two major problems in achieving this goal: (i) SRM requires a measure of the capacity of a particular learning machine or, at least, an upper bound on this measure; (ii) an algorithm to select the desired learning machine according to SRM's goal is needed.

To address these two problems Vapnik and Chervonenkis [2] proposed the concepts of Vapnik Chervonenkis (VC) confidence and SVM.

It is possible to choose a function that classifies well training data, but does not generalize well on test or real data, i.e., therefore the real *Risk* (see 4) will not be minimized. In VC theory, Vapnik and Chervonenkis prove that it is necessary to restrict the class of functions that f is chosen from to one with the *capacity* suitable for the amount of training data. VC theory provides bounds on the test error, circumventing the generalization problems presented earlier. Minimizing these bounds leads to the principle of *Structural Risk Minimization*. A function's capacity can take the form of VC-Dimension, defined as the largest number h of points that can be separated in all possible ways, using functions of the given class. If $h < l$ is the VC-Dimension of the class of functions that the learning machine can implement, then for all the functions of that class, with a probability of at least $1 - \eta$, the bound (6):

$$R(\lambda) \leq R_{emp}(\lambda) + \phi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right) \tag{6}$$

holds, where the *confidence term* ϕ is defined as (7):

$$\phi\left(\frac{h}{l}, \frac{\log(\eta)}{l}\right) = \sqrt{\frac{h(\log\frac{2l}{h} + 1) - \log(\frac{\eta}{4})}{l}} \tag{7}$$

b) Support Vector Classification: Although text categorization is a multi-class, multi-label problem, it can be broken into a number of binary class problems without loss of generality. This means that instead of classifying each document into all available categories, for each pair $\{document, category\}$ we have a two class problem: the document either belongs or does not to the category.

Although there are several linear classifiers that can separate both classes, only one, the Optimal Separating Hyperplane, maximizes the margin, i.e., the distance to the nearest data point of each class, thus presenting better generalization potential.

The output of a linear SVM is $u = \mathbf{w} \times \mathbf{x} - b$, where $\mathbf{w}$ is the normal vector to the hyperplane and $\mathbf{x}$ is the input vector. Maximizing the margin can be seen as an optimization problem:

$$\begin{aligned} &minimize \quad \frac{1}{2}||\mathbf{w}||^2, \\ &subjected \ to \ \ y_i(\mathbf{w}.\mathbf{x} + b) \geq 1, \forall i, \end{aligned} \tag{8}$$

where x_i is the training example and y_i is the correct output for the ith training example, as represented in figure 1.

Intuitively the classifier with the largest margin will give low expected risk, and hence better generalization.

To deal with the constrained optimization problem in (8) Lagrange multipliers $\alpha_i \geq 0$ and the Lagrangian (9) can be introduced:

$$L_p \equiv \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i(y_i(\mathbf{w}.\mathbf{x} + b) - 1). \tag{9}$$

The Lagrangian has to be minimized with respect to the primal variables $\mathbf{w}$ and b and maximized with respect to the dual variables α_i (i.e. a saddle point has to be found) [15].

SVM are universal learners. In their basic form, shown so far, SVM learn linear threshold functions. However, using an appropriate kernel function, they can be used to learn polynomial classifiers, radial-basis function networks and three layer sigmoid neural networks.

E. Committee classification approaches

Classifier committees or ensembles are based on the idea that, given a task that requires expert knowledge, k experts may perform better than one, if their individual judgments are appropriately combined. A classifier committee is then characterized by (i) a choice of k classifiers, and (ii) a choice of a combination function, usually denominated a voting algorithm. The classifiers should be as independent as possible to guarantee a large number of inductions on the data. By using different classifiers to exploit diverse patterns of errors to make the ensemble better than just the sum (or average) of the parts, we can obtain a gain from potential synergies existing between the different ensemble classifiers [16].

An ensemble is started by creating base classifiers with necessary accuracy and diversity. Unlike the traditional approach of choosing the best performing learning machine, an ensemble strategy compares the performance of the combined output with the selection and use of the best one, in terms of classification performance. Thus, our goal is to improve classification performance, which is possible when the base classifiers show different patterns of errors, since the errors made by one of them can be compensated by the correct output of others. There exist several methods to create the set of elements in an ensemble, such as, different training samples, different preprocessing methods or different learning parameters. The conjugation of their results can also be accomplished in a number of ways, like weighted average or majority voting.

F. Text classification

The goal of text classification is the automatic assignment of documents to a fixed number of semantic categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples, which assign categories automatically. This is usually considered a supervised learning problem. To facilitate effective and efficient learning, each category is treated as a separate binary classification problem. Each of such problems answers the question of whether or not a document should be assigned to a particular category.

Documents, which typically are strings of characters, have to be transformed into a suitable representation for both the learning algorithm and the classification task. The most common representation is known as the *Bag of Words* and represents a document by the words occurring in it. Usually the irrelevant words are filtered using a stopword list and the word ordering is not deemed relevant for most applications. Information retrieval investigation proposes that instead of words, the units of representation could be word stems. A word stem is derived from the occurrence form of a word by removing case and inflection information. For example "viewer", "viewing", and "preview" are all mapped to the same stem "view".

This leads to an attribute-value representation of text. Each distinct word w_i corresponds to a feature $TF(w_i, x)$, representing the number of times word w_i occurs in the document x. Refining this basic representation, it has been shown that scaling the dimensions of the feature vector with their inverse document frequency $IDF(w_i)$ leads to an improved performance. $IDF(w_i)$ (10) can be calculated from the document frequency $DF(w_i)$, which is the number of documents the word w_i occurs in.

$$IDF(w_i) = log\left(\frac{D}{DF(w_i)}\right) \tag{10}$$

Here, D is the total number of documents. The inverse document frequency of a word is low if it occurs in many documents and is highest if the word occurs in only one. To disregard different document lengths, each document feature vector $\mathbf{x}_i$ is normalized to unit length.

III. The ensemble model adopted

This section presents an ensemble of a TAT-based model with an SVM implementation for text classification.

In what follows we present the proposed AIS-SVM ensemble structure. There are several methods to create the set of elements in an ensemble, such as, different training samples, applying diverse preprocessing methods or using various learning parameters. The conjugation of their results can also be accomplished in a number of ways, like weighted average or majority voting. Having in this case two radically different approaches to structure an ensemble framework, we defined a two-level hybrid model illustrated in Figure 2 that joins the predictions of both SVM and TAT-based models. During the training phase the models are dealt with separately,

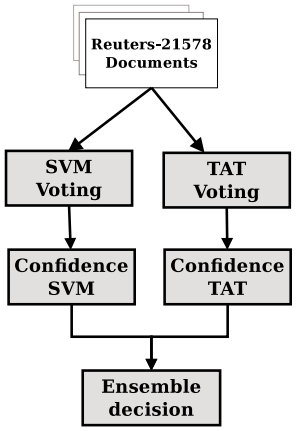


Figure 2: TAT based and SVM hybrid model for text classification.

i.e. a number n of classifiers is generated by varying SVM parameters and a number m of classifiers is generated varying the TAT parameters. On the other hand, for the testing phase, first each model is called to independently classify a testing example, and then two sets are constructed, one for each type of model (SVM and TAT). We then apply a majority voting strategy to each set to define its decision, i.e. if the document is a positive or negative example of the class.

When both SVM and TAT sets agree on the classification of the testing example the two-level model outputs directly their consensus decision. However, if both sets majority voting disagree or tie (ties can happen when n or m are even), a different algorithm must be in place. We defined a heuristic voting rule based on the strength D of the confidence of each set decision. To determine D for SVM one follows these steps:

1) sum SVM outputs
2) when all agree, we have a baseline 100%, otherwise 50%
3) when the sum is less than 3, we say that prediction is 50%, otherwise 100%
4) D is the product of baseline by prediction

And for TAT D is determined according to:

1) when all agree, we have a confidence of 100%
2) when we have maximum disagreement we have a confidence of 0%
3) confidence is proportional to the degree of agreement between TAT classifiers

The set with higher confidence will define the output of the two-level hybrid model in Figure 2. Note that the value of D must be the same for both sets of models. In our experiments, detailed in Section IV, we used $n = 3$, $m = 4$ and $D = 4$.

IV. Experimental evaluation and results

This section describes the results obtained by processing the Reuters-21578 data set with the AIS-SVM ensemble framework previously described in Section III. We start by analysing the data set and its appropriateness to be processed by the TAT-based AIS. We then proceed by describing the experimental setup and conclude by evaluating the results obtained.

A. Reuters-21578 data set

The widely used Reuters-21578 benchmark was used in the experiments. It is a financial corpus with news articles documents averaging 200 words each. Reuters-21578 is publicly available [1] and its corpus has 21,578 documents classified into 118 categories. It is a very heterogeneous corpus, since the number of documents assigned to each category is very variable. There are documents not assigned to any of the categories and documents assigned to more than 10 categories. On the other hand, the number of documents assigned to each category is also not constant. There are categories with only one assigned document and others with thousands of assigned documents. Figure 3 presents an example of a document in the collection. The *ModApte split* was used, using 75% of the articles (9603 items) for training and 25% (3299 items) for testing. Table II presents the 10 most frequent categories and the number of positive training and testing examples. These 10 categories are widely accepted as a benchmark, since 75% of the documents belong to at least one of them. In the TAT model the activation threshold of each T-cell is adjusted in a

[1] http://kdd.ics.uci.edu/databases/reuters-21578/reuters21578.html.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET"
OLDID="5552" NEWID="9">
<DATE>26-FEB-1987 15:17:11.20</DATE>
<TOPICS><D>earn</D></TOPICS>
<PLACES><D>usa</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;F
&#22;&#22;&#1;f0762&#31;reute
r f BC-CHAMPION-PRODUCTS-&lt;CH   02-26 0067</UNKNOWN>
<TEXT>&#2;
<TITLE>CHAMPION PRODUCTS &lt;CH> APPROVES STOCK SPLIT</TITLE>
<DATELINE>    ROCHESTER, N.Y., Feb 26 - </DATELINE><BODY>Champion
Products Inc said its
board of directors approved a two-for-one stock split of its
common shares for shareholders of record as of April 1, 1987.
    The company also said its board voted to recommend to
shareholders at the annual meeting April 23 an increase in the
authorized capital stock from five mln to 25 mln shares.
 Reuter
&#3;</BODY></TEXT>
</REUTERS>
```

Figure 3: Reuters-21578 document.

Category	Train	Test
Earn	2715	1044
Acq	1547	680
Interest	313	121
Money-fx	496	161
Ship	186	89
Grain	395	138
Wheat	194	66
Crude	358	176
Corn	164	52
Trade	346	113

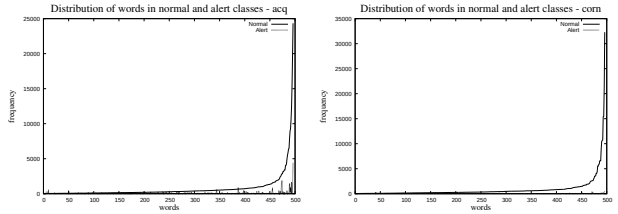
Table II: Number of positive training and testing documents for the Reuters-21578 most frequent categories.

temporal basis and its value reflects the historical iterations with the environment, measured by signal intensity. When applied to text classification, this signal intensity reflects the concentration of words in each document presented in a timely ordered data set. Thus, a data set for which we may expect a good performance with TAT should be two-fold. It has to have a comprehensive set of words that appear recurrently through time thus inducing a subset of the T-cells repertoire to become quiescently; and it also has to have another set of words that appear sporadically but with a high concentration, thus allowing a group of T-cells in the repertoire to be activated in the presence of such a received strong signal.

Figure 4 clearly illustrates the peptides distribution among the various classes of documents presented in the data set. From the ten data sets of Reuters-21578, only in the data set related to the *earn* category we are able to find a clear distinction between those two classes (Figure 4(a)). On the remaining data sets the normal behavior is dominant, in that their representative words appear on a much larger amount when compared with such representative of anomalous behavior. Figure 5 stress this fact by depicting the occurrences of each word in both classes, for all the categories.

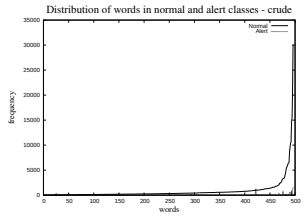
B. Performance Metrics

In order to evaluate a binary decision task we first define a contingency matrix representing the possible outcomes of the classification, namely the True Positive (TP - positive

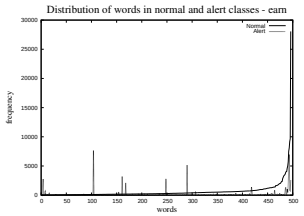


(a) Category=*acq*

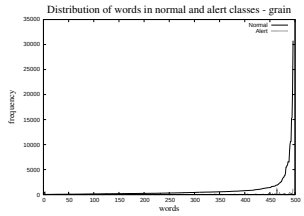
(b) Category=*corn*



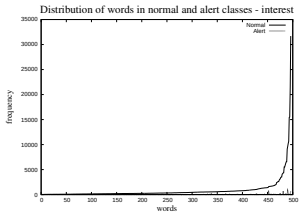
(c) Category=*crude*



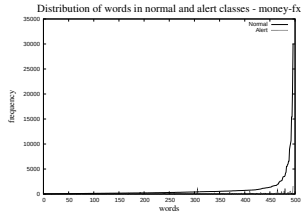
(d) Category=*earn*



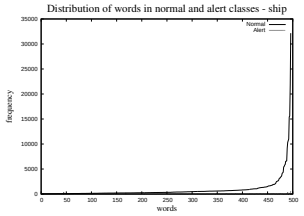
(e) Category=*grain*



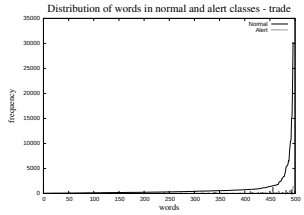
(f) Category=*interest*



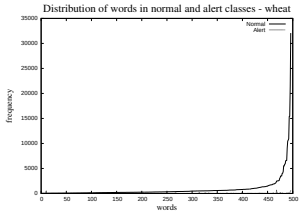
(g) Category=*money-fx*



(h) Category=*ship*



(i) Category=*trade*



(j) Category=*wheat*

Figure 4: Words distribution by class in the Reuters-21578 data set.

examples classified as positive), the True Negative (TN - negative examples classified as negative), False Positive (FP - negative examples classified as positive) and False Negative (FN - positive examples classified as negative).

Several measures have been defined based on this contingency table, such as, error rate ($\frac{FP+TN}{Tp+TN+FP+FN}$), recall ($\frac{TP}{TP+FN}$), and precision ($\frac{TP}{TP+FP}$), as well as combined measures, such as, the van Rijsbergen F_β measure, which combines recall and precision in a single score, $F_\beta = \frac{(\beta^2+1)P \times R}{\beta^2 P+R}$. The latter is one of the best suited measures for text classification used with $\beta = 1$, i.e. F_1, and thus the results reported in this paper are macro-averaged F_1 values.

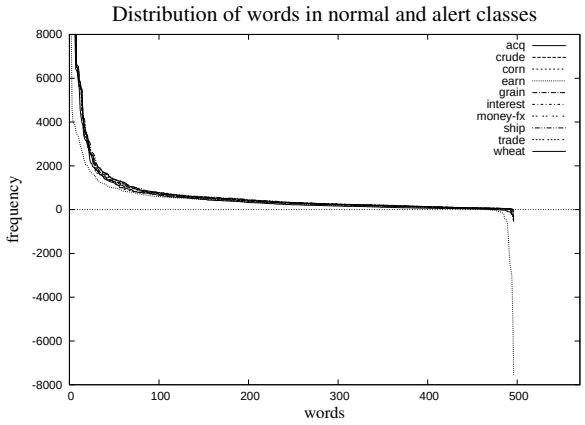


Figure 5: Concentration of words representative of normal and anomalous classes in all the categories.

C. Experimental setup

In this case, our working hypothesis is that an AIS-SVM ensemble model is able to produce a better text classification than each one isolated. According to TAT, this is achieved by a self/non-self distinction process based on the temporal historic frequencies of patterns presented in past documents. Through time, the T-cells that recognise frequent patterns become inactive and evolve to a quiescent state, while those that detect sporadic patterns within APCs with a reasonable concentration, become reactive thus initiating an immune response. We have conducted experiments with the *earn* data set using the processing parameters and criteria illustrated in the following. For SVM we also explored different parameters[2], resulting in three different learning machines:

- SVM_1: Linear default kernel
- SVM_2: Linear kernel with trade-off C, training error *vs* margin, set to 100
- SVM_3: Linear kernel with the cost-factor (by which training errors in positive examples outweigh errors in negative examples) set to 2

For TAT we used a set of fixed values for LS, Ct and Inc, together with a Latin Hypercube (LHC) sampling generator to obtain the multidimensional squares for the remaining parameters ϕ, τ and t. We then run each parameters set against the training data set, being the following those that achieved the best performance:

- TAT_1: $\phi = 0.038$; $\tau = 0.939$; $t = 0.00774$; $LS = 5$; $Ct = 0.05$; $Inc = 0.005$
- TAT_2: $\phi = 0.038$; $\tau = 0.939$; $t = 0.00774$; $LS = 15$; $Ct = 0.05$; $Inc = 0.005$
- TAT_3: $\phi = 0.031$; $\tau = 0.921$; $t = 0.00890$; $LS = 5$; $Ct = 0.05$; $Inc = 0.005$

[2]http://svmlight.joachims.org

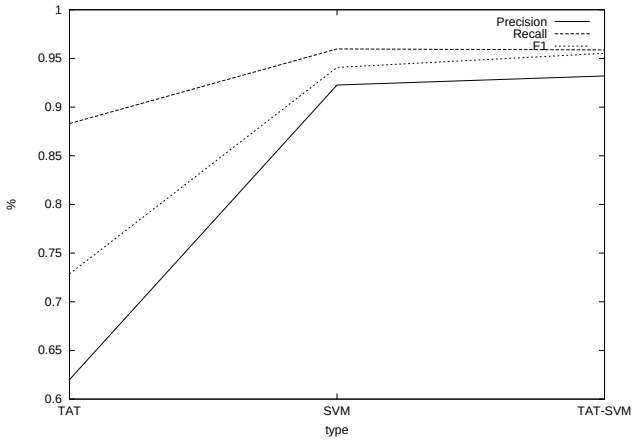


Figure 6: Performance analysis.

- TAT_3: $\phi = 0.062$; $\tau = 0.942$; $t = 0.00730$; $LS = 5$; $Ct = 0.05$; $Inc = 0.005$

D. Evaluation results

Figure 6 shows the results obtained with the AIS-SVM hybrid model described in Section II. The performances attained by each model are presented, as well as the conjugated performance obtained with the ensemble model.

From the evaluation of the experimental results we may observe an improvement of the results previously achieved by the standalone processing of the ensemble models. Although with a slight margin, the ensemble model was able to outperform the previous global results of $F1$ achieved only with the SVM processing, mainly due to the decreasing of false positives.

Despite their differences, we also observed that the union of such paradigms may bring substantial benefits to the final classification decision, by taking advantage of the individual features of each approach. From one side, SVM is currently the state-of-the art performance algorithm for text classification. On the other side, the temporal self/non-self discrimination carried out by the immune system strongly inspires the use of AIS for such dynamic environments where the meaning of self and non-self changes throughout time, like text classification and spam detection.

V. Conclusions

In this paper we proposed a hybrid approach for text classification, based on the ensemble of two distinct classification paradigms: a non adaptive machine learning SVM implementation and an immune-inspired adaptive approach based on the tunable activation thresholds of immune cells, particularly the T-cells. Although they are grounded on different learning fundamentals, both approaches individually revealed distinctive features suitable to be used in text classification. Regarding the generic TAT based AIS framework previously deployed [13], [17], [18], it was also possible to confirm its flexibility on accomplishing the Reuters-21578 training and testing data sets processing, by converting the text classification into a binary classification problem.

The preliminary results obtained thus far with this ensemble approach were very encouraging to proceed with this line of research. Further developments will be directed towards the enhancements that should be made to the preprocessing phase, since we are confident that this hybrid model may also produce satisfactory results in the classification of the other yet uncovered Reuters-21578 document classes. We also intend to apply this hybrid model to other contextual environments, for example those related to spam filtering.

References

[1] F. Sebastiani, "Classification of text, automatic," in *The Encyclopedia of Language and Linguistics*, K. B. (ed), Ed., vol. 14. Elsevier, 2006, pp. 457–462.

[2] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 1999.

[3] K. Murphy, K. Murphy, P. Travers, M. Walport, and C. Janeway, *Janeway's immunobiology*. Garland Pub, 2008.

[4] L. de Castro and J. Timmis, *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer, 2002.

[5] J. Kim, P. Bentley, U. Aickelin, J. Greensmith, G. Tedesco, and J. Twycross, "Immune system approaches to intrusion detection - a review," *Natural Computing*, vol. 6, no. 4, pp. 413–466, 2007.

[6] A. Abi-Haidar and L. Rocha, "Biomedical article classification using an agent-based model of t-cell cross-regulation," in *ICARIS*, S. L. 6209, Ed., 2010, pp. 237–249.

[7] M. Pöllä, "A generative model for self/non-self discrimination in strings," in *ICANNGA*. Springer, 2009, pp. 293–302.

[8] J. Greensmith and S. Cayzer, "An artificial immune system approach to semantic document classification," in *ICARIS*, S. L. 2787, Ed., 2003, pp. 136–146.

[9] F. Burnet, "A modification of Jerne's theory of antibody production using the concept of clonal selection," *Aust J Sci*, vol. 20, pp. 67–69, 1967.

[10] P. Matzinger, "The Danger Model: A Renewed Sense of Self," *Science's STKE*, vol. 296, no. 5566, pp. 301–305, 2002.

[11] Z. Grossman and W. Paul, "Adaptive cellular interactions in the immune system: The tunable activation threshold and the significance of subthreshold responses," *Proc.National Academy of Sciences*, vol. 89, no. 21, pp. 10 365–10 369, 1992.

[12] J. Carneiro, T. Paixão, D. Milutinovic, K. Sousa, J.and Leon, R. Gardner, and J. Faro, "Immunological self-tolerance: Lessons from mathematical modeling," *J. Computational and Applied Mathematics*, vol. 184, no. 1, pp. 77–100, 2005.

[13] M. Antunes and M. Correia, "Self tolerance by tuning t-cell activation: an artificial immune system for anomaly detection," in *Bionetics*, S. LNICST, Ed., 2010.

[14] C. Silva and B. Ribeiro, *Inductive Inference for Large Scale Text Classification: Kernel Approaches and Techniques*. Springer Verlag, 2009.

[15] B. Schölkopf, C. Burges, and A. Smola, *Advances in Kernel Methods: Support Vector Machines*. Cambridge, MIT Press, 1998.

[16] L. Kuncheva, *Combining Patt Classifiers, Methods and Algorithms*. Wiley, 2004.

[17] M. Antunes and M. Correia, "Temporal Anomaly Detection: an Artificial Immune Approach Based on T-cell Activation, Clonal Size Regulation and Homeostasis," *Advances in Computational Biology - Book series*, vol. 680, pp. 291–298, 2010.

[18] ——, "TAT-NIDS: an immune-based anomaly detection architecture for network intrusion detection," *IWPACBB'08 - Advances in Soft Computing (Springer)*, pp. 60–67, 2008.