A hybrid AIS-SVM ensemble approach
for text classification

Mário Antunes\textsuperscript{1,3}, Catarina Silva\textsuperscript{1,2}, Bernardele Ribeiro\textsuperscript{2}, and Manuel Correia\textsuperscript{3}

\textsuperscript{1} School of Technology and Management, Polytechnic Institute of Leiria; Portugal
\{mario.antunes, catarina\}@ipleiria.pt
\textsuperscript{2} Department of Informatics Engineering, Center for Informatics and Systems of the
University of Coimbra (CISUC); Portugal
\{catarina, bribeiro\}@dei.uc.pt
\textsuperscript{3} Faculty of Science, University of Porto; Center for Research in Advanced
Computing Systems (CRACS); Portugal
mcc@dcc.fc.up.pt

Abstract. In this paper we propose and analyse methods for expanding state-of-the-art performance on text classification. We put forward an ensemble-based structure that includes Support Vector Machines (SVM) and Artificial Immune Systems (AIS). The underpinning idea is that SVM-like approaches and AIS approaches, while having radically different genesis, and probably because of that, can cooperate in a committee setting, using an heterogeneous ensemble to improve overall performance, including the confidence on each system classification as the differentiating factor.

Results on the well-known Reuters-21578 benchmark are presented, showing promising classification performance gains, resulting classification that improves all baseline contributors of the ensemble committee.

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

1 Introduction

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 [3].

On the other hand, a bubbling field of research are the Artificial Immune Systems (AIS) [4]. AIS takes advantage of the Vertebrate Immune System (IS) cognitive features to defend the body from external agents (pathogens). These features are expressed by two temporal scales: one corresponding to the somatic experience of each individual throughout their life and another related to the germ-line history of the species [5]. The former assumes that the capacity to detect open-ended abnormal behavior (anomalies) has been developed by some natural selection during the evolution of the IS, tuning its innate functions of defence to optimal values, similar in all individuals of the same species. The later is related to the fact that each one of us is continuously exposed to a myriad of unseen pathogens, relying on our own IS the ability to distinguish, at each given moment in time, pathogens that belong to the organism’s own healthy cells and tissues (self), from those that may correspond to an harmful pathogen (non-self).

From the aforementioned, we may say that IS provides a rich source of inspiration for the development of innovative detection systems applied to dynamic real world environments, like network intrusion detection [6] and spam filtering [7,8]. These are clear examples in which the detection system is obliged to continuously adjust itself according to the temporal events it processes.

There are also a few examples of AIS applied to text classification [9–11]. In [11] an artificial immune system approach to semantic document classification is presented, centering the goals on semantic interpretation rather than text classification. In [9] an agent-based model to classify biomedical articles is presented, but results are still far from state-of-the-art. In [10] a statistical model is presented to detect anomalies based in self/non-self discrimination in strings.

The underpinning idea of the framework proposed in this paper is that SVM-like approaches and AIS approaches, while having radically different genesis, and probably because of that, can cooperate in a committee setting, using an heterogeneous ensemble to improve overall performance. In this paper we present a framework where SVM and AIS share data and participate as equals inputting classifications and confidence levels to provide a resulting classification that improves all baseline contributions of the ensemble committee.

The rest of the paper is organized as follows. We start by presenting in Section 2 the fundamentals of the baseline AIS and SVM learning systems. We then proceed in Section 3 by describing the proposed hybrid AIS-SVM Ensemble framework. Then, we show and discuss the results obtained on processing Reuters-21578 data set. Finally, in Section 5 we discuss the conclusions of our work and delineate some future work.

2 Background

Here we describe the fundamentals of the learning systems used throughout the later proposed framework, including AIS, SVM and committee-based learning.
2.1 An immune model inspired on tunable activation thresholds

The two most popular immunological theories that are being used on AIS deployment for anomaly detection are Negative Selection (NS) and Danger Theory (DT). Despite the promising results achieved, they proved to have documented drawbacks to deal with real world problems [6, 12]. However, more recently a new branch of immunological models have been applied for AIS deployment for anomaly detection. One of such theories is Tunable Activation Threshold (TAT), which postulates that self tolerance and non-self discrimination are made by the tunable adjustment of activation thresholds for the immune cells [13, 14].

Generally speaking, in such a model immune cells (like T-cells) tune up and update their responsiveness according to the stimuli received from the environment over the time. Each antigen undergoes a phagocytosis process which generate a set of corresponding peptides identified by a pattern representative (ligand). These peptides are presented to the T-cells repertoire by a specific kind of cells named Antigen Presenting Cell (APC). For each presented peptide, the stimulus, or signal, is going to provoke a perturbation that is measured as a function of its concentration in the APC and the affinity between its ligand and the T-cell pattern representative (T-cell Receptor (TCR)). Thus, higher the concentration of a peptide and/or its affinity with the TCR, higher the perturbation received by the T-cell.

We adopted a minimal TAT model derived from [14] in which the activation threshold of a cell is tunable by the activity of two specific enzymes that respond to antigenic signals ($S$): Kinase ($K$) and Phosphatase ($P$). Assuming $\{P_0, K_0\}$ as the basal values, in each iteration $i$, the values for $K$ and $P$ are given by the linear equations 1 and 2:

\[
\begin{align*}
K_i = \begin{cases} 
\min((S + S_0) \cdot \tau K, K_{i-1}^+ = \phi K \cdot t); & \text{if } (S + S_0) \cdot \tau K > K_{i-1} \\
\max((S + S_0) \cdot \tau K, K_{i-1}^- = \phi K \cdot t); & \text{otherwise}
\end{cases}
\end{align*}
\tag{1}
\]

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

Generally, if a T-cell receives a signal ($S > 0$), $K$ and $P$ should increase linearly until reach a turnover point ($\tau K$ and $\tau P$). The slope for $K$ and $P$, as well as the rate of growth are defined by $\phi K$, $\phi P$ and $t$ respectively. Similarly, on signaling absence, $K$ returns to the basal level at a faster rate than $P$. It is also assumed that T-cell activation is a switch-type response that requires that $K$ supersedes $P$, at least transiently. Thus, for the same signal, $K$ increases faster than $P$ ($\phi K > \phi P$), but if the signal persists $P$ will supersede $K$ and reach a higher plateau ($\tau P > \tau K$). According to the model, those auto-reactive T-cells that are continuously stimulated by self antigen end up by adapt its level of responsiveness and thus prevent from mounting an immune response. On the other side, those that are sporadically stimulated with a strong stimulus become activated and may be able to start an immune response [13].
In order to strengthen the recent temporal events a T-cell is exposed to, \( S \) is calculated as a function of the affinity between TCR and peptides ligand that exists in the APC lifespan (LS). This means that, for each T-cell, \( S \) reflects not only the signal sent by the bound peptides in the APC, but also by others such recently processed and memorised APC whose lifetime didn’t yet expire [15].

The immune response is supposed to be populational instead of being a simple consequence of the activation of just one single cell [13, 14]. Thus, in the TAT model, the classification of each APC is decided by a committee of T-cells that become active (with \( K > P \)) in each APC processing, being its threshold termed \( Ct \). This parameter starts with a predefined reasonable value and it is adjusted in run time, by a fixed value \( Inc \), according to the observed evidences.

The TAT processing was conducted in a generic and context-independent TAT simulator [15]. The text classification was adapted to this generic framework in the following way. An APC corresponds to a text document and its peptide ligands are the words on it. The T-cells repertoire correspond the list of words managed by the system that tries to bind those presented on each document. For the sake of simplicity, the affinity between strings representative of T-cells and peptides is equal to 1 if the strings are equal and zero otherwise.

### 2.2 Support Vector Machines

SVM are a learning method introduced by Vapnik [2] based on his Statistical Learning Theory and Structural Risk Minimization Principle. When using SVMs 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.

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 \( u = w \times x - b \), where \( w \) is the normal weight vector to the hyperplane and \( x \) is the input vector. Maximizing the margin can be seen as an optimization problem:

\[
\begin{align*}
\text{minimize} \quad & \frac{1}{2}||w||^2, \\
\text{subjected to} \quad & y_i(w \cdot x + b) \geq 1, \forall i,
\end{align*}
\]
where $x$ is the training example and $y_i$ is the correct output for the $i$th training example. Intuitively, the classifier with the largest margin will give low expected risk, and hence better generalization.

To deal with the constrained optimization problem in (3) Lagrange multipliers $\alpha_i \geq 0$ and the Lagrangian (4) can be introduced:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i (y_i(w \cdot x + b) - 1).$$  \hfill (4)

The Lagrangian has to be minimized with respect to the primal variables $w$ and $b$ and maximized with respect to the dual variables $\alpha_i$ (i.e. a saddle point has to be found) [16].

SVM are universal learners. In their basic form, 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.

### 2.3 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. 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 synergies between the ensemble classifiers [17].

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. When using the same learning algorithm, different classifiers are generated by manipulating the training set, manipulating the input features, manipulating the output targets or injecting randomness in the learning algorithm.

### 3 Proposed approach

This section presents the ensemble structure defined. 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 combination of their results can also be accomplished in a number of ways, like weighted average or majority voting. Having two radically different approaches to structure an ensemble framework, we defined a two-level hybrid model illustrated in Figure 1 that joins the predictions of both SVM and TAT-based models.
During the training phase the models are dealt with separately, 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). Each set then applies a majority voting strategy to define its decision, i.e. if the document is a positive or negative example of the class.

We have then two votes and, as easily concluded, hardly a decision-making ensemble. To reach a decision we compute the confidence level of each set to determine which should affect the ensemble output. Given the differences in SVM and TAT models, confidences were also calculated differently:

- SVM: The real value of the output was taken into account as a confidence level. Their absolute value was added and linearly mapped into a value between 0 and $P$, where $P$ is any integer.

- TAT: The $m$ TAT classifiers output a binary decision, thus the confidence level was defined as maximum when all models in the set agree with the same classification, and minimum when there is the most disagreement. The resulting confidence mapped between 0 and $P$, where $P$ is any integer.

Note that the value of $P$ must be the same for both sets of models. In our experiments, detailed in Section 4, we used $n = 3$, $m = 4$ and $P = 4$.

4 Experimental evaluation and results

4.1 Reuters-21578 Benchmark

The widely accepted 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 in 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. The \textit{ModApte split} was used.

<table>
<thead>
<tr>
<th>Category</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earn</td>
<td>2715</td>
<td>1044</td>
</tr>
<tr>
<td>Acq</td>
<td>1547</td>
<td>680</td>
</tr>
<tr>
<td>Money-fx</td>
<td>496</td>
<td>161</td>
</tr>
<tr>
<td>Grain</td>
<td>395</td>
<td>138</td>
</tr>
<tr>
<td>Crude</td>
<td>358</td>
<td>176</td>
</tr>
</tbody>
</table>

Table 1: Number of positive training and testing documents for the Reuters-21578 most frequent categories

using 75\% of the articles (9603 items) for training and 25\% (3299 items) for testing. Table 1 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.

4.2 Data set analysis

In TAT model the activation threshold of each T-cell is adjusted in a temporal basis and its value reflects the historical iterations with the environment, measured by a 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 with which we may expect a good performance with TAT should be two-fold. From one side, 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. On the other side, 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 strong signal.

Figure 2 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 \textit{earn} category we are able to find a clear distinction between those two classes (Figure 2(a)). On the remaining data sets the shape is similar to those shown in Figures 2(b) and 2(c). In these cases the normal behavior is dominant, in that their representative words appear on a much

\(^1\) \url{http://kdd.ics.uci.edu/databases/reuters-21578/reuters21578.html}.
larger amount when compared with such representative of anomalous behavior (class “Alert”). Figure 2(d) emphasises this fact by depicting the occurrences of each word in both classes, for all the categories.

![Distribution of words in normal and alert classes - Dataset earn](image1)

![Distribution of words in normal and alert classes - Dataset grain](image2)

![Distribution of words in normal and alert classes - Dataset wheat](image3)

![Distribution of words in normal and alert classes](image4)

Fig 2: Words distribution by class in the Reuters-21578 data set.

4.3 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 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 Rijssbergen \( F_1 \) measure [18], 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.
4.4 Results and analysis

Our working hypothesis is that a AIS-SVM ensemble model is able to produce a better classification of 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 *eurn* data set using the processing parameters and criteria illustrated in the following. For SVM we explored different parameters [19], resulting in three different learning machines:

- *SVM*$_1$: Linear default kernel
- *SVM*$_2$: Linear kernel with trade-off $C$ between training error and 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

<table>
<thead>
<tr>
<th>TAT$_1$</th>
<th>TAT$_2$</th>
<th>TAT$_3$</th>
<th>TAT$_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi = 0.038$</td>
<td>$\phi = 0.038$</td>
<td>$\phi = 0.031$</td>
<td>$\phi = 0.0619$</td>
</tr>
<tr>
<td>$\tau = 0.939$</td>
<td>$\tau = 0.939$</td>
<td>$\tau = 0.921$</td>
<td>$\tau = 0.942$</td>
</tr>
<tr>
<td>$t = 0.00774$</td>
<td>$t = 0.00774$</td>
<td>$t = 0.0089$</td>
<td>$t = 0.0073$</td>
</tr>
<tr>
<td>$LS = 5$</td>
<td>$LS = 15$</td>
<td>$LS = 5$</td>
<td>$LS = 5$</td>
</tr>
<tr>
<td>$Ct = 0.05$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Inc = 0.005$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 shows the TAT processing parameters for each model, namely $\phi$, $\tau$ and $t$, which were generated using a Latin Hypercube (LHC) sampling approach. We used an LHC generator to obtain the corresponding multidimensional squares for the parameters sets and then run each one against the training data set, thus obtaining the outperform one that will be further used in testing processing.

Table 3 shows the results obtained with the immune-SVM hybrid model described in Section 3. The performances attained by each model are presented, as well as the conjugated performance obtained with the ensemble model.

5 Conclusions

We presented a hybrid approach for text classification, based on the ensemble of two rather different classification paradigms: a non adaptive machine learning SVM implementation and an immune-inspired approach based on the tunable activation thresholds of immune cells. Although they are grounded on different
Table 3: Results obtained with immune-SVM hybrid model.

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM1</td>
<td>997</td>
<td>1728</td>
<td>69</td>
<td>47</td>
<td>93.53%</td>
<td>95.50%</td>
<td>94.06%</td>
</tr>
<tr>
<td>SVM2</td>
<td>1002</td>
<td>1713</td>
<td>84</td>
<td>42</td>
<td>92.27%</td>
<td>95.98%</td>
<td>94.08%</td>
</tr>
<tr>
<td>SVM3</td>
<td>1044</td>
<td>0</td>
<td>1797</td>
<td>0</td>
<td>36.75%</td>
<td>100%</td>
<td>53.75%</td>
</tr>
<tr>
<td>Ensemble SVM</td>
<td>1002</td>
<td>1713</td>
<td>84</td>
<td>42</td>
<td>92.27%</td>
<td>95.98%</td>
<td>94.08%</td>
</tr>
<tr>
<td>TAT1</td>
<td>898</td>
<td>1284</td>
<td>513</td>
<td>146</td>
<td>63.64%</td>
<td>86.02%</td>
<td>73.16%</td>
</tr>
<tr>
<td>TAT2</td>
<td>879</td>
<td>1275</td>
<td>522</td>
<td>165</td>
<td>62.74%</td>
<td>84.20%</td>
<td>71.90%</td>
</tr>
<tr>
<td>TAT3</td>
<td>898</td>
<td>1281</td>
<td>516</td>
<td>146</td>
<td>64.00%</td>
<td>86.69%</td>
<td>73.64%</td>
</tr>
<tr>
<td>TAT4</td>
<td>905</td>
<td>1288</td>
<td>509</td>
<td>139</td>
<td>63.51%</td>
<td>86.02%</td>
<td>73.07%</td>
</tr>
<tr>
<td>Ensemble TAT</td>
<td>922</td>
<td>1232</td>
<td>565</td>
<td>122</td>
<td>62.00%</td>
<td>88.31%</td>
<td>72.86%</td>
</tr>
</tbody>
</table>

learning fundamentals, both approaches individually revealed distinctive features suitable to be used in text classification.

From the evaluation of the experimental results we may conclude the following. Firstly, we observed 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 in such dynamic environments where the meaning of self and non-self changes through time, like text classification and spam detection. Finally, it was also possible to confirm the flexibility of the generic TAT based AIS framework deployed [15], by converting the text classification in a binary classification problem and thus being able to successfully accomplish the training and testing data sets processing.

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 rather, as those related to spam filtering.
References