A Hybrid AIS-SVM Ensemble Approach for Text Classification

Mário Antunes[1,3], Catarina Silva[1,2], Bernardete Ribeiro[2], and Manuel Correia[3]

[1] Computer Science Communication and Research Centre,
School of Technology and Management, Polytechnic Institute of Leiria, Portugal
{mario.antunes,catarina}@ipleiria.pt
[2] Department of Informatics Engineering, Center for Informatics and Systems of the University of Coimbra (CISUC), Portugal
{catarina,bribeiro}@dei.uc.pt
[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 can be enhanced with AIS approaches which can capture dynamics in models. While having radically different genesis, and probably because of that, SVM and AIS can cooperate in a committee setting, using a heterogeneous ensemble to improve overall performance, including a 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 in a classification that improves upon 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

A. Dobnikar, U. Lotrič, and B. Šter (Eds.): ICANNGA 2011, Part II, LNCS 6594, pp. 342–352, 2011.

implement the principle of structural minimization and different solutions based on committees of kernel-based machines, such as boosting.

On the other hand, a bubbling field of research are Artificial Immune Systems (AIS) [3]. 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* [4].The former is related to the fact that each one of us is continuously exposed to a myriad of *unseen* pathogens, relying on our own IS 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*). The latter assumes that the capacity to detect open-ended *abnormal behavior* (anomalies) has been developed by natural selection during the evolution of the IS, tuning its *innate* functions of defence to appropriate values, similar in all individuals of the same species.

The IS provides thus a very appealing and rich source of inspiration for the development of innovative detection systems applied to dynamic real world environments, like network intrusion detection [5] and spam filtering [6,7]. 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 some examples of AIS applied to text classification [8, 9, 10]. In [10] an artificial immune system approach to semantic document classification is presented, centering the goals on semantic interpretation rather than text classification. In [8] an agent-based model to classify biomedical articles is introduced, but results are still far from state-of-the-art. In [9] a statistical model is described to detect anomalies based in self/non-self discrimination in strings.

In this work, the underpinning idea for the proposed framework 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. SVM cutting-edge performance is enhanced with AIS capabilities of grasping dynamics in concepts present in real data sets. We introduce a framework where SVM and AIS share data and participate as equals partners, providing classifications and confidence levels to obtain a resulting classification that improves on all baseline contributors 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 terminate by delineating some future work.

2 Background

Here we describe the fundamentals of 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 thus far, they proved to have some well documented difficulties in dealing with real world problems [5]. More recently a new branch of immunological theories have been applied on new AIS deployments for anomaly detection. One of such theories is the Tunable Activation Threshold (TAT), which postulates that self tolerance and non-self discrimination are made by the tunable adjustment of immune cells activation thresholds [11, 12].

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 throughout time. Each antigen undergoes a *phagocytosis* process which generates 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 cell, named the 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, the higher the perturbation received by the T-cell. We adopted a minimal TAT model derived from [12] 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, for each time iteration i, the values for K and P are given by the linear equations 1 and 2:

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

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

Generally, if a T-cell receives a signal ($S > 0$), K and P should increase linearly until a turnover point (τK and τP) is reached. The slope for K and P, as well as the rate of growth are defined by ϕK, ϕP and t respectively. Similarly, during 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 TAT model, those auto-reactive T-cells that are continuously stimulated by self antigens end up adapting its level of responsiveness and thus preventing from mounting an immune response. On the other side, those that are sporadically stimulated with a strong stimulus become activated and start an immune response [11].

In order to strengthen the recent temporal events a T-cell as been exposed to, S is calculated as a function of the affinity between the TCR and the 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 as those that have been *recently* processed and *memorised* in the APCs whose lifetime has not yet expired [13].

The immune response is populational based, instead of being a simple consequence of the activation of just one single cell [11, 12]. Thus, in the TAT model, the *classification* of each APC is decided by a *committee* of T-cells that become active (with $K > P$) for each processed APC, with 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.

TAT behavior is reproduced by a generic and context-independent TAT simulator for anomaly detection [13]. In order to cope with the text *classification* as being an anomaly *detection* task, for each category of the Reuters-21578 data set we label the positive examples as "Alert" and the remaining as corresponding to the "normal" behaviour. In this way, a trigger should thus be raised on the presence of an example of the category we are looking for. In the text classification an APC corresponds to a text document and its peptide ligands are the words on it. The T-cells repertoire correspond to 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

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

$$\begin{aligned} &minimize \quad \frac{1}{2}||\mathbf{w}||^2, \\ &subjected\ to\ \ y_i(\mathbf{w}.\mathbf{x} + b) \geq 1, \forall i, \end{aligned} \tag{3}$$

where $\mathbf{x}$ is the training example and y_i is the correct output for the ith 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:

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

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) [14].

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. 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 [15].

3 Proposed Approach

This section presents 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 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). 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.

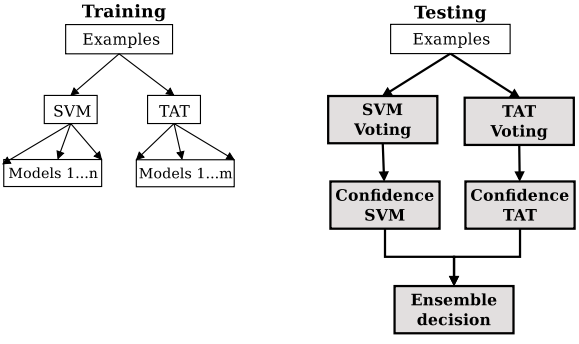


Fig. 1. TAT based and SVM hybrid model for text classification

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 a maximum confidence factor, D, of each set decision, as described in Algorithm 1.The set with higher confidence will define the output of the two-level hybrid model in Figure 1. To linearly scale the confidence, D must be same for both sets of models. In our experiments, detailed in Section 4, we used $n = 3$, $m = 4$ and $D = 4$.

Algorithm 1. Heuristic voting rule

SVM:
$sum = \sum_{i=1}^{n} SVM_i$
IF all SVM agree $base = 1$ **ELSE** $base = 0.5$
IF sum<3 $pred = 0.5$ **ELSE** $pred = 1$
$SVM\_Confidence = base * pred$
$linear\_scale(SVM\_Confidence, 0, D)$
TAT:
IF all TAT agree $TAT\_Confidence = 1$
IF maximum TAT disagree $TAT\_Confidence = 0$
IF some TAT agree $TAT\_Confidence = linear\_scale(\frac{agree}{disagree}, 0, 1)$
$linear\_scale(TAT\_Confidence, 0, D)$

4 Experimental Evaluation and Results

4.1 Reuters-21578 Benchmark

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

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

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

Category	Train	Test	Category	Train	Test
Earn	2715	1044	Trade	346	113
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

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 *ModApte split* was used, 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.

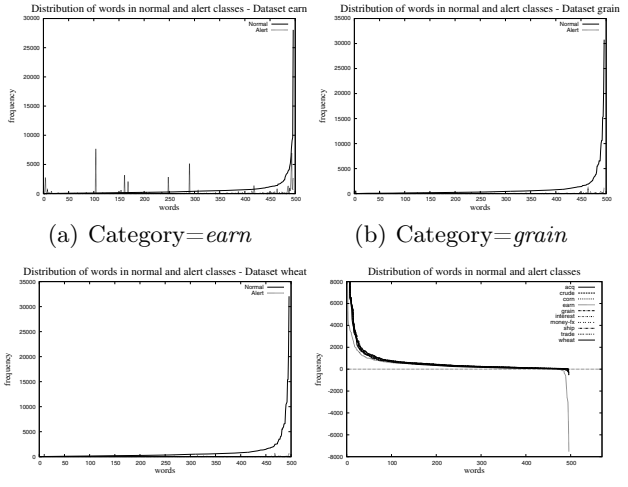
4.2 Data Set Analysis for TAT Processing

In the 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 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 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 *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 larger amount when compared with such representative of anomalous behavior (class "Alert"). Figure 2(d) stress this fact by depicting the occurrences of each word in both classes, for all the categories.

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).



(a) Category=*earn* (b) Category=*grain*

(c) Category=*wheat* (d) All categories.

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

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.

4.4 Results and Analysis

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:

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

- 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. TAT training phase has two distinct data sets. The *validation* data set that has only examples of the *earn* class and the *calibration*, which contains examples of all the classes, is used to test the parameters set suggested by the LHC sampling generator. 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: $\phi = 0.031$; $\tau = 0.921$; $t = 0.00890$; $LS = 5$; $Ct = 0.05$; $Inc = 0.005$
- TAT_4: $\phi = 0.062$; $\tau = 0.942$; $t = 0.00730$; $LS = 5$; $Ct = 0.05$; $Inc = 0.005$

Table 2 shows the results obtained with the AIS-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. 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

Table 2. Results obtained with immune-SVM hybrid model

	TP	TN	FP	FN	P	R	F1
SVM_1	997	1728	69	47	93.53%	95.50%	94.00%
SVM_2	1002	1713	84	42	92.27%	95.98%	94.08%
SVM_3	1044	0	1797	0	36.75%	100%	53.75%
Ensemble SVM	1002	1713	84	42	92.27%	95.98%	94.08%
TAT_1	898	1284	513	146	63.64%	86.02%	73.16%
TAT_2	879	1275	522	165	62.74%	84.20%	71.90%
TAT_3	898	1281	516	146	64.00%	86.69%	73.64%
TAT_4	905	1288	509	139	63.51%	86.02%	73.07%
Ensemble AIS	922	1232	565	122	62.00%	88.31%	72.86%
AIS-SVM Ensemble	1001	1724	73	43	93.20%	95.88%	95.52%

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.

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 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], 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. Sebastiani, F.: Classification of text, automatic. In: Brown, K. (ed.) The Encyclopedia of Language and Linguistics, vol. 14, pp. 457–462. Elsevier, Amsterdam (2006)
2. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1999)
3. de Castro, L., Timmis, J.: Artificial Immune Systems: A New Computational Intelligence Approach. Springer, Heidelberg (2002)
4. Cohen, I.: Tending Adam's Garden: evolving the cognitive immune self. Academic Press, San Diego (2004)
5. Kim, J., Bentley, P., Aickelin, U., Greensmith, J., Tedesco, G., Twycross, J.: Immune system approaches to intrusion detection - a review. Natural Computing 6(4), 413–466 (2007)
6. Abi-Haidar, A., Rocha, L.: Adaptive Spam Detection Inspired by the Immune System. In: Proc. of the 11th Int. Conference on the Simulation and Synthesis of Living Systems, vol. 11, pp. 1–8 (2008)
7. Oda, T., White, T.: Immunity from spam: An analysis of an artificial immune system for junk email detection. In: Jacob, C., Pilat, M.L., Bentley, P.J., Timmis, J.I. (eds.) ICARIS 2005. LNCS, vol. 3627, pp. 276–289. Springer, Heidelberg (2005)
8. Abi-Haidar, A., Rocha, L.: Biomedical article classification using an agent-based model of T-cell cross-regulation. In: Hart, E., McEwan, C., Timmis, J., Hone, A. (eds.) ICARIS 2010. LNCS, vol. 6209, pp. 237–249. Springer, Heidelberg (2010)

9. Pöllä, M.: A generative model for self/Non-self discrimination in strings. In: Kolehmainen, M., Toivanen, P., Beliczynski, B. (eds.) ICANNGA 2009. LNCS, vol. 5495, pp. 293–302. Springer, Heidelberg (2009)
10. Greensmith, J., Cayzer, S.: An artificial immune system approach to semantic document classification. In: Timmis, J., Bentley, P.J., Hart, E. (eds.) ICARIS 2003. LNCS, vol. 2787, pp. 136–146. Springer, Heidelberg (2003)
11. Grossman, Z., Paul, W.: Adaptive cellular interactions in the immune system: The tunable activation threshold and the significance of subthreshold responses. Proc. National Academy of Sciences 89(21), 10365–10369 (1992)
12. Carneiro, J., Paixão, T., Milutinovic, D., Sousa, J., Leon, K., Gardner, R., Faro, J.: Immunological self-tolerance: Lessons from mathematical modeling. J. Computational and Applied Mathematics 184(1), 77–100 (2005)
13. Antunes, M., Correia, M.: Self tolerance by tuning t-cell activation: an artificial immune system for anomaly detection. In: Lnicst, S. (ed.) Bionetics (2010)
14. Schölkopf, B., Burges, C., Smola, A.: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge (1998)
15. Kuncheva, L.: Combining Patt Classifiers, Methods and Algorithms. Wiley, Chichester (2004)