Citation: A. Abi-Haidar and L.M. Rocha [2010]."Biomedical Article Classification Using an Agent-Based Model of T-Cell Cross-Regulation". In: Artificial Immune Systems: 9th International Conference, (ICARIS 2010). E. Hart et al (Eds.) Lecture Notes in Computer Science. Springer-Verlag, **6209**: 237-249.

Biomedical Article Classification Using an Agent-Based Model of T-Cell Cross-Regulation

Alaa Abi-Haidar and Luis M. Rocha

School of Informatics and Computing, Indiana University, Bloomington IN 47401, USA
FLAD Computational Biology Collaboratorium, Instituto Gulbenkian de Ciência, Oeiras, Portugal
{aabihaid,rocha}@indiana.edu

Abstract. We propose a novel bio-inspired solution for biomedical article classification. Our method draws from an existing model of T-cell cross-regulation in the vertebrate immune system (IS), which is a complex adaptive system of millions of cells interacting to distinguish between harmless and harmful intruders. Analogously, automatic biomedical article classification assumes that the interaction and co-occurrence of thousands of words in text can be used to identify conceptually-related classes of articles—at a minimum, two classes with relevant and irrelevant articles for a given concept (e.g. articles with protein-protein interaction information). Our agent-based method for document classification expands the existing analytical model of Carneiro et al. [1], by allowing us to deal simultaneously with many distinct T-cell features (epitomes) and their collective dynamics using agent based modeling. We already extended this model to develop a bio-inspired spam-detection system [2, 3]. Here we develop our agent-base model further, and test it on a dataset of publicly available full-text biomedical articles provided by the BioCreative challenge [4]. We study several new parameter configurations leading to encouraging results comparable to state-of-the-art classifiers. These results help us understand both T-cell cross-regulation and its applicability to document classification in general. Therefore, we show that our bio-inspired algorithm is a promising novel method for biomedical article classification and for binary document classification in general.

Keywords: Artificial Immune System, Bio-medical Document Classification, T-cell Cross-Regulation, Bio-inspired Computing, Artificial Intelligence, BioCreative.

1 Introduction

With faster genome sequencing [5] and microarray analysis [6], the last decade has witnessed an exponential growth of metabolic, genomic and proteomic documents (articles) being published [7]. Pubmed [8] encompasses a growing collection of more than 18 million biomedical articles. Manually classifying these articles as relevant or irrelevant to a given topic of interest is very time consuming and inefficient for curation of new published articles [9]. A few conferences

E. Hart et al. (Eds.): ICARIS 2010, LNCS 6209, pp. 237–249, 2010.

have been dedicated to literature or text mining offering challenges to address biomedical document classification. BioCreative is a community-wide effort for assessing bio-literature mining [4] . Machine Learning has offered a plethora of solutions to this problem [9, 10], but even the most sophisticated of solutions often overfit to the training data and do not perform as well on real-world data such as that provided by BioCreative—in this case, articles for curation selected from FEBS Letters.[11–13].

The immune system is a complex biological system made of millions of cells all interacting to distinguish between harmless and harmful intruders, to ultimately attack the latter [14]. In analogy, relevant biomedical articles for a given concept need to be distinguished from irrelevant ones which should be discarded in topical queries. To employ computational intelligence to automatically implement this topical classification, we can use the occurrence and co-occurrence of thousands of words in a document describing an approach, an experiment, a result or a conclusion. In this sense, words can be seen as interacting in a text in such a way as to allow us to distinguish between relevant and irrelevant documents. Recent advances in artificial immune systems [15] have offered a few immune-inspired solutions to document classification in general, though none to our knowledge has been applied to biomedical article classification. Our aim is not to explore the applicability of existing immune inspired solutions on biomedical article classification [16], but to propose a new solution and compare it with state-of-art classifiers.

We extend an existing model of T-cell cross-regulation [1] to deal with multiple features simultaneously using agent based modeling. We applied a first version of our agent-based model to a similar document classification problem dealing with spam detection. On that task, we obtained encouraging results, which were comparable to state-of-art text classifiers [2, 3]. However, our preliminary implementation did not explore all parameter configurations such as T-cell death rates, different training scenarios, and lacked extensive parameter search for optimized performance [2, 3]. In the work reported here, we test variations of our agent-based model to understand the effect of T-cell death and of training exclusively on relevant articles. We also test our agent-based model on full-text biomedical data from BioCreative and compare it with state-of-art classifiers to understand the model's applicability to real-World biomedical classification specifically, and to document classification in general. This more extensive study allows us to establish the capability of T-cell cross-regulation dynamics to classify data. It also leads to a competitive, novel bio-inspired text classification algorithm.

In section 2, we describe the original T-cell cross-regulation model [1]. In section 3, we describe the expanded agent-based implementation of the cross-regulation model and explain its parameters. In section 4, we discuss the biomedical data from BioCreative and the feature selection process. In section 5, we report our results on biomedical document classification and compare them to those obtained by Naive Bayes [17] and SVM [18].

2 The Cross-Regulation Model

The T-cell Cross-Regulation Model (CRM) [1] is a dynamical system that aims to distinguish between harmless and harmful protein fragments (antigens) using only four possible interactions of three cell-types: Effector T-cells (E), Regulatory T-cells (R) and Antigen Presenting Cells (APC). As their name suggests, APC present antigens for the other two cell-types, E and R, to recognize and bind to them. Effector cells (E) proliferate upon binding to APC, unless adjacent to regulatory cells (R), which regulate E by inhibiting their proliferation. For simplicity, proliferation of cells is limited to duplication in quantity in contrast to having a proliferation rate. T-cells that do not bind to APC die off with a certain death rate. The four possible interactions, illustrated in Fig. 1, can be simply expressed by the following equations:

$$E \underset{d_E}{\rightarrow} \{\} \text{ and } R \underset{d_R}{\rightarrow} \{\} \tag{1}$$

$$A + R \rightarrow A + R \tag{2}$$

$$A + E \rightarrow A + 2E \tag{3}$$

$$A + E + R \rightarrow A + E + 2R \tag{4}$$

The first equation (1) expresses E and R cell death with the corresponding death rates d_E and d_R. The last three proliferation equations express (2) the maintenance of R, (3) the duplication of E, and (4) the maintenance of E and duplication of R.

Carneiro et al. [1] developed the analytical CRM to study the dynamics of a population of T-cells and APC that recognize a single antigen. In [3, 2], we adapted the original CRM model to deal with multiple populations of textual features using agent-based modeling. Our basic implementation of the model yielded encouraging results when applied to spam detection, a binary document classification problem. More recently, Sepulveda [21, pp 111-113] extended the original CRM to study multiple populations of T-cells that can be recognized by APC, each capable of recognizing at most two distinct T-cell populations. In our preliminary model [3, 2], we have used APC that are capable of recognizing hundreds of T-cells of different populations, simultaneously, using the same four interaction rules of the CRM. In the following section, we explain in more details our agent-based model adapted for document classification.

3 The Agent Based Cross-Regulation Model

In order to adapt CRM to an Agent-Based Cross-Regulation Model (ABCRM) for text classification, one has to think of documents as analogous to the organic substances that upon entering the body are broken into constituent pieces. These pieces, known as epitopes, are presented on the surface of Antigen Presenting Cells (APC) as antigens. In the ABCRM, antigens are textual features (e.g. words, bigrams, titles, numbers) extracted from articles and presented by artificial APC such that they can be recognized by a number of artificial Effector

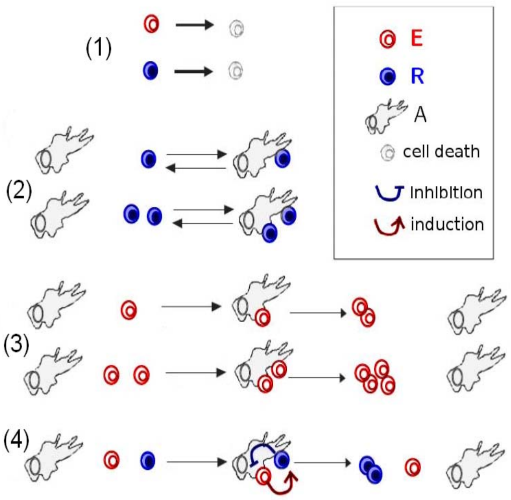

Fig. 1. The diagram illustrates the CRM interactions underlying the dynamics of APC, E and R as assumed in the model where APC can only form conjugates with a maximum of two T-cells

T-cells (E) and artificial Regulatory T-cells (R). In other words, individual E and R have receptors for a single, specific textual feature: they are *monospecific*. E proliferate upon binding to antigens presented by APC unless suppressed by R; R suppress E when binding in adjacent locations on APC. Individual APC present various document features: they are *polyspecific*. Each APC cell is produced when documents enter the cellular dynamics, by breaking the latter into constituent textual features. Therefore we can say that APC are representative of specific documents whereas E and R are representative of specific features.

A document d contains a set of features F_d; An artificial APC A_d that represents d, presents antigens/features $f_i \in F_d$ to artificial E and R T-cells. E_i and R_i bind to a specific feature f_i on **any** APC that contains it; if $f_i \in F_d$, then either E_i or R_i may bind to A_d as illustrated in figure 2. In biology, antigen recognition is a more complex process than mere polypeptide sequence matching but for simplicity we limit our feature recognition to string matching. Once T-cells bind to an APC A_d, every pair of adjacent T-cells on A_d proliferates according to the last three interaction rules of equations (2-4). APC are organized as a sequence of pairs of "slots" of textual features, where T-Cells, specific for those features, can bind. We use this simplified antigen/feature presentation scheme of pairs of "slots" to simplify our algorithm. In future work we will study alternative feature presentation scenarios. In summary, each T-cell population is specific to and can bind to only one feature presented by APC. Implementing the algorithm as an Agent-based model (ABM) allows us to deal with recognition of many features simultaneously, rather than a single one as the original mathematical model does.

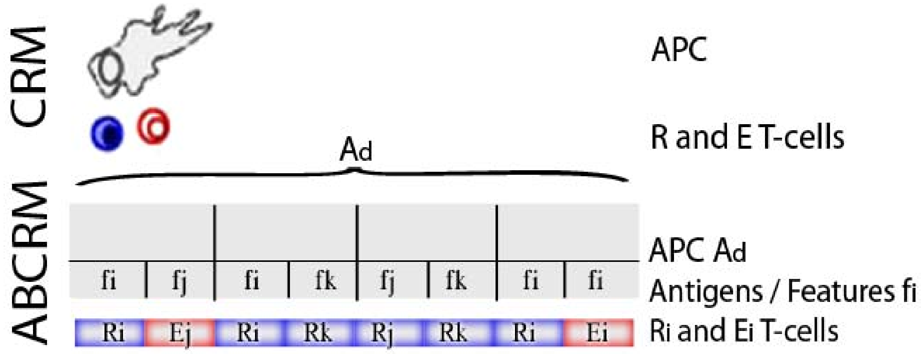

Fig. 2. To illustrate the difference between the CRM and the ABCRM, the top part of the figure represents a single APC of the CRM which can bind to a maximum of two T-Cells. The lower part represents the APC for a document d in the ABCRM, which contains many pairs of antigen/feature "slots" where pairs of T-cells can bind. In this example, the first pair of slots of the APC A_d presents the features f_i and f_j; in this case, a regulatory T-cell R_i and an effector T-cell E_j bind to these slots, which will therefore interact according to reaction (4)—R_i inhibits E_j and in turn proliferates by doubling. The next pair of slots leads to the interaction of T-cells R_i,R_k, etc.

The ABCRM uses incremental learning to first train on N labeled documents (relevant and irrelevant), which are ordered sequentially (typically by time signature) and then test on M unlabeled documents that follow in time order. The sequence of articles is assumed to be of importance to our model [2] but is outside the scope of this study. Carneiro et al. [1] show that both E and R T-cells co-exist in healthy individuals assuming enough APC exists. R T-cells require adequate amounts of E T-cells to proliferate, but not too many that can out-compete R for the specific features presented by APC. "Healthy" T-cell dynamics is identified by observing the co-existence of both E and R features with $R \geq E$. "Unhealthy" T-cell dynamics is identified by observing $E \gg R$, and should result when encountering many irrelevant features in a document. In other words, features associated with relevant documents should have E and R T-cell representatives in comparable numbers in the artificial cellular dynamics (with slightly more R). In contrast, features associated with irrelevant documents should have many more E than R T-cells. Therefore, when a document d contains features F_d, that bind mostly to E rather than R cells, we can classify it as irrelevant—and relevant in the opposite situation.

The ABCRM is controlled by 6 parameters:

- E_0 is the initial number of Effector T-cells generated for all new features
- R_0^- is the initial number of Regulatory T-cells generated for all new features in irrelevant and unlabeled documents
- R_0^+ is the initial number of Regulatory T-cells generated for all new features in relevant documents
- d_E is the death rate for Effector T-cells that do not bind to APC

- d_R is the death rate for Regulatory T-cells that do not bind to APC
- n_A is the number of slots in which each feature f_i is presented on an APC

When the features of a document d are encountered for the first time, a fixed initial number of E_0 and R_0 , for every new feature f_i, is generated. These initial values of T-cells vary for relevant and irrelevant documents in training and in testing stages. More Regulatory T-cells (R_0^+) than Effector T-cells are generated for features that occur for the first time in documents that are labeled relevant in the training stage ($R_0^+ > E_0$), while fewer Regulatory T-Cells (R_o^-) than Effector T-cells are generated in the case of irrelevant documents ($R_0^- < E_0$). Features appearing in unlabeled documents for the first time during the testing stage are treated as features from irrelevant documents, assuming that new features are foreign until neutralized by co-occurrence with relevant ones. Of course, relevant features might occur in irrelevant documents and *vice versa*. However, the assumption is that relevant features tend to co-occur more frequently with other relevant features in relevant documents and similarly for irrelevant features thus correcting the erroneous initial bias. The following pseudocode **highlights** the minor differences between the training and validation/testing stages of the algorithm:

```
TRAINING:
$\forall d$ generate $A_d$ presenting each $f_i$ at $n_A$ slots, where $f_i \in F_d$
  Let $C_t$ be the set of all $E_k$ and $R_k$ for all features $f_k$ in the cellular dynamics
  $\forall f_i \in F_d$, if $E_i \notin C_t$ and $R_i \notin C_t$ then,
    $E_i = E_0$ (generate $E_0$ Effector T-cells for feature $f_i$)
    if $d$ is labeled relevant
      $R_i = R_0^+$ (generate $R_0^+$ Regulatory T-Cells for feature $f_i$)
    otherwise
      $R_i = R_0^-$ (generate $R_0^-$ Regulatory T-Cells for feature $f_i$)
  Let $E_i$ and $R_i$ bind specifically to matching $f_i$ presented on $A_d$:
  $\forall$ pair of adjacent $(f_i, f_j)$ on $A_d$ apply the last three interaction rules:

      Ri+Rj+Ad->Ri+Rj
      Ei+Ej+Ad->2Ei+2Ej
      Ei+Rj+Ad->Ei+2Rj

$\forall R_i$ and $E_i$ that bind to $A_d$, update total number of $E_i$ and $R_i$
$\forall R_k, E_k \in C_t$ that do not bind to $A_d$, cull $E_k$ and $R_k$ via death rates $d_E$ and $d_R$
```

```
TESTING:
$\forall d$ generate $A_d$ presenting each $f_i$ at $n_A$ slots, where $f_i \in F_d$
  Let $C_t$ be the set of all $E_k$ and $R_k$ for features $f_k$ in the cellular dynamics
  $\forall f_i \in F_d$, if $E_i \notin C_t$ and $R_i \notin C_t$ then,
    $E_i = E_0$ (generate $E_0$ Effector T-cells for feature $f_i$)
    $R_i = R_0^-$ (generate $R_0^-$ Regulatory T-Cells for feature $f_i$)
  Let all the $E_i$ and $R_i$ bind specifically to matching $f_i$ presented on $A_d$:
  $\forall$ pair of adjacent $(f_i, f_j)$ on $A_d$ apply the last three interaction rules:
        Ri+Rj+Ad->Ri+Rj
        Ei+Ej+Ad->2Ei+2Ej
        Ei+Rj+Ad->Ei+2Rj
$\forall R_i$ and $E_i$ that bind to $A_d$, update total number of $E_i$ and $R_i$
        and compute for $d$: $R(d) = \sum_{\forall f_i \in F_d}(R_i)$ and $E(d) = \sum_{\forall f_i \in F_d}(E_i)$
    Hence the normalized score for $d$ is $S(d) = (R(d) - E(d))/\sqrt{R^2(d) + E^2(d)}$
    If $S(d) > 0$ then classify $d$ as relevant, otherwise irrelevant
$\forall E_k, R_k \in C_t$ that do not bind to $A_d$, cull $E_k$ and $R_k$ via death rates $d_E$ and $d_R$
```

According to the original CRM model [1], T-cells that do not bind to a presented antigen die at a certain death rate determined by d_E and d_R. Cell death is supposed to help the algorithm forget old features and focus on more recently encountered ones. Cell death was not fully explored in our previous application of this model [2, 3] and therefore in section 5 we test the effect of cell death in the dynamics of the ABCRM.

Negative selection in the adaptive immune system is thought to help discrimination between harmless and harmful antigens by eliminating immature Effector T-cells that bind to harmless or self antigens in the thymus—thus helping to prevent auto-immunity. Mature Effector T-cells that did not bind to harmless antigens are released from the thymus to recognize harmful antigens [14]. Therefore, Effector T-cells are trained to discriminate between harmless and harmful antigens avoiding autoimmunity, by preliminary "training" on harmless or self antigens. In the context of machine learning, this is known as positive unlabeled (PU) training, which we test here against training on both relevant (positive) and irrelevant (negative) documents.

4 Data and Feature Selection

The BioCreative (BC) challenge aims to assess state-of-art in bio-literature mining— in particular, biomedical document classification. More recently, the article classification task of BC2.5 [4] was based on a training data set comprised of 61 full-text articles relevant to protein-protein interaction (P_T) and 558 irrelevant ones (N_T). This imbalance between the relevant and irrelevant instances can be very challenging. In order to assess our bio-inspired algorithm as a biomedical text classifier, we first identify optimal parameters on samples of training that are balanced in the numbers of relevant and irrelevant documents since we cannot predict if the validation data will be imbalanced. We assume that the adaptive nature of our algorithm, will adapt well to unpredictable imbalance by adjusting the proportions between the populations of E and R T-cells automatically. For the purpose of identifying optimal parameters, we chose the first 60 relevant and sampled 60 irrelevant articles that were published around the same date (uniform distribution between Jan and Dec 2008) as illustrated in figure 3.

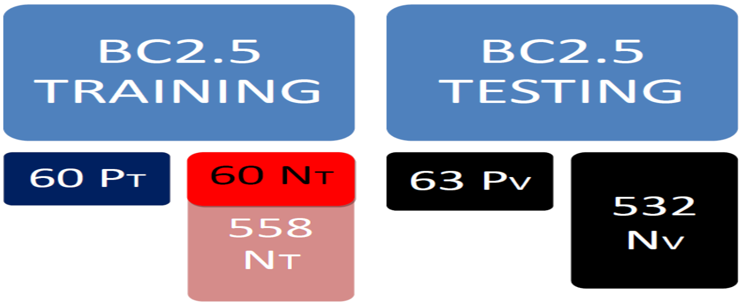

Fig. 3. Numbers of relevant (P) and irrelevant (N) documents in the training (T) and testing (V) data sets of the Biocreative 2.5 challenge. In the parameter search stage, we use a balanced set of 60 P_T (blue) and 60 N_T (red) randomly selected articles from the training data set. In the testing stage we use the unbalanced validation set containing 63 P_V (black) and 532 N_V (black) documents. Notice that the validation data was provided to the participants in the classification task of Biocreative 2.5 unlabeled, therefore participants had no prior knowledge of class proportions.

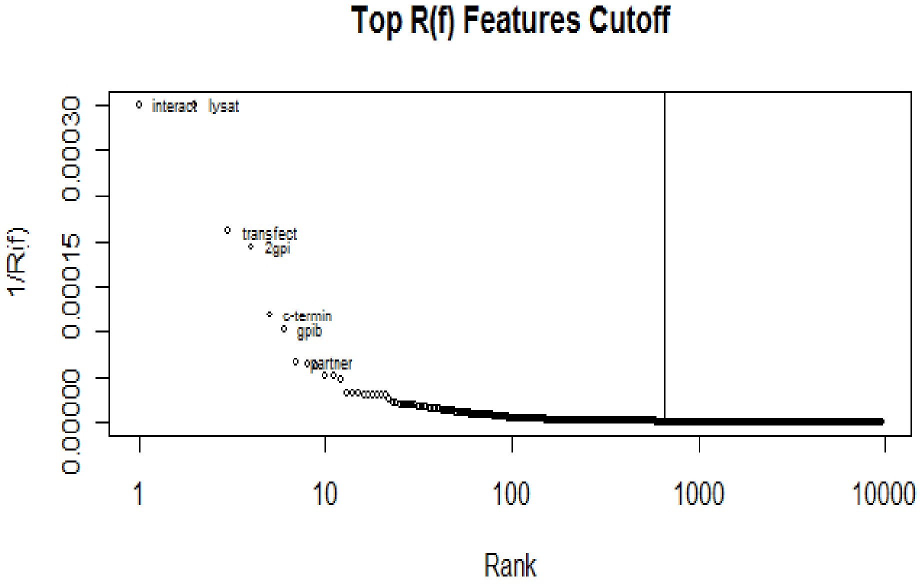

Fig. 4. We choose the top 650 ranked features according to the rank product R(f) = TF.IDF(f) × S(f). The y-axis represents $\frac{1}{R(f)}$ and the x-axis represents the index of R(f) for the sorted features. Features ranked below the 650th feature have a similar score $\frac{1}{R(f)} < 0.00001$.

We compared our fine-tuned algorithm with Naive Bayes classifier (NB) [17] and support vector machine (SVM) [18]. For testing and validation we used the Biocreative 2.5 testing data set consisting of 63 full-text articles relevant to protein- protein interaction (P_V) and 532 irrelevant ones (N_V) as shown in figure 3.

We pre-processed all articles by filtering out stop words[1] and porter stemming [22] the remaining words/features. We then ranked features f extracted from BC2.5 training articles in addition to BC2[2], according to two scoring methods. The first one is the average TF.IDF[3] per feature over all documents [10] and the second one is according to the separation score $S(f) = |p_{Relevance}(f) - p_{Irrelevance}(f)|$ where $p_{relevance}$ is the probability of a feature occurring in a relevant article and $p_{irrelevance}$ is the probability of it occurring in an irrelevant one [19, 11, 20, 13]. The final rank for every feature f_i is defined by the rank product R(f) = TF.IDF(f) × S(f). We only use the top 650 ranked features as shown in figure 4 to represent each document d as a vector of these top 650 features for optimization purposes.

5 Results

5.1 Parameter Search

We performed an exhaustive parameter search by training the ABCRM on 60 balanced full-text articles (30 P_T and 30 N_T from BC2.5 training) and testing it on the remaining 60 balanced ones (30 P_T and 30 N_T from BC2.5 training) as illustrated in figure 3. Each run corresponds to a unique configuration of 6 parameters. The explored parameter ranges are listed in table 1 and they sum up to a total of 192500 unique parameter configurations for each experiment. In the case of the PU learning experiment, we only trained on 30 relevant articles (30 P_T from BC2.5 training) and tested on 60 balanced ones (30 P_T and 30 N_T from BC2.5 training) as illustrated in figure 3. Finally, the parameter configurations were sorted with respect to the resulting F-scores[4] and the top 6 results are reported in table 2 for the four possible outcomes of two different experiments. The F-score is a fair measure between precision and recall when applied to balanced data [23]. Therefore, we use it to evaluate the performance of the ABCRM for all parameter configurations of each of the following two experiments: comparing a range of T-cell death rated to no cell death and comparing training on P_T and N_T with PU learning.

[1] The list of stop words includes 33 of the most common English words from which we manually excluded the word "with", as we know it to be of importance to protein interaction.

[2] The BC2 challenge offered Pubmed abstracts for the classification task. We downloaded some of the full-articles and used a balanced data set of 558 relevant and 558 irrelevant only for the feature selection process. We also used these features in [20, 13].

[3] TF.IDF is a common feature weighting measure to evaluate the importance of a feature/word to a document in a certain corpus. TF stands for term frequency and IDF for inverse document frequency. [10]

[4] F-score = $\frac{2.Precision.Recall}{Precision+Recall}$ where Precision = $\frac{TP}{TP+FP}$ and Recall = $\frac{TP}{TP+FN}$. True Positives (TP) and False Positives (FP) are our positive predictions while True Negatives (TN) and False Negatives (FN) are our negative predictions.

Table 1. The parameter ranges used for the parameter search for fine-tuning the algorithm

Parameter	Range	Step
E_0	[1,7]	1
R_0^-	[3,12]	1
R_0^+	[3,12]	1
d_E	[0.0,0.4]	0.1
d_R	[0.0,0.4]	0.1
n_A	[2,22]	2

In the **first** experiment we compare the top 50 parameter configurations obtained using cell death to those with no cell death. We choose only the top 50 configurations to study the algorithm at its best performance that is robust to parameter changes. We conclude that cell death, which helps in the forgetting of useless features, improves the classification performance of the algorithm regardless of whether the algorithm is trained on both P_T and N_T or not.

In the **second** experiment we compare the top 50 parameter configurations according to F-score obtained using training on both positive and negative to those obtained using training on positive only (PU learning). We conclude that training on both classes gives a better overall performance regardless of cell death.

Table 2. Top 6 parameter configurations of the ABCRM in terms of F-score. The top 50 parameter configurations of the four possible outcomes are plotted in figure 5. The highlighted parameter configuration has the highest F-score and is selected for the ABCRM to test on a different set of unbalanced articles in the following subsection.

	CELL DEATH		NO CELL DEATH	
	F-score	$[E_0,R_0^-,R_0^+,d_R,d_E,n_A]$	F-score	$[E_0,R_0^-,R_0^+,d_R,d_E,n_A]$
TRAINING on P_T and N_T	**0.85**	**[2 11 10 0.3 0.2 18]**	0.83	[1 4 7 0.0 0.0 18]
	0.84	[1 11 10 0.3 0.1 22]	0.81	[1 4 6 0.0 0.0 16]
	0.84	[1 8 6 0.1 0.1 22]	0.78	[5 7 6 0.0 0.0 10]
	0.84	[1 12 6 0.4 0.1 22]	0.78	[2 5 6 0.0 0.0 16]
	0.83	[1 9 8 0.3 0.2 22]	0.77	[2 7 5 0.0 0.0 16]
	0.83	[1 8 7 0.1 0.1 22]	0.77	[1 3 3 0.0 0.0 8]
TRAINING on P_T	0.85	[1 12 8 0.1 0.0 8]	0.75	[2 12 6 0.0 0.0 18]
	0.84	[1 8 8 0.3 0.2 16]	0.75	[2 9 6 0.0 0.0 18]
	0.82	[1 12 9 0.1 0.0 8]	0.75	[2 8 6 0.0 0.0 18]
	0.81	[1 7 10 0.2 0.1 16]	0.75	[2 11 6 0.0 0.0 18]
	0.81	[1 11 12 0.4 0.1 18]	0.74	[2 10 6 0.0 0.0 18]
	0.80	[3 7 10 0.2 0.3 18]	0.73	[2 6 6 0.0 0.0 18]

We confirm our comparisons statistically using the paired student t-test with the null hypothesis being that the two samples were drawn from the same distribution. We reject the null hypothesis for p-values less than 0.01. The top 6 configurations are listed with their corresponding F-score measure in table 2 and the 50 top-ranked configurations of each of the experiments are plotted in figure 5:

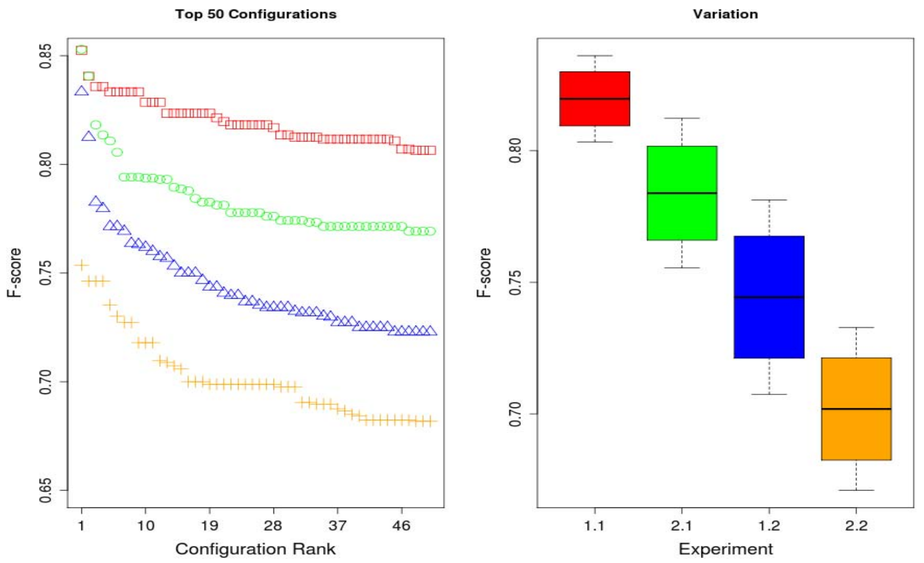

Fig. 5. The two experiments resulting in four possible outcomes: 1.1) training on both sets with cell death (red squares), 2.1) PU learning with cell death (green circles), 1.2) training on both sets with no cell death (blue triangles) and PU learning with no cell death (orange pluses) are clearly distinguishable for the top 50 configurations of each experiment. On the right, the horizontal lines represent the mean, the boxes represent 95%CI, and the whiskers represent standard deviation of F-scores from the top 50 parameter configurations.

5.2 Classification Performance

We finally adopt the parameter configuration from the experiment resulting with the best F-score (highlighted in table 2) and test our algorithm on a larger set of imbalanced full-text articles obtained from BC2.5 as illustrated in figure 3. We then compare our algorithm with the multinomial Naive Bayes (NB) with boolean attributes, explained in [17], and the publicly available SVM^{light} implementation of support vector machine applied to normalized feature counts [18]. All classifiers were tested on the same features obtained from the same data set.

The F-score metric is not very reliable for evaluating imbalanced classification [23], therefore we also use the Area Under the interpolated precision and recall Curve (AUC) to evaluate the performance of the algorithms on the imbalanced BC2.5 testing data. The AUC was the preferred performance measure of the Biocreative 2.5 challenge [4]. Table 3 lists the results in contrast to the central tendancy of the results submitted by all Biocreative 2.5 teams participating in the article classification task. However, the ABCRM, NB, and SVM classifiers, used only single-word features in order to establish the feasibility of the method, while most classifiers submitted to the Biocreative 2.5 challenge used more sophisticated features such as n-grams. Therefore, it is not surprising that the

Table 3. F-Score and AUC performance of various classifiers when training on the balanced training set of articles and testing on the full unbalanced Biocreative 2.5 testing set. Also shown is the mean performance values for all systems submitted to Biocreative 2.5.

	ABCRM	NB	SVM	BC2.5 Mean
Precision	0.22	0.14	0.24	0.38
Recall	0.65	0.71	0.94	0.68
F-score	0.33	0.24	0.36	0.39
AUC	0.34	0.19	0.46	0.43

performance of these methods was below the average. Nevertheless, when we compare the performance of the ABCRM to NB and SVM on the exact same single-words, the results are encouraging. Hence, we establish the ABCRM as a new bio-inspired text classifier to be further improved in the future with more sophisticated features.

6 Conclusion

We adapted a simple and novel mathematical model of T-cell cross-regulation in the adaptive immune system to recognize multiple textual features and classify biomedical articles using agent based modeling. We tested several variations of our algorithm to classify full-text articles according to their relevance to protein interaction. We obtained encouraging results comparable to state-of-art text classifiers. In summary, we have shown that our novel bio-inspired algorithm is promising for biomedical article classification, and for binary document classification in general.

References

1. Carneiro, J., Leon, K., Caramalho, Í., van den Dool, C., Gardner, R., Oliveira, V., Bergman, M., Sepúlveda, N., Paixão, T., Faro, J., et al.: When three is not a crowd: a Crossregulation Model of the dynamics and repertoire selection of regulatory CD4 T cells. Immunological Reviews 216(1), 48–68 (2007)
2. Abi-Haidar, A., Rocha, L.: Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift. In: Bentley, P.J., Lee, D., Jung, S. (eds.) ICARIS 2008. LNCS, vol. 5132, p. 36. Springer, Heidelberg (2008)
3. Abi-Haidar, A., Rocha, L.: Adaptive spam detection inspired by the immune system. In: Bullock, S., Noble, J., Watson, R., Bedau, M.A. (eds.) Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems, pp. 1–8. MIT Press, Cambridge (2008)
4. Krallinger, M., et al.: The BioCreative II. 5 challenge overview. In: Proc. the BioCreative II. 5 Workshop 2009 on Digital Annotations, pp. 7–9 (2009)
5. Myers, G.: Whole-genome DNA sequencing. Computing in Science & Engineering [see also IEEE Computational Science and Engineering] 1(3), 33–43 (1999)

6. Schena, M., Shalon, D., Davis, R., Brown, P., et al.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (Washington) 270(5235), 467–470 (1995)
7. Hunter, L., Cohen, K.: Biomedical Language Processing: What's Beyond PubMed? Molecular Cell 21(5), 589–594 (2006)
8. Pubmed
9. Jensen, L.J., Saric, J., Bork, P.: Literature mining for the biologist: from information retrieval to biological discovery. Nat. Rev. Genet. 7(2), 119–129 (2006)
10. Feldman, R., Sanger, J.: The Text Mining Handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, Cambridge (2006)
11. Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Rechtsteiner, A., Verspoor, K., Wang, Z., Rocha, L.: Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. Genome Biology 9(2), S11 (2008)
12. Krallinger, M., Valencia, A.: Evaluating the detection and ranking of protein interaction relevant articles: the BioCreative challenge interaction article sub-task (IAS). In: Proceedings of the Second Biocreative Challenge Evaluation Workshop (2007)
13. Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A., Rocha, L.: Classication of protein-protein interaction documents using text and citation network features (in press)
14. Hofmeyr, S.: An Interpretative Introduction to the Immune System. In: Design Principles for the Immune System and Other Distributed Autonomous Systems (2001)
15. Timmis, J.: Artificial immune systems today and tomorrow. Natural Computing 6(1), 1–18 (2007)
16. Twycross, J., Cayzer, S.: An immune system approach to document classification. Master's thesis, COGS, University of Sussex, UK (2002)
17. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam Filtering with Naive Bayes–Which Naive Bayes? In: Third Conference on Email and Anti-Spam, CEAS (2006)
18. Joachims, T.: Learning to classify text using support vector machines: methods, theory, and algorithms. Kluwer Academic Publishers, Dordrecht (2002)
19. Abi-Haidar, A., Kaur, J., Maguitman, A., Radivojac, P., Retchsteiner, A., Verspoor, K., Wang, Z., Rocha, L.: Uncovering protein-protein interactions in the bibliome. In: Proceedings of the Second BioCreative Challenge Evaluation Workshop, pp. 247–255 (2007) ISBN 84-933255-6-2
20. Kolchinsky, A., Abi-Haidar, A., Kaur, J., Hamed, A., Rocha, L.: Classification of protein-protein interaction documents using text and citation network features. In: BioCreative II.5 Workshop 2009: Special Session on Digital Annotations, Madrid, Spain, October 7-9, p. 34 (2009)
21. de Sepulveda, N.H.S.: How is the t-cell repertoire shaped (2009)
22. Porter, M.: An algorithm for suffix stripping. In: Program 1966-2006: Celebrating 40 Years of ICT in Libraries, Museums and Archives (2006)
23. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: Sattar, A., Kang, B.-h. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 1015–1021. Springer, Heidelberg (2006)