AN ADAPTIVE DOCUMENT CLASSIFIER INSPIRED BY T-CELL CROSS-REGULATION IN THE IMMUNE SYSTEM

Alaa Abi Haidar

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in Informatics

Indiana University

May 2011

Accepted by the Graduate Faculty, Indiana University, Bloomington, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Luis M. Rocha
(Principal Advisor)

Doctoral

Committee

Filippo Menczer

Alessandro Flammini

April 29 2011.

Predrag Radivojac

To my predecessors and successors in this stream of life

Acknowledgements

This dissertation was partially funded by grants from the School of Informatics and Computing at Indiana University and the FLAD collaboratorium at the Instituto Gulbenkian de Ciência, Oeiras, Portugal.

Albeit I am the only author of this dissertation, many people have contributed to its successful completion vis-a-vis my unforgettable graduate experience.

My foremost and deepest gratitude goes to Luis M. Rocha. I have been very fortunate to have him as an advisor who encouraged me to explore my own interests but never hesitated to help whenever I struggled. His flexibility and generosity extended beyond academia and offered me a plethora of opportunities, including several estival visits to the Instituto Gulbenkian de Ciência in Portugal, ones which resulted in many fruitful collaborations—needless to mention Portugal's ineffable beauty, delicious food, and heartwarming people. If you find the dissertation sound and clear, it is thanks to my advisor's patience; otherwise, it is for the lack of mine.

Alessandro Flammini, has always been there to listen and offer advice. I am deeply grateful for his readiness to help.

Filippo Menczer and Predrag Radivojac have invited me several times to present my work to their research groups and provided me with constructive criticism that cemented my thoughts and helped me sculpt my dissertation. I am truly grateful to their indispensable feedback.

Jorge Carneiro has motivated a huge deal of this work. I am grateful for all the inspirational discussions and insights about the Immune system and the cross-regulation model, without which none of this would have been possible.

My co-advisee (and fellow sufferer), Artemy Kolchinsky, has been an extremely resourceful friend whose company enlightened me with many thought-provoking perspectives on research and life. I am very grateful to have him as a friend.

I am particularly grateful to Rabih Sultan, Ahmad Nasri and Peter Ortoleva for preparing me and supporting me throughout my predoctoral years and directing me towards this doctoral path.

I am grateful to Douglas Hofstadter for the insightful walks we shared that seamlessly connected Woodlawn Ave. to Ballantine Rd., anti-analogously to how his law, *Hofstadter's Law*[1], connected the abstract of this dissertation to its conclusion.

I would like to acknowledge James Costello for availing a convenient LaTeXtemplate, and many recent PhDs that reminded me of how treacherous time is, namely Nicola Perra, Xiao Dong, Chris McEwan, Sebastian Von Mammen, Dania El-Khechen and many others. Special thanks to Francesco Catania and Rossano Schifanella for offering me prospective insights about postdoctoral vicissitudes, and Souheil Haddad for his unconditional positivity and generosity. Many thanks to the Nussmeier and Pecorelli families for their endless support. Many thanks to my friends from l'hebdofrancofolie,

[1] Hofstadter's Law: It always takes longer than you expect, even when you take into account Hofstadter's Law.

il circolo italiano, cafezinho and the three Cypriots. I am very thankful to all my friends without whom my doctoral life would have been unbearable and although their names will not fit here, they shall be remembered forever.

I am infinitely indebted to the numerous resources of Indiana University and Bloomington, not only for the completion of my dissertation, but also for their endless social, linguistic, cultural, artistic and musical clubs and events, and above all, their extremely amiable people; I will always cherish them. I am equally indebted to the School of Informatics and Computing and all its staff for facilitating my research with the latest technologies and prompt technical support. I cannot be any less grateful to the Instituto Gulbenkian de Ciência for its resourcefulness and beautiful ambiance.

Most importantly, none of this would have been feasible or worthy without the patience and love of my family, to whom I dedicate this dissertation. Notwithstanding their infinite love, they let me traverse the hemisphere believing it will be for a good cause and I hope that this manuscript *ipsum* will be evidence thereof.

Last but not least, I thank my other two thirds, the polyglot and the artist, for their patience and sacrifices during the course of this dissertation.

Alaa Abi Haidar

AN ADAPTIVE DOCUMENT CLASSIFIER INSPIRED BY T-CELL CROSS-REGULATION IN THE IMMUNE SYSTEM

Over millions of years, the vertebrate immune system has evolved into one of the most complex and intelligent biological systems. The immune system's function is to protect the body from harmful intruders. Several mathematical models have been proposed to understand the adaptive immune system and its functional subsystems. We develop a novel agent-based model of T-Cell cross-regulation in the adaptive immune system, and we apply it to binary classification problems analogous to those faced by the immune system.

We expect our study to help immunologists better understand the general mechanism behind T-cell cross-regulation and also to raise questions about the behavior of the adaptive immune system in general such as immune memory, cell death and homeostasis. However, our chief aim is to show that cross-regulation dynamics can be used to classify textual documents in changing corpora. We validate the model on real-world data from biomedical articles and personal e-mails, and we compare our algorithm with other machine learning classifiers. Finally, we discuss how the guiding of T-cell self-organizing dynamics can be seen as a general system of classification, the study of which is helpful for complex systems, text classification, and theoretical immunology.

Principal Advisor

Committee Member

Committee Member

Committee Member

Contents

List of Abbreviations

95%CI	95% Confidence Interval
Ab	Antibody
ABCRM	Agent-Based Cross Regulation Model
ABM	Agent-Based Model
Ag	Antigen
AIS	Artificial Immune System
APC	Antigen Presenting Cells
BC	BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology)
BDC	Biomedical Document Classification
CRM	Cross-Regulation Model
DM	Data Mining
E	Effector T-cells

Fig.	Figure
FN	False Negatives
FP	False Positives
iAUC	interpolated Area Under Curve
IS	Immune System
MCC	Matthew's Correlation Coefficient
ML	Machine Learning
N	Negatives
NB	Naive Bayes
P	Positives
PPI	Protein-Protein Interaction
PU	Positive Unlabeled
R	Regulatory T-Cells
StDev	Standard Deviation
SVM	Support Vector Machines
TCR	T-Cell Receptor
TF.IDF	Term Frequency Inverse Document Frequency

TN	True Negatives
TP	True Positives
VTT	Variable Trigonometric Threshold

List of Tables

List of Figures

Chapter 1

Introduction

Aut viam inveniam aut faciam

I will either find a way or make one

Hannibal

1.1 Motivation and Background

The number of documents of all types available online has been expanding at an astounding rate. In this dissertation we deal with text classification, though our bio-inspired binary classifier is generalizable to other types of documents or data sets.

One ubiquitous type of text document we look at is e-mail. We are all aware of this particular type of information pollution resulting from massive product advertisement (often fraudulent) sent through the cheapest and fastest of communication methods: unsolicited e-mail (spam). Spam is irritating and time and money consuming. One solution to the spam problem is anti-spam filtering that is aggressive enough to detect and eliminate spam but not legitimate e-mail. Spam and anti-spam filters have been co-evolving[1] over the last couple of decades, thus hindering the performance of traditional classifiers [GC06]. Analogously, the vertebrate immune system has been co-evolving with harmful viruses and bacteria for millions of years. The immune system can be understood as a binary decentralized classifier capable of learning about new intruders while avoiding autoimmune responses and thus discriminating between self and nonself substances. We interchangeably refer to self and nonself as harmless and harmful respectively, though much of our biomass is comprised of genetically non-self which is harmless, such as bacteria. An alternative is simply to regard all harmless substances as self, even if originating from genetically different organisms. For the sake of clarity and integrity, we assume that only potentially harmful substances are nonself for the model presented in this dissertation.

Biomedical document classification is another application area. It aims at extracting relevant information from huge amounts of textual documents in biomedical literature and databases. Given the technological advances in life sciences, such as faster genome sequencing [Mye99] and microarray analysis [SSD$^+$95], the last decade has witnessed an exponential growth [HC06] among metabolic, genomic and proteomic documents (articles) being published. Nevertheless, biomedical research should be facilitated rather than hindered by the tremendous abundance of bioliterature data. The field of literature mining aims to analyze and categorize millions of biomedical articles and records [Hea99]. A solution to identify relevant literature is automatic article recommendation or classification [JSB$^+$06b]. The immune system is a complex biological system made of millions of cells all interacting to collectively distinguish

[1] The spam detection problem is an arms race between spammers and spam-filtering techniques

between self and nonself substances, to ultimately attack the latter [Hof01]. In analogy, relevant textual documents for a given concept need to be distinguished from irrelevant ones which should be discarded in topical queries. We explore this analogy in this dissertation.

When thousands of cells interact to discriminate between self and nonself, they do so collectively in a *self-organized* manner. Our approach is based on the interaction of cells without access to a central controller [SC01]. Therefore, its classification ability needs to result from a *collective classification* process, defined as the ability of decentralized systems of many components to classify situations that require global information or coordinated action [Mit06]. Nature is full of examples of collective classification, for example, the dynamics of stomata cells on leaf surfaces are known to be statistically indistinguishable from the dynamics of automata that are capable of performing nontrivial classification [PWMM04]. Biochemical intracellular signal transduction networks are capable of emergent classification [HKHR08]. Quorum sensing in bacteria [WS06] and social insects [Pra05] are other examples of systems performing collective classification. We can study collective classification in general models of complex systems such as Cellular Automata, namely by identifying regular patterns in the dynamics that store, transmit and process information [CM95, RH05, SHR$^+$06]. In this dissertation, we study the self-organized behavior of T-cells and their ability to collectively classify documents.

1.2 Challenges

Machine Learning (ML) aims at creating automatic methods that can solve problems often by learning from available training data [FS06]. Chief among these problems is that of *classification*, in which instances are categorized into predefined classes, and in particular, *binary classification*, where the predefined class labels are limited to two: relevant and irrelevant. However, there is evidence that the state-of-the-art in ML has not been answering modern classification challenges found in *data streams* [Han06], where the sequential order of instances is an important factor for data analysis as we discuss below. A common problem, more specifically in textual data streams, is known as *concept drift* and is described by the gradual or sudden change of the underlying data distribution over time [Tsy04]. In the context of biomedical document classification, we observe concept drift in the appearance and disappearance of new terms that result from new discoveries, diseases, experiments or emergence of a new subfield [CBH04, TPCP06]. In spam detection, we observe it in illegitimate advertisement of new products and novel e-mail obfuscation techniques [MFRI^{+}06, DCTC05b]. In concept drift, the proportions of data instances associated with the classes to be predicted may also vary over time. This *dynamic class imbalance* makes the prior probabilities for the classes change between training and testing data, rendering classification a hard endeavor for supervised learners [KHA99, Kun04]. Therefore, in the presence of dynamic class imbalance and concept drift, an adaptive classifier that is resilient to class balance variation and capable of generalizing concepts that drift, is necessary [Tsy04].

Our challenge is thus to:

1. Design a self-organized, bio-inspired, binary document classifier using agent-based modeling.
2. Deal with dynamic class imbalance.
3. Track concept drift and distinguish it from noise.
4. Understand how a decentralized immune system can achieve binary classification.

1.3 Contribution

Recently, Carneiro et al. [CLC^{+}07] proposed an analytical model of T-cell cross-regulation. This model shows that the interaction dynamics of three cell-types is capable of discriminating between self and nonself substances. We develop this model as a discrete agent-based model (ABM) and generalize it into a general purpose bio-inspired algorithm for document classification, that we call the Agent-Based Cross-Regulation Model (ABCRM). With ABM, we are able to deal with recognition of many antigens/features simultaneously, rather than the single one as the original mathematical model. We show that this model performs well in binary document classification problems, especially in the presence of concept drift and dynamic class imbalance.

Our work establishes a bio-inspired algorithm of T-cell cross-regulation for document classification that is comparable to state-of-the-art classification methods. In this dissertation, we explore alternative model parameters and experimental setups in order to optimize the classification performance and understand the collective and decentralized behavior of T-cell cross-regulation dynamics and more generally, the vertebrate adaptive immune system, namely by better understanding various configurations of our model and parameter ranges that lead to optimal classification results on textual documents. We also validate our model against other classifiers on two datasets.

The following list summarizes our contributions:

1. An original ABM of T-cell cross-regulation dynamics.
2. Insights about the behavior of T-cell cross-regulation as modeled mainly concerning the collective classification of T-cells, the self-organized behavior of the immune system, the role of cell death in immune memory and the importance of the sequential order in which proteins are introduced to the immune system.
3. The application of the ABCRM to spam detection in the presence of dynamic class imbalance and concept drift.
4. The application of the ABCRM to biomedical document classification in the presence of dynamic class imbalance and concept drift.

1.4 Outline

The structure of this dissertation is outlined as follows:

- Chapter 2 provides a background on binary document classification and text mining to familiarize the reader with traditional feature processing techniques, data classification and verification methods, and advances in text classification and concept drift in particular. This chapter also provides an overview of two applications to binary document classification, namely spam detection and biomedical document classification.

- Chapter 3 provides an overview of the vertebrate immune system and specifically the adaptive immune system, which is detailed within the context of artificial immune systems. We focus on T-cell cross regulation dynamics.

- Chapter 4 introduces our agent-based cross-regulation model (ABCRM). This chapter explains algorithm parameters and possible configurations adapted from relevant publications [AHR08b, AHR08a, AHR10a, AHR10b, AHR].

- Chapter 5 discusses the application of our model to spam detection. In addition, this chapter addresses some of the challenges brought about by concept drift and class imbalance. This chapter is adapted from relevant publications [AHR08b, AHR08a].

- Chapter 6 discusses the application of our model to the binary classification of biomedical documents. This chapters also studies the collective behavior of T-cell dynamics focusing on its robustness. This chapter is adapted from relevant publications [AHR10a, AHR10b, AHR].

- Chapter 7 presents insights about T-cell cross-regulation, gained by our immune model, such as the role of cell death in immune memory, the collective classification of T-cells in a self-organized system and the effects of the sequential order of proteins on the immune system's adaptive ability to discriminate between self and nonself.

- The last chapter, chapter 7, refers to the conclusion of this dissertation.

Chapter 2

Text Classification

$\Pi\alpha\nu\tau\alpha\, P\epsilon\iota$

The only constant is change

Heraclitus

In this chapter, we give an overview of text classification and some of its major challenges such as class imbalance and concept drift. We also review traditional and state-of-art techniques for document classification, as well as applications thereof, such as spam detection and biomedical document classification.

2.1 Introduction

Huge amounts of textual data are constantly being generated automatically or manually by millions of human beings and computational and analytical institutes. Fayyad and Uthurusamy [FU02] report:

"The capacity of digital data storage worldwide has doubled every nine months for at least a decade, at twice the rate predicted by Moore's Law for the growth of computing power during the same period...Our ability to capture and store data far outpaces our ability to process and exploit it."

Much of this data is in the form of textual documents, such as logs, blogs, social networking sites, scientific articles and e-mails. When documents are collectively analyzed, they can result in useful information that cannot be conveyed from a single document, such as age range, common interests, average number of protein interactions, and much more [McC05, JSB$^+$06b]. The integration of useful information, which we call *knowledge* often requires the analysis of thousands of documents, well beyond the capabilities of manual human labor [FU02].

2.2 Machine Learning and Data Mining

Machine Learning (ML) is a subfield of artificial intelligence which aims at developing algorithms that are capable of automatically solving problems. ML is used to automatically solve classification, clustering and prediction problems and can be applied to data mining, speech recognition, computer vision, bioinformatics and many other problem areas [FS06]. For instance, text mining aims to aid the problem of extracting knowledge from large-scale corpora of documents very efficiently.

Data mining uses ML for automating the analysis of raw data in order to extract high quality information or discover new knowledge. Text mining focuses on extracting information from textual documents whereas literature mining adds to text mining the linking to databases and ontologies [SF03, JSB$^+$06b]. Web mining focuses on the extraction of high quality information from web documents [Men03].

Some common text mining applications are text categorization, text summarization, document recommendation, entity tagging and information extraction. In the course of this dissertation, we will focus on document classification for the purpose of understanding the applicability of our bio-inspired method and comparing it with traditional text classifiers. Nonetheless, any classification algorithm can be generalized to classify nontextual data, which we leave for future work.

Text data can be unstructured (purely textual) or structured (when tagged with markers that relate textual tokens with other sources such as ontologies and databases). We are interested in classifying unstructured text data, which is the most common and challenging [McC05]. Unstructured text is available in corpora with huge numbers of documents such as e-mail, scientific articles, blog posts and web pages.

ML methods can be supervised, unsupervised or semi-supervised. In **Supervised learning**, algorithms are trained on class-labeled data to predict on unlabeled data. Unsupervised algorithms predict unlabeled data with no prior training or knowledge of the labels. Semi-supervised algorithms use a combination of both learning methods. We are more interested in supervised and semi-supervised learning for assessing T-cell cross-regulation as a classification algorithm.

Classification is the act of assigning a set of unlabeled *instances*[1] to the corresponding class labels. In document classification, instances are documents and *attributes*, which are fragments of text such as words occurring in the document or combinations thereof, such as bigrams and phrases. Documents are classified based on their constituting features and the relevance of these features to classes. We are

[1] In machine learning, instances, also known as examples, are the items (such as documents) that the algorithm is supposed to classify

particularly interested in binary document classification, which aims at classifying documents into two predefined labels/classes, namely relevant and irrelevant documents, also denoted as positive (P) and negative (N) documents. Binary classification is the most basic form of classification, which can be always extended to multi-class classification simply by *bottom-up* combination of binary classes [GL03]. Traditionally, a supervised binary classifier is trained on a *corpus*, a set of labeled (relevant and irrelevant) documents, then tested on a distinct corpus of unlabeled documents that are to be classified into the two categories. We are also interested in some aspects of semi-supervised binary classification, mainly Positive Unlabeled (PU) learning, in which the classifier is trained only on Positive instances and tested on unlabeled ones [FS06].

Several **representation models** are used for describing textual documents. Amongst the most common ones is the vector-space model [Sal91] in which documents/instances are represented as vectors of *features*. Features are text attributes, usually words or more complex sets of words such as bigrams, trigrams ... n-grams. For example, an instance $\vec{x} = [1, 0, 0, 1, 0, 0...0]$ represents a document composed of features f_1 and f_4 (marked with 1) from a set of features $f_i \in F_k$, where F_k is the set of all features in a corpus k. The vector-space model can represent feature occurrence with binary vectors, feature frequency (also known as term frequency) with numeric vectors or more sophisticated feature-document scores such as TF.IDF, that we discuss in section 2.3.3.

2.3 Text Processing

The literature contains various approaches to text processing and feature selection. We list the most traditional ones in text mining and discuss in detail the ones we use.

2.3.1 Stop Words

Stop words are common dictionary words that occur very frequently in unstructured text and are usually filtered out because they have no discriminative power [Fox89]. For each language and application we can produce a specific list of stop words. The most common stop word list (or negative dictionary) for English text is based on the Brown corpus, a broad collection of literature in English [*ibid*]. Notwithstanding their frequency in English text, some terms could be important in the domain of the application and useful for discrimination. For example, the terms "with" and "between" are specific to the domain of protein-protein interaction, often describing "interactions between" proteins or "interactions with" proteins, and therefore excluding them from the stop list could help with the extraction of more meaningful features [AHKM$^+$08].

2.3.2 Stemming

In text mining and natural language processing, stemming is used to reduce a word to its stem (also known as morphological form or root) such that every conjugation of a word is then tokenized and recognized similarly by the system. For example "stemming", "stemmed" and "stems" are automatically stemmed to their morphological root "stem". The most widely used stemmer is the Porter stemming algorithm [Por80b], which has been implemented in many platforms and for many languages [Por06]. We use the Porter stemming algorithm for stemming unstructured English text.

2.3.3 Feature Selection and Reduction

A common challenge in machine learning is that of dealing with huge numbers of potential features (or attributes), which can be computationally expensive in document classification. Text classification typically includes the selection and reduction of features (also known as feature extraction). In other words, the features are selected to serve the machine learning problem adequately, while minimizing loss in the classification or learning performance of the algorithm. Several solutions have been developed for feature selection in document classification, such as latent semantic indexing (LSI) based on singular value decomposition [Dum90, WRR03], Information Gain (IG) and Mutual Information (MI) [Seb02]. Experiments on real-data show that the feature selection scores, IG and *Document Frequency*[2] (DF), are strongly correlated when classified with a k-nearest neighbors [YP97]. We use two measures based on document frequency for ranking and selecting features: TF.IDF [Sal89] and the *S Score*, that we define in the following sections. Both measures were proven to be good in the domain of biomedical document classification [AHKM$^+$08, KAHK$^+$10]

[2] Document frequency of a feature is the number of documents having at least an occurrence for this feature

using a simple linear classifier.

TF·IDF

TF $\cdot$ IDF (term frequency-inverse document frequency) is a common measure used to evaluate the relevance of a feature f to a document, which is proportional to its frequency in the document (TF) and inversely proportional to the number of distinct documents in a corpus it occurs in (IDF) [Sal89]. Let n_f be the number of times a feature f occurs in a document. The relevance of a term f to a document d, where $f \in F_d$ (all features constituting document d), is defined by $TF(f, d) \times IDF(f)$:

where the *Term Frequency* is:

$$TF(f, d) = \frac{\{n_f : f \in d\}}{\sum_{k \in d} n_k} \tag{1}$$

The *Inverse Document Frequency* is:

$$IDF(f) = \log \frac{|D|}{|\{d \in D : f \in d\}|}, \tag{2}$$

where D is the set of documents or the *corpus*.

S Score

The S score is a measure that we used for feature selection in binary document classification [AHKM$^+$07, AHKM$^+$08, KAHK$^+$10, KAHK$^+$09]. A similar variation of it has also been used in spam filtering [FRID$^+$07b]. The S score of a feature f in a corpus D is $S(f) = |p_P(f) - p_N(f)|$ where p_P is the probability of f occurring in an article labeled as relevant (positive) and p_N is the probability of it occurring in one labeled as irrelevant (negative) [AHKM$^+$07, AHKM$^+$08, KAHK$^+$10, KAHK$^+$09, RODV04]. The higher the S score of a feature, the more this feature is discriminatory or helpful for document classification, in which case, the feature is very relevant or irrelevant to one of the classes.

2.4 Text Classification Methods

Topical categorization of text started in the late 60's with probabilistic text indexing [Mar61] but was not used widely until the beginning of the 90's [Seb02]. Several methods have been developed for document classification such as Rocchio's classifiers [Roc71], k Nearest Neighbors (k-NN) [YL99], Naive Bayes [Lew98], decision trees [DPHS98], support vector machines [Joa98], etc . The Rocchio method finds a prototype vector for every class and then computes document cosine similarity between the prototype and target document vectors and it subsequently assigns the documents to the class with the maximum similarity. k-NN is an *example based* classifier where classification is based on neighboring examples. k-NN is considered a *lazy learner* since all classification is performed simply by looking at nearby examples without prior training [YL99]. Decision Trees (DT) are used for classification and decision making, where a decision tree has internal nodes denoting simple logical rules and leafs denoting decisions or class labels [FHK$^+$91, CH98, LJ98]. Artificial Neural Networks (ANN) are composed of interconnecting artificial neurons inspired by structural and functional aspects of biological neural networks for solving artificial intelligence problems such as text classification [LL99]. In the following sections we discuss state-of-art classifiers that we have chosen to validate our method against, namely Naive Bayes Classifier and Support Vector Machines, which have been widely used for text classification [YL99]. In addition, we discuss the Variable Trigonometric Threshold, a lightweight linear classifier that we developed for biomedical document classification [AHKM+07, AHKM+08, KAHK+10, KAHK+09].

2.4.1 Naive Bayes

The Naive Bayes (NB) Classifier is a simple probabilistic classifier based on Bayes' theorem [Lew98]. Given a document represented as a feature vector, NB classifies it into the most likely class assuming that features in the vector are independent, given the class.

Given two classes of documents: relevant (P) and irrelevant (N), and a document d, we form the conditional probabilities $p(d|P)$ and $p(d|N)$ from the labeled training corpus. Next, we apply Bayes' theorem to calculate the *a posteriori* probabilities $p(P|d)$ and $p(N|d)$ to predict the class of d.

The probabilities $p(d|P)$ and $p(d|N)$, assuming the features are independent, are $p(d|\{P, N\}) = \prod_{i=1}^{n} p(f_i|\{P, N\})$, where a document d is a vector of features $\langle f_1 \dots f_n \rangle$.

Various Bayes classification rules are used to maximize the difference between $p(P|d)$ and $p(N|d)$ in order to classify the document d. Metsis et al. [MAP06] discusses these classification rules and compares their performance on e-mail classification to show that multinomial NB with boolean attributes is very reliable. Therefore we use it in our implementation of the NB classifier.

2.4.2 Support Vector Machines

A Support Vector Machine (SVM) is a supervised learning algorithm that is commonly used in text classification. In binary classification, a SVM is used to draw a linear boundary with optimal separation of instances between the two classes with respect to the given features [Vap95, Joa02]. Assuming that the two classes are linearly separable, one can draw infinitely many linear boundaries to separate between them but only one optimal boundary with the maximum-separation margin (see fig. 2.1).

In high-dimensional space, the optimal boundary is represented by a decision hyperplane, which is satisfied by the equation $\langle \vec{w}, \vec{d} \rangle + b = 0$, where $\vec{w}$ is the vector that is normal to the optimal hyperplane, $\vec{d}$ is the feature-vector document and b is the bias. The optimal hyperplane is surrounded by two secondary equidistant hyperplanes, $\langle \vec{w}, \vec{d} \rangle + b = -1$ and $\langle \vec{w}, \vec{d} \rangle + b = +1$, respectively, and the distance between them is called *the margin* (see fig. 2.1).

The computation of the optimal hyperplane is a quadratic optimization of the function of (w, b) that maximizes the margin, $\frac{2}{||w||}$, such that all documents $\vec{d_i} \in D$ fall on (as support vectors) or beyond the secondary hyperplanes (see fig. 2.1). Finally, the function $f(\vec{d_i}) = sign(\vec{w}^T \vec{d_i} + b)$ predicts the class label for each document $\vec{d_i}$ as relevant, if $f(\vec{d_i}) = 1$, or irrelevant, if $f(\vec{d_i}) = -1$.

For non-linearly separable classes, or soft classification, Cortes and Vapnik introduced a slack variable that corresponds to margin errors due to misclassification[CV95].

The only drawback of SVM is their training time which was estimated as $\mathcal{O}(N^x)$

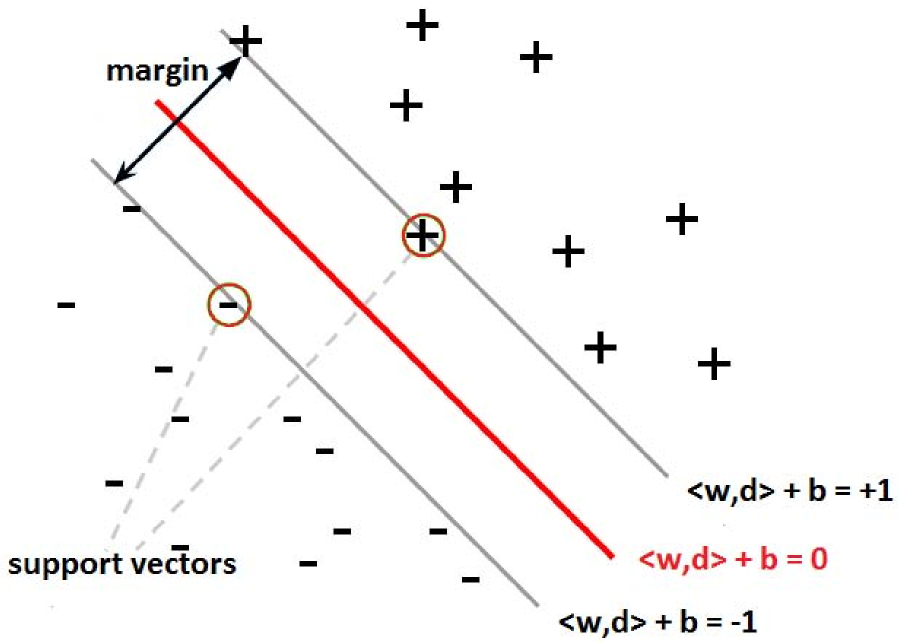

Figure 2.1: An illustration of binary classification with SVM showing the two classes, the optimal and secondary hyperplanes, the support vectors and the margin.

where N is the number of documents and $x \in [1.8, 2.1]$ [CRS03].

We use the publicly available SVM^{light} [Joa99]. The documents are represented as feature vectors, with feature counts in documents mapped to the range [-1,1] such that they can be processed by SVM^{light}.

2.4.3 Variable Trigonometric Threshold

The Variable Trigonometric Threshold (VTT) is a linear classifier that we developed for biomedical document classification [AHKM^{+}08, AHKM^{+}07, KAHK^{+}10, KAHK^{+}09].

The ideal features in the p_P/p_N boundary plane (see Figure 2.2) are those closest

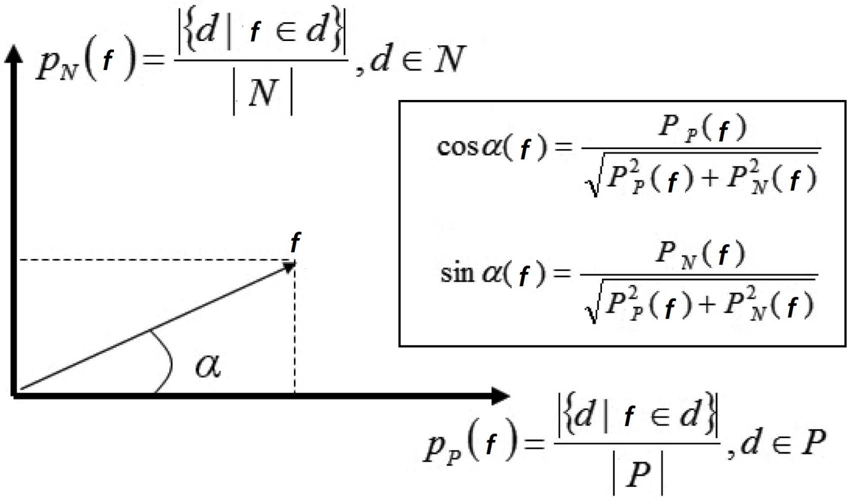
$p_N(f) = \frac{|\{d \mid f \in d\}|}{|N|}, d \in N$
$\cos\alpha(f) = \frac{P_P(f)}{\sqrt{P_P^2(f) + P_N^2(f)}}$
$\sin\alpha(f) = \frac{P_N(f)}{\sqrt{P_P^2(f) + P_N^2(f)}}$
f
α
$p_P(f) = \frac{|\{d \mid f \in d\}|}{|P|}, d \in P$

Figure 2.2: Trigonometric measures of term relevance in the p_P/p_N plane; p_P and p_N computed from labeled documents d in training data.

to either one of the axes. Any feature f is a vector on this plane (see Fig. 2.2), and therefore feature relevance to each of the classes can be measured with the traditional trigonometric measures of the angle α between this vector and the p_P axis: $\cos(\alpha)$ is a measure of how strongly features are associated with positive/relevant documents, and $\sin(\alpha)$ with negative/irrelevant ones in the training data. Then, for every document d, we compute the sum of all feature contributions for a positive (P) and negative (N) decision:

$$
\begin{aligned}
P(d) &= \sum_{f \in d} \cos(\alpha(f)) = \sum_{f \in d} \frac{p_P(f)}{\sqrt{p_P^2(f) + p_N^2(f)}}, \\
N(d) &= \sum_{f \in d} \sin(\alpha(f)) = \sum_{f \in d} \frac{p_N(f)}{\sqrt{p_P^2(f) + p_N^2(f)}}
\end{aligned}
\tag{3}
$$

The decision of whether a document d is classified as positive or negative is then computed according to:

$$\begin{cases} d \in P, & \text{if } \frac{P(d)}{N(d)} \geq \lambda_0 + \frac{\beta - np(d)}{\beta} \\ d \in N, & \text{otherwise} \end{cases} \tag{4}$$

where λ_0 is a constant threshold that is optimized for deciding whether a document is relevant or irrelevant. This threshold is subsequently adjusted for each document d with the factor $(\beta - np(d))/\beta$, where β is another constant, and $np(d)$ is the number of entities in a document d that can be discriminatory based on the domain of the problem. For example, in the domain of protein-protein interaction, $np(d)$ could represent the number of unique protein names tagged in a document d by a biological entity recognition tool such as ABNER [Set05].

We first employed VTT for the protein-interaction abstract classification task of the second BioCreative Challenge [AHKM$^+$08, AHKM$^+$07] and our method was ranked among the top performing classifiers [KV07]. Our team then participated in the full-text classification tasks of BioCreative 2.5 [KAHK$^+$10, KAHK$^+$09] and our method was deemed best classifier [LMK$^+$10] with respect to all F-score, Accuracy and AUC.

2.4.4 Overfitting

The main challenge in text classification is to have an algorithm that is capable of learning from a training corpus to generalize and predict document class labels without *overfitting*. Overfitting occurs when a classifier is too fine-tuned to the training data or the noise therein, that it cannot generalize on testing data [FS06]. Community wide efforts, such as the Biocreative challenge [KV07], try to address this issue of overfitting by manually annotating a validation data set that is distinct from the training one. Machine learning solutions try to address the overfitting problem differently [FS06] — for example, decision trees use pruning to cut random decision subtrees [Qui87]. Several methods have been proposed for this problem including bagging [Bre96], boosting [FS95], and stacking [Wol92]. We propose an adaptive bio-inspired solution based on the interaction of decentralized and self-organized T-cells that are capable of learning collectively and adapting to changes between the training and the validation data.

2.4.5 Dynamic Class Imbalance

Another challenge in data mining is that of **class imbalance**, in which we are given many more instances from one class (usually the negative or irrelevant) than from the other [CJK04]. Class imbalance is often due to lack of information about one of the classes [Abe03, VR05, KKP06]. Several attempts have dealt with its challenges [CJK04, Wei04] by either oversampling instances from the smaller class, under-sampling instances from the larger class or combinations of both [CJK04]. The problem becomes even more challenging when class imbalance changes dynamically or has different properties between the training data and the validation data [KHA99, Kun04]. For example, in spam detection the number of desired or spam e-mails that a person receives in a fixed time interval varies all the time and depends on many factors that are not easy to predict. Therefore, training on a fixed proportion of relevant-to-irrelevant instances is not always accurate in real-world data. Hence, we propose an adaptive bio-inspired solution that is robust to these changes and we validate our hypothesis by training our method on a set of balanced data and testing it on a distinct set of imbalanced data.

2.4.6 Concept Drift

Data can be static or temporal (data stream), where the time dimension and the sequence in which the data is analyzed are of importance to the data analysis. In data streams, instances can be studied over fixed time intervals—for example, e-mails received in the last month or biomedical articles published in the last year. Data stream mining has gained a lot of attention over the last few years through conferences, workshops and journals [GZK05]. A major challenge in data stream mining is that of *concept drift*.

Concept drift is the (gradual or sudden) change of underlying distributions of features in classes over time in unforeseen ways [Tsy04]. It is often assumed in ML that these distributions remain unchanged, however they are constantly changing due to changes in *hidden context*. The hidden context is not given explicitly as a predictive feature but the assumption is that it ultimately affects the proportions and values of the predictive features [Tsy04]. For example, the text feature "yeast" is relevant to the concept of protein-protein interaction (PPI) only when a hidden context of the document such as "yeast two-hybrid" relates to a PPI extraction method.

Concept drift is very common in textual stream data such as e-mail data and biomedical articles [MFRI$^+$06, DCTC05b, TPCP06]. In supervised learning, the target concepts simply represent the predefined class labels [DCTC05b]. Hence, in text document classification, we can define concept drift as the change in feature distribution underlying the relevant class, which we assume changes with linguistic context through time. From a Bayesian perspective, the drift may occur in three ways [Kun04, KHA99]: (i) The prior probabilities $p(c)$ of classes/concepts c, may change over time. (ii) The probability distribution $p(x|c)$ of features x given class c, may change over time. (iii) The third case is when the posterior distribution of class memberships $p(c|x)$ change, which can be inferred from the previous two cases. Frequent re-training can be expensive and therefore a *one-pass* adaptive learner is desirable to make more accurate predictions in the presence of concept drift. In supervised learning, concept drift can be identified and measured by the drop in classification performance over time [Tsy04].

In spam detection, a textual feature such as 'Rolex' becomes relevant to the concept/class of spam after many unsolicited and undesirable Rolex e-mails. However, users involved in the Rolex business might be interested in Rolex relevant e-mails. Some users might even become interested in buying a Rolex and in turn in Rolex relevant e-mails, or *ham*. From a Bayesian perspective, $p('ham'|'Rolex')$ and $p('spam'|'Rolex')$ may fluctuate between 0 and 1 (see Fig. 2.3). The feature 'Rolex' is not the only feature drifting as all its contextual features such as 'replica', 'order', or even 'unsubscribe' may also drift. Needless to say, the numerous spelling variations of 'Ro1ex' can be made relevant to the spam concept by spammers. Therefore, it is not enough to train the algorithm on recent instances, a *personalized* algorithm is required to treat the concept of spam differently for every user depending on the dynamic change in their legitimate e-mails (ham). The personal aspect of e-mail makes the Artificial Immune Systems (AIS) approach all the more compelling, as each user can generate personalized "spam immunity" in analogy to "personalized" natural immunity. However, first we need to test if our method works as a classifier.

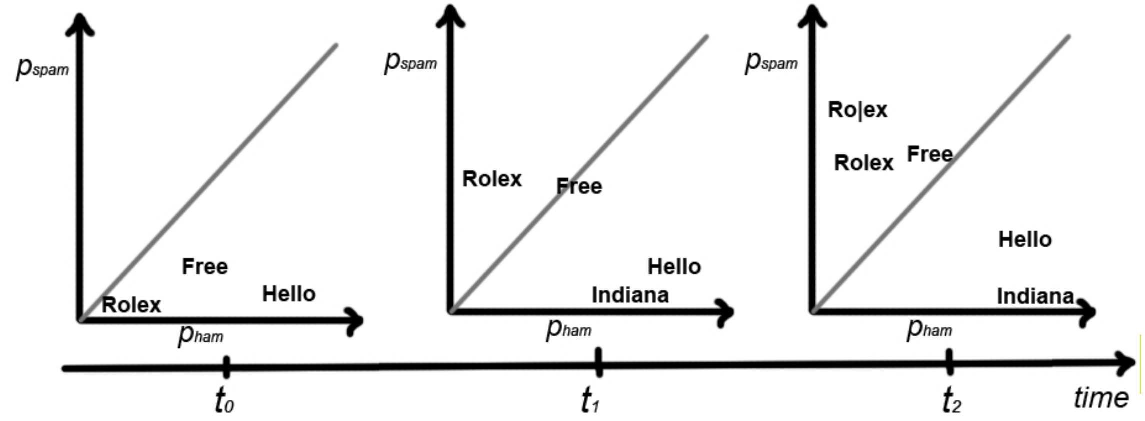

Figure 2.3: An example of a concept drift in spam where the feature 'Rolex' becomes more relevant to spam with $p('spam'|'Rolex')$ increasing over time.

Three approaches distinguish algorithms handling concept drift: instance selection, instance weighting and ensemble learning [Tsy04]. *Instance selection* focuses on the most recent instances that can identify the current concept. Time windows (of fixed and varying sizes) are effective in doing so [WK96, MM99]. For example, FLORA is a window based memory system that was later equipped with varying window size (FLORA2) and then improved to store concepts for later use when context change is detected (FLORA3) [WK96]. Finally, FLORA4 was designed to additionally deal with noise and distinguish it from concept drift [WK96]. *Instance weighting* is sometimes used to weight more significant and recent instances and age others. For example, Syed [SLS99] use block-by-block incremental SVM, by training on the current block (batch) of instances and the support vector from the previous training block. Instance weighting resembles instance selection, except that it is not limited to a window of instances and therefore it can weigh older instances more than more recent ones. *Ensemble learning* combines predictions from more than one system for decision making [KM07, SK01, TPCP08]. For example, SEA is an ensemble of separate decision tree classifiers that are constantly replaced by newer ones that improve the performance of the ensemble [SK01]. DWM-NB and DWM-ITI are also ensembles of weighted naive bayes and incremental tree inducer systems that classify based on a weighted majority vote [KM07]. Other models used for tracking concept drift include instance-base reasoning, such as Spamhunting [FRID$^+$07b, FRID$^+$07a] and case-base reasoning, such as ECUE [DCS06], both for tracking concept drift in spam.

Many of these systems were tested solely on artificial data benchmarked for studying concept drift, such as STAGGER concepts [SG86]. Others were tested with real life data, such as sequential and time series data from the UCI machine learning repository [AN07], having a fixed number of features in contrast to boundless textual features from documents.

The adaptive immune system has aspects similar to the aforementioned approaches naturally built-in and evolved rather than designed. For example, natural cell death in general and the positive and negative selection of T-cells in the thymus are analogous to feature selection. We plan to address the problem of concept drift for document classification from a new bio-inspired perspective.

2.5 Application

Text can be categorized to identify conceptually-related classes of documents—at a minimum, two classes with relevant and irrelevant documents for a given concept. The relevance can be with respect to a personal interest such as desirable e-mail in contrast to spam, or with respect to a certain topical query, such as biomedical articles that are relevant to protein-protein interaction. In this section we offer an overview on both application areas.

2.5.1 Spam Detection

Spam detection is a binary document classification problem in which e-mail is classified as either ham (legitimate e-mail) or spam (illegitimate or unsolicited e-mail). Spam is very dynamic in terms of advertising new products and finding new ways to defeat anti-spam filters. Spam detection has recently become an important problem with the ubiquity of e-mail and the rewards of no-cost advertisement that can reach the largest audience possible. With millions of users, spam becomes not only annoying but also expensive, costing businesses around $130 billion, as estimated[3] for 2009. The challenge in spam detection is to obtain the smallest possible number of misclassifications, especially of legitimate e-mail (false negatives). To avoid confusions, we label ham and spam as positives and negatives respectively, although the opposite labeling is also practiced. Spam detection often focuses on e-mail headers (e.g. sender, receiver, relay servers...) or textual content (e.g. subject, body), however we are only

[3] http://www.ferris.com/research-library/industry-statistics/

interested in the textual contents of e-mail for assessing our method as a document classifier.

Support vector machines [KA01], Naive Bayes classifiers [SDHH98, MAP06] and other classifiers such as those using Case-Based Reasoning [FRID$^+$07b] have been very successful in detecting spam in the past. However, they generally lack the ability to track concept drift that is common in this domain with new advertisement themes in spam. Bayesian poisoning, a technique used by spammers to surpass spam filters based on Naive Bayes classifiers, is also difficult to escape. Ideally, a spam detection system should be capable of handling concept drift, distinguishing it from noise [Tsy04]. Research in spam detection has recently been focusing on detecting concept drifts in spam, with very promising results [DCS06, MFRI$^+$06]. Other areas of intense development in spam-detection include *social-based* spam detection models [BR05, CDN05, TSC10], in which spam-detection relies on feedback about spam e-mails and addresses from multiple users. *Network-based* detection models [HSF$^+$09] process and analyze network packets from e-mails to detect patterns similar to those sent by spammers.

AIS models are inspired by various sub-systems and theories of the natural immune system [Hof01], such as negative selection, clonal selection, danger theory and the immune network theory. Oda's Masters thesis was based on using clonal selection of AIS for spam detection and the preliminary results were encouraging [Oda05]. AIS algorithms based on ABNET, an AntiBody Network [BB06] and incremental clustering Immune Networks [YAC$^+$07] were also used for spam detection. Nevertheless, our bio-inspired model is based on a novel and simple model of T-cell cross-regulation

[CLC$^+$07] that we test and discuss in the following chapters.

2.5.2 Biomedical Document Classification

Much of modern biomedical research relies on the induction of correlations and interactions from all types of data. Indeed, in the last decade, fueled by the production of large biomedical databases, particularly those containing genomic data, as well as the widespread use of high-throughput technology, we have witnessed the emergence of a more data-driven paradigm for biological research, which in turn created new analysis challenges. Since we ultimately want to increase our knowledge of the bio-chemical and functional roles of genes and proteins in organisms, there is an obvious need to integrate the associations and interactions amongst biological entities, which have been reported and accumulate in literature and databases. Such integration can provide a comprehensive perspective of presently accumulated experimental knowledge, and even uncover new relationships and interactions induced from global information but unreported in individual experiments.

Literature mining [SF03, JSB06a] is expected to help with such integration and inference; its objective is to automatically sort through huge collections of literature and databases (the "bibliome") and suggest the most relevant pieces of information for a specific analysis task, e.g. the annotation of proteins [HYBV05]. Given the size of the bibliome, it is no surprise that literature mining has become an important component of bioinformatics. But this success has raised the important issue of validation and comparison of the inferences uncovered via "bibliome informatics". While it is difficult to develop a "gold standard" for all literature mining approaches, it is important to provide a means to test and validate different algorithms and electronic sources of biological knowledge. Thus, researchers in this field have focused on testing algorithms and resources on specific tasks, e.g. protein annotation[HYBV05] and protein family [MRV$^+$06] and structure [RLRS06] prediction.

As for the specific task of binary biomedical document classification concerning our research, community wide efforts such as KDD Cup 2002, TREC Genomics and BioCreative dedicated tasks for the classification of biomedical published articles. The KDD Cup at 2002 provided 862 journal articles curated by FlyBase and the challenge was to determine whether 213 test articles contained experimental evidence about gene products [YHM02]. The TREC Genomics conference for 2004 provided 10 years of completed citations from the MEDLINE database inclusive from 1994 to 2003 (4,591,008 articles) and dedicated a subtask for determining whether an article has experimental evidence warranting GO annotation (relevant) or not (irrelevant) [HBC04]. The goal of this triage process was to reduce the number of articles sent to human curators for further analysis. The top performing submissions used various domain-specific techniques for identifying gene names and a Bayesian classifier for additional weighting [*ibid*]. The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge evaluation is precisely an effort to enable comparison of various approaches to literature mining. The BioCreAtIvE challenges have provided labeled and annotated articles and dedicated a task for biomedical document classification in terms of relevance for annotations concerning proteomics and to protein-protein interaction [KV07].

In the Article Classification Task (ACT) of BioCreative II.5 participants were provided with 619 full-text articles for training (out of which 61 articles were curation-relevant), and 595 full-text articles for testing (out of which 63 articles were curation-relevant) [LMK$^+$10]. Our team (team 9) was among 15 participants and used several methods, especially VTT (see 2.4.3), that were ranked amongst top performing methods for the classification task in BCII.5 [LMK$^+$10] and previous Biocreative challenges [AHKM$^+$08, KAHK$^+$10, VCJ$^+$05]. Bio-inspired methods were never employed for such competitions and therefore we test our immune-inspired method on biomedical data from the Biocreative challenge and compare our results with traditional classifiers and performances of other participants (see chapter 6).

2.6 Validation

Validation is used to assess the quality of a given classifier, especially in terms of its ability to generalize from the training data. In order to answer the ability of a classifier to generalize, the labeled training documents are partitioned into training and validation documents [FS06]. The labels of the latter predicted by the classifier are then compared to the correct labels using standard measures to assess the quality of the classifier. Finally, these measures are compared to those obtained by state-of-art classifiers to assess the relative performance of the classifier. In this section, we discuss data partitioning techniques, performance metrics for the assessment of classification quality, and statistical tests used for the comparison of algorithms in terms of their classification performance.

2.6.1 Data Partitioning

K-fold cross-validation is commonly used by splitting the training data into K parts, training the classifier on the $(K-1)$ partition ($K = 10$, leaving 90% of the data for training, or $K = 4$, leaving 75% of the data for training, or $K = 2$ with validation and training portions equal) and testing on the remaining part [FS06]. Each of the K testing sets is then evaluated using a performance metric, that are averaged and reported in addition to other statistical variations.

2.6.2 Stream Data Partitioning

In stream data, the order of the documents can be of importance to the algorithm and therefore the algorithm is trained on a set of documents that are ordered by time of creation or publishing, and then tested on a distinct set of documents that follows in time order [SG86].

2.6.3 Performance Metrics

The classifier can be tested on the aforementioned partitions to compute statistics of predictions of false negatives (FN), false positives (FP), true negatives (TN) and true positives (TP). From these, one can calculate:

- *Error rates* $\%FN$ and $\%FP$
- *Precision* ($\frac{TP}{TP+FP}$) and *recall* ($\frac{TP}{TP+FN}$)
- *F-score* $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

- *Accuracy* $\frac{TP+TN}{TP+TN+FP+FN}$
- *Area Under the ROC Curve* (AUC) represents the trade off between TP rates (TPR) and FP rates (FPR) and should not be confused with the Area Under the interpolated precision and recall Curve (iAUC) . We use the latter since it is the recommended metric for the evaluation of unbalanced biomedical document classification [LMK$^+$10].
- Mathew's Correlation Coefficient (MCC) is often used in document classification and it is calculated as $\frac{TP \cdot TN - FP \cdot FN}{\sqrt{((TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN))}}$

We use all the above measures to assess the classification performance of our algorithm, especially F-score for balanced data and additionally iAUC for imbalanced data.

2.6.4 Statistical Tests for Comparison

We compare classification performance of algorithms using the paired student t-test. Let H_0 be the null hypothesis that any two samples are drawn from the same distribution. A p-value < 0.05 rejects (p-value < 0.01 strongly rejects) H_0, establishing a statistical distinction between the two samples —in our case, the performance measures drawn from the two classification algorithms. Hence, the alternative hypothesis (H_1) that one classifier outperforms the other is proven [FS06]. We also use 95% confidence intervals for the purpose of comparing the distribution of classification performance [*ibid*].

2.7 Summary

In this chapter, we have defined text classification in terms of data mining and machine learning. We have introduced and discussed traditional techniques for feature processing, feature selection, document classification, validation and performance evaluation. We also gave an overview of the current challenges in text classification in general, and in two applications of binary text classification, namely, spam detection and biomedical document classification.

Chapter 3

Natural and Artificial Immune Systems

Atento ao que sou e vejo, torno-me eles e não eu.

Attentive to what I am and see, I become them and stop being myself.

Fernando Pessoa in *Himself*

In this chapter, we give an overview of natural immune systems focusing on adaptive immunity and T-cell dynamics in vertebrate organisms. We also review the recent field of artificial immune systems, especially the topics that are most relevant to text classification.

3.1 Introduction

The word 'immunity' derives from the Latin *immunus*, which referred to an exemption from tax or burden that was applied to all Roman citizens. The first written records of adaptive immunity can be traced back to Thucydides' notes about the plague of Athens in 430 BC. Thucydides reported that patients recovering from a disease could nurse the sick without risking a second infection with the same disease. This adaptive phenomenon was witnessed by many other societies but it was not until the

19th century that it was scientifically studied.

3.2 The Natural Immune System

The Immune System (IS) has a multi-layered structure of defense (see figure 3.1). The organism's skin forms the first barrier of defense against infection. Following is another physiological layer that is equipped with temperature and pH conditions that can be unbearable for most intruders. However, some intruders bypass the elementary layers of defense to then face the innate and the adaptive immune systems. For many years both systems were separate areas of study until evidence recently showed that both consist of millions of cells of various types that interact in a complex fashion to recognize and eliminate pathogens [Hof01]. The process of recognition and elimination is accomplished by chemical bonding directly between cell *receptors*[1] and pathogens with matching patterns, or indirectly between cells that communicate by emitting signals to mediate immune responses.

3.2.1 The Innate Immune System

The "innate" immune system response is mostly inherited at birth, in contrast to the response of the adaptive immune system which is acquired over time. The innate IS is brought to recognize generic targets in pathogens, while the adaptive immune system learns to recognize specific targets [Rob07]. The innate IS is mainly divided into the *complement system*, and the *endocytic* and *phogocytic* systems. In the last

[1] Cells are equipped with polypeptide structures called receptors that can specifically recognize fragments of pathogens (or patterns) by chemically binding to them

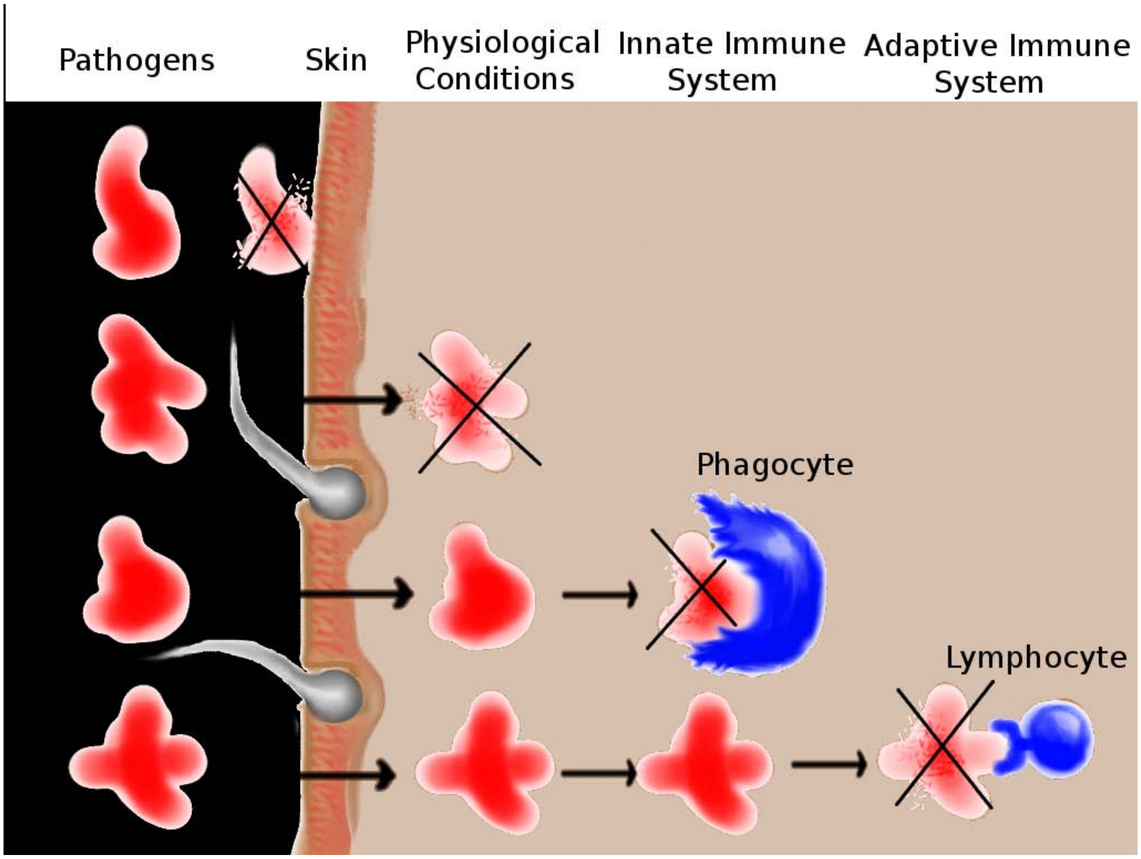

Figure 3.1: The three lines of defense of the immune system: physiological barriers, the innate immune system and the adaptive immune system.

two systems, roaming cells such as *macrophages* detect and engulf foreign antigens and other extracellular molecules [Hof01].

The complement system consists of small proteins and protein fragments that circulate in the plasma to bind and detect pathogens (mostly bacteria) and help eliminate them through *lysis* or *opsonization*. In lysis, the complement molecules destroy bacteria by rupturing their membranes, while in opsonization, the bacteria is coated by complement molecules (or antibodies) that help the detection of bacteria by macrophages. Macrophages are activated to engulf and destroy pathogens—they "eat" them. They also secrete signaling proteins called *cytokines* that induce an inflammatory response, characterized by fever, among other things. Fever increases body temperature and in turn speeds up the blood flow to help circulate recruited immune cells to sites of infection. One type of immune cells called *interferons*, when infected by viruses, produce proteins to inhibit viral replication and activate *natural killer cells* to kill virus-infected cells. Activated natural killer cells release chemicals in infected cells to trigger *apoptosis*, which is programmed cell death.

3.2.2 The Adaptive Immune System

The adaptive or acquired immune system, is a complex network of cells in vertebrate organisms, that learns to distinguish between harmless and harmful substances or *antigens*—usually fragments of proteins and certain types of polymers that can be recognized by the immune system. When harmful antigens from pathogens are discovered, an immune response to eliminate them is set in motion. Perhaps the key molecular components of the adaptive IS are *T-cells* (lymphocytes). It is in the thymus that T-cells develop and mature; only T-cells that are capable of recognizing antigens and those that fail to bind to self antigens (to prevent autoimmunity) are released from the thymus (maturation), while the rest of the T-cells are culled by *positive and negative selection* respectively [Cou80, PT93]. Mature T-cells are allowed out of the thymus to detect harmful nonself antigens. They do this by binding to *antigen presenting cells* (APC), typically macrophages and dendritic cells that collect and present substances through (Major Histocompatibility Complexes) MHC after breaking them chemically into antigen fragments (or epitopes). Specific T-cells are able to bind to the presented antigens when the *affinity* between them is high enough. The affinity is the binding strength between lymphocytes and antigens. Once the T-cells bind to the APC, they stimulate (recruit) another type of lymphocytes, *B-cells*, that are produced in the bone marrow. B-cells start a cascade of events, including a Darwinian process of B-cell generation via *somatic hypermutation*, leading to more *antibody* production. Antibodies are protein fragments similar to T-cell receptors produced by the immune system to bind and destroy pathogens or tumors linked to the antigens by macrophages and dendritic cells. However, it is possible that T-cells and B-cells mature before being exposed to all self antigens and therefore cause threat to some self antigens outside the thymus. Even more problematic is somatic hypermutation in clonal selection [Bur59] that ensues in lymph nodes, outside the thymus, after the activation of B-cells. At this stage, it is possible to generate many mutated lymphocyte clones that could bind to self antigens—(auto-immunity). One way around this is by a process called *co-stimulation* which involves the co-verification of B-cells by helper T-cells (Th cells) before the antigen is released from lymph nodes.

To insure that the T-cells do not also attack self, another type of T-cells known as *T regulatory cells*, are formed in a distinct cell lineage in the thymus where they mature to recognize self antigens [Sak04]. These regulatory T-cells have recently attracted a lot of attention after years in the "wilderness" [Sak04, JC06]. Regulatory T-cells were originally named *suppressor T cells* but their function was not fully understood and the field was discredited until their crucial role as regulatory cells in the immune system was better understood [M"ol88]. The detailed mechanisms in which regulatory T cells administer adaptive learning are still being studied [Rob07] and will be investigated in the following chapters. Nonetheless, there is enough evidence to show that regulatory T-cells are responsible for preventing autoimmunity by suppressing other T-cells that might bind to self antigens and thus recognize and attack harmless self cells [Sak04, CLC$^+$07]. The anatomy and functionality of the immune system is very briefly illustrated in Figure 3.2, described by Dasgupta [DN08] and Janeway [JTWS96] and revisited in section 3.3.2.

3.3 Artificial Immune Systems

Artificial Immune Systems (AIS) are adaptive systems, inspired by theories and observed principles of the immune system, and applied towards solving computational problems [DCT02b]. Common AIS techniques are based on specific theoretical models explaining the behavior of the vertebrate adaptive immune system such as: Negative Selection, Clonal Selection, Immune Networks and Danger Theory [Tim07].

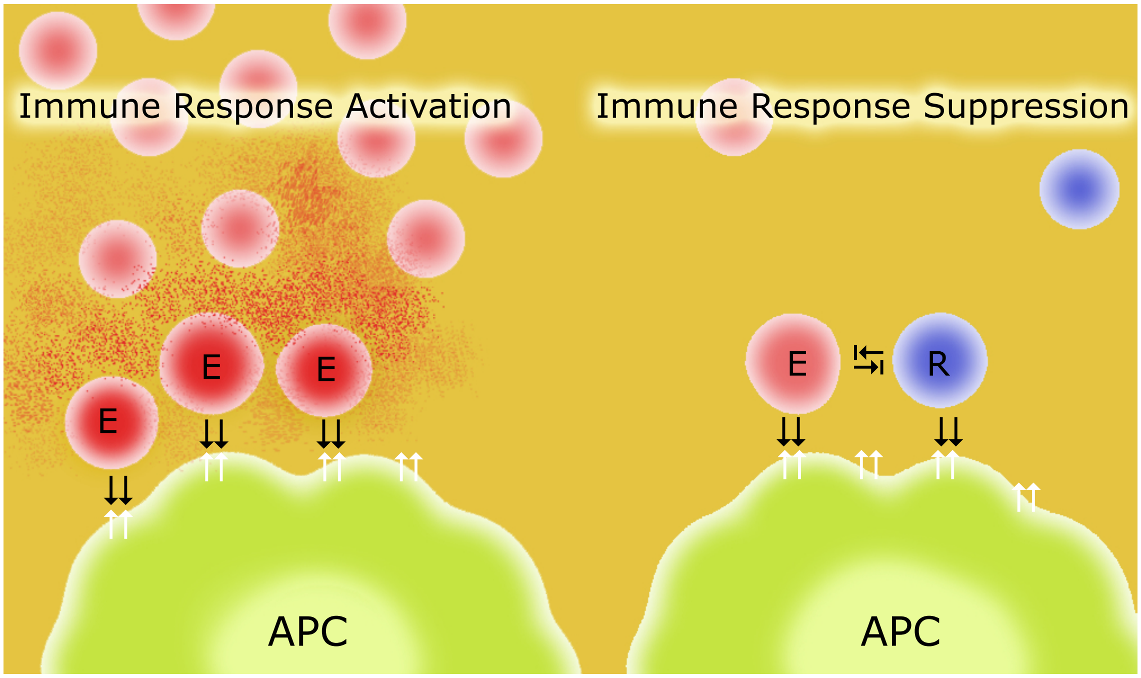

Figure 3.2: APC ingest protein fragments and present them as antigens to be recognized by T-cell receptors when the affinity between them is high enough. Effector (E) T-cells that bind to APC initiate several immune responses that may lead to autoimmunity if the recognized antigens are "self", as illustrated on the left. To lower the chances of autoimmune diseases, Effector T-cells are suppressed by another type of autoreactive T-cells (blue), know as Regulatory (R) T-cells, that are trained to recognize self antigens, as shown on the right.

AIS can be classified into two categories: (i) mathematically modeling the immune system (ii) a metaphorical inspiration to engineer algorithms that are capable of learning and solving a variety of machine learning problems such as classification, clustering and regression analysis. In this dissertation, we develop an agent based model of T-cell cross-regulation in the immune system (an example of the first category), and its application to text classification (an example of the second category).

3.3.1 A Brief History

AIS is a relatively new field which began in the mid 80's with the modeling and refinement of Jerne's immune network theory [Jer74] by Farmer et al. [FPP86], and later by Varela [VC91]. However, it wasn't until the mid-90's that AIS became a subject area when negative selection was used by Forrest [FPAC94] for protecting computer networks from viruses. Cooke and Hunt [CH95, HC96] adapted immune networks for classification and Timmis [TN01] further improved it while De Castro et al [dCT02a, dCVZ02] worked on aiNet for multimodal function optimization and data analysis. The first book on AIS was edited by Dasgupta [Das98].

3.3.2 Framework

According to Timmis [Tim07], AIS are defined by a framework of three basic elements: A representation of the system's **components**, a set of **mechanisms** that define the interaction of components with each other and the environment, and a set of procedures that govern the dynamics and behavior of the entire system, and that we call **model**.

Components

Traditionally, **components** are the AIS equivalents of antigens (Ag) and antibodies (Ab), and they are represented as vectors. Ab and Ag can be conceived as a key and a lock that need to have a matching pattern to bind. Ag and Ab can be binary (e.g. $\vec{Ab} = (0, 1, 1, 0, 0)$), integer ((e.g. $\vec{Ag} = (3, 14, 2, 0, 4)$), real ((e.g. $\vec{Ag} = (0.2, 0.6, 0.2, 0.0, 0.4)$) or string (e.g. $\vec{Ab} = (hello, this, is, an, example)$) vectors.

Mechanisms

Most of the cellular-interaction mechanisms are inspired by a key-and-lock interaction between the Antibody (Ab) and the Antigen (Ag) that can identify Ag's adversity based on the affinity between Ab and Ag. Ab can be though of as the system's trained army that can recognize Ag enemies. Perelson's shape-space model [PO79, Per89] presents the Ab as circles that can recognize circumscribed Ag based on their affinity. It is more common to use complementarity (or distance) measures for affinity rather than similarity measures. Following are examples of distance measures: i) the *euclidean distance* $D(\vec{Ab}, \vec{Ag}) = \sqrt{\sum_i (Ab_i - Ag_i)^2}$, ii) the *manhattan* or city-block distance $D(\vec{Ab}, \vec{Ag}) = \sum_i |Ab_i - Ag_i|$ or (iii) the *Hamming distance*, which counts the number of matches between $\vec{Ab}$ and $\vec{Ag}$. For example, let $\vec{Ab} = (0, 1, 1, 0, 0)$ and $\vec{Ag} = (1, 1, 1, 1, 0)$, then the Hamming distance between $\vec{Ab}$ and $\vec{Ag}$, $D(\vec{Ab}, \vec{Ag}) = 3$. $D(\vec{Ab}, \vec{Ag}) = 2$ for the previous example. There are several variations of the commonly used Hamming distance, for instance, *r-contiguous bits* [PPP93] give more importance to contiguous matched bits, whereas the *r-chunk bits* [BEFG02] is concerned with the matching of a window of a fixed number of bits.

Models

The most established theoretical **models** explored by AIS approaches are: *clonal selection* [Bur59] and *immune networks* [Jer74, FPP86, Par90, SC98], which are both B-cell inspired. T-cell inspired models, such as *negative selection* [Cou80, FPAC94] and *danger theory* [Mat94] have gained popularity, while the more recent cross-regulation model [CLC$^+$07] has only been proposed as a theoretical model with no applications developed other than the work presented in this dissertation. Many variations of these models (described below) and a plethora of other AIS models have been used for specific applications such as text classification.

- **negative selection** is the process of T-cell maturation that takes place in the thymus gland where T-cells are exposed to a repertoire of self antigens. T-cells that recognize and bind to self antigens are culled, whereas others mature and leave the thymus to recognize nonself antigens and participate in their elimination [Cou80, FPAC94].

- The **clonal selection** model is the oldest immune model first described by Burnet [Bur59] and inspired by Darwinian B-cell hypersomatic mutation and its ability to evolve new receptors or antibodies, *Ab*, capable of recognizing antigens, *Ag*. The initial population of *Ab* is randomly generated. *Ab* that recognize *Ag* with a certain affinity proliferate while the others that are not capable of recognizing any *Ag* slowly die out until the population of Ab is fine-tuned or trained to recognize as many *Ag* as possible [Das98].

- The **immune network** model was initially proposed to achieve immunological memory through a network of B-cells [Jer74]. The network is constructed by matching between its binding regions (known as *idiotopes* and *paratopes* and are comparable to *Ab* and *Ag*). The network reaches a stable state that is capable of recognizing nonself when its self-recognizing regions have all been regulated. The immune network model was later enhanced to focus more on systemic properties for understanding the maintenance of memory and repertoire selection [FPP86, VC91].

- Matzinger's **danger theory** [Mat94] offers an alternative perspective to the traditional self/nonself discrimination to detect dangerous intruders. The danger theory measures cell damage indicated by distress signals when cells die unnaturally (in contrast to natural cell death or *apoptosis*). The dendritic cell algorithm, which attracted attention recently, is based on danger theory where dendritic cells distinguish between safe signals represented by*apoptosis*, and danger signals represented by *Pathogen Associated Molecular Patterns* (PAMP) signals, which are biological signatures or motifs of potential intrusion [Gre03, GAT].

3.3.3 AIS in Text Classification

AIS has been applied in many areas such as robotics, control, optimization, anomaly detection, learning, and data mining [dCVZ, DCT02c], however we are specifically interested in AIS applications to text classification. A significant amount of AIS algorithms have been used for classification in general but few have focused on text mining and classification [WT04, KT01, TC02]. Even fewer AIS algorithms have attempted to address the challenges of concept drift.

For example, AISEC is a B-cell inspired AIS that is based on the clonal selection model. AISEC was tested on spam detection and showed to be comparable with Naive Bayes but not on a publicly available e-mail dataset [SFT03]. AISEC is the only example of an AIS dealing with concept drift to our knowledge. STRABNET tracks concept drift using antigen networks for spam detection but was tested using k-fold cross-validation disregarding time order [BB06]. In this dissertation we also study the importance of the order in which documents are presented to the classifier.

A fuzzy B-cell inspired AIS was used to detect session profiles from a web access log data. However, it was concerned with clustering and not classification, and it was limited to a fixed number of non-textual attributes in contrast to more numerous and unbounded textual attributes[2] [NGD02]. Another B-cell inspired algorithm, AIRS, was applied specifically to deal with class imbalance and predict the cause of "power distribution outage" using three major causes (tree, animal and lightning) as prototypes [Tay07]. Although AIRS outperformed artificial neural networks, the data it was applied to was limited to a fixed number of attributes (5) and did not address the dynamics of class imbalance.

We propose an ABM T-cell cross-regulation algorithm to address head-on the issues of concept drift and dynamic class imbalance for the classification of textual documents.

In the last couple of years, several review papers have discussed the slow advances in AIS and proposed improvement strategies through novel and simpler AIS models

[2] Numerous textual attributes are assumed to be better suited for studying concept drift since they are more adequate for describing concepts [Tsy04]

(inspired by the vertebrate innate immune systems and immune systems of plants) as well as the development of a unified architecture for integration of existing models [Das06, Tim07, TA07, HT08]. Dasgupta [DN08] concluded that negative selection algorithms suffer from scalability (for binary representation) and dimensionality issues (for real-valued representation), while algorithms inspired by clonal selection and artificial immune networks have been shown to be equivalent and very similar to evolutionary algorithms, with antibody somatic hypermutation instead of genetic variation [Gar03]. Hence, a novel and simple AIS model for text classification, in imbalanced and dynamic scenarios, would be very welcome by the AIS community.

While we are left with many challenges regarding the fundamental understanding of the immune system and its potential application to similar problems, we show that a novel ABM of T-cell cross-regulation, which uses only three interacting cellular components, is very promising for document classification in noisy and changing environments. We base our assessment not only on the novelty and simplicity of the model and the adaptability, decentralization and robustness of the vertebrate immune system, but also on the encouraging results we report here on spam detection and biomedical document classification.

3.4 The T-Cell Cross-Regulation Model

The T-cell Cross-Regulation Model (CRM) [CLC$^+$07] 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. Effector T-cells (E) proliferate upon binding to APC, unless adjacent to regulatory T-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. 3.3, can be simply expressed by the following interaction rules:

$$
\begin{aligned}
&(1)E_{\overrightarrow{d_E}}\{\} \text{ and } R_{\overrightarrow{d_R}}\{\}\\
&(2)A+R \rightarrow A+R\\
&(3)A+E \rightarrow A+2E\\
&(4)A+E+R \rightarrow A+E+2R
\end{aligned}
$$

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

Carneiro et al. [CLC$^+$07] developed the analytical CRM to study the dynamics of a population of T-cells and APC that recognize a single antigen. In [AHR08a, AHR08b],

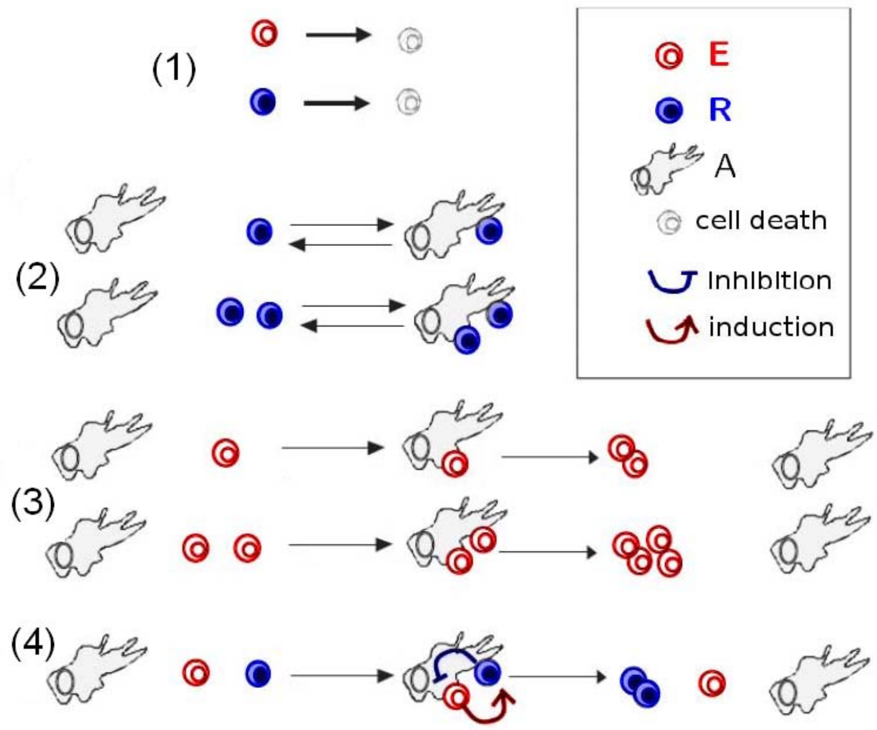

Figure 3.3: The diagram illustrates CRM interactions underlying the dynamics of APC, E and R as assumed in the model where APC can bind to a maximum of two T-cells.

we pushed the original CRM model into an agent based framework, to deal with multiple populations of antigens or features. Our agent-based model yielded encouraging results when applied to spam detection, a binary document classification problem. More recently, Sepulveda [Sep09, pp 111-113] extended the original analytical 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 contrast, our model [AHR08a, AHR08b] uses 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 chapter, we explain in more detail our agent-based model of T-cell cross-regulation.

3.5 Summary

In this chapter, we have provided an overview of the natural and artificial immune systems and more specifically the adaptive immune system and an analytical model of T-cell dynamics, which is of particular interest to our study.

Chapter 4

Agent Based Cross-Regulation Model

Ce qui m'intèresse, c'est qu'on vive et qu'on meure de ce qu'on aime.
What interests me is that we live and die of what we love.
Albert Camus in *The Plague*

In this chapter we present and study an agent-based model of T-cell cross regulation, which we apply to binary text classification in the following chapters. Adapting the CRM to an *Agent-Based Cross-Regulation Model* (ABCRM) for text classification is a nontrivial and risky task. The CRM is an analytical model that studies the T-cell dynamics of only one population of antigens and T-cells [CLC$^+$07]. The challenges range from expanding the simple model to deal with multiple populations of T-cells and antigens, mapping the cell-types in the CRM to corresponding elements in text mining using Agent-Based Modeling (ABM), and applying the four reaction rules (eq 1-4) while maintaining a bi-stable behavior of "healthy" and "unhealthy" states that translate to "relevant" and "irrelevant" classes respectively in text classification. Much of the algorithm design work that goes into these choices is unreported here, but the creation of a working ABM of T-cell cross-regulation is one of the key contributions of this dissertation.

4.1 Representation

In order to adapt the CRM to an ABCRM one can think of documents as analogous to the organic substances that upon entering the body are broken into constituent fragments. These fragments, known as epitopes, are presented on the surface of Antigen Presenting Cells (APC) as antigens. In the current application of 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 T-cells (E) and artificial Regulatory T-cells (R). Individual E and R have receptors for a single, specific (textual) feature: they are *monospecific*. E proliferate[1] 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 is produced when documents enter the artificial cellular dynamics, by breaking the former into constituent textual features. Therefore we can say that APC are representative of specific documents whereas E and R are representative of specific features.

More specifically, a document d contains a set of features F_d; an artificial APC $\mathbf{A_d}$ that represents d, presents a subset of antigens/features $A_d \subseteq F_d$ to artificial E and R T-cells. T-cell populations, E_f and R_f, bind to a specific feature f on any APC that presents it; if $f \in A_d$, then any available T-cell, E_f or R_f, in the cellular dynamics may bind stochastically to $\mathbf{A_d}$[2], as illustrated in figure 4.1.

[1] The simplification of proliferation to mere duplication adopted in the canonical CRM model is maintained in our agent-based model to minimize the number of parameters (excluding proliferation rates) and the parameter search space

[2] The probability of T-cells, E_f or R_f, binding to the APC that presents feature f strictly depends on the proportion of $|E_f|$ to $|R_f|$ T-cells available.

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. Affinity in our model is simply a binary function of whether the T-cells exactly match the presented antigens or not. This enables us to better understand the bare behavior of T-cell dynamics and leaves the more complex affinity measures to be explored in future work. APC are organized as a list of pairs of "slots" of (textual) features, where T-cells, specific for those features, can bind. We use this antigen/feature presentation scheme of pairs of "slots" to simplify our algorithm. In future work we will study alternative feature presentation scenarios. An APC is modeled as a list of "slots" of pairs of features: $\mathbf{A_d} = \mathbf{s_1} \cdots \mathbf{s_{n_S}}$, where a generic slot is defined as $\mathbf{s} = \langle f, g \rangle$, $f, g \in A_d$, and $n_S = \frac{n_A \times |A_d|}{2}$, where n_A is the number of times in which a feature is presented in slots. Features f (and g) are sampled without repetition from set A_d and randomly distributed exactly n_A times over the available slots that make up the APC (see Figure 4.1). Features are treated as a *bag of words*–i.e. the sequence of words in the document is not maintained [FS06]. Once T-cells bind to an APC, every pair of T-cells that bind to features on to the same slot s proliferate according to interaction rules (2-4) (see Section 3.4).In summary, each T-cell population is specific to and can bind to only one feature which can be presented by any APC.

The original analytical model (see Section 3.4) used differential equations to understand the behavior of T-cell dynamics for a single T-cell population. Implementing the algorithm as an ABM allows us to deal with the recognition and co-recognition (co-occurrence in the same document) of many features simultaneously, rather than a single one as the original CRM does. In our model, each artificial cell (APC or T-cell)

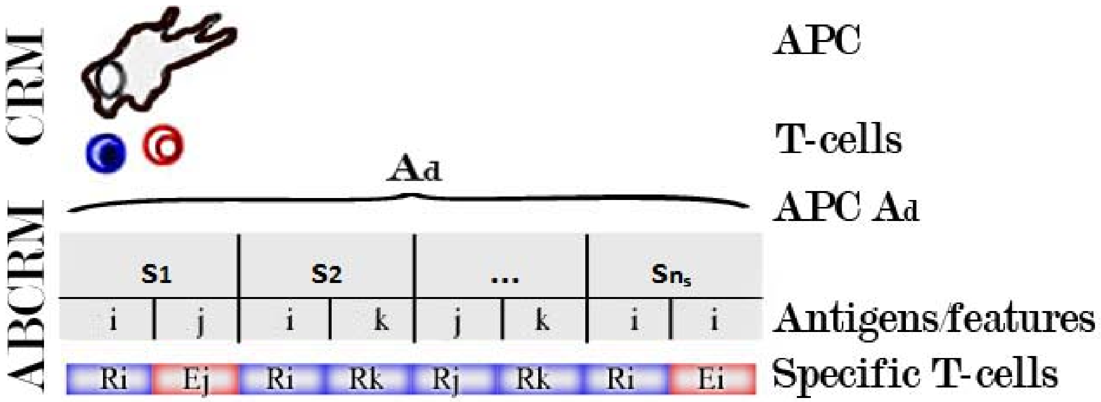

Figure 4.1: To illustrate the difference between the CRM and the ABCRM, the top part of the figure represents a single APC of the CRM which can bind to a maximum of two T-Cells. The lower part represents the APC for a document d in the ABCRM, which contains many pairs of antigen/feature "slots" where pairs of T-cells can bind. In this example, the first pair of slots of the APC $\mathbf{A_d}$ presents the features i and j; 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 regulatory T-cells R_i,R_k that duplicate via reaction (2), etc.

is considered an agent. T-cells of the same population are identical and specific to a unique feature that they can recognize and bind to when presented on an APC. R and E are distinguished in the same population by how they behave dynamically and interact with other cell types; specifically by having different death rates and *via* reaction rules (2-4) (see Section 3.4).

Agent-based modeling allows us to easily extend T-cell agents to have additional properties such as importance and affinity, which we will explore in future work. The importance factor can be based on whether a feature comes from a specific section (e.g. title or body) and the affinity could vary within the same population to include similar or synonymous features.

4.2 Training and Testing

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. Incremental learning is advantageous over traditional methods that require time-consuming re-training when dealing with live stream data or large data sets [SLS99]. It is from the dynamics defined by the ABCRM that the algorithm does not require the testing data to be preceded by the training data as done conventionally in supervised machine learning. A few labeled documents can be sufficient at the beginning for training the algorithm to classify unlabeled documents while capable of learning from additional labeled documents that may follow in order as we discuss in chapter 6. In both training and testing stages, documents are processed the same way, however, features that first occur in a labeled (relevant or irrelevant) or unlabeled document are treated differently as we discuss in the following section. The sequence in which documents are received affects the artificial cellular dynamics, as incoming APC and T-cells face a T-cell dynamics that depends on the specific documents previously encountered. Therefore, we use publication-time as the default ordering for incoming documents, and study if there is an advantage to preserving the original temporal sequence of articles (see Chapter 6). We explore alternative training regimes in future work. Figure 4.2 illustrates this stream of labeled documents (blue for relevant and red for irrelevant) followed by unlabeled grayed documents. A relevant (blue) document d is shown producing a polyspecific APC, $\mathbf{A_d}$ with features sampled from d. Monospecific T-cells bind subsequently stochastically to the features presented on $\mathbf{A_d}$. Newly introduced features f are biased with more R_f or more E_f based on whether they first occur in a relevant or irrelevant document respectively (see Figure 4.2). New features f occurring in unlabeled documents are treated as irrelevant features and biased with more E_f. Once the T-cells bind to the APC, they interact according to the last three reaction rules (see Section 3.4) and all remaining T-cells that do not bind to the APC undergo cell death with the death rates d_R and d_E. Finally, document d is classified as relevant if the majority of its features f have more R_f than E_f, and irrelevant otherwise (see Figure 4.4).

4.3 Cellular Dynamics and Parameters

Carneiro et al [CLC$^+$07] showed in the CRM that the cellular dynamics simulated *in silico* can lead to a bistable attractor of "healthy" and "unhealthy" states. Evidence of the bistable states *via* interactions between Effector and Regulatory T-cells underlying the crossregulation model are also relevant *in vivo* [APABD$^+$01].

The resulting state was shown to depend on the initial population proportions of E and R T-cells when the APC size is fixed [CLC$^+$07]. 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. Thus, "healthy" T-cell dynamics is identified by observing the co-existence of both E and R T-cells with $R \geq E$. "Unhealthy" T-cell dynamics, on the other hand, is identified by observing $E >> R$ or the disappearance of R T-cells. This can be the result of larger populations of E than R T-cells, or sufficiently large APC, in which case, E have higher chances of proliferating than R T-cells. That is because E T-cells can proliferate independently of neighboring T-cells according to interaction rule 3, however, R rely on E T-cells in on order to proliferate according to interaction rule 4.

In the text classification context, features associated with relevant documents should have more R than E T-cells in the artificial cellular dynamics whereas features associated with irrelevant documents should have many more E than R T-cells (see Figure 4.2). Therefore, in the *training phase*, we bias features that occur for the first time in relevant documents with more R T-cells whereas those from irrelevant documents we bias with more E T-cells. This initial bias might be erroneous as a relevant feature might first occur in an irrelevant document and *vice versa*, however, we assume that the collective dynamics of T-cells and feature co-occurrence in documents will automatically correct the bias, as we discuss and test in chapter 6.

In the *testing phase*, when an unlabeled document d contains features A_d that are specific mostly to E rather than R T-cells, we can classify it as irrelevant—and relevant in the opposite situation. Fig. 4.4 illustrates the classification example of the relevant document in Fig. 4.3 where APC is exemplified.

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 (testing) 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 maximum number of slots in which each feature f is repeatedly presented on APC

In the immune system, millions of novel T-cells are randomly generated in the thymus every day to attempt to predict future/unseen antigens. In our algorithm, in contrast, we generate T-cells only for features (e.g. words) occurring in the document corpus. This is reasonable because the space of meaningful words in a language

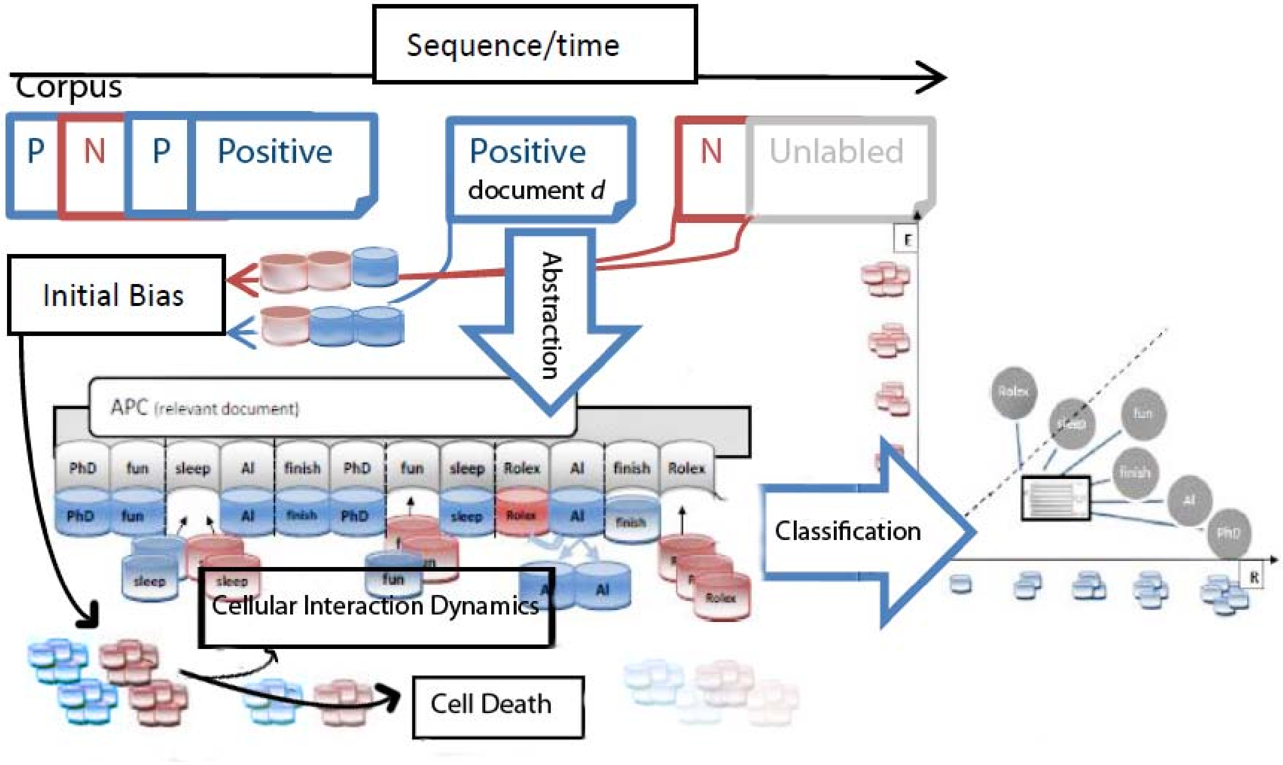

Figure 4.2: A stream of ordered labeled documents (blue for relevant and red for irrelevant) followed by ordered unlabeled grayed documents is introduced to the system. Each document d is represented by a polyspecific APC $\mathbf{A_d}$ that arbitrarily presents the features f of d as antigens such that the monospecific E_f (red cells) and R_f (blue cells) T-cells can recognize and bind to them. Newly introduced features f are biased with more R_f or more E_f based on whether they first occur in a relevant or irrelevant document respectively. New features f occurring in unlabeled documents are treated as irrelevant features and biased with more E_f. Once the T-cells bind to the APC, they interact according to the last three reaction rules (see Section 3.4) and all remaining T-cells that do not bind to the APC undergo cell death with the death rates d_R and d_E. Finally, document d is classified as relevant if the majority of its features f have more R_f than E_f, and irrelevant otherwise.

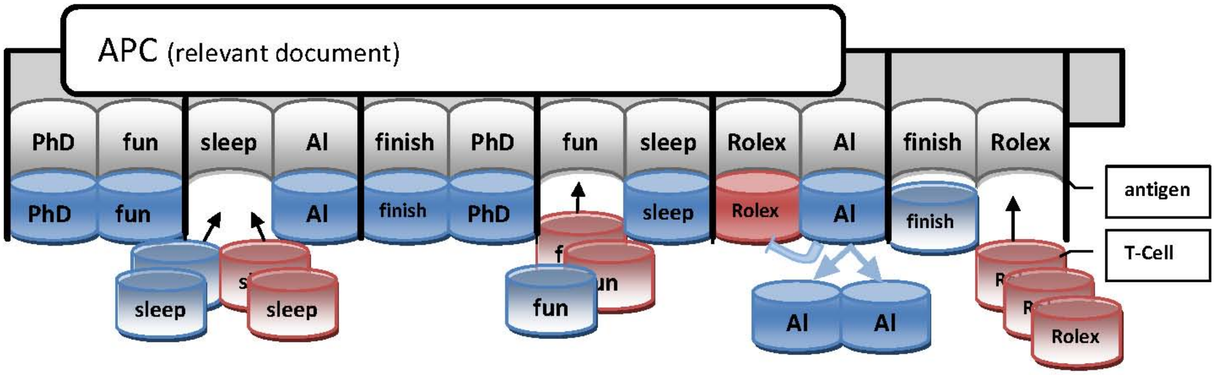

Figure 4.3: A polyspecific artificial antigen presenting cell $\mathbf{A_d}$ (representing document d) presents antigens/features $f \in F_d$ such that only monospecific artificial E_f and R_f T-cells can bind to it and proliferate according to equations (2-4) for every pair of antigens/features. In this example of a legitimate e-mail document, all features f are relevant (self) except for f_{Rolex}, which has relatively more E than R T-cells, since it tends to be associated with spam (nonself).

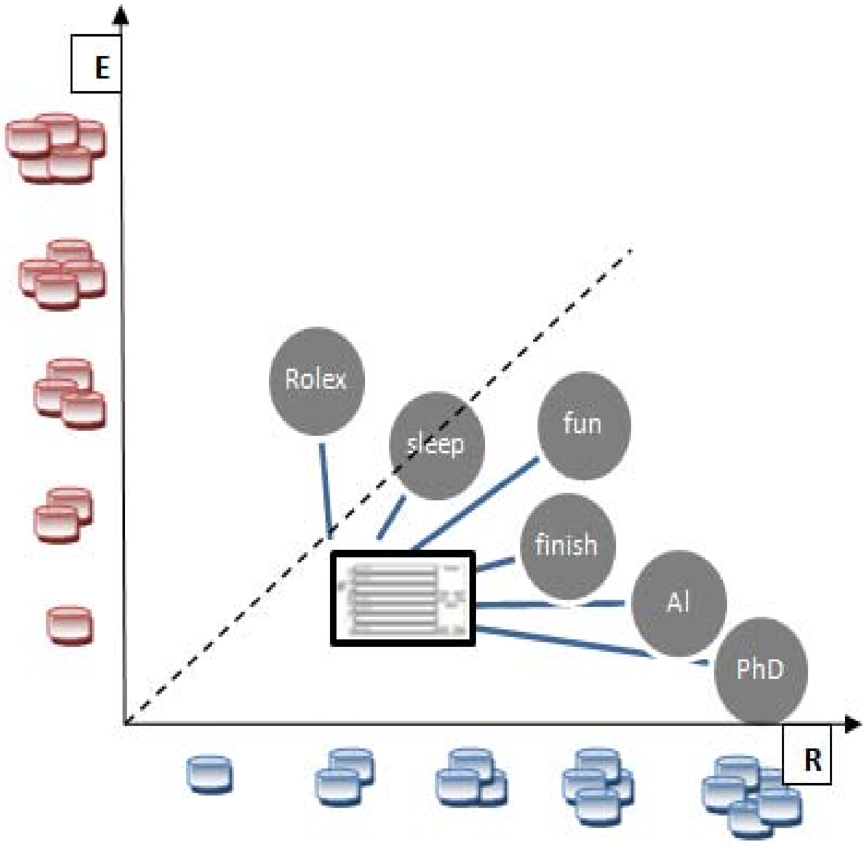

Figure 4.4: A document is assessed based on the E-to-R ratios for all its features. The axes represent the number of R and E cells. In this example, the document has most of its features with higher R and is therefore classified as relevant.

is largely fixed and much smaller than the space of possible polypeptide epitopes in biology. Nonetheless the space of possible word combinations, features, such as bigrams and n-grams quickly grows. In future work, we aim to study methods to generate new feature combinations for a much larger repertoire of antigens/T-cells.

When (textual) features are encountered for the first time, a fixed initial number of E_0 effector T-Cells and R_0 regulatory T-Cells is generated for every new feature f. Many factors such as APC size (determined by n_A) and death rates (d_E and d_R) play a huge role in the T-cell dynamics for each feature. However, in order to train on the basis of labeled documents, we rely only on the initial numbers of E and R T-cells by varying them for features occurring in relevant and irrelevant documents in the training phase, and unlabeled documents in the testing phase. More Regulatory than Effector T-cells are generated for features that occur in documents that are labeled relevant ($R_0^+ > E_0$), while fewer Regulatory than Effector T-cells are generated in the case of irrelevant or unlabeled documents ($R_0^- < E_0$) (see Fig. 4.2). Features appearing in unlabeled documents for the first time are treated as features from irrelevant documents, assuming that new features are irrelevant (nonself) until neutralized by the collective dynamics given their co-occurrence with relevant ones.

Naturally, relevant features will 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. Therefore, the proliferation dynamics defined by the 4 reactions and guided by co-binding to APC slots is expected to correct the erroneous initial bias, a hypothesis we will test in chapter 6.

In the original CRM model [CLC$^+$07], 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. In chapter 6, we also test the effect of cell death in the dynamics of our model when applied to biomedical document classification.

Finally, the decentralized and self-organized adaptive immune system is supposed to adapt to varying intrusions of self and nonself antigens and still be able to discriminate between them. Therefore, in chapter 5 we study the capability of our method to handle classification in changing environments such as varying class imbalance (by varying the class proportions of the validation data) and concept drift when applied to spam detection.

4.4 Algorithm

We implemented the ABCRM (using PHP 5.0) and ran many different experimental setups. The cellular interaction dynamics defined by the interaction rules (1-4) for a sequence of documents and its subsequent classification based on E-to-R ratios of its features is illustrated in Fig. 4.2. A detailed pseudocode of the algorithm follows:

Algorithm Pseudocode:

```
Input: Stream of labeled and unlabeled documents
Output: Labels for unlabeled documents
foreach document $d$ at time $t = t + 1$ do
    Generate APC $\mathbf{A_d}$ presenting each $f \in A_d$ at $n_A$ randomly distributed slot positions.
    Let $C_t$ contain $|E_k|$ and $|R_k|$ T-cells for all features $k$ in the cellular dynamics at time $t$.
    foreach $f \in A_d$ do
        if T-cells $E_f \not\subset C_t$ and T-cells $R_f \not\subset C_t$ then
            $|E_f| = E_0$ (i.e. generate $E_0$ Effector T-cells for $f$)
            if $d$ is labeled relevant then
                $|R_f| = R_0^+$ (i.e. generate $R_0^+$ Regulatory T-cells for $f$)
            end
            else
                $|R_f| = R_0^-$ (i.e. generate $R_0^-$ Regulatory T-cells for $f$)
            end
            Update $C_t$ with $E_f$ and $R_f$
            Let T-cells $E_f$, $R_f$ bind specifically to matching $f$ on $\mathbf{A_d}$
        end
    end
    foreach pair $(f, g)$ on $\mathbf{A_d}$ do
        Apply the interaction rules and update $C_{t+1}$ with the total number of $E$, $R$ T-cells:
        $(R_f, R_g) \to R_f + R_g$
        $(E_f, E_g) \to 2.E_f + 2.E_g$
        $(E_f, R_g) \to E_f + 2.R_g$
    end
    foreach T-cells $R_h, E_h \subset C_t$ that do not bind to $\mathbf{A_d}$ do
        Cull $E_h$ and $R_h$ according to death rates $d_E$ and $d_R$
    end
    if $d$ is unlabeled then
        Let $R(d) = \sum_{f \in A_d}(\frac{R_f}{\sqrt{R_f^2 + E_f^2}})$ and $E(d) = \sum_{f \in A_d}(\frac{E_f}{\sqrt{R_f^2 + E_f^2}})$
        if $R(d) \geq E(d)$ then
            Classify $d$ as relevant
        end
        else
            Classify $d$ as irrelevant.
        end
    end
end
```

The ABCRM runs with a complexity of $\mathcal{O}(N \cdot m + N \cdot |A_d| \cdot n_A)$ per iteration[3], where N is the number of documents, m is the maximum number of features in the corpus, $|A_d|$ is the number of features sampled per document and n_A is the number of positions in slots in which a feature is presented on the APC. The first part of the equation $N \cdot m$ corresponds to the cell death over all documents N for all inactive (not binding) T-cells specific to features m. The rest of the equation $N \cdot |A_d| \cdot n_A$ corresponds to the presentation of $|A_d|$ features each in n_A slots on $\mathbf{A_d}$ over all documents N.

Our parameter search for training the ABCRM is non-optimal as it exhaustively searches all parameter space requiring millions of iterations, however, this can be improved using heuristic search strategies in the future. Let n_A be a constant, where $n_A < 30$ (see Chapter 6 for the ranges of n_A explored). Let $|A_d|$ be another constant, where $|A_d| \leq 50$ features are sampled from every document (see Chapter 5). Therefore, the complexity grows as $\mathcal{O}(N \cdot m)$ which is faster than the quadratic optimization of SVM but slower than the linear computation of NB — hence, $\mathcal{O}_{\mathcal{NB}}(N) < \mathcal{O}_{\mathcal{ABCRM}}(N \cdot m) < \mathcal{O}_{\mathcal{SVM}}(N^x)$ where $x \in [1.8,2.1]$ [CRS03] for a fixed corpus size of many documents, assuming that the number of iterations needed to optimize the ABCRM is a constant. Note that m can also be fixed to a constant number of features, where $m = 650$ (see Chapter 6) includes the most important features for discriminating between protein-protein interaction relevant and irrelevant biomedical articles (see Figure 4.5).

[3] A number of iterations is initially required to optimize our algorithm by searching for the best configuration of parameters. For an exhaustive search on all 6 parameters, this number can be in the order of hundreds of thousands as we discuss in chapter 6

4.5 Application to Binary Document Classification

Our chief aim is to first establish a prototype of the ABCRM as a document classifier and then explore various parameters and representation scenarios to improve its classification performance and better understand the underlying mechanisms of T-cell cross-regulation dynamics and collective classification.

In chapter 5, we first test a prototype of the ABCRM on the binary classification problem of spam detection. The goal is to test if the ABCRM can classify. In other words, whether the collective behavior of T-cell cross-regulation dynamics as currently understood can function as a classifier. Since no model which includes multiple populations of T-cells has been previously studied, just showing that the system can classify is an advance in our understanding of T-cell cross-regulation dynamics as a general principle of classification in bio-complexity. Since we show that the ABCRM achieves promising results comparable to those obtained by a Naive Bayes classifier, the results are interesting from a bio-inspired computing viewpoint.

In chapter 6, we test the ABCRM on bio-medical document classification, another binary classification problem that is more relevant to our research interest. Moreover, we explore various parameters and configurations of our algorithm that optimize the ABCRM and provide insights about immune memory, cell death, concept drift, collective and self-organized behavior.

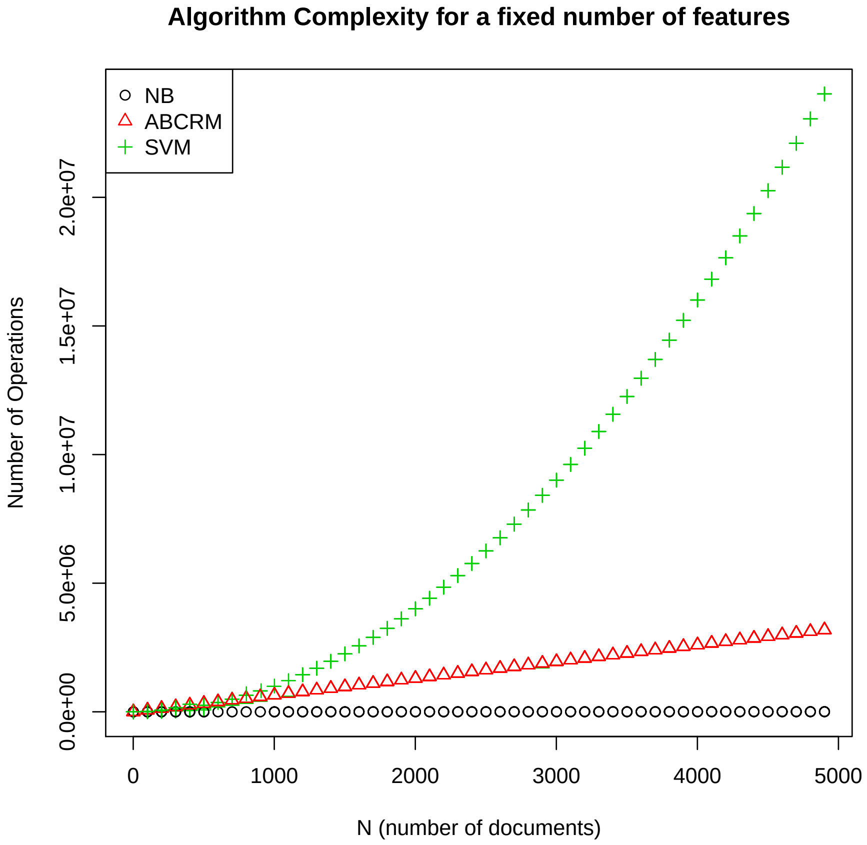

Figure 4.5: The complexities per iteration of NB, ABCRM and SVM for a fixed number of features $m = 650$, as done in chapter 6, and for the first 1000 documents leads to the inequality $\mathcal{O}_{\mathcal{NB}}(N) < \mathcal{O}_{\mathcal{ABCRM}}(650 \cdot N) < \mathcal{O}_{\mathcal{SVM}}(N^x)$ where $x \in [1.8{,}2.1]$, in this example $x = 2$.

Chapter 5

Application to Spam Detection

Los ordenadores son inùtiles. Sólo pueden darte respuestas.

Computers are useless. They can only give you answers

Pablo Picasso

In this chapter we test a prototype of our agent-based cross-regulation model on the publicly accessible Enron e-mail datasets (http://www.iit.demokritos.gr/ ionandr/publications/). The goal is to test our method's ability to classify, therefore we compare its performance with that of a Naive Bayes classifier. We also study the ability of our method to track concept drift, its resilience to spam-to-ham ratio variations and its ability to generalize on new data. This chapter is adapted from relevant published articles [AHR08b, AHR08a].

5.1 Introduction

Spam detection is a binary classification problem in which e-mail is classified as either ham (legitimate e-mail) or spam (illegitimate or fraudulent e-mail). Spam is very dynamic in terms of advertising new products and finding new ways to defeat anti-spam filters. The challenge in spam detection is to find the appropriate threshold between ham and spam leading to the smallest number of misclassifications, especially of legitimate e-mail (false negatives).

The vertebrate adaptive immune system learns to discriminate between self and nonself substances (known as pathogens) such as viruses and bacteria that intrude the body. These pathogens often evolve new mechanisms to attack the body and its immune system, which in turn adapts and evolves to deal with changes in the repertoire of pathogen attacks. A weakly responsive immune system is vulnerable to attacks while an aggressive one can be harmful to the organism itself, causing autoimmunity, described by reactions against self antigens. Given the conceptual similarity between the problems of spam detection and immunity, we investigate the applicability of our agent-based cross-regulation model (ABCRM) to spam detection.

Machine learning techniques such as support vector machines [KA01], Naive Bayes classifiers [SDHH98, MAP06] and other classification rules have excelled in textual e-mail classification. However, most of these algorithms generally lack the ability to detect *concept drift* since they rely on fixed training data [Tsy04] as described earlier (see Chapter 2).

Concept drift is very common in spam due to new advertisements and spam obfuscation techniques such as Bayesian poisoning [MFRI$^+$06, GC06]. Research in spam detection is now focusing on detecting concept drift, for instance, ICBC is a cluster-based classification method that uses incremental learning mechanisms to adapt to concept drift [DCS06]. An improved SpamHunting system uses instance-based reasoning based on a tunable instance retrieval network for e-mail selection to outperform classical machine learning techniques [FRID$^+$07a]. Another challenge in spam detection is marked by the inconsistency of spam-to-ham ratios between training and testing data: dynamically imbalanced classification (see Chapter 2). Algorithms based on Artificial Immune System (AIS) [Oda05, BB06] are another area of exciting development. AIS models are inspired by diverse models of the adaptive immune system [Hof01] such as negative selection, clonal selection, danger theory and the immune network theory (see Chapter 3).

Since our ABCRM is quite compelling in the simplicity by which it achieves discrimination between self and nonself antigens (see Chapter 4), we expect our method to be useful for e-mail binary classification. Moreover, its dynamic nature, in principle, makes it a good candidate algorithm to deal with concept drift in e-mail classification. Therefore, we study the ability of our model to track concept drift by measuring the drop in classification performance over time. We also study the resilience of our adaptive model to spam-to-ham ratio variation in comparison to a Naive Bayes classifier. In summary, our goal is to establish an immune-inspired document classifier and in future work, we plan to compare our method with other classifiers that deal specifically with concept drift.

5.2 ABCRM Configuration for Spam Detection

Here we discuss some of the techniques used for feature selection and parameter configuration of our ABCRM when applied to spam detection.

5.2.1 Text Processing and Feature Selection

Spam detection often focuses on e-mail headers (e.g. sender, receiver, relay servers...) or textual content (e.g. subject, body). However, we are interested only in the latter for understanding the applicability of our method to text document classification. All words constituting the e-mail subject and body are lowercased and stemmed using Porter's algorithm [Por80b] after filtering out common English stop words and words of length less than 3 characters (see Chapter 3). A maximum of $|A_d|$ processed unique features are sampled and presented by the artificial APC that corresponds to the e-mail document (see Chapter 4). These antigen presenting cells have n_A slots per feature f (that specific E_f and R_f can bind to). Examples of APC representing ham and spam e-mails are illustrated in Figure 5.1. The breaking up of the e-mail message into constituent portions (features) is inspired by the natural process in biology, but is simplified in this model to select the first and last $\frac{|A_d|}{2}$ unique features in the e-mail. The assumption is that the most indicative information is in the beginning (e.g. subject) and the end of the e-mail (e.g. signature), especially concerning ham e-mails.

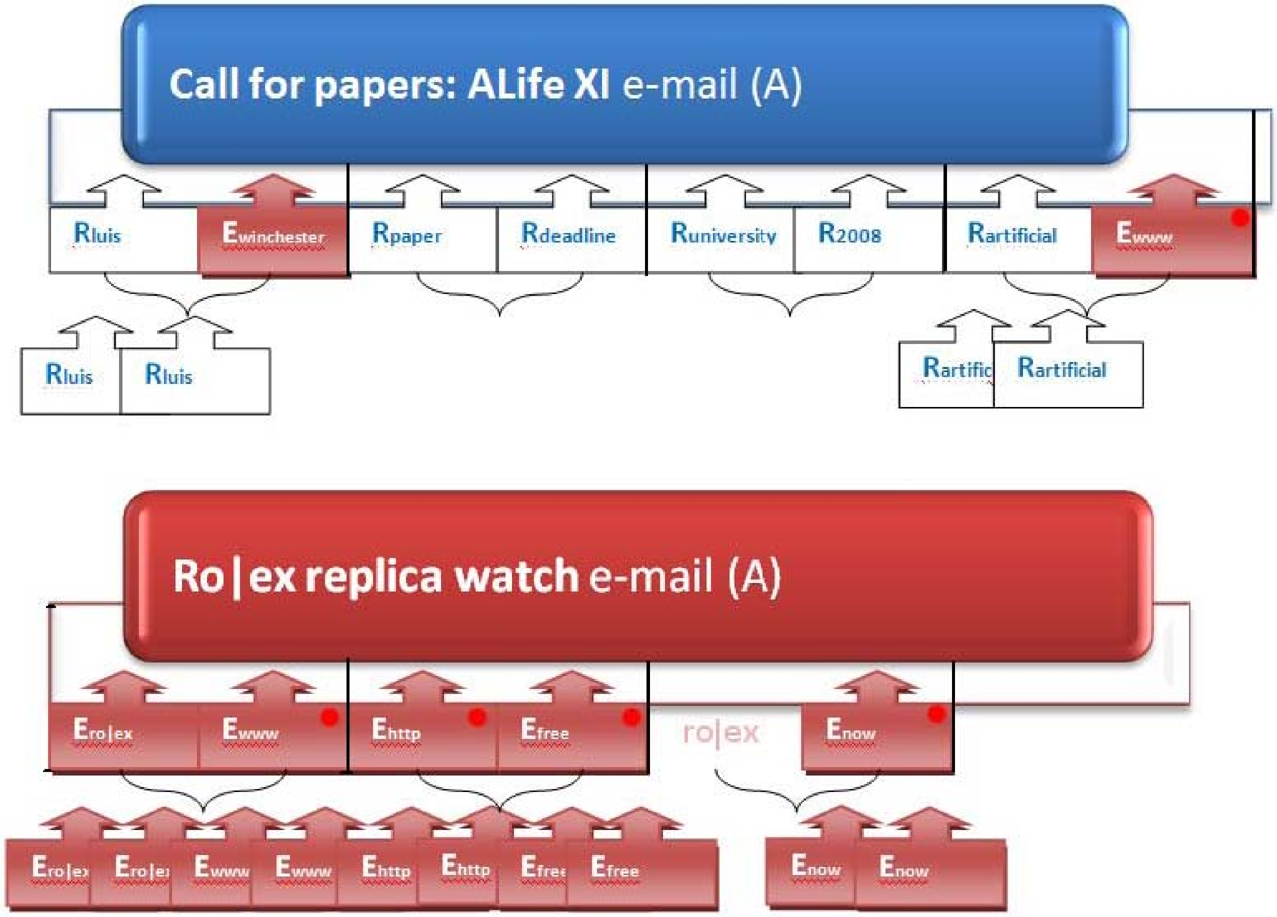

Figure 5.1: An example of APC representing ham (upper) and spam (lower) e-mails with T-cells binding specifically to features presented on APC. In the case of the ham e-mail, new features with relatively large populations of E T-cells (e.g. "winchester") become relevant by co-occurring with relevant features with relatively large populations of regulatory T-cells, that suppress the proliferation of neighboring E T-cells (e.g. $E_{winchester}$ and E_{www}). However, spam features such as "www", have their effector T-cells (e.g. E_{www}) proliferate in spam e-mails with many neighboring (on the APC) E T-cells from co-occurring spam features.

5.2.2 Settings and Parameters

Let $|A_d| \leq 50$ be the first and last non-overlapping unique features selected from each e-mail. Let $n_A = 10$ be the number of APC slots per feature and Let $E_0 = 6$, $R_0^+ = 12$, and $R_0^- = 5$ (see Chapter 4 for definitions of these parameters). These initial E and R populations for features occurring for the first time, are based on the initial ratios used in the CRM [CLC$^+$07, p. 6], given a fixed APC size. Different parameter values can lead to similar, better or worse classification performances, therefore our choice of parameters is not necessarily optimal. While our main goal here is to establish an original agent-based bio-inspired model and test its applicability to text classification, we explore various configurations and parameter ranges in the following chapter on another application. We use the same cellular dynamics and classification criteria described in Chapter 4 to classify an e-mail document as either spam or ham (see Fig. 5.2 for a graphical description).

5.3 Validation

In order to validate our method as a classifier, we train it on a batch of 200 e-mails and then test it on a distinct batch of e-mails that varies in size and class imbalance based on validation forms in order to test for our method's robustness to dynamic class imbalance and it's ability to track concept drift.

Three forms of validation are conducted: *random partition validation*, that is similar to a 2-fold cross-validation, *sequential partition validation* for evaluating the importance of the order in which e-mails are presented to the system as well as

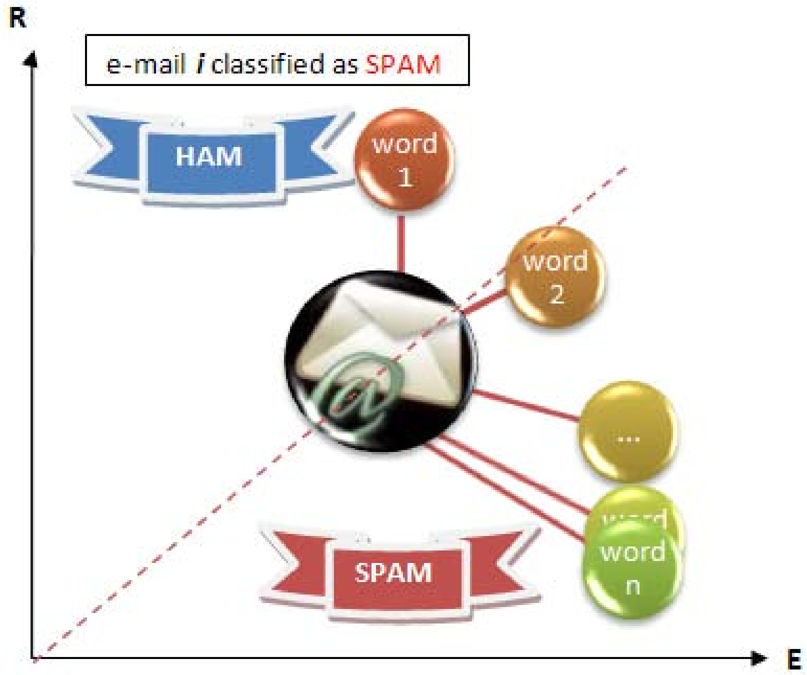

Figure 5.2: Spam classification is based on R-to-E ratios for all features sampled from the e-mail document as described in Chapter 4.

unbalanced validation scenarios, and *dynamic validation* using a sliding window, that is particularly useful to study concept drift in spam and ham over time. For every validation scenario, we do the following:

For each of the six Enron sets, we obtain 10 independent partitions. Each partition is balanced and consists of 200 training (50% spam) and 200 validation e-mails (50% spam) .

5.3.1 Random Partition Validation

In this case, we do not use time-sequence information: the 10-subset partitions are randomly sampled and therefore we can sample many variations of these partitions. We produce five different 10-subset partitions and compute performance statistics (e.g. mean and standard deviation).

5.3.2 Sequential Partition Validation

In this case, the partitions are chosen consecutively according to time-stamp (see Fig. 5.3). Each partition consists of 200 training and 200 validation e-mails that follow in the order of time the email was received. The training data is balanced with 50% spam whereas the validation data is studied for balanced and unbalanced scenarios as follows:

Balanced Scenario

Validation data is balanced (50% spam).

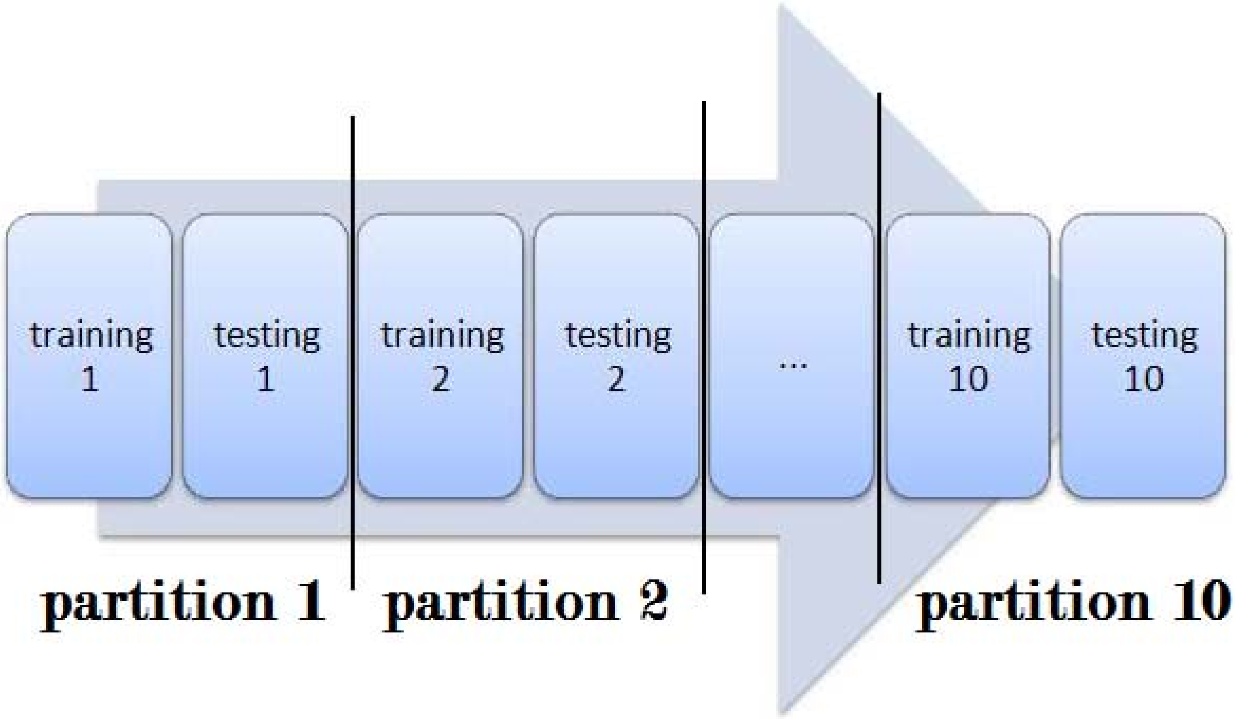

Figure 5.3: Sequential partition validation uses a partition of documents for training and a distinct partition of e-mails that follow in time-stamped order for testing. The partitions are non-overlapping. The size of each training and testing partition is 200 e-mails.

Unbalanced Scenario

Validation data is unbalanced in two different cases: 30% and 70% spam. In the case of 70% spam or 70% ham, the additional e-mails are from the partition of e-mails that follows in time order. Since each partition is independent, training and testing in one partition does not affect performance in the one that follows.

5.3.3 Dynamic Validation

In this validation scenario, for each of the six Enron sets, we train each classifier on the first 200 e-mails (50% spam) and then test on overlapping windows of 200 testing (50% spam) e-mails that follow in the order of time the email was received. The sliding shift is 10 e-mails and the range is between e-mail 201 and e-mail 2800 resulting in 260 slides (see Fig. 5.4). We report classification performance for each slide and the average performance over all slides for each Enron dataset as discussed next.

5.4 Performance Evaluation

We compute variation statistics (mean, standard deviation and 95% CI) of the F-score and Accuracy measures for each Enron data set by averaging the performance over all partitions [FS06]. We use the paired student t-test to compare the classification performances of two methods in order to accept or reject the null hypothesis H_0 that assumes that the two performance results were drawn from the same distribution. The rejection of H_0 depends on the p-value (p) and can mean that one method

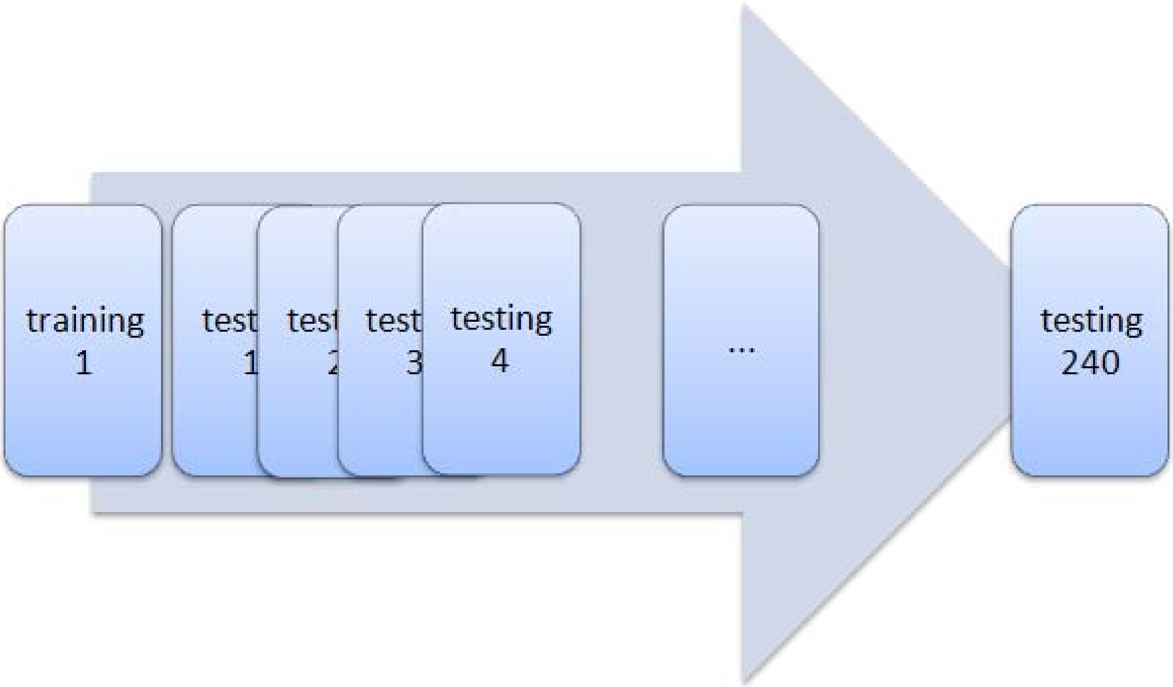

Figure 5.4: Dynamic evaluation of stream data using a sliding window for the first training set and testing sets that follow in time-stamped order. The size of each training and testing set is 200 e-mails.

moderately outperforms the other with $p < 0.05$, or strongly outperforms the other with $p < 0.01$. Otherwise, both methods are statistically indistinguishable (tied). Our statistical comparisons of $p < 0.05$ and $p < 0.01$ are presented in the last two columns of the classification performance tables. Furthermore, our statistical comparison for $p < 0.05$ is confirmed with visual comparisons using the 95%CI, that is illustrated in Figures for some of the experiments.

5.4.1 Sequential Partition Validation

In the case of unbalanced data, we evaluate only a balanced selection of the results. In other words, after classification, we under-sample the documents from the class with more instances (70%) to make it equivalent to the class with fewer instances (30%), since F-score and Accuracy can be biased in unbalanced scenarios. In the following chapter, we use better metrics such as AUC for measuring performance in unbalanced classes.

5.4.2 Dynamic Validation

In order to measure the decline in classification performance, which is a measure of concept drift, we also perform a linear regression of accuracy and F-score, using least squares and we compute the slope coefficients and the coefficient of determination R^2 for each.

5.5 Data

Given the assumption that personal e-mails (i.e. e-mails sent or received by one specific user) are more representative of a writing style, signature and themes, it would be preferable to test the ABCRM on e-mails from a personal mailbox. Unfortunately, this is not offered by the most common spam corpus of *spamassasin*[1] and similarly for *ling-spam*[2]. In addition, the ABCRM algorithm requires time-stamped e-mails, since order of arrival affects final E and R populations. Time-stamped data is also important for analyzing concept drift over time, thus we cannot use the *PU1*[3] data described by Androutsopoulos et al. [AKCS00]. Delany's e-mail data set[4], introduced by Delany et al. [DCTC05a], meets the requirements in terms of time-stamped and personal ham and spam for two personal mailboxes and spam, however its features are hashed and therefore it is not easy to make tangible conclusions based on their semantics. In this chapter, we show trajectories of "healthy" and "unhealthy" T-cell populations for ham and spam features over time that are only meaningful using original words in lieu of hash numbers.

The *enron-spam*[5] preprocessed data perfectly meets the requirements in terms of being personalized, time-stamped, and in comparison to the data publicized by Delany et al. [DCTC05a], it is a larger data set and it keeps the original words. The enron-spam data is composed of six personal mailboxes that were made public after the Enron scandal. The ham mailboxes belong to the following Enron employees:

[1] http://spamassassin.apache.org/publiccorpus/

[2] http://www.aueb.gr/users/ion/publications.html

[3] http://www.iit.demokritos.gr/skel/i-config/downloads/enron-spam/

[4] http://www.comp.dit.ie/sjdelany/Dataset.htm

[5] http://www.iit.demokritos.gr/˜ionandr/publications/

Dataset	ham + spam	time-stamp range
Enron1	farmer-d + g	[12/99, 06/00], [12/03, 01/05]
Enron2	kaminski-v + sh	[12/99, 05/00], [05/01, 07/05]
Enron3	kitchen-l + b	[2/01, 06/01], [08/04, 03/05]
Enron4	williams-w3 + g	[4/01, 01/02], [12/03, 06/04]
Enron5	beck-s + sh	[1/00, 11/00], [05/01, 03/05]
Enron6	lokay-m + b	[6/00, 7/01], [08/04, 10/04]

Table 5.1: The Enron-spam preprocessed datasets

farmer-d, kaminski-v, kitchen-l, williams-w3, beck-s and lokay-m. Combinations of four spam datasets were added to the ham data from *spamassassin* (s), *HoneyProject* (h), *Bruce Guenter* (b) and *Georgios Paliousras'* (g) spam corpora and then all six datasets were tokenized [MAP06]. In practice, some spam e-mails can be personalized, which unfortunately cannot be captured in this dataset since the spam data comes from different sources. The 6 Enron data sets vary in the number of e-mails they contain, therefore we trim them such that only the first 1500 e-mails of every Enron dataset are used to facilitate comparison in the following experiment.

5.6 Results

We report on the classification performance of a simple prototype of the ABCRM and compare it with a multinomial NB classifier. We compare both classifiers when trained on balanced data and tested on balanced and unbalanced data in order to understand their resilience to unforeseen class imbalance. Then, using the dynamic evaluation, we compare the classification performance of both classifiers over time knowing that a drop in the performance indicates the presence of concept drift. Finally, we test for the generalization of both methods on data distinct from the training data.

Dataset		**ABCRM**	**Naive Bayes**	p<0.01	p<0.05
Enron 1	F-score	0.87 ± 0.02	0.94 ± 0.02	NB	
	Accuracy	0.86 ± 0.02	0.94 ± 0.02	NB	
Enron 2	F-score	0.88 ± 0.04	0.94 ± 0.01	NB	
	Accuracy	0.86 ± 0.05	0.94 ± 0.01	NB	
Enron 3	F-score	0.87 ± 0.01	0.94 ± 0.02	NB	
	Accuracy	0.86 ± 0.01	0.94 ± 0.02	NB	
Enron 4	F-score	0.88 ± 0.03	0.93 ± 0.05	tie	NB
	Accuracy	0.88 ± 0.03	0.93 ± 0.04	NB	
Enron 5	F-score	0.92 ± 0.02	0.95 ± 0.02	tie	NB
	Accuracy	0.92 ± 0.02	0.95 ± 0.02	NB	
Enron 6	F-score	0.85 ± 0.05	0.91 ± 0.04	NB	
	Accuracy	0.82 ± 0.08	0.92 ± 0.04	NB	
Average	**F-score**	**0.88 ± 0.04**	**0.93 ± 0.03**	**NB**	
	Accuracy	**0.87 ± 0.05**	**0.93 ± 0.03**	**NB**	

Table 5.2: Results for the random partition validation. F-score and Accuracy mean ± sdev of 10 randomly sampled partitions for 50% spam ratio Enron data sets for ABCRM and Naive Bayes.

In the first validation experiment, NB outperforms the ABCRM statistically for all Enron data sets as shown in the last two columns of Table 5.2. However, in the second experiment, which preserves the time order in which e-mails were received, NB and the ABCRM are more competitive, where NB outperforms the ABCRM in only three Enron data sets and ties with the ABCRM in the remaining three as shown in the last two columns of Table 5.3. When comparing NB with the ABCRM, the two-tailed 95%CI illustrated as boxes in Figure 5.5 show similar results to the paired student t-test comparisons presented in table 5.3. Figure 5.5 shows that NB outperforms the ABCRM in Enron 3 in terms of F-score and in Enron 2 in terms of accuracy, whereas the two classifiers are comparable for the remaining Enron sets.

Moreover, when testing on unbalanced data, the ABCRM outperforms NB in three Enron data sets for 30% spam and ties with NB in the remaining three, while it outperforms NB in five and loses to two Enron data sets for 70% spam as shown in the last two columns of Tables 5.4 and 5.5 respectively. Therefore the results suggest that the ABCRM can be more resilient to spam-to-ham ratio variations than NB. The two-tailed 95%CI values illustrated as boxes in Figure 5.6 confirm most of our statistical comparisons with the ABCRM outperforming NB for Enron sets 1 and 6 in terms of F-score and Enron 6 in terms of accuracy for 30% spam. Similarly, ABCRM outperforms NB in Enron 1 in terms of F-score and Enron sets 1 and 6 in terms of accuracy for 70% spam. In the remaining case ABCRM and NB are comparable.

The improved performance of the ABCRM, in the sequential over the random partition validation, unveils the important role of the sequential order of documents to which only dynamical systems are sensitive to. In the light of the above, we assume that the order in which the documents are processed can provide our method valuable information to track concept drift. We further investigate problem of concept drift in the dynamic validation, by measuring the drop of classification performance over time, and in Chapter 6, by studying various document-shuffling scenarios and comparing them to that of time-stamp ordered documents.

While the overall performance of both algorithms was comparable for the sequential validation data with an advantage for NB for 50% spam for all Enron datasets (see Figure 5.7), the average performance of NB drops for 30% spam ratio (5% lower F-score than ABCRM on average) and 70% spam ratio (9% less accurate than ABCRM on average) as reported in Tables 5.4 and 5.5, however, the ABCRM maintains a relatively good performance.

Arguably, the performance of NB could be increased, in the unbalanced spam-to-ham ratio experiments, by unbalancing the Naive Bayes equation (see Chapter 2). But this would imply that, in real situations, one would need to know *a priori* the spam-to-ham ratio for a given user. The ABCRM model, on the other hand, does not need to adjust any parameter for different or varying spam-to-ham ratios—it is automatically reactive to whatever ratio it encounters. It has been shown that spam-to-ham ratios indeed vary widely [MW04, DCTC05a, DCT06], hence we conclude from the ability of the ABCRM to automatically handle spam-to-ham ratio variation that our method is more beneficial for dynamic data classification, and this should be further investigated (see Chapter 6).

Natural cell death is supposed to play an important role in immune memory, however, here we implement the prototype of the ABCRM without cell death (i.e. $d_E = 0$ and $d_R = 0$). Nevertheless, we explore ranges of cell death rates and study their effect on immune memory and learning in Chapters 6 and 7 respectively.

5.6.1 Dynamic Evaluation Results.

In this experiment, we study the performance of our method over time after training on the first 200 e-mails of each Enron data set. In figure 5.8, we observe T-cell dynamics leading to “healthy” and “unhealthy” states for the features ‘rolex’, which is obviously spam and ‘fyi’ (i.e. “for your information”), which appears to be one of the most ham words besides the word ‘enron’. These trajectories of T-cell populations are similar to the ones reported in the CRM with $E >> R$ for “unhealthy” and $R \geq E$ for “healthy”.

Dataset		**ABCRM**	**Naive Bayes**	$p<0.01$	$p<0.05$
Enron1	F-score	0.9 ± 0.03	0.89 ± 0.04	tie	tie
	Accuracy	0.9 ± 0.03	0.87 ± 0.05	tie	tie
Enron2	F-score	0.86 ± 0.06	0.92 ± 0.07	NB	
	Accuracy	0.85 ± 0.06	0.93 ± 0.05	NB	
Enron3	F-score	0.88 ± 0.04	0.93 ± 0.03	NB	
	Accuracy	0.87 ± 0.05	0.92 ± 0.04	tie	NB
Enron4	F-score	0.92 ± 0.05	0.92 ± 0.05	tie	tie
	Accuracy	0.92 ± 0.05	0.91 ± 0.06	tie	tie
Enron5	F-score	0.92 ± 0.03	0.94 ± 0.04	tie	NB
	Accuracy	0.91 ± 0.03	0.95 ± 0.03	NB	
Enron6	F-score	0.89 ± 0.04	0.91 ± 0.02	tie	tie
	Accuracy	0.88 ± 0.05	0.9 ± 0.03	tie	tie
Average	**F-score**	**0.9 ± 0.05**	**0.92 ± 0.04**	**tie**	**tie**
	Accuracy	**0.89 ± 0.05**	**0.91 ± 0.05**	**tie**	**tie**

Table 5.3: Results for the balanced sequential partition validation. F-score and Accuracy mean ± sdev of 10 sequential partitions for 50% spam ratio Enron data sets for ABCRM and Naive Bayes.

Dataset		**ABCRM**	**Naive Bayes**	$p<0.01$	$p<0.05$
Enron1	F-score	0.93 ± 0.02	0.88 ± 0.04	ABCRM	
	Accuracy	0.89 ± 0.03	0.85 ± 0.04	tie	ABCRM
Enron2	F-score	0.88 ± 0.02	0.88 ± 0.03	tie	tie
	Accuracy	0.82 ± 0.04	0.85 ± 0.03	tie	tie
Enron3	F-score	0.9 ± 0.02	0.87 ± 0.03	tie	ABCRM
	Accuracy	0.85 ± 0.04	0.84 ± 0.04	tie	tie
Enron4	F-score	0.93 ± 0.04	0.89 ± 0.12	tie	tie
	Accuracy	0.9 ± 0.07	0.88 ± 0.12	tie	tie
Enron5	F-score	0.9 ± 0.03	0.88 ± 0.04	tie	tie
	Accuracy	0.85 ± 0.05	0.86 ± 0.04	tie	tie
Enron6	F-score	0.9 ± 0.02	0.77 ± 0.04	ABCRM	
	Accuracy	0.85 ± 0.03	0.75 ± 0.04	ABCRM	
Average	**F-score**	**0.91 ± 0.03**	**0.86 ± 0.07**	**tie**	**tie**
	Accuracy	**0.86 ± 0.05**	**0.84 ± 0.07**	**tie**	**tie**

Table 5.4: Results for the unbalanced sequential partition validation (30% spam). F-score and Accuracy mean ± sdev of 10 partitions for 30% spam for ABCRM and NB.

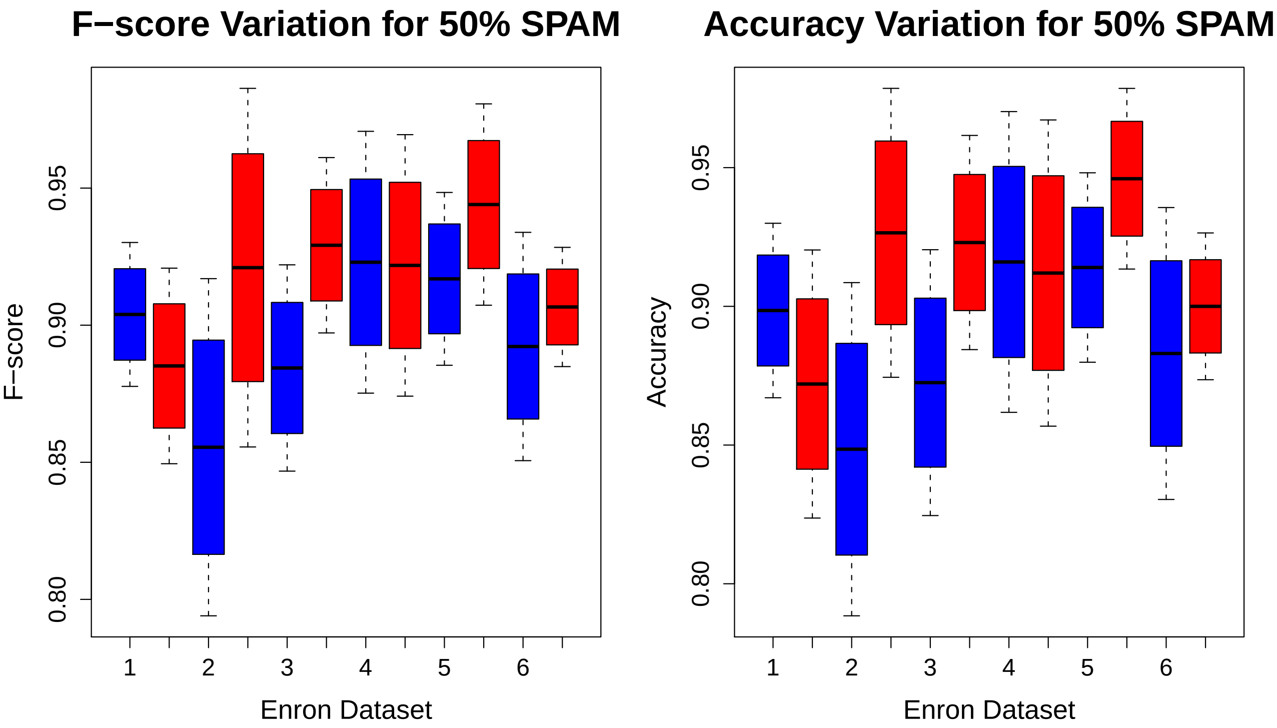

Figure 5.5: Results for the balanced sequential partition validation. F-score and Accuracy variation plots for ABCRM (blue) and NB (red) for each of the six Enron datasets. The black horizontal bars indicate the mean, the boxes indicate the 95%CI and the whiskers indicate the standard deviation. These results are for the static evaluation scenario. The values are reported in Table 5.3.

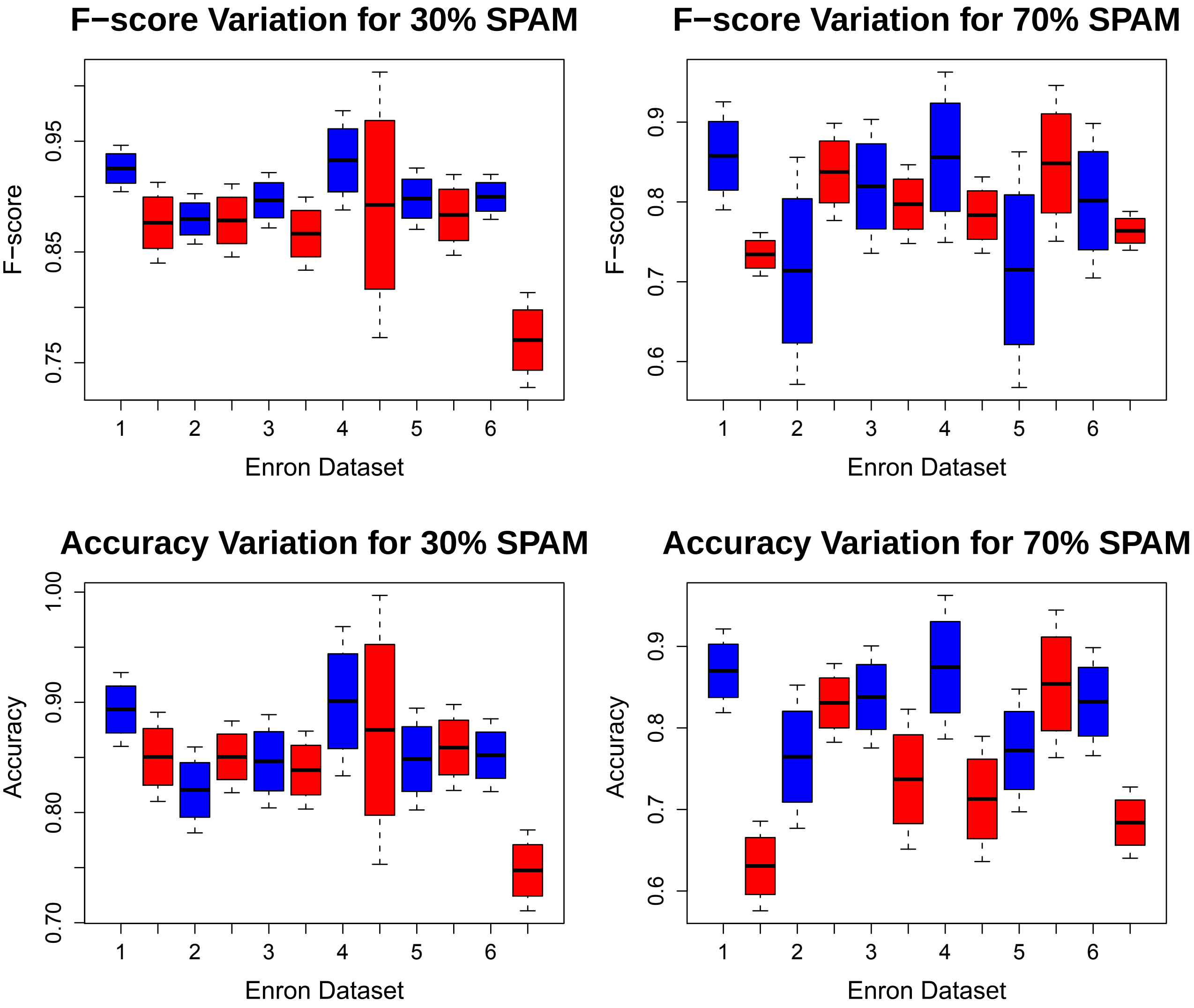

Figure 5.6: Results for the unbalanced sequential partition validation. F-score and accuracy variation plots for ABCRM (blue) and NB (red) for each of the six Enron datasets with varying spam-to-ham ratios: 30% and 70% spam. The black horizontal bars indicate the mean, the boxes indicate the 95%CI and the whiskers indicate the standard deviation. The averages of these values over all Enron datasets are reported in Tables 5.4 and 5.4.

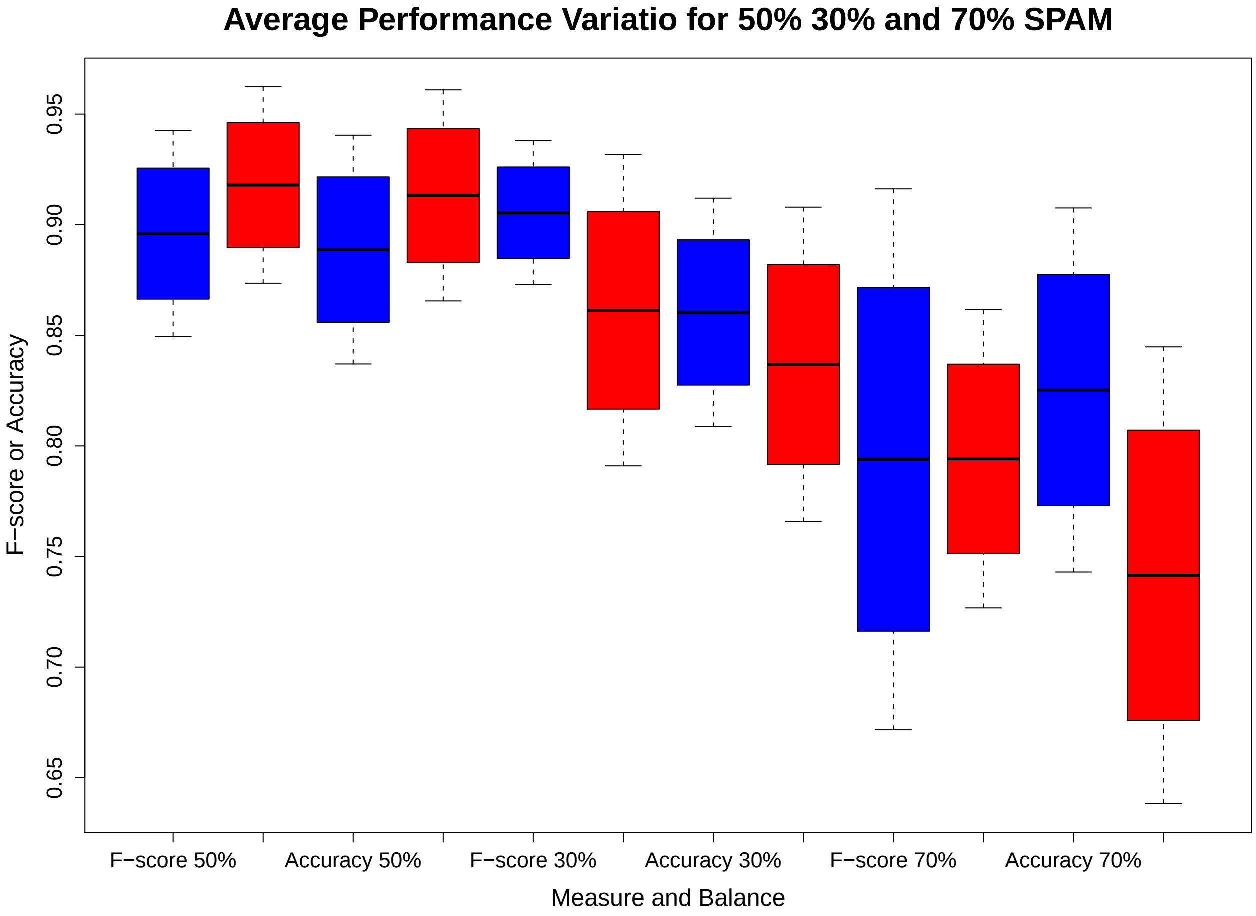

Figure 5.7: Results for the balanced and unbalanced sequential partition validation. F-score and Accuracy average variation statistics (mean, sdev and 95%CI) over all Enron sets for ABCRM (blue) and NB (red) for the partition evaluation.

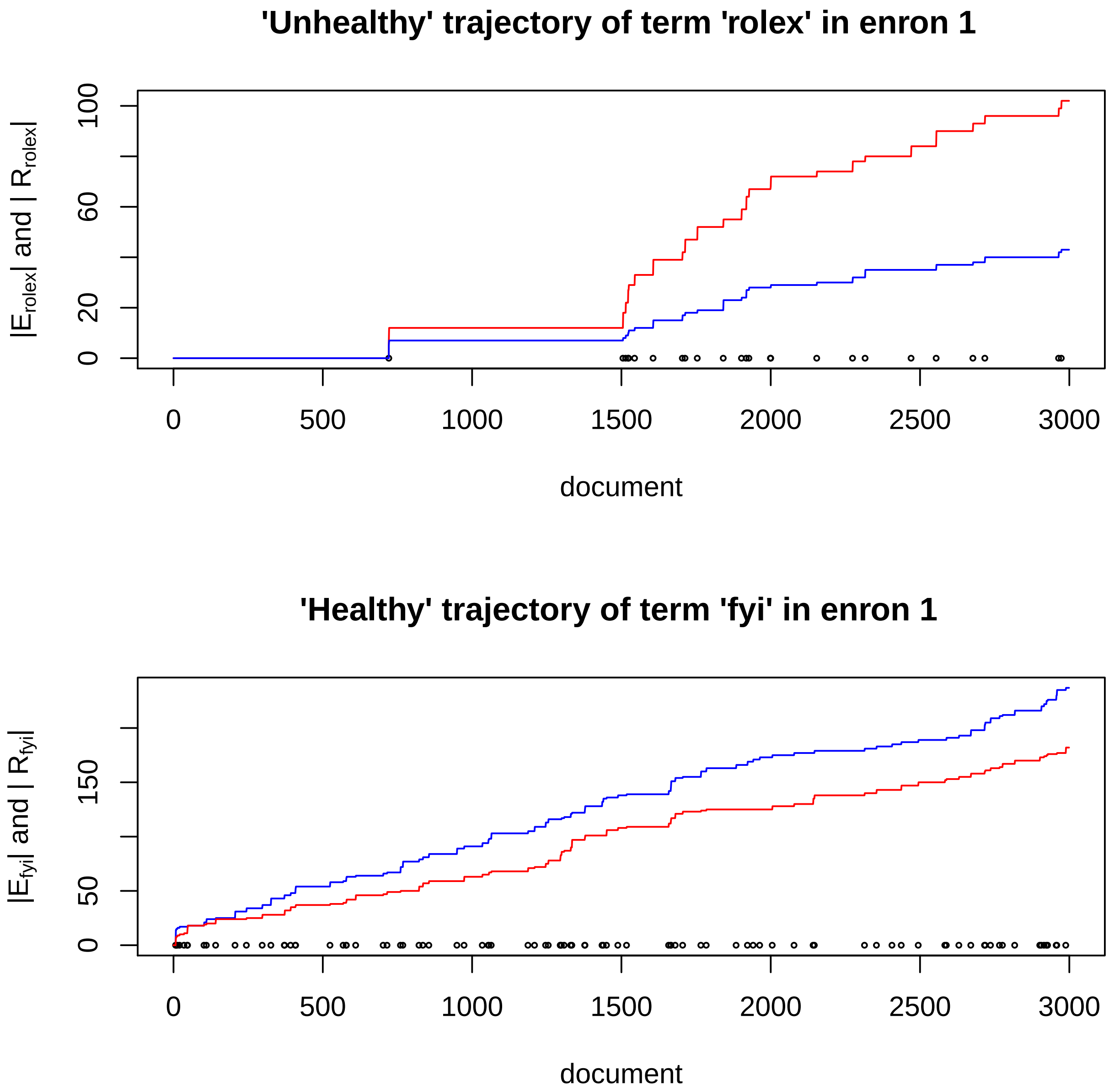

Figure 5.8: “Healthy” and “unhealthy” trajectories of T-cell population sizes ($|E_f|$ in red and $|R_f|$ in blue) for f={‘fyi’, ‘rolex’} respectively. Circles on the *x-axis* indicate occurrences of f in documents x.

Dataset		**ABCRM**	**Naive Bayes**	p<0.01	p<0.05
Enron1	F-score	0.86 ± 0.07	0.73 ± 0.03	ABCRM	
	Accuracy	0.87 ± 0.05	0.63 ± 0.05	ABCRM	
Enron2	F-score	0.71 ± 0.14	0.84 ± 0.06	tie	NB
	Accuracy	0.76 ± 0.09	0.83 ± 0.05	tie	tie
Enron3	F-score	0.82 ± 0.08	0.8 ± 0.05	tie	tie
	Accuracy	0.84 ± 0.06	0.74 ± 0.09	tie	ABCRM
Enron4	F-score	0.86 ± 0.11	0.78 ± 0.05	tie	tie
	Accuracy	0.87 ± 0.09	0.71 ± 0.08	ABCRM	
Enron5	F-score	0.72 ± 0.15	0.85 ± 0.1	tie	tie
	Accuracy	0.77 ± 0.08	0.85 ± 0.09	tie	NB
Enron6	F-score	0.8 ± 0.1	0.76 ± 0.02	tie	tie
	Accuracy	0.83 ± 0.07	0.68 ± 0.04	ABCRM	
Average	**F-score**	**0.79 ± 0.12**	**0.79 ± 0.07**	**tie**	**tie**
	Accuracy	**0.83 ± 0.08**	**0.74 ± 0.1**	**tie**	**tie**

Table 5.5: Results for the unbalanced sequential partition validation (70% spam). F-score and Accuracy mean ± sdev of 10 partitions for 70% spam for ABCRM and NB.

A feature f that is once relevant but then becomes irrelevant (and *vice versa*) over time, according to its E_f-to-R_f ratio, is either the result of the automatic correction of the possibly erroneous initial bias, or an example of concept drift, in which the relevance of f to one class, changes over time. in Figure 5.9 we see a log-log trajectory of a feature that is initially assumed spam since it first occurs in a spam e-mail, but then after co-occurring with ham features in ham e-mails it maintains "healthy" T-cell dynamics. We leave the issue of automatic correction of T-cell dynamics for the following chapter and address concept drift here.

We assume that a drop in the classification performance is evidence of concept drift in either ham or spam. In order to track concept drift, we compute the slope coefficients, $\alpha_{Accuracy}$, α_{Fscore} and their corresponding R^2 for the least square linear fit of Accuracy and F-score. Declined slopes with high slope coefficients indicate

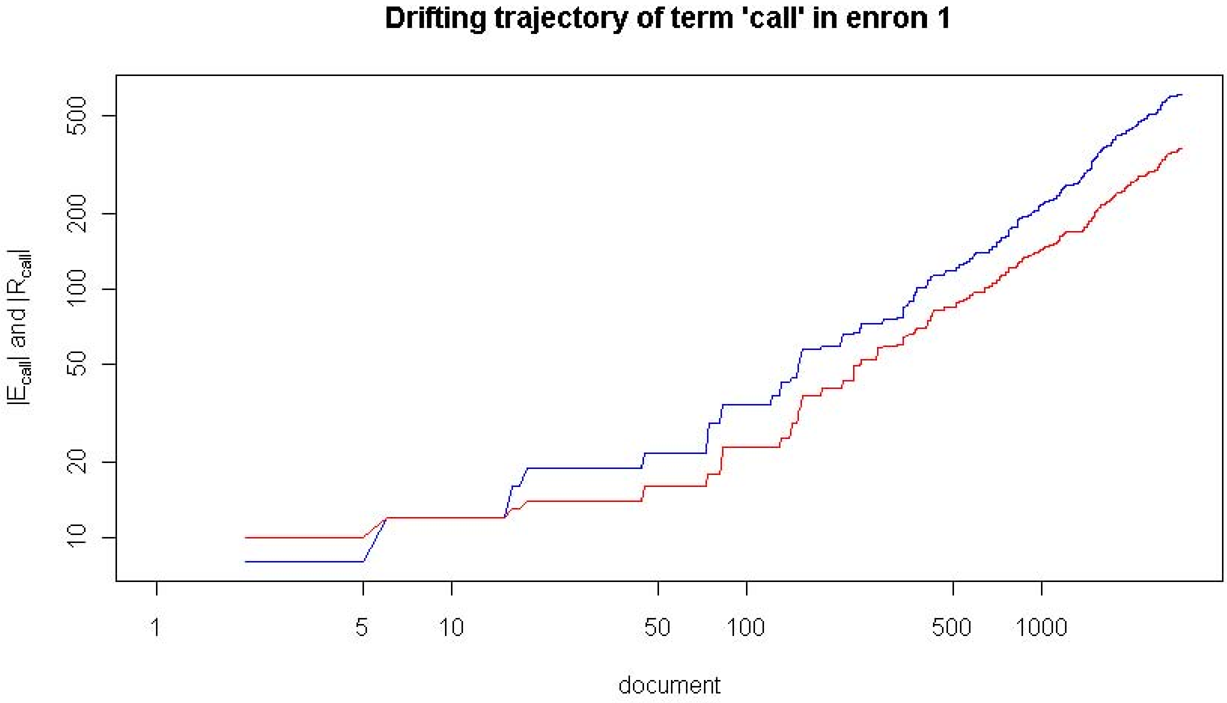

Figure 5.9: The feature ‘call’ is initially biased with $|R_{call}| < |E_{call}|$ since the feature ‘call’ initially occurs in an irrelevant document, however it is then automatically corrected by the T-cell dynamics leading to follow a “healthy” trajectory with $|R_{call} > |E_{call}|$ as shown in this log-log plot.

the presence of concept drift. However, the results in the penultimate two columns of Table 5.6 and the plots in Figure 5.10 show negligible drop in performance for both methods in all Enron data sets. Since both classifiers maintain a relatively high performance, it is unclear whether there is no concept drift or both methods are capable of tracking it effectively. The results are inconclusive. We use different techniques to detect concept drift in the following chapter.

In terms of performance, both the ABCRM and NB are competitive with NB outperforming ABCRM in 3.5[6] Enron sets and ABCRM outperforming NB in 2 Enron sets.

Dataset		**ABCRM**	**NB**	$\alpha_{perf}ABCRM$,R^2	$\alpha_{perf}NB$, R^2	$p < 0.01$	$p < 0.05$
Enron1	F-score	0.95 ± 0.01	0.93 ± 0.01	0.00,0.06	0.00,0.28	ABCRM	
	Accuracy	0.95 ± 0.01	0.93 ± 0.01	0.00,0.11	0.00,0.36	ABCRM	
Enron2	F-score	0.92 ± 0.01	0.95 ± 0.01	0.00,0.02	0.00,0.00	NB	
	Accuracy	0.92 ± 0.01	0.94 ± 0.01	0.00,0.00	0.00,0.04	NB	
Enron3	F-score	0.93 ± 0.02	0.92 ± 0.03	0.00,0.60	0.00,0.00	ABCRM	
	Accuracy	0.94 ± 0.02	0.92 ± 0.02	0.00,0.67	0.65,0.63	ABCRM	
Enron4	F-score	0.92 ± 0.03	0.92 ± 0.01	0.00,0.42	0.00,0.59	tie	tie
	Accuracy	0.92 ± 0.03	0.93 ± 0.01	0.00,0.43	0.00,0.58	NB	
Enron5	F-score	0.90 ± 0.02	0.96 ± 0.03	0.00,0.42	0.00,0.55	NB	
	Accuracy	0.90 ± 0.02	0.96 ± 0.03	0.00,0.45	0.00,0.55	NB	
Enron6	F-score	0.93 ± 0.01	0.95 ± 0.01	0.00,0.60	0.00,0.02	NB	
	Accuracy	0.93 ± 0.02	0.95 ± 0.01	0.00,0.75	0.00,0.00	NB	
Average	**F-score**	**0.92 ± 0.02**	**0.94 ± 0.02**			**NB**	
	Accuracy	**0.93 ± 0.03**	**0.94 ± 0.02**			**NB**	

Table 5.6: ABCRM vs NB F-score, accuracy, their slope coefficients (α_{Fscore} and $\alpha_{Accuracy}$), and their corresponding R^2 for all Enron sets over time.

5.6.2 Generalization and Overfitting

In this last experiment, we study the ability of our method to generalize on new data with respect to NB. More specifically, we compare the classification performance of our

[6] Assuming each Enron consists of comparing both F-score and Accuracy

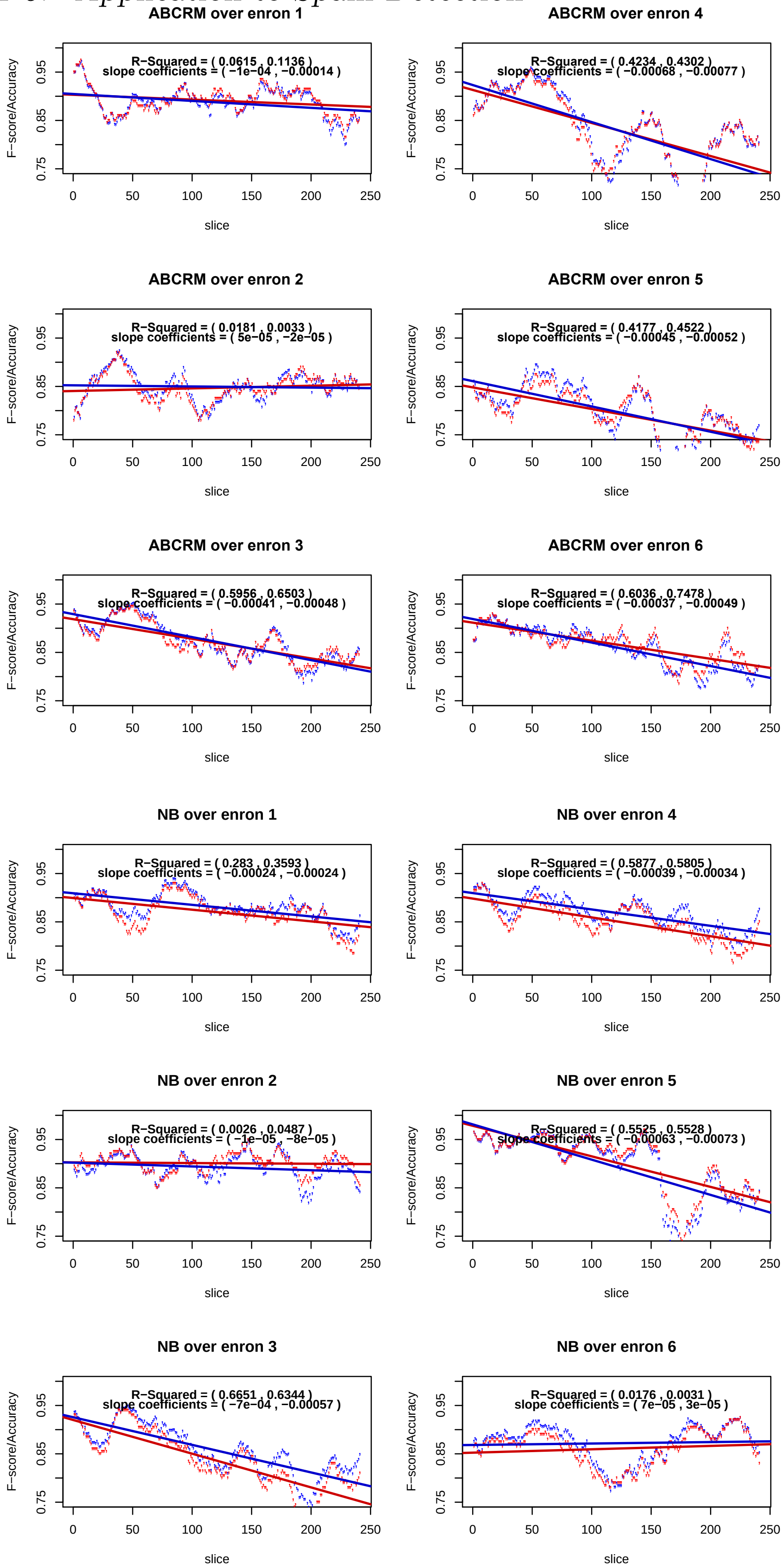

Figure 5.10: Negligible drop in both F-score and accuracy over time for ABCRM and NB, showing no signs of concept drift.

	ABCRM Testing	**ABCRM Training**	**NB Testing**	**NB Training**
F-score	0.9 ± 0.05	0.92 ± 0.04	0.92 ± 0.04	0.99 ± 0.01
Accuracy	0.89 ± 0.05	0.90 ± 0.08	0.91 ± 0.05	0.99 ± 0.01
Fscoredrop	2.17%		7.07%	
Accuracydrop	1.11%		8.08%	

Table 5.7: F-score and Accuracy mean ± sdev of 10 runs for 50% spam ratio Enron data sets using ABCRM and NB when tested on the same training data set and on a distinct one showing the % drop in classification performance.

method to that of NB when tested on the same data set it has been trained on *versus* a separate data set since a substantial drop in performance is evidence of overfitting. We use the aforementioned "sequential partition validation" for 10 partitions for all Enron sets when testing on a distinct validation set, and we compare the average performance with that of testing on the same sequential 10 partitions that we trained on.

The drop in classification performance is measured using the following scoring methods:

$$AccuracyDrop = \frac{Accuracy(Training) - Accuracy(Testing)}{Accuracy(Training)} \tag{5}$$

$$FscoreDrop = \frac{Fscore(Training) - Fscore(Testing)}{Fscore(Training)} \tag{6}$$

The results in Table 5.7 show that the ABCRM generalizes better with a 1% and 2% drop in classification performance in comparison to NB with a 8% and 7% drop in classification performance.

Note that it is more common to compare the performance of k-fold cross-validation on the optimized parameters with that of testing on a distinct data set, however, we do not search the parameter space here. in Chapter 6, we optimized the classification performance of the ABCRM and then test it on a distinct validation set.

5.7 Summary

In this chapter, we test a prototype of the ABCRM on the Enron data and we conclude that it is capable of classifying. We also compare it with NB and show that it is competitive in general and resilient to spam-to-ham ratio variations in particular. The classification results, even though not stellar, seem quite promising especially in the areas of spam-to-ham ratio variation. We also observe the importance of timestamp sequential order for the ABCRM, especially when comparing the first two validation experiments, 5.3.1 and 5.3.2. In the following chapter, we analyze the effect of sequential order by exploring the parameter space comprehensively.

Chapter 6

Application to Biomedical Document Classification

I don't express myself in my paintings. I express my non-self.

Mark Rothko

6.1 Introduction

In this chapter we give an introduction to biomedical document classification and discuss benchmarks, in particular, the article classification task of the Biocreative challenge. We fine-tune our algorithm and study its robustness under various experimental setups. We report our classification results on the Biocreative dataset and compare them with state of art classifiers such as Naive Bayes and Support Vector Machines. This chapter is adapted from relevant published articles [AHR10a, AHR10b, AHR].

Biomedical document classification (BDC) is a binary classification problem in which biomedical articles are classified as either relevant or irrelevant to a certain topic or query. BDC has received a lot of attention in the last few years given the ever increasing number of publications and the need to retrieve only a small relevant portion of the data. Several organizations have dedicated conferences, competitions and workshops to address this challenge: Critical Assessment of Information Extraction systems in Biology (Biocreative)[1], Natural language processing of biology text (BioNLP)[2], Knowledge Discovery and Data Mining (KDD)[3] and GENIA[4].

6.2 Data

The Biocreative (Critical Assessment of Information Extraction systems in Biology) challenge is an effort to enable comparison of various approaches to bioliterature mining [AHKM$^+$07, AHKM$^+$08, KV07]. The various Biocreative challenges have dedicated subtasks to the binary classification of PubMed articles, for instance, as to whether they are relevant or irrelevant to Protein-Protein Interaction (PPI). Interestingly in the second Biocreative challenge, the validation data did not relate much to the training data possibly due to the publication time gap between the training (since 1976) and validation (mostly from 2006) data [AHKM$^+$08]. This could be an example of abrupt concept drift. The order in which the articles were published may be of importance to our model [AHR08b] and thus may help deal with the concept drift especially in validation data consisting of more recent articles. Another issue in BDC is that of class imbalance, especially since the number of relevant articles is often much smaller than irrelevant ones. Medline stores more than 19 million articles

[1] http://Biocreative.sourceforge.net
[2] http://www.bionlp.org
[3] sigkdd.org
[4] http://www-tsujii.is.s.u-tokyo.ac.jp/

of which a small subset is relevant to a particular concept, given the wide diversity in bioliterature. Recently, Biocreative II.5 and III have tried to raise the class imbalance challenge by offering around 10 times fewer relevant articles than to irrelevant ones. To date, no AIS has been used for biomedical document classification. Therefore, in this chapter we plan to test our original AIS classifier on both abstract and full-text biomedical articles addressing the challenges of concept drift and imbalanced classification

The article classification task of Biocreative II.5 [Kra09] was based on a training data set (T) comprised of 61 full-text articles relevant (P_T) to the topic of *protein-protein interaction* (PPI) and 558 irrelevant ones (N_T). The realistic imbalance between the relevant and irrelevant instances is very challenging for common machine learning techniques, since there are few instances of the topical category of interest to generalize from. Because in a general scenario we cannot predict how imbalanced the validation set will be, we first search for optimal ABCRM parameters on a smaller sample of the training that is balanced in the numbers of relevant and irrelevant documents. For this purpose, 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 6.1. For final validation we used the entire Biocreative II.5 testing data set (V) consisting of 63 full-text articles relevant to PPI (P_V) and 532 irrelevant ones (N_V) as also shown in Figure 6.1.

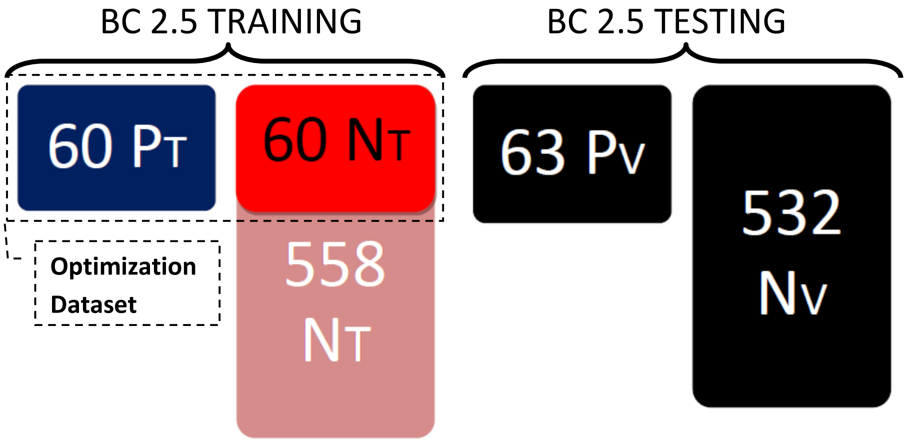

Figure 6.1: Numbers of relevant (P) and irrelevant (N) documents in the training (T) and testing (V) data sets of the Biocreative II.5 challenge. In the parameter search stage, we use a balanced set of 60 P_T (blue) and 60 N_T (red) randomly selected articles from the training data set. In the testing 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 II.5 unlabeled, therefore participants had no prior knowledge of class proportions.

6.3 Configuration for Biomedical Document Classification

We have shown in Chapter 4 how our agent-based model can be used to recognize features and classify textual documents. In this chapter, we test our method on BDC, which is a binary document classification problem, to understand its applicability to document classification in general and to BDC in particular. We discuss some of the techniques used for feature selection, parameter configuration and classification performance evaluation.

6.3.1 Text Processing and Feature selection

We pre-processed all articles by filtering out common words[5] and porter stemming [Por80a] the remaining words. We then ranked words/features f extracted from training articles (T)[6] according to two scores: the first one is the average TF.IDF (see Chapter 2), and the second one is the separation score $S(f) = |p_P(f) - p_N(f)|$ (see Chapter 2) [AHKM$^+$08, KAHK$^+$10]. The final rank $R(f_i)$ for every feature f_i is given by the product of the ranks obtained from both scores; we used only the 650 top ranked features according to $R(f_i)$. Features such as "interact", "lysat" and "transfect" were top ranked as shown in Figure 6.3.1. See [KAHK$^+$10] for more details about the feature extraction procedure.

[5] The list of common (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 PPI

[6] For feature extraction we used both the training data of Biocreative II.5 and Biocreative II as described in [KAHK$^+$10]; all classifiers used the exact same feature set.

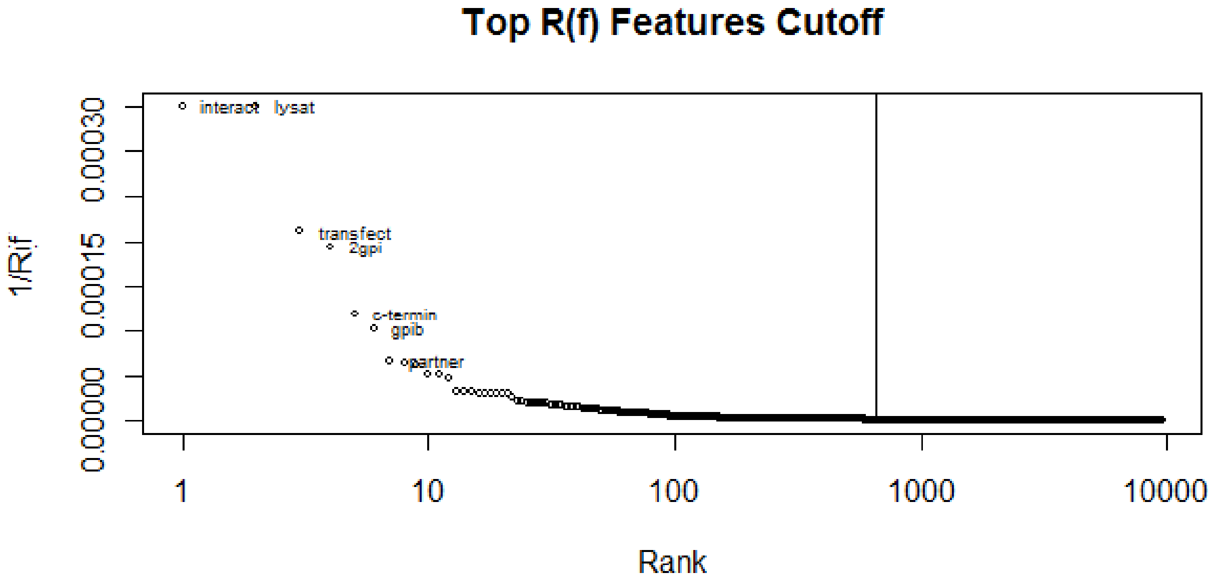

Figure 6.2: We choose the top 650 ranked features according to the rank product R(f) = TF.IDF(f) × S(f). We use $\frac{1}{R(f)}$ in the y-axis for display purposes. Features ranked below the 650th feature have negligible variation. These same features have been used successfully with a linear classifier [KAHK$^+$10].

6.3.2 Parameter Search Settings

We performed an exhaustive parameter search by training the ABCRM on 60 balanced full-text articles (30 P_T and 30 N_T from BCII.5 training) and testing it on the remaining 60 balanced ones (also 30 P_T and 30 N_T from BCII.5 Training) as illustrated in Figure 6.1[7]. Each run corresponds to a unique configuration of the 6 parameters of

[7] Notice that this parameter search on the provided labeled training data uses only the information available to the teams participating in Biocreative II.5 challenge, and none of the testing data whose labels were revealed post-challenge.

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

Table 6.1: Parameter ranges used for optimizing the ABCRM

the ABCRM. The explored parameter ranges are listed in Table 6.1 which result in a total of 192500 unique parameter configurations for each experiment. Finally, the parameter configurations were sorted with respect to the resulting F-score measure of performance (discussed below), which is a good measure combining precision and recall when applied to balanced data [SJS06].

6.4 Robustness

6.4.1 Cell Death

This experiment aims to study the effect of cell death on immune memory and classification performance. In this experiment we compare the top 50 parameter configurations according to F-score obtained using cell death (exp 1.1) to those with no cell death (exp 1.2)—while training on both self and nonself documents. The average performance for the top 50 parameter configurations shows the robustness of the classification performance of the algorithm for each experimental setup. We observe a notable difference in classification performance that we validate statistically (according to the criteria in Chapter 2) to show that using cell death improves the performance (see Fig. 6.3)—regardless of whether the algorithm is trained on just relevant or on both relevant and irrelevant documents (see below). Therefore we conclude that cell death, which helps in the forgetting of useless features and focuses on more recent and frequent ones, improves classification performance. This suggests that cell death is important for immune memory in the T-Cell cross-regulation model.

6.4.2 Training Sets

This experiment is conducted to show if we can rely solely on the positive set for classification, or if the performance can be improved by training on both positive and negative sets. We compare the top 50 parameter configurations according to F-score obtained using training on positive only, also known as PU learning (experiments 2.1 and 2.2), to the previous experiments (1.1 and 1.2). This way we compare training on positive documents only, with and without cell death. The results show that using both training sets always (significantly) improves the robustness of classification performance (see Fig. 6.3). Although the top performance obtained for 1.1 (training on both classes with cell death) and 2.1 (training on positive documents with cell death) is equivalent with F-Score=0.85 (see Table 6.2), the robustness as measured by the performance of the top 50 parameter sets is significantly lower for experiment 2.1.

Exp.	F-Score	E_0	R_0^+	R_0^-	d_R	d_E	n_A
1.1	0.85	2	11	10	0.3	0.2	18
1.2	0.83	1	4	7	0.0	0.0	18
2.1	0.85	1	12	8	0.1	0.0	8
2.2	0.75	2	12	6	0.0	0.0	18

Table 6.2: Performance and parameters of top classifiers in experiments 1 and 2.

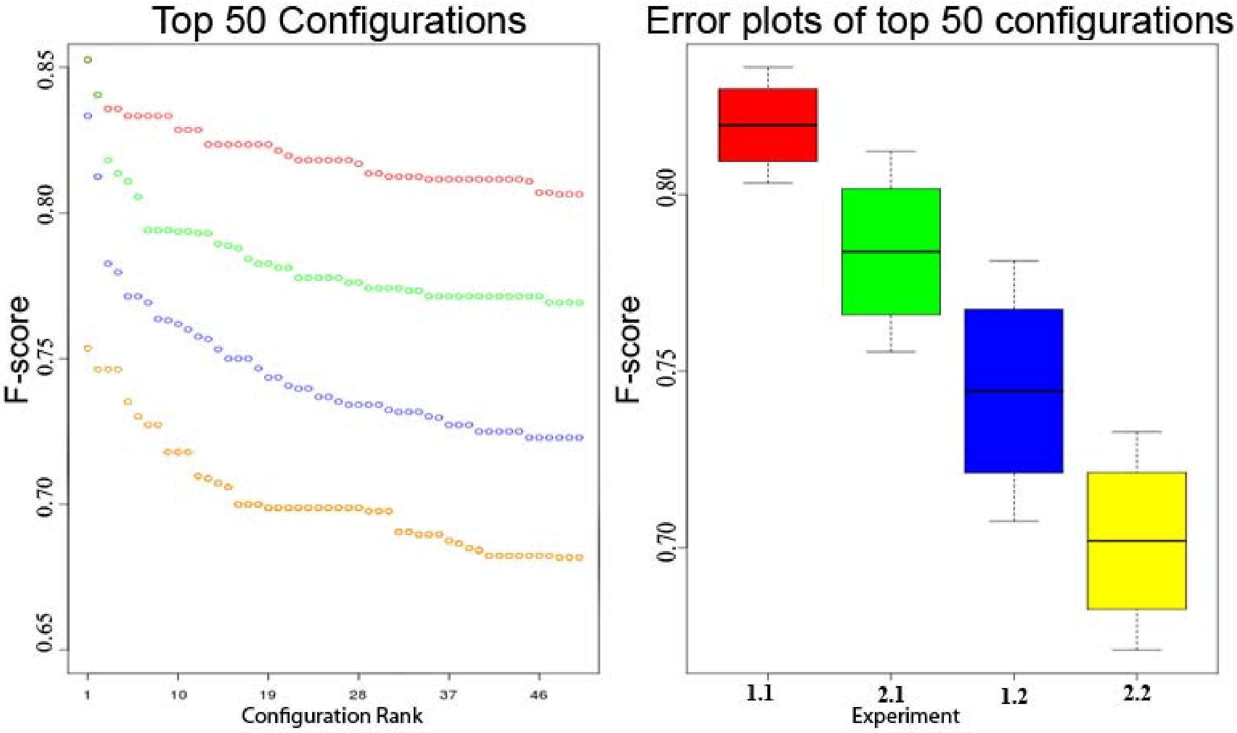

Figure 6.3: The first two experiments result in four experimental setups: 1.1) training on both sets with cell death (red), 2.1) learning with cell death (green), 1.2) training on both sets with no cell death (blue) and 2.2) PU learning with no cell death (yellow) are clearly distinguishable for the top 50 configurations of each experiment on the plot on the left. 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

6.4.3 Sequential Order

This experiment aims to establish how much the sequence order of processing documents impacts performance. In particular, we test if preserving the original temporal order of biomedical documents results in better performance, as this would indicate that the ABCRM can use its sequence-dependent dynamics to track the natural concept or topical drift and thus improve classification. Therefore, we compared the performance of the ABCRM when tested on a sequence of biomedical articles ordered by the original publication, against randomly shuffling the articles. We tested four distinct experimental setups in order to fully explore the influence of document order:

3.1 Ordered training set ⇒ ordered test set

3.2 Ordered training set ⇒ shuffled test set

3.3 Shuffled training set ⇒ shuffled test set

3.4 Shuffled training set ⇒ ordered test set

In the case of shuffled sets, we produced 8 runs with distinct random document orderings; in those cases, performance is represented by central tendency. For this and the following experiment we use training on both classes and using cell death (exp 1.1) that resulted in the best F-score results from the previous two experiments. Therefore exp 1.1 is equivalent to exp. 3.1.

The results of this experiment are summarized in Figure 6.4. The robustness of performance of the first experimental setup (preserving temporal order of articles)

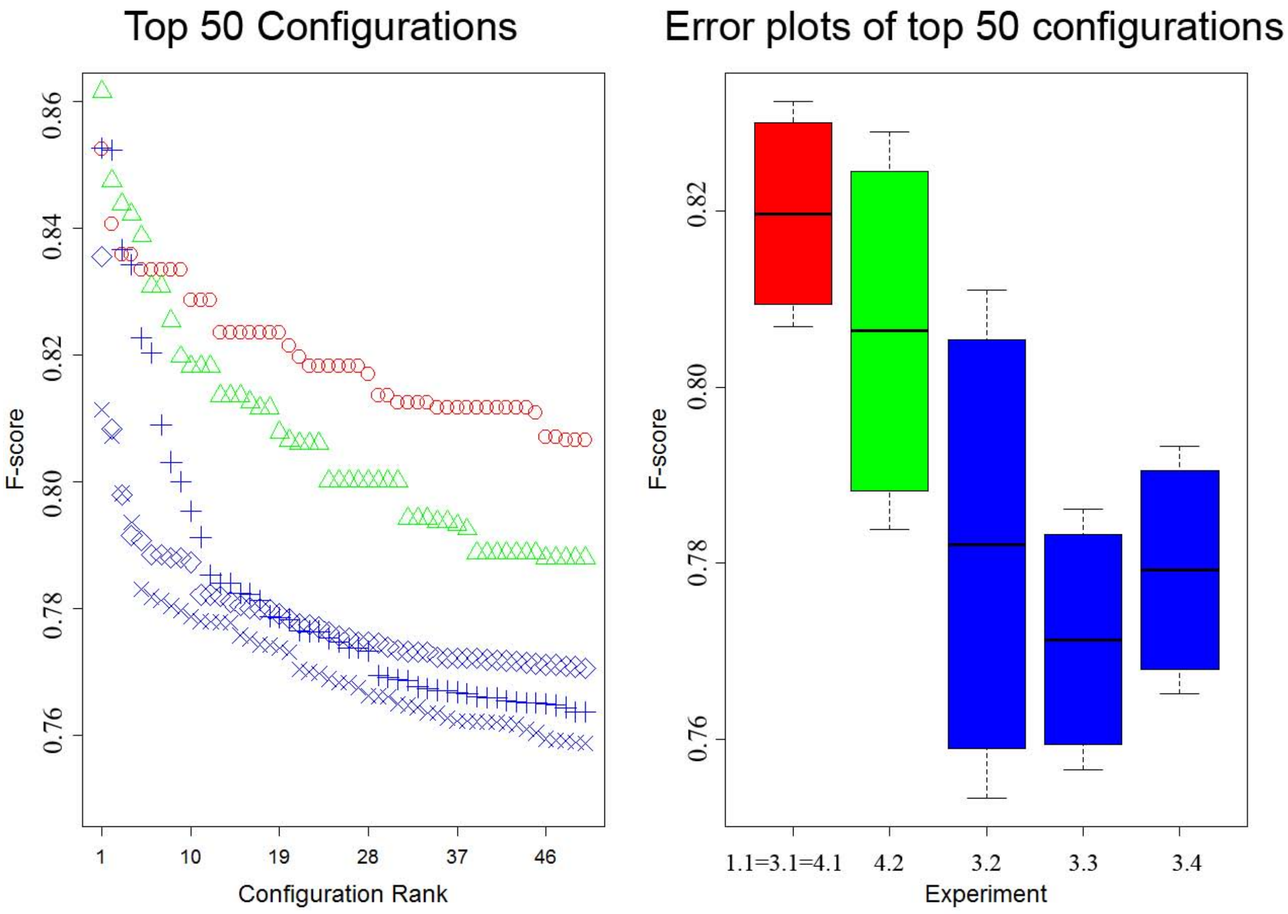

Figure 6.4: The second two experiments result in 5 experimental outcomes. To the left we show the top 50 parameter configurations ranked in terms of F-score for experimental setups 1.1=3.1=4.1 (red circles), 3.2 (blue pluses), 3.3 (blue crosses), 3.4 (blue diamonds), and 4.2 (green triangles). To the right we show the mean (line), 95%CI (boxes), and standard deviation (whiskers) of F-scores for the top 50 parameter configurations.

Exp.	F-Score	E_0	R_0^+	R_0^-	d_R	d_E	n_A
$1.1 = 3.1 = 4.1$	0.85	2	11	10	0.3	0.2	18
3.2	0.85	2	7	6	0.0	0.0	20
4.2	0.86	3	8	7	0.2	0.1	14

Table 6.3: Performance and parameters of top classifiers in experiments 1.1=3.1=4.1, 3.2 and 4.2. Experiments 1.1, 3.1 and 4.1 are equivalent.

is significantly above the other setups. Using the paired student t-test as described in Chapter 2, we conclude that the ABCRM is sensitive to article order—i.e. if the articles are shuffled, the performance is worse. While the performance of the best classifier obtained via experimental setup 3.2 is equivalent to the best one obtained for experimental setup 3.1 (F-Score = 0.85, see Table 6.3 and Figure 6.4), that setup is very sensitive to parameter changes and the performance quickly and significantly decreases for subsequent best classifiers (see Figure 6.4). Indeed, the performance of the top 50 classifiers for experimental setups 3.2, 3.3, and 3.4 is statistically indistinguishable from each other, but is significantly lower than the performance of the top 50 classifiers for experimental setup 3.1. This means that there is indeed a conceptual drift in the Biocreative II.5 article data stream, and the ABCRM can track it better (and in a more robust manner) when publication date is used as the sequence for processing articles than when the temporal order of articles is shuffled. This also suggests that the process of T-Cell cross-regulation in the IS, as modeled here, can track changing nonself environments.

It should be noted that in this experiment, the partitioning of training and test data was done according to the time-stamp of documents. The documents in the test set were published after all documents in the training set. Therefore, even in the shuffled training and test sets (experimental setup 3.3), there is some preservation of temporal order. In future work we will explore experimental setups where the training and test sets are drawn from the same time-stamp distribution to better understand the effects of concept drift and how well our model can track it.

6.4.4 Initial Bias

In this experiment, we test for the effect of the initial biases introduced when features are first encountered. The initial biases of regulatory T-cells injected in the dynamics for a new feature f_i, depend on whether the first document d where the feature is encountered is labeled irrelevant/unknown (R_0^-) or relevant (R_0^+). Since features will occur in both relevant and irrelevant articles, this initial bias for a feature could be detrimental, as a feature most associated with one class could be first encountered on a document of the opposite class. Therefore, it is important to test if the dynamics of the four reactions and APC feature co-presentation that define the ABCRM can self-correct such erroneous biases. To perform this test, we altered the ABCRM algorithm such that T-cells are incremented appropriately every time a feature occurs in a document, and not just the first time the feature occurs (as the canonical algorithm does). Specifically, every time a feature f_i occurs in a document d, we increment $E_i = E_i + E_0$ and $R_i = R_i + R_0^+$ if d is labeled relevant and $R_i = R_i + R_0^-$ if d is labeled irrelevant or unlabeled. We label this experimental set up 4.2, which was conducted with cell death and training on both positive and negative documents. The results of this experiment are also summarized in Figure 6.4. The performance of top classifiers obtained for experimental setups 4.1 (same as 1.1 and 3.1 that are trained on both training sets using cell death) and 4.2 (incremental experimental setup) is shown in Table 6.3. While the best overall classifier is obtained with experimental setup 4.2, the performance of both setups is statistically indistinguishable. Indeed, using the paired student t-test as described above, we cannot reject the null hypothesis claiming that both distributions of F-scores were drawn from a similar distribution. Therefore, we conclude that this modification does not improve the performance of the ABCRM on the Biocreative data set, thus showing that the initial bias can be corrected by the ABCRM collective dynamics and does not require incrementing T-cells for all new features. Because features most associated with a given class tend to co-occur in text with other features most associated with the same class, they will also tend to be co-presented in APC and thus the relevant T-cells will proliferate with similar rates. Therefore, the dynamics of the ABCRM can self-correct initial erroneous biases from the natural textual co-occurrence of features. This shows that T-Cell cross-regulation as modeled here can self-correct initial antigen misclassification by the IS, assuming that antigens from one class (self/nonself) tend to co-occur with antigens from the same class.

6.5 Validation and Conclusions

To test the ABCRM on the full, unbalanced testing set of the Biocreative challenge (see figure 6.1), thus establishing its merit as a bio-inspired biomedical literature mining classifier, we adopted the best parameter configuration from the canonical ABCRM (experimental setup 1.1=3.1=4.1, see Table 6.3) obtained from the parameter search described above. We compared the ABCRM classifier with the multinomial Naive Bayes (NB) with boolean attributes [MAP06], and the publicly available SVM^{light} implementation of SVM applied to normalized feature counts [Joa02]. All classifiers were tested on the same features obtained from the same data.

Since the F-score and Accuracy are not very reliable for evaluating unbalanced

	ABCRM	NB	SVM	Mean	StDev.	Median
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	0.14	0.38
Accuracy	0.71	0.52	0.74	0.67	0.30	0.84
AUC	0.34	0.19	0.46	0.43	0.17	0.44
MCC	0.24	0.13	0.31	0.31	0.19	0.33

Table 6.4: F-Score, Accuracy, AUC and MCC performance of various classifiers when training on the balanced training set of articles and testing on the full unbalanced Biocreative II.5 testing set. Also shown is the central tendency and variation of all systems submitted to Biocreative II.5.

classification [SJS06], we also use the Area Under the interpolated precision and recall Curve (AUC) and Matthew's Correlation Coefficient (MCC) that we defined in Chapter 2. The results are listed in Table 6.4, which also includes the central tendency of the results of all systems submitted by all Biocreative II.5 participating teams [Kra09, KAHK$^+$10]. It should be noted that the ABCRM, NB, and SVM classifiers we tested here, used only single-word features because we wish to establish the feasibility of the method. In contrast, most classifiers submitted to the Biocreative II.5 challenge (including another method from our group which was the top-performing classifier [KAHK$^+$10]) used more sophisticated features such as bigrams and problem-specific entities. Therefore, it is not surprising that these methods as tested here performed under the mean of the challenge.

Our goal was to establish the ABCRM as a new bio-inspired text classifier to be further improved in the future with more sophisticated features. When we compare its performance to NB and SVM on the exact same single-word features, the results are encouraging. Indeed, based on the given measures, while SVM out-performed the ABCRM, the latter out-performed NB. Therefore, the dynamics of T-Cell cross-regulation lead to a competitive collective classification of biomedical articles, which we intend to develop further.

Our dynamical method offers a new perspective in machine learning that traditional classifiers such as NB lack. In Figure 6.5 we show how R (blue) and E (red) T-cell population sizes are in transient states from the first document till the last one for the 100 most discriminant features according to the S-score (see Chapter 2). The parameter configuration used was that optimized for experiment 1.1 (see Table 6.2 for parameter values) that includes cell death. However, in Figure 6.5 we illustrate the feature frequencies for each of the relevant (blue) and irrelevant (red) classes and we observe a fast convergence to one solution, especially after document 60, that is the last training document.

Nevertheless, in Figure 6.5 we observe “less'' cellular dynamics when deactivating cell death and using optimized parameter configuration from experiment 1.2. Cell death plays a huge role not only for the immune memory to focus on recent and frequent features but also in making this model more dynamical and robust. While this experiment offers only a qualitative comparison between our dynamical method and Naive Bayes, we leave quantitative analyses for future work.

In conclusion, we observe that cell death is useful for immune memory as it helps forget old features/antigens and focus on more frequent or recent ones and training on both labeled sets helps improve the classification results. We also observed that our algorithm adapts to the initial bias of T-cell populations generated for new features, and it performs best when tested on a sequence of articles ordered by publication

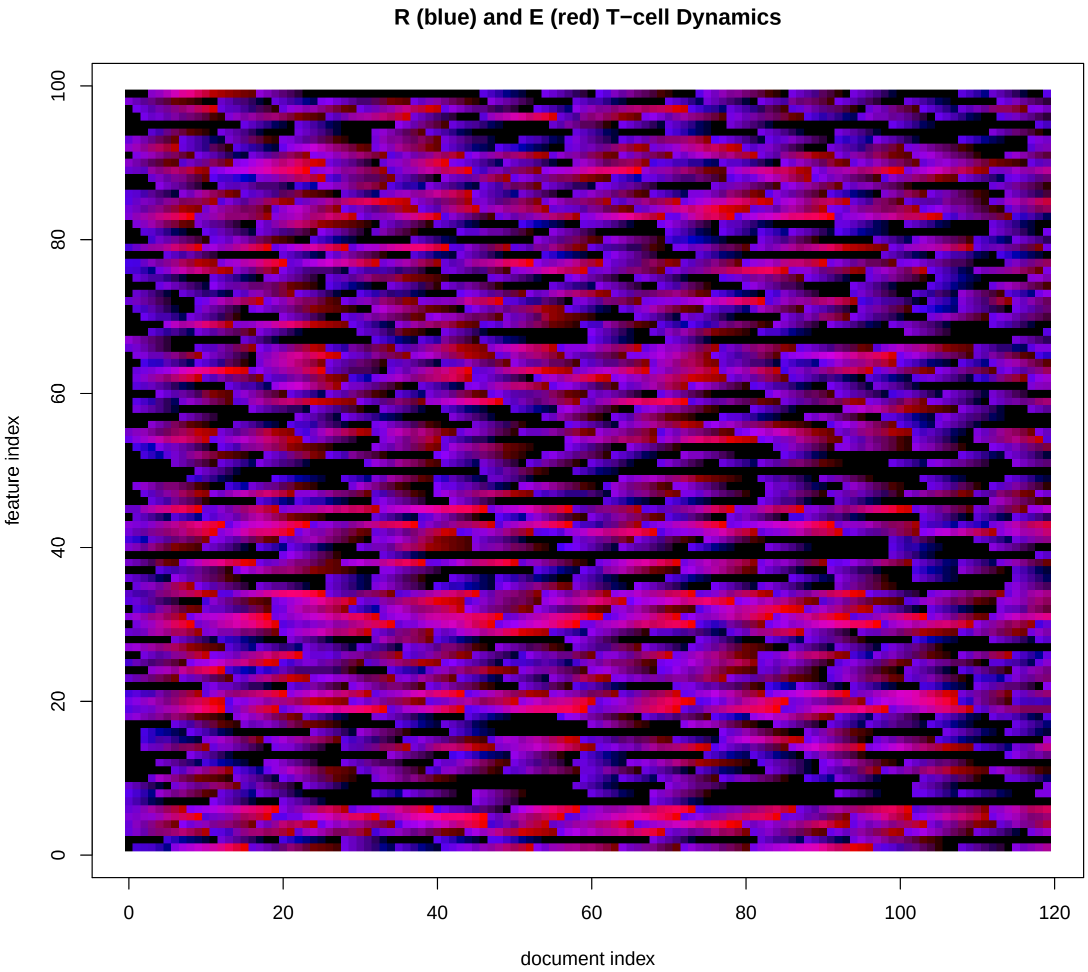

Figure 6.5: R (blue) and E (red) T-cell population size from experiment 1.1 over 120 documents showing constantly changing dynamics

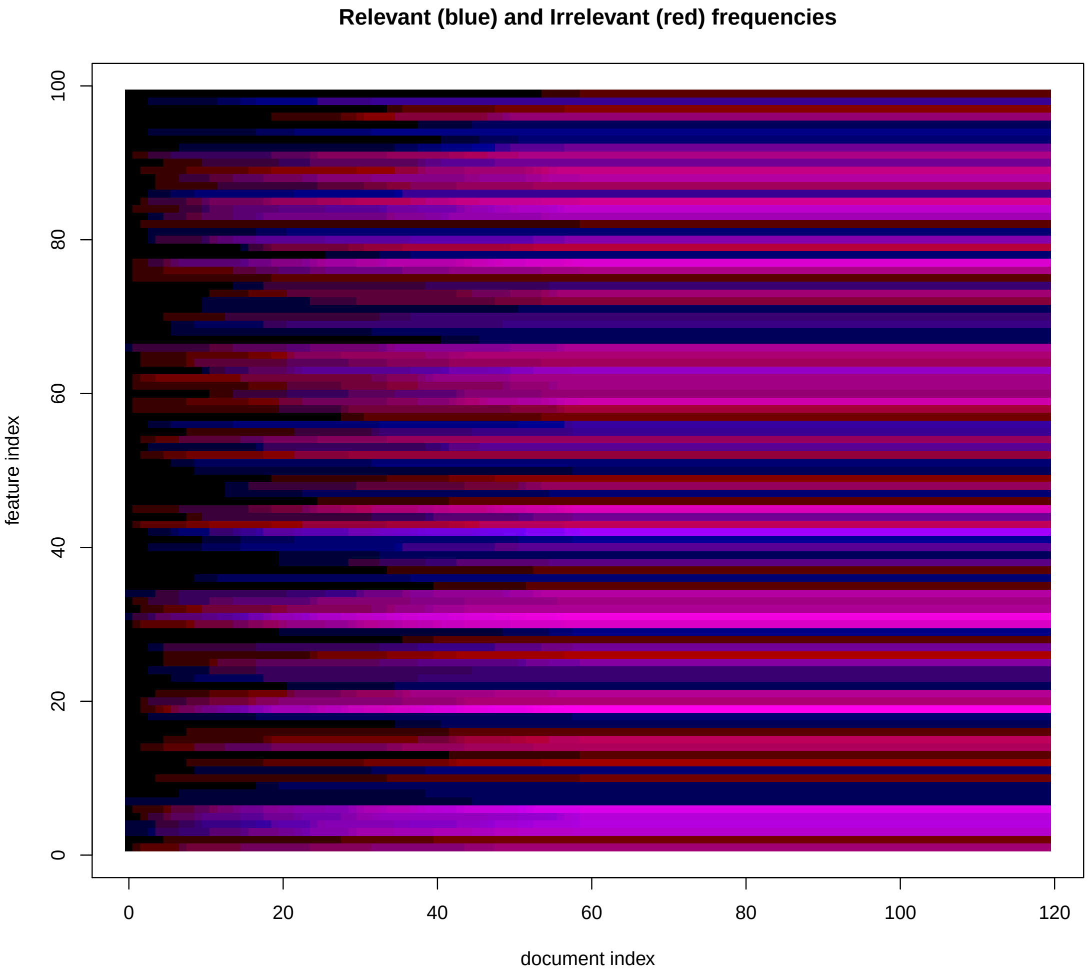

Figure 6.6: Frequencies of features in Relevant (blue) and Irrelevant (red) documents over 120 documents converging to one solution

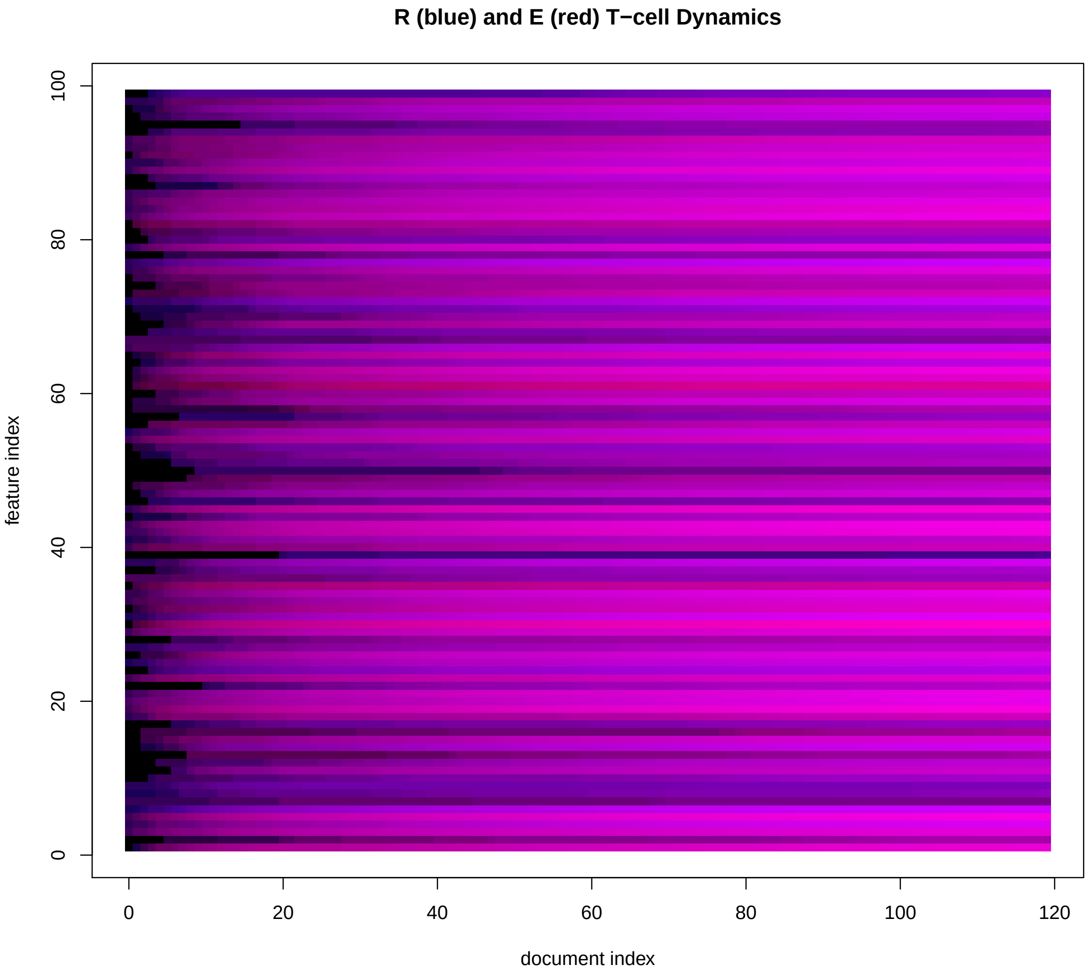

Figure 6.7: R (blue) and E (red) T-cell population size from experiment 1.2 over 120 documents converging to one solution

date—showing that it can track concept drift in the biomedical literature. These properties of our Artificial Life model also show that T-Cell cross regulation is capable of efficient collective classification of nonself antigens and suggest that T-Cell cross-regulation can naturally respond to drift in the pathogen population. Therefore T-Cell cross-regulation defined by the 4 reaction rules and co-presentation of features in APC can be seen as an effective general principle of collective classification available to populations of cells. Clearly, there is still much to do to improve the model. For biomedical literature mining applications, we need to test it with more sophisticated features (as top classifiers in the field do). For our goal of understanding T-Cell cross-regulation in the IS, we need to understand better how memory is sustained in the collective cellular dynamics; for instance, how to sustain regulatory T-Cells, which keep memory of self, in the dynamics even in the presence of very unbalanced scenarios where there are many more nonself instances.

This original work [AHR10a, AHR10b, AHR] should be regarded not only as a promising bio-inspired method that can be further developed and even integrated with other methods but also as a model that could help us better understand the behavior of the natural immune system.

Chapter 7

Conclusions and Future Work

Adde parvum parvo magnus acervus erit

Add a little to a little and there will be a great heap

Ovid

7.1 Contributions to the Immune System

Here we discuss some of the aspects of the immune system that may be of interest to immunologists. We compare observations drawn from our bio-inspired model of document classification (previous two chapters) to aspects of the immune system. We hope to raise interesting questions and insights about T-cell dynamics in particular, and the immune system in general.

7.1.1 Cell-Death and Immune Memory

The immune system possesses memory of previous infections and uses it to respond to similar ones more effectively in the future [SL31]. However, the mechanism behind immune memory is still poorly understood. There are several theories, each supported by experimental evidence.

The most established theory is that of hyper-sensitive *memory cells* that come in two varieties, memory B-cells and memory T-cells [MWMW05]. Infections form two types of long-term memory: humoral immunity, in which B-cells produce antibodies that recognize and bind to nonself antigens, and cellular immunity, in which activated T-cells proliferate and lead to the death of the infected cells. The memory of an infection is retained for several years [SL31] and is measured by the population size of memory cells even in the absence of the antigen of the infection.

It has been shown that the total number of memory cells is roughly constant and any increase in the population is followed by a return to the constant concentration [TR95]. This indicates that a homeostasis mechanism is able to regulate and maintain the population size of memory cells. Evidently, no organism can accommodate the infinite increase of cells and therefore *apoptosis* or programmed cell death takes care of the elimination of some cells. Our experiments in Section 6.4.1 have shown how cell death is useful for the immune memory to focus on recent and frequent features or antigens. Moreover, our illustrations in Figures 6.5 and 6.5 show a substantial difference in the T-cell population dynamics between using and not using cell death. Cell death offers a much more dynamical system with constantly changing populations of E and R T-cells.

The question of how memory cells are formed and maintained remains unanswered with several theories trying to explain it:

The **long-lived memory cell theory** claims that the highly responsive B-cells and T-cells differentiate into long lived memory cells and do not undergo any cell death or cell division for many years. However, there is no convincing evidence to this especially given the short life span of these cells and indeed a series of experiments on mice showed that T-cells continue to divide after primary response [TS94] and another experiment shows that plasma cells in mice have a life span of only months [SAWA98].

The **emergent memory theory** suggests that highly specific effector T-cells are preserved from cell death by an enzyme such as telomerase. Telomerase increases the length of telomeres, which are made of DNA sequences that protect the tips of chromosomes from being shortened during cell reproduction [WPL$^+$97]. Each cell can reproduce a certain number of times that is predefined by the length of its telomeres, that is shortened with every reproduction. Therefore, telomerase can establish long term-memory for highly specific effector T-cells by allowing them to reproduce more.

The **residual antigen theory** suggests that antigens themselves can be stored in the lymph node [PW97] and could keep the immune system active to sustain the homeostasis of memory cell populations. This theory remains widely accepted by immunologists [AGA05].

The **immune network theory** defined in chapter 3, is based on the assumption that a network of B-cells with idiotopes and paratopes becomes capable of recognizing nonself when all of its self-recognizing idiotopes have been regulated by other self-recognizing paratopes. This network undergoes cycles of excitation and suppression leading to a homeostatic memory pool. However, this theory received no sufficient evidence to observe this behavior *in vitro* or *in vivo*.

In the T-cell cross-regulation model, Sepulveda [Sep09] argues that initial populations of the effector and regulatory T-cells are of similar sizes for an antigen. However, the antigen population size diverges when these T-cells become long lived. In other words, after many series of suppression and proliferation, T-cell populations eventually become long lived to be either “healthy'' with more regulatory T-cells or “unhealthy'' with more effector T-cells. In our simulations (see figure 5.8) we observe trajectories of T-cell population sizes that diverge into either “healthy'' (with more regulatory T-cells) or “unhealthy'' (with more effector T-cells) states, that are maintained over time. Moreover, we observe the emergent self-maintenance of features.. Figure 5.8 shows how regulatory T-cells are overwhelmed by constantly increasing, self-maintained effector T-cells for the irrelevant feature of “rolex”.

Studies have demonstrated the dynamic nature of T-cell homeostatis that is maintained through constant competition and flux [Jam05]. In the original analytical cross-regulation model, Carneiro et al. [CLC^{+}07] claim that regulatory T-cells depend on effector T-cells to maintain their population size since they cannot proliferate independently. However, recent *in vivo* experiments report that the stability of the regulatory T-cell lineage is maintained through self-renewal [RNJ^{+}10]. Self-renewing regulatory cells can result in interesting dynamics that we leave for future work. Our ABCRM expands on a mathematical model of T-cell cross-regulation to deal with multiple populations. In it, we observe that regulatory T-cells of a population can be maintained and excited to proliferate by effector cells of distinct populations in a homeostatic network of cellular interactions. This dependence of regulatory cells on

T-cells from other populations was not possible in the original model due to its limitation to only one population of T-cells. For example, figure 5.8 (below) shows the trajectory of a healthy population size of T-cells, in which R T-cells are maintained through their "interaction" with other populations of T-cells *via* APC. Moreover, R T-cells need to bind to APC in order for this interaction with other populations of T-cells to happen, and that is possible with the occurrence (illustrated as circles in figure 5.8) of relevant features.

Therefore, we conclude that cell death is useful for immune memory and plays a huge role in T-cell dynamics. In addition, we show how the expansion of a simple cross-regulation model to deal with multiple antigens and populations of T-cells can achieve binary classification *via* R T-cells that are maintained by other populations of T-cells.

7.1.2 Negative Selection and beyond

Negative selection in the adaptive immune system is known to eliminate naive effector T-cells that recognize and bind to self antigens in the thymus. This prevents the randomly generated T-cells, with various T-cell receptors, from recognizing and attacking self antigens (autoimmunity) when later matured and released from the thymus [Hof01]. Therefore, effector T-cells are trained to discriminate between self and nonself antigens by "training" on a repertoire of self antigens in the thymus.

In the context of machine learning, this is similar to a situation known as positive unlabeled (PU) learning. In section 6.4.2, we tested if we can rely solely on the positive set for classification, as done in negative selection and PU learning, and whether the performance can be improved by training on both positive and negative sets. The results showed that using both training sets could improve the classification performance although training solely on relevant documents can be reliable.

As discussed in chapter 3, matured effector T-cells are still prone to escaping the thymus without being trained on all self antigens. Self-recognizing effector cells may lead to auto-immune diseases unless regulated by self-recognizing regulatory T-cells that play a huge roll in the cross-regulation model. Nevertheless, both negative selection and the cross-regulation model represent only a minute part of the adaptive immune system, which itself represents only a part of the entire immune system. While our aim was to study if such a subsystem of the adaptive immune system is capable of classifying, our experiments also show that the immune system can benefit from exposure to nonself antigens. Indeed, evidence shows that the ability for adult mice to recognize grafts of foreign skin depends on earlier exposure to nonself antigens [PPSO04, p.361].

7.2 Contributions to Complex Systems and Machine Learning

Our experiments from the previous chapters address challenges in machine learning, such as dynamic class imbalance and concept drift, and raise interesting questions about self-organized and adaptive systems that can classify using a collective behavior.

7.2.1 Decentralized control and Robustness to dynamic class imbalance

The immune system is a complex system of thousands of interacting cells that interact to defend our body from malicious intruders. More specifically, the vertebrate adaptive immune system consists of decentralized B-cells and T-cells that interact for the purpose of discriminating between self and nonself antigens. The thymus gland, bone marrow and spleen play a huge role in the differentiation and training of cells, however, immune responses are orchestrated in a decentralized fashion by the interaction of millions of cells [Hof01]. The decentralized control exhibited by the immune system, in which not a single cell or organ controls the classification of intruders, makes the binary classification problem of discriminating between self and nonself all the more compelling.

One of the biggest challenges of the immune system is adapting to the constantly changing ratios between harmless and harmful intruders. The immune system is capable of discriminating between self and nonself antigens in healthy (only self antigens) and unhealthy scenarios (unbalanced ratios between self and nonself antigens).

In machine learning, this same problem is characterized by changes in the ratios of class instances. In binary document classification, a corpus may have more relevant than irrelevant documents in the training set, but this may change in the validation set in unpredictable ways. For example, in spam detection, a user is prone to being bombarded by undesired e-mails at any time. This results in an unpredictable spam-to-ham ratio variation, and makes it hard for traditional machine learners to deal with this accurately. We call this problem "dynamically unbalanced classification". Therefore, a decentralized adaptive system that is capable of dealing with the dynamically changing unbalance is needed.

In chapter 4, we used agent-based modeling to implement T-cell dynamics in a decentralized fashion. In chapter 5, we have shown how our adaptive, decentralized, agent-based model is capable of dealing with spam-to-ham ratio variations between training and testing. Although the classification results in the balanced scenario were in favor of NB for three of the cross-validation subsets and statistically indistinguishable between NB and the ABCRM for the remaining three, they were in favor of the ABCRM in the imbalanced ones (see chapter 5), where NB did not cope as well with balance changes. In chapter 6, we have also shown how our decentralized model is capable of adapting to these changes when trained on a balanced set of articles and tested on an imbalanced one. Note that NB can be tweaked for imbalanced classification but the imbalance ratio cannot be known ahead of time. However, the ABCRM is adaptive and auto-reactive to changes in class imbalance through its decentralized control.

7.2.2 Collective Behavior of T-Cell Dynamics

When thousands of cells interact to discriminate between self and nonself, they do so collectively in a self-organized manner. Our approach is based on the idea that the immune system is a distributed collection of molecular constituents with no central controller [SC01]. Therefore, its classification ability needs to result from a *collective* *classification* process, defined as the ability of decentralized systems of many components to classify situations that require global information or coordinated action [Mit06]. Nature is full of examples of collective classification such as the dynamics of stomata cells on leaf surfaces [PWMM04], biochemical intracellular signal transduction networks [HKHR08], quorum sensing in bacteria [WS06] and social insects [Pra05], etc. For example, colonies of the ant *Temnothorax albipennis* collectively choose a site for their nest based on quorum rules that quantify the rate of direct encounters with their nest mates [*ibid*]. We can study collective classification in general models of complex systems such as Cellular Automata by identifying regular patterns in the dynamics that store, transmit and process information [CM95, RH05, SHR^{+}06]. But or approach here is based on a more realistic agent-based modeling of cellular dynamics. In analogy with the interactions among T-cells and antigens that lead to self/nonself discrimination in the immune system, words co-occurring in documents can be seen as interacting in text in such a way as to allow us to distinguish between relevant and irrelevant documents. Figure 7.1 illustrates the idea of "interacting" words from one document to the other *via* their corresponding T-cells that interact by adjacently binding to APC.

In chapter 4, we described our initial biases of E_0 and $R_0^{\pm}$ for features that are newly introduced to the cellular dynamics. Features f first occurring in a relevant document are biased with more R_f than E_f ($R_0^{+} > E_0$) whereas features first occurring in an irrelevant or unlabeled document are biased with less R_f than E_f ($R_0^{-} < E_0$). Nonetheless, this bias can be erroneous since a more relevant feature can first occur in an irrelevant document and *vice versa*.

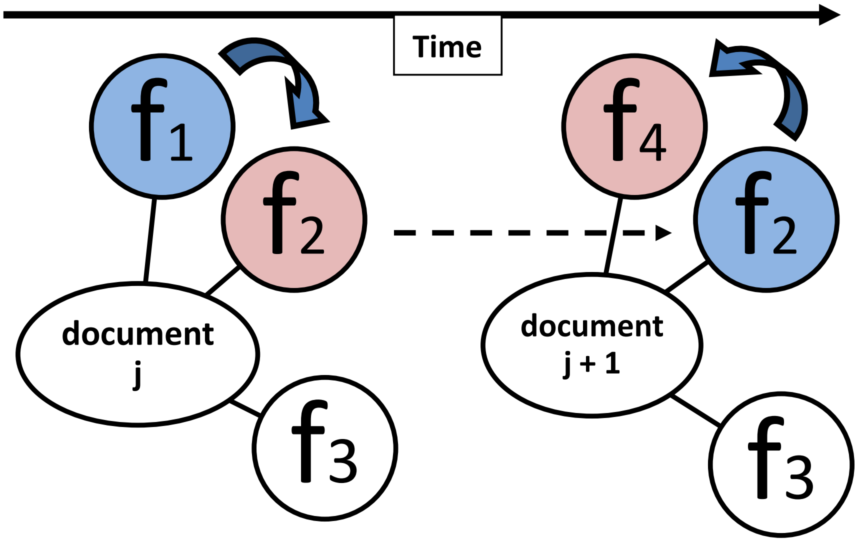

Figure 7.1: An illustration of collective learning from document j to document $j+1$ that follows in time order. In document j, relevant feature f_1 (with $R_1 \geq E_1$) "regulates" irrelevant or new feature f_2 (with $R_2 << E_2$). Consequentially, in document $j+1$, feature f_2 that is now relevant is capable of regulating another irrelevant or new feature f_4, etc.

We tested the self-correcting ability of T-cell dynamics in chapter 6 by comparing the classification performance with the erroneous initial bias to that of a minor variation of our model in which T-cells are added incrementally every time a feature occurs in a document and not only the first time. The performance under the second condition was statistically indistinguishable from the original scenario, thus proving that the initial bias is automatically corrected by collective T-cell dynamics.

7.2.3 Sequential Order and Concept Drift

The adaptive nature of the vertebrate immune system can be compared to incremental learners in machine learning, which in contrast to batch learners, learn incrementally or instance-by-instance [FS06]. While, incremental learning depends on the set of instances previously encountered, it is not typically sensitive to the sequential order in which the instances are presented. The ABCRM is novel in that it is based on a dynamical system whose behavior depends on sequence presentation, not simply the set of previously encounterd instances. This sequentially ordered data, known as stream data, is very common in real-world application and is of particular interest to modern machine learning.

A very common problem in stream data is concept drift, which we have previously defined as the (gradual or sudden) change of underlying data distributions over time.

In section 6.4.3, we tested for concept drift in biomedical document classification by comparing the classification performance on articles ordered by publication date against that of randomly shuffled articles. The former setup outperformed the latter one. This indicates that some useful information is available in the publication order, and our method was able to benefit from it. Indeed, the ABCRM is a classifier based on a dynamical system framework, which makes it dependent on sequence of items to learn from and classify. This establishes an alternative bio-inspired approach to binary classification. It also highlights how much the immune system seems to be dependent on the history of pathogens and primings it encounters.

7.3 Concluding Remarks and contribution

In this thesis, we have established an original bio-inspired classification method inspired by T-cell cross regulation of the adaptive immune system. We tested our method on two binary document classification problems, using publicly available benchmark: the enron dataset for spam detection and the Biocreative dataset for biomedical document classification. From the first application we concluded that our bio-inspired model is able to classify. We also obtained encouraging results that are comparable to traditional classification methods. We studied the robustness of our method to dynamically changing class imbalance and concept drift. Our method was particularly promising in terms of resilience to dynamic class imbalance between training and testing documents as experiments suggested it is more resilient to balance changes than NB. Also, our dynamical system has shown to benefit from the order in which the documents were presented to track concept drift. This was evident from the drop of classification performance on documents that are shuffled in time-stamped order.

The cross-regulation model is based a theory of the *cellular immunity* of the adaptive immune system. deriving from cellular interactions. However, cellular immunity is known to be complemented with a *humoral response* that is mediated by secreted antibodies [Hof01]. Moreover, the adaptive immune system is complemented by the innate immune system in order to discriminate between self and nonself antigens effectively. Therefore, we can hope for huge improvements in terms of classification performance simply by aggregating our cross-regulation model to other artificial immune systems that model other subsystems of the immune system. Moreover, our immune-inspired binary classifier raises many questions about the generalization of the model in terms of dealing with multiple-class classification and more complex features that we leave for future work.

Our contribution was not limited to the machine learning but also to complex systems and immunology:

- We developed a bio-inspired classifier based on T-cell cross-regulation using agent-based modeling.
- We tested a prototype of our model on spam detection to show that T-cell collective classification works.
- We optimized our model for bio-medical document classification to study the robustness of our method and raise insights about the immune system and T-cell dynamics:
    - We concluded that cell death is useful for immune memory to focus on more recent and frequent antigens. Also, regulatory T-cells are maintained in

the memory pool by interacting with T-cells from other populations *via* APC.

- We have shown that training on both self and nonself is more effective than training on self alone and therefore AIS models such as the cross-regulation model and negative selection can benefit from training on nonself antigens as well.
- We have demonstrated how a dynamical system can be more robust to dynamic class imbalance with spam-to-ham ratio variation.
- We have shown how the sequential order of documents is important for our method to track concept drift.

7.4 Future Work

Our work addresses many questions about the applicability of our cross-regulation agent-based model to binary document classification—specifically on real-world data from spam detection and biomedical document classification— and the behavior of T-cell dynamics. However, it also raises many new questions of whether it can be extended for multi-class classification and whether it can be improved using more complex features such as bigrams and trigrams. Testing our method on artificial data can provide us and immunologists with better insights about T-cell dynamics and the immune system, that is still in many perspectives, poorly understood.

7.4.1 Artificial Data

So far, we have only tested our ABCRM on real-world data, namely enron data for spam detection and Biocreative data for biomedical document classification. Many artificial immune systems have been tested on artificial data, such as STAGGER concepts [SG86]. However, none of these datasets characterize textual data. Our future aim is to develop a benchmark of artificial textual data having similar *zipfian* distributions, describing frequency distributions of words in corpora [TH03]. This artificial data would allow us to accurately study various cases of concept drift and class imbalance. We leave such an approach for future work.

Our benchmark can be useful for testing other classifiers with the presence of concept drift and comparing results easily.

7.4.2 ABCRM N-gram Feature Selection

In terms of feature selection, our preliminary ABCRM samples single words from documents and uses them as features. We plan to extend our algorithm to allow more complex n-gram (unigrams, bigrams, trigrams...etc) features to be presented as antigens. The selection of n-gram features can be evolutionary in such a way to simulate the evolution of an immune system. The evolutionary selection of n-grams can be guided by a feature selection method such as mutual information, information gain or χ^2 that are common in data mining [FS06]. However, we plan to use the S score (see 2.3.3), which simply measures the absolute difference between the normalized number of regulatory (R_f) and effector (E_f) T-cells specific to an n-gram feature f. Highly discriminant n-gram features f have relatively high E_f or R_f, exclusively. The initialization of an n-gram f is based on the score of its components. For example, a bigram feature f_{bi}, constituting of the unigrams f_1 and f_2, is initialized if $|score(f_1)|$ or $|score(f_2)|$ are above a certain threshold. When initialized, the n-gram initial R_0 and E_0 values are set to the mean values of the R and E of the n-gram components. N-grams can solve the *cross-affinity* problem by allowing T-cells specific to "similar" features, f_1 and f_2, to bind to the antigen complex $f_{bi} \equiv f_1 f_2$. The n-gram can be extended to synonymous compounds (if f_1 synonymous to f_2), features with short edit distance (very useful to account for Bayesian poisoning with f_1 and f_2 visually similar with short edit distance), or co-occurring features (f_1 and f_2 co-occur in the same document). Also, the combinatorial explosion of n-gram features can be handled in a bio-inspired fashion by modeling T-cell turnover using clonal selection (see chapter 3) in the adaptive immune system.

7.4.3 ABCRM Generalization to Multi-classification

So far, we have only used ABCRM for binary document classification. Eventually, we aim to have a general multi-classifier for documents. The immune system's various responses are orchestrated by several types of cells (e.g. T-helper cells, T-regulatory cells, B-cells) [DN08]. Effector (E) and Regulatory (R) T-cells could sufficient for binary classification, however, additional classes (for other responses) would require additional cell-types.

Multi-classification can serve as a useful extension to the binary biomedical document classification where some articles are neither strictly relevant nor strictly irrelevant but related. One of the second BioCreAtIvE tasks focused on protein-protein interaction (PPI) pair retrieval in bioliterature [KV07], however the real challenge was in mapping extracted protein names to species-specific UNIPROT ids and therefore a multi-species document classifier was necessary. The BioCreAtIvE II.5[1] raises the same multi-species classification challenge for PPI retrieval. We plan to address the problem of multi-species document classification with ABCRM by having species-relevant cell-types that are also PPI-relevant and therefore by the differentiation of the Regulatory cell-type.

For example, in a document, a "human" relevant feature f_h, would have a relatively high RH_{f_h} (i.e. numerous human regulatory T-cells). Since regulatory cell-types do not regulate each other, f_h can be equally "mouse" relevant with high RM_{f_h}; such a feature examples many protein names shared between human and mice. However, f_h could have high RH_{f_h} and RM_{f_h}, and still be PPI irrelevant with high E_{f_h}, in which case, f_h would be classified as irrelevant. Species relevant regulatory cells are capable of suppressing PPI relevant effector cells and therefore species relevant regulatory cells are also PPI relevant.

Similarly for spam detection, ABCRM can be extended to multi-classification specifically for legitimate e-mail or ham (e.g. urgent, mailing lists, family, work, school). For example, "urgent" e-mails would have many features f_u with high RU_{f_u} and low E_{f_u}, however they could also relate to other categories such as "family" and therefore have high RF_{f_u} as well. Spam can be categorized as well with spam-relevant effector cell types that by pairing can only lead to more E proliferation (more spam). The various effector cell types can be relevant to obfuscation techniques, advertised

[1] http://www.biocreative.org/

products or intelligible foreign languages.

7.4.4 Spatio-Temporal Model of ABCRM

We aim to implement the ABCRM on a new agent-based framework known as Bit-Bang [BMC06]. We will use an environment similar to Floriano's [FMMK07] poison-food (or irrelevant-relevant) model. In this model, polyspecific antigen presenting cells (A) are allowed to bind to a maximum of two T-cells as assumed in the original CRM [CLC$^+$07]. A spatially sample pairs of horizontally, vertically or diagonally neighboring features (e.g. words, bigrams) to present them as pairs of antigens. Antigens attract monospecific Effector (E) and Regulatory (R) T-cells that are within vicinity. T-cells that do not bind to A approach A such that they have higher chances of binding to it next time. This would eventually allow for conclusive features to gather around the bottom of the document (e.g. signatures in e-mail and footers in other documents in general) and similarly for introductive features at the beginning (e.g. "hi" and "hello" in e-mail and document titles or headers in general). On the other hand, E that bind to A undergo proliferation unless suppressed by R according to the previously described interaction rules. Newly proliferated T-cells are generated around their parent T-cells while T-cells for features occurring are randomly scattered around the document space. This application not only studies the importance of spatial interaction but is also flexible to many variations that can help us study A antigen preferential presentation (if there is any preference for some more informative antigens to be presented over others) and T-cell preferential attachment (if there is preference for some T-cells to bind to the A over others). Answering these two questions could provide immunologists with insights about A preferential antigen presentation and T-cell preferential binding to A in biology.

Bibliography

[Abe03] N. Abe. Invited talk: Sampling approaches to learning from imbalanced datasets: active learning, cost sensitive learning and beyond. In *Proc. of ICML Workshop: Learning from Imbalanced Data Sets*, 2003. 22

[AGA05] R. Antia, V.V. Ganusov, and R. Ahmed. The role of models in understanding CD8+ T-cell memory. *Nature Reviews Immunology*, 5(2):101–111, 2005. 118

[AHKM$^+$07] A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Retchsteiner, K. Verspoor, Z. Wang, and L.M. Rocha. Uncovering protein-protein interactions in the bibliome. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, volume ISBN 84-933255-6-2, pages 247–255, 2007. 15, 16, 17, 19, 21, 97

[AHKM$^+$08] A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Rechtsteiner, K. Verspoor, Z. Wang, and L.M. Rocha. Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks. *Genome Biology*, 9(2):S11, 2008. 13, 14, 15, 16, 17, 19, 21, 31, 97, 100

[AHR] A. Abi-Haidar and L.M. Rocha. Collective Classification of Textual Documents Using Self-Organized Cross-Regulatory T-cells in the Adaptive Immune System. *Evolutionary Intelligence*. 7, 96, 115

[AHR08a] A. Abi-Haidar and L. Rocha. Adaptive spam detection inspired by the immune system. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Artificial Life XI: Proceedings of the Eleventh International Conference on the Simulation and Synthesis of Living Systems*, pages 1–8. MIT Press, Cambridge, MA, 2008. 7, 48, 50, 68

[AHR08b] A. Abi-Haidar and L.M. Rocha. Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift. In *Artificial Immune Systems: 7th International Conference, Icaris 2008, Phuket, Thailand, August 10-13, 2008, Proceedings*, page 36. Springer, 2008. 7, 48, 50, 68, 97

[AHR10a] Alaa Abi-Haidar and Luis M. Rocha. Biomedical article classification using an agent-based model of t-cell cross-regulation. In Hart et al., editor, *ICARIS 2010: Proc. of the 9th Int. Conf. on Artificial Immune Systems*, LNCS, page In Press., 2010. 7, 96, 115

[AHR10b] Alaa Abi-Haidar and Luis M. Rocha. Collective classification of biomedical articles using t-cell cross-regulation. In S. Rasmussen et al (Eds.), editor, *Artificial Life XII: Twelfth International Conference on the Simulation and Synthesis of Living Systems.* ., LNCS, Springer-Verlag, page In Press., 2010. 7, 96, 115

[AKCS00] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, and C.D. Spyropoulos. *An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages.* ACM Press New York, NY, USA, 2000. 80

[AN07] A. Asuncion and D.J. Newman. UCI machine learning repository. *School of Information and Computer Sciences. University of California, Irvine, California, USA*, 2007. 26

[APABD$^+$01] O. Annacker, R. Pimenta-Araujo, O. Burlen-Defranoux, T.C. Barbosa, A. Cumano, and A. Bandeira. CD25+ CD4+ T cells regulate the expansion of peripheral CD4 T cells through the production of IL-10. *The Journal of Immunology*, 166(5):3008, 2001. 57

[BB06] G.B. Bezerra and T.V. Barra. An Immunological Filter for Spam. *International Conference on Artificial Immune Systems (ICARIS 2006), LNCS*, pages 446–458, 2006. 28, 46, 70

[BEFG02] J. Balthrop, F. Esponda, S. Forrest, and M. Glickman. Coverage and generalization in an artificial immune system. In *Proceedings of GECCO*, pages 3–10, 2002. 43

[BMC06] T. Baptista, T. Menezes, and E. Costa. BitBang: A Model and Framework for Complexity Research. In *Proc. of the European Conference on Complex Systems 2006*, 2006. 133

[BR05] P.O. Boykin and V.P. Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, 2005. 28

[Bre96] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 22

[Bur59] F.M. Burnet. *The Clonal Selection Theory of Acquired Immunity*. Vanderbilt University Press, 1959. 39, 44

[CBH04] AM Cohen, RT Bhupatiraju, and WR Hersh. Feature generation, feature selection, classifiers, and conceptual drift for biomedical document triage. In *The Thirteenth Text Retrieval Conference: TREC*, 2004. 4

[CDN05] P.A. Chirita, J. Diederich, and W. Nejdl. MailRank: using ranking for spam detection. *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 373–380, 2005. 28

[CH95] DE Cooke and JE Hunt. Recognising promoter sequences using an artificial immune system. *Proc Int Conf Intell Syst Mol Biol*, 3:89–97, 1995. 42

[CH98] W.W. Cohen and H. Hirsh. Joins that generalize: Text classification using WHIRL. In *Proceedings of KDD-98, 4th International Conference on Knowledge Discovery and Data Mining*, pages 169–173, 1998. 16

[CJK04] N.V. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004. 22

[CLC$^+$07] J. Carneiro, K. Leon, Í. Caramalho, C. van den Dool, R. Gardner, V. Oliveira, M.L. Bergman, N. Sepúlveda, T. Paixão, J. Faro, 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. 5, 29, 40, 44, 47, 48, 51, 57, 63, 73, 119, 133

[CM95] James Crutchfield and Melanie Mitchell. The evolution of emergent computation. *PNAS*, 92(23), 1995. 3, 124

[Cou80] A. Coutinho. The self-nonself discrimination and the nature and acquisition of the antibody repertoire. *Ann Immunol (Paris)*, 131(3):235–53, 1980. 39, 44

[CRS03] S. Chakrabarti, S. Roy, and M.V. Soundalgekar. Fast and accurate text classification via multiple linear discriminant projections. *The VLDB journal*, 12(2):170–185, 2003. 19, 65

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995. 10.1007/BF00994018. 18

[Das98] D. Dasgupta. *Artficial Immune Systems and Their Applications.* Springer-Verlag New York, Inc. Secaucus, NJ, USA, 1998. 42, 44

[Das06] D. Dasgupta. Advances in artificial immune systems. *Computational Intelligence Magazine, IEEE*, 1(4):40–49, 2006. 47

[DCS06] Sarah Jane Delany, Padraig Cunningham, and Barry Smyth. Ecue: A spam filter that uses machine leaming to track concept drift. In Gerhard Brewka, Silvia Coradeschi, Anna Perini, and Paolo Traverso, editors, *ECAI 2006, 17th European Conference on Artificial Intelligence, August 29 - September 1, 2006, Riva del Garda, Italy, Including Prestigious Applications of Intelligent Systems (PAIS 2006), Proceedings*, pages 627–631. IOS Press, 2006. 26, 28, 70

[dCT02a] LN de Castro and J. Timmis. An artificial immune network for multimodal function optimization. *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, 1, 2002. 42

[DCT02b] L.N. De Castro and J. Timmis. *Artificial immune systems: a new computational intelligence approach.* Springer Verlag, 2002. 40

[DCT02c] L.N. De Castro and J. Timmis. *Artificial Immune Systems: A New Computational Intelligence Approach.* Springer, 2002. 45

[DCT06] S. J. Delany, P. Cunningham, and A. Tsymbal. A comparison of ensemble and case-base maintenance techniques for handling concept drift in spam filtering. In G. Sutcliffe and R. Goebel, editors, *Proceedings of the 19th International Conference on Artificial Intelligence (FLAIRS 2006)*, pages 340–345. AAAI Press, 2006. 84

[DCTC05a] S. J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle. A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*, 18(4–5):187–195, 2005. 80, 84

[DCTC05b] S.J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle. A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems*, 18(4-5):187–195, 2005. 4, 24

[dCVZ] L.N. de Castro and F.J. Von Zuben. Artificial Immune Systems: Part II–A SURVEY OF Applications. 45

[dCVZ02] L.N. de Castro and F.J. Von Zuben. aiNet: An Artificial Immune Network for Data Analysis. *Data Mining: A Heuristic Approach*, 2002. 42

[DN08] D. Dasgupta and F. Nino. *Immunological Computation: Theory and Applications*. AUERBACH, 2008. 40, 47, 131

[DPHS98] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 148–155. ACM, 1998. 16

[Dum90] S. Dumais. Enhancing performance in latent semantic indexing. Technical Report TM-ARH-017527, Bellcore, 1990. 14

[FHK$^+$91] N. Fuhr, S. Hartmann, G. Knorz, G. Lustig, M. Schwantner, and K. Tzeras. AIR/X–a rule-based multistage indexing system for large subject fields. In *Proceedings of RIAO-91, 3rd International Conference Recherche dInformation Assistee par Ordinateur*, pages 606–623. Citeseer, 1991. 16

[FMMK07] D. Floreano, S. Mitri, S. Magnenat, and L. Keller. Evolutionary Conditions for the Emergence of Communication in Robots. *Current Biology*, 17(6):514–519, 2007. 133

[Fox89] C. Fox. A stop list for general text. In *ACM SIGIR Forum*, volume 24, page 21. ACM, 1989. 13

[FPAC94] S. Forrest, A.S. Perelson, L. Allen, and R. Cherukuri. Self-nonself discrimination in a computer. *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*, 212, 1994. 42, 44

[FPP86] J.D. Farmer, N.H. Packard, and A.S. Perelson. The immune system, adaptation, and machine learning. *Physica*, 22(2):187–204, 1986. 42, 44, 45

[FRID$^+$07a] F. Fdez-Riverola, EL Iglesias, F. Díaz, JR Méndez, and JM Corchado. Applying lazy learning algorithms to tackle concept drift in spam filtering. *Expert Systems With Applications*, 33(1):36–48, 2007. 26, 70

[FRID$^+$07b] F. Fdez-Riverola, EL Iglesias, F. Díaz, JR Méndez, and JM Corchado. SpamHunting: An instance-based reasoning system for spam labelling and filtering. *Decision Support Systems*, 43(3):722–736, 2007. 15, 26, 28

[FS95] Y. Freund and R. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer, 1995. 22

[FS06] R. Feldman and J. Sanger. *The Text Mining Handbook: advanced approaches in analyzing unstructured data.* Cambridge University Press, 2006. 4, 10, 12, 22, 31, 32, 33, 53, 77, 126, 130

[FU02] U. Fayyad and R. Uthurusamy. Evolving data into mining solutions for insights. *Communications of the ACM*, 45(8):31, 2002. 9, 10

[Gar03] SM Garrett. A paratope is not an epitope: Implications for immune networks and clonal selection. In *2nd International Conference in Artificial Immune Systems*, pages 217–228, 2003. 47

[GAT] J. Greensmith, U. Aickelin, and J. Twycross. Articulation and Clarification of the Dendritic Cell Algorithm. *Proc. of the 5th International Conference on Artificial Immune Systems, LNCS*, 4163:404–417. 45

[GC06] J. Graham-Cumming. Does Bayesian poisoning exist. *Spam Bulletin*, 2006. 2, 69

[GL03] G. Guo and S.Z. Li. Content-based audio classification and retrieval by support vector machines. *IEEE Transactions on Neural Networks*, 14(1):209–215, 2003. 12

[Gre03] J. Greensmith. New Frontiers For An Artificial Immune System. *Digital Media Systems Laboratory HP Laboratories Bristol HPL-2003-204, October 7th*, 2003. 45

[GZK05] M.M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining data streams: a review. *ACM Sigmod Record*, 34(2):18–26, 2005. 23

[Han06] D.J. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–14, 2006. 4

[HBC04] William Hersh, Ravi Teja Bhupatiraju, and Sarah Corley. Enhancing access to the bibliome: the trec genomics track. *Medinfo*, 11(Pt 2):773–777, 2004. 30

[HC96] J.E. Hunt and D.E. Cooke. Learning using an artificial immune system. *Journal of Network and Computer Applications*, 19(2):189–212, 1996. 42

[HC06] L. Hunter and K.B. Cohen. Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell*, 21(5):589–594, 2006. 2

[Hea99] M.A. Hearst. Untangling text data mining. *Proceedings of the 37th conference on Association for Computational Linguistics*, pages 3–10, 1999. 2

[HKHR08] Tomás Helikar, John Konvalina, Jack Heidel, and Jim A Rogers. Emergent decision-making in biological signal transduction networks. *Proc Natl Acad Sci U S A*, 105(6):1913–1918, Feb 2008. 3, 124

[Hof01] S.A. Hofmeyr. An Interpretative Introduction to the Immune System. *Design Principles for the Immune System and Other Distributed Autonomous Systems*, 2001. 3, 28, 36, 38, 70, 120, 122, 128

[HSF$^+$09] Shuang Hao, Nadeem Ahmed Syed, Nick Feamster, Alexander G. Gray, and Sven Krasser. Detecting spammers with snare: spatio-temporal network-level automatic reputation engine. In *Proceedings of the 18th conference on USENIX security symposium*, SSYM'09, pages 101–118, Berkeley, CA, USA, 2009. USENIX Association. 28

[HT08] E. Hart and J. Timmis. Application areas of AIS: The past, the present and the future. *Applied Soft Computing Journal*, 8(1):191–201, 2008. 47

[HYBV05] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005. 29, 30

[Jam05] S.C. Jameson. T cell homeostasis: keeping useful T cells alive and live T cells useful. In *Seminars in immunology*, volume 17, pages 231–237. Elsevier, 2005. 119

[JC06] H. Jiang and L. Chess. Regulation of immune responses by T cells. *New England Journal of Medicine*, 354(11):1166, 2006. 40

[Jer74] NK Jerne. Towards a network theory of the immune system. *Ann Immunol (Paris)*, 125(1-2):373–89, 1974. 42, 44, 45

[Joa98] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, pages 137–142, 1998. 16

[Joa99] T. Joachims. SVMLight: Support Vector Machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 1999. 19

[Joa02] T. Joachims. *Learning to classify text using support vector machines: methods, theory, and algorithms*. Kluwer Academic Publishers, 2002. 18, 109

[JSB06a] Lars Juhl Jensen, Jasmin Saric, and Peer Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7(2):119–129, Feb 2006. 29

[JSB$^+$06b] L.J. Jensen, J. Saric, P. Bork, et al. Literature mining for the biologist: from information retrieval to biological discovery. *NATURE REVIEWS GENETICS*, 7(2):119, 2006. 2, 10

[JTWS96] C.A. Janeway, P. Travers, M. Walport, and M. Shlomchik. *Immunobiology*. Garland Pub New York, 1996. 40

[KA01] A. Kolcz and J. Alspector. SVM-based filtering of e-mail spam with content-specific misclassification costs. *Proceedings of the TextDM*, 1, 2001. 28, 69

[KAHK$^+$09] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. Hamed, and L. M. Rocha. 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, 2009*, page 34, 2009. 15, 16, 17, 19, 21

[KAHK$^+$10] A. Kolchinsky, A. Abi-Haidar, J. Kaur, A. Hamed, and L. M. Rocha. Classification of protein-protein interaction documents using text and citation network features. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.*, page In Press, 2010. xxii, 14, 15, 16, 17, 19, 21, 31, 100, 101, 110

[KHA99] M.G. Kelly, D.J. Hand, and N.M. Adams. The impact of changing populations on classifier performance. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 367–371. ACM New York, NY, USA, 1999. 4, 22, 24

[KKP06] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas. Handling imbalanced datasets: a review. *GESTS International Transactions on Computer Science and Engineering. v30 i1*, pages 25–36, 2006. 22

[KM07] J.Z. Kolter and M.A. Maloof. Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007. 26

[Kra09] M Krallinger. The biocreative ii. 5 challenge overview. In *Proc. the BioCreative II.5 Workshop 2009 on Digital Annotations*, page 19, 2009. 98, 110

[KT01] T. Knight and J. Timmis. AINE: An immunological approach to data mining. In *Proceedings IEEE International Conference on Data Mining, 2001. ICDM 2001*, pages 297–304, 2001. 45

[Kun04] L.I. Kuncheva. Classifier ensembles for changing environments. *Lecture Notes in Computer Science*, 3077:1–15, 2004. 4, 22, 24

[KV07] M. Krallinger and A. Valencia. Evaluating the Detection and Ranking of Protein Interaction Relevant Articles: the BioCreative Challenge Interaction Article Sub-task (IAS). In *BioCreAtIvE II Workshop, Madrid*, pages 29–39, 2007. 21, 22, 30, 97, 132

[Lew98] D. Lewis. Naive (Bayes) at forty: The independence assumption in information retrieval. *Machine Learning: ECML-98*, pages 4–15, 1998. 16, 17

[LJ98] YH Li and AK Jain. Classification of text documents. *The Computer Journal*, 41(8):537, 1998. 16

[LL99] S.L.Y. Lam and D.L. Lee. Feature reduction for neural network based text categorization. In *Proceedings of DASFAA-99, 6th IEEE International Conference on Database Advanced Systems for Advanced Application*, pages 195–202, 1999. 17

[LMK$^+$10] F. Leitner, S.A. Mardis, M. Krallinger, G. Cesareni, L.A. Hirschman, and A. Valencia. An Overview of BioCreative II. 5. *IEEE IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 385–399, 2010. 21, 31, 33

[MAP06] V. Metsis, I. Androutsopoulos, and G. Paliouras. Spam Filtering with Naive Bayes–Which Naive Bayes? *Third Conference on Email and Anti-Spam (CEAS)*, 2006. 17, 28, 69, 81, 109

[Mar61] ME Maron. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417, 1961. 16

[Mat94] P. Matzinger. Tolerance, Danger and the Extended Family. *Annual Review of Immunology*, 12(1):991–1045, 1994. 44, 45

[McC05] Andrew McCallum. Information extraction: Distilling structured data from unstructured text. *Queue*, 3(9):48–57, 2005. 10, 11

[Men03] F. Menczer. Complementing search engines with online web mining agents. *Decision Support Systems*, 35(2):195–212, 2003. 10

[MFRI$^+$06] JR Méndez, F. Fdez-Riverola, EL Iglesias, F. Díaz, and JM Corchado. Tracking Concept Drift at Feature Selection Stage in SpamHunting: an Anti-Spam Instance-Based Reasoning System. *Proceedings of the 8th European Conference on Case-Based Reasoning, ECCBR-06*, pages 504–518, 2006. 4, 24, 28, 69

[Mit06] Melanie Mitchell. Complex systems: Network thinking. *Artificial Intelligence*, 170(18):1194–1212, 2006. 3, 124

[MM99] M.A. Maloof and R.S. Michalski. AQ-PM: A System for Partial Memory Learning. In *Proceedings of the Eighth Workshop on Intelligent Information Systems*, page 7079, 1999. 25

[M"ol88] G. M
"oller. Do suppressor T cells exist? *Scandinavian journal of immunology*, 27(3):247–250, 1988. 40

[MRV$^+$06] AG Maguitman, A. Rechtsteiner, K. Verspoor, CE Strauss, and LM Rocha. Large-scale testing of bibliome informatics using Pfam protein families. *Pac Symp Biocomput*, 76:87, 2006. 30

[MW04] T.A. Meyer and B. Whateley. SpamBayes: Effective open-source, Bayesian based, email classification system. In *Proceedings of the First Conference on Email and Anti-Spam (CEAS)*, volume 98. Citeseer, 2004. 84

[MWMW05] L.J. McHeyzer-Williams and M.G. McHeyzer-Williams. Antigen-specific memory B cell development. *Immunology*, 23(1):487, 2005. 117

[Mye99] G. Myers. Whole-genome DNA sequencing. *Computing in Science & Engineering [see also IEEE Computational Science and Engineering]*, 1(3):33–43, 1999. 2

[NGD02] O. Nasaroui, F. Gonzalez, and D. Dasgupta. The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling. In *IEEE International Conference on Fuzzy Systems*, pages 711–716, 2002. 46

[Oda05] T. Oda. A Spam-Detecting Artificial Immune System. Master's thesis, Carleton University, 2005. 28, 70

[Par90] G. Parisi. A Simple Model for the Immune Network. *Proceedings of the National Academy of Sciences*, 87(1):429–433, 1990. 44

[Per89] AS Perelson. Immune network theory. *Immunol Rev*, 110:5–36, 1989. 43

[PO79] A.S. Perelson and G.F. Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol*, 81(4):645–670, 1979. 43

[Por80a] MF Porter. An algorithm for suffix stripping. *Program*, 13(3):130–137, 1980. 100

[Por80b] MF Porter. An algorithm for suffix stripping (1980). *Program*, 14:130–137, 1980. 14, 71

[Por06] MF Porter. An algorithm for suffix stripping. *Program 1966-2006: Celebrating 40 Years of ICT in Libraries, Museums and Archives*, 2006. 14

[PPP93] JK Percus, OE Percus, and AS Perelson. Predicting the Size of the T-Cell Receptor and Antibody Combining Region from Consideration of Efficient Self-Nonself Discrimination. *Proceedings of the National Academy of Sciences*, 90(5):1691–1695, 1993. 43

[PPSO04] W.K. Purves, W.K. Purves, D. Sadava, and G.H. Orians. *Life: The Science of Biology: Volume III: Plants and Animals*, volume 3. WH Freeman & Co, 2004. 121

[Pra05] Stephen C. Pratt. Quorum sensing by encounter rates in the ant temnothorax albipennis. *Behav. Ecol.*, 16(2):488–496, 2005. 3, 124

[PT93] W.E. Paul and I.O. Technologies. *Fundamental immunology.* Raven Press New York, 1993. 39

[PW97] A.S. Perelson and G. Weisbuch. Immunology for physicists. *Reviews of Modern Physics*, 69(4):1219–1268, 1997. 118

[PWMM04] David Peak, Jevin D. West, Susanna M. Messinger, and Keith A. Mott. Evidence for complex, collective dynamics and distributed emergent computation in plants. *PNAS*, 101(4):918–922, 2004. 3, 124

[Qui87] J.R. Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987. 22

[RH05] L.M. Rocha and W. Hordijk. Material representations: From the genetic code to the evolution of cellular automata. *Artificial Life*, 11(1-2):189–214, 2005. 3, 124

[RLRS06] A Rechtsteiner, J Luinstra, L M Rocha, and C Strauss. Use of text mining for protein structure prediction and functional annotation in lack of sequence homology. volume Joint BioLINK and Bio-Ontologies Meeting 2006 (ISMB Special Interest Group), 2006. 30

[RNJ$^+$10] Y.P. Rubtsov, R.E. Niec, S. Josefowicz, L. Li, J. Darce, D. Mathis, C. Benoist, and A.Y. Rudensky. Stability of the Regulatory T Cell Lineage in Vivo. *Science*, 329(5999):1667, 2010. 119

[Rob07] A. Robins. Innate and Adaptive Immunity. *In Silico Immunology*, pages 11–21, 2007. 36, 40

[Roc71] JJ Rocchio. Relevance feedback in information retrieval in The SMART Retrieval System-Experiments in Automatic Document Processing. *Salton ed*, pages 313–323, 1971. 16

[RODV04] P. Radivojac, Z. Obradovic, A.K. Dunker, and S. Vucetic. Feature selection filters based on the permutation test. *Machine Learning: ECML 2004*, pages 334–346, 2004. 16

[Sak04] S. Sakaguchi. Naturally arising CD4 regulatory T cells for immunologic self-tolerance and negative control of immune responses. *Annu. Rev. Immunol*, 22:531–562, 2004. 40

[Sal89] G. Salton. Automatic text processing: the transformation. *Analysis and Retrieval of Information by Computer*, 1989. 14, 15

[Sal91] G. Salton. The Smart document retrieval project. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, page 358. ACM, 1991. 12

[SAWA98] M.K. Slifka, R. Antia, J.K. Whitmire, and R. Ahmed. Humoral immunity due to long-lived plasma cells. *Immunity*, 8(3):363–372, 1998. 118

[SC98] J. Stewart and J. Carneiro. The Central and the Peripheral Immune Systems: What is the Relationship? *Artificial Immune Systems and Their Applications*, page 47, 1998. 44

[SC01] L.A. Segel and I. Cohen. *Design Principles for the Immune System and Other Distributed Autonomous Systems.* Oxford University Press, 2001. 3, 123

[SDHH98] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A Bayesian approach to filtering junk e-mail. *Learning for Text Categorization: Papers from the 1998 Workshop*, 62, 1998. 28, 69

[Seb02] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002. 14, 16

[Sep09] Nuno H. Sepulveda. *How is the T-cell repertoire shaped.* PhD thesis, Instituto Gulbenkian de Ciencia, 2009. 50, 119

[Set05] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text, 2005. 21

[SF03] Hagit Shatkay and Ronen Feldman. Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–856, 2003. 10, 29

[SFT03] A. Secker, AA Freitas, and J. Timmis. AISEC: an artificial immune system for e-mail classification. In *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 1, 2003. 46

[SG86] J.C. Schlimmer and R.H. Granger. Incremental learning from noisy data. *Machine Learning*, 1(3):317–354, 1986. 26, 32, 130

[SHR$^+$06] Cosma Shalizi, Rob Haslinger, Jean-Baptiste Rouquier, Kristina Klinkner, and Cristopher Moore. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Phys.Rev.E*, 73, 2006. 3, 124

[SJS06] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021, 2006. 102, 110

[SK01] W.N. Street and Y.S. Kim. A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382. ACM New York, NY, USA, 2001. 26

[SL31] W.A. Sawyer and W. Lloyd. The use of mice in tests of immunity against yellow fever. *The Journal of Experimental Medicine*, 54(4):533, 1931. 116, 117

[SLS99] N.A. Syed, H. Liu, and K.K. Sung. Handling concept drifts in incremental learning with support vector machines. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 317–321. ACM New York, NY, USA, 1999. 26, 55

[SSD$^+$95] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science(Washington)*, 270(5235):467–470, 1995. 2

[TA07] J. Twycross and U. Aickelin. Biological Inspiration for Artificial Immune Systems. *LECTURE NOTES IN COMPUTER SCIENCE*, 4628:300, 2007. 47

[Tay07] L.S. Taylor. Power distribution outage cause identification with imbalanced data using artificial immune recognition system (AIRS) algorithm. *IEEE Transactions on Power Systems*, 22(1), 2007. 46

[TC02] J. Twycross and S. Cayzer. An immune system approach to document classification. *Master's thesis, COGS, University of Sussex, UK*, 2002. 45

[TH03] C. Tullo and J.R. Hurford. Modelling Zipfian distributions in language. In *Proceedings of Language Evolution and Computation Workshop/-Course at ESSLLI*, pages 62–75. Citeseer, 2003. 130

[Tim07] J. Timmis. Artificial immune systems today and tomorrow. *Natural Computing*, 6(1):1–18, 2007. 40, 42, 47

[TN01] J. Timmis and M. Neal. A resource limited artificial immune system for data analysis. *Knowledge-Based Systems*, 14(3-4):121–130, 2001. 42

[TPCP06] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Handling local concept drift with dynamic integration of classifiers: domain of antibiotic resistance in nosocomial infections. In *Proceedings* *of the 19th IEEE Symposium on Computer-Based Medical Systems*, pages 679–684. IEEE Computer Society Washington, DC, USA, 2006. 4, 24

[TPCP08] Alexey Tsymbal, Mykola Pechenizkiy, Pdraig Cunningham, and Seppo Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1):56 – 68, 2008. Special Issue on Applications of Ensemble Methods. 26

[TR95] C. Tanchot and B. Rocha. The peripheral T cell repertoire: independent homeostatic regulation of virgin and activated CD8+ T cell pools. *European journal of immunology*, 25(8):2127–2136, 1995. 117

[TS94] D.F. Tough and J. Sprent. Turnover of naive-and memory-phenotype T cells. *The Journal of experimental medicine*, 179(4):1127, 1994. 118

[TSC10] C.Y. Tseng, P.C. Sung, and M.S. Chen. Cosdes: A Collaborative Spam Detection System with a Novel Email Abstraction Scheme. *IEEE Transactions on Knowledge and Data Engineering*, 2010. 28

[Tsy04] A. Tsymbal. The problem of concept drift: definitions and related work. *Informe técnico: TCD-CS-2004-15, Departament of Computer Science Trinity College, Dublin, https://www. cs. tcd. ie/publications/techreports/reports*, 4:1–15, 2004. 4, 23, 24, 25, 28, 46, 69

[Vap95] V. Vapnik. The nature of statistical learning, 1995. 18

[VC91] FJ Varela and A. Coutinho. Second generation immune networks. *Immunol Today*, 12(5):159–66, 1991. 42, 45

[VCJ$^+$05] Karin Verspoor, Judith Cohn, Cliff Joslyn, Sue Mniszewski, Andreas Rechtsteiner, Luis Mateus Rocha, and Tiago Simas. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, 6 Suppl 1:S20, 2005. 31

[VR05] S. Visa and A. Ralescu. Issues in mining imbalanced data sets-A review paper. In *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference, MAICS*, pages 16–17, 2005. 22

[Wei04] G.M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1):7–19, 2004. 22

[WK96] G. Widmer and M. Kubat. Learning in the Presence of Concept Drift and Hidden Contexts. *Machine Learning*, 23(1):69–101, 1996. 25

[Wol92] D.H. Wolpert. Stacked generalization*. *Neural networks*, 5(2):241–259, 1992. 22

[WPL$^+$97] N. Weng, L.D. Palmer, B.L. Levine, H.C. Lane, C.H. June, and R.J. Hodes. Tales of tails: regulation of telomere length and telomerase activity during lymphocyte development, differentiation, activation, and aging. *Immunological reviews*, 160(1):43–54, 1997. 118

[WRR03] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In D.P. Berrar, W. Dubitzky, and M. Granzow, editors, *A Practical Approach to Microarray Data Analysis*, pages 91–109. Kluwer, Norwell,MA, 2003. 14

[WS06] Matthew Walters and Vanessa Sperandio. Quorum sensing in escherichia coli and salmonella. *Int. Journal of Medical Microbiology*, 296(2-3):125 – 131, 2006. 3, 124

[WT04] A. Watkins and J. Timmis. Artificial Immune Recognition System (AIRS): Revisions and Refinements. *AISB 2004 Convention*, 2004. 45

[YAC$^+$07] X. Yue, A. Abraham, Z.X. Chi, Y.Y. Hao, and H. Mo. Artificial immune system inspired behavior-based anti-spam filter. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, 11(8):729–740, 2007. 28

[YHM02] A. Yeh, L. Hirschman, and A. Morgan. Background and overview for KDD Cup 2002 task 1: Information extraction from biomedical articles. *ACM SIGKDD Explorations Newsletter*, 4(2):87–89, 2002. 30

[YL99] Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 1999. 16, 17

[YP97] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*, pages 412–420. Citeseer, 1997. 14

Alaa E. Abi-Haidar

Department of Informatics
School of Informatics and Computing
Indiana University
919 E. 10th Street, Room 330
Bloomington, IN 47408

Phone: (812) 856-5754
Fax: (812) 856-4764

alahay@alahay.org
www.alahay.org

Education

Ph.D. in Informatics, Complex Systems Track (Major), Indiana University, 2011

Ph.D. in Computer Science (Minor), Indiana University, 2011

M.S. in Computer Science, Indiana University, 2005

B.S. in Computer Science, American University of Beirut, 2003

Teaching Experience

Assistant Instructor 2008–2011
Indiana University Indiana University
Assistant Instructor for the following courses: Simplicity of Complexity (I400), Informatics Management (I502), Introduction to Programming (I210) and Information Representation (I308)

Research Experience

Research Assistant 2006–2008
Luis M. Rocha Indiana University
Immune-inspired spam detection, bioliterature mining and classification, and social dynamics at CASCI.

Research Assistant 2006–2008
Luis M. Rocha Instituto Gulbenkian de Ciência
The BioCreAtIvE bioliterature tex mining challenge and T-cell cross-regulation at the FLAD Collaboratorium .

Research Assistant 2005
Katy Börner Indiana University
Information visualization of taxonomy data.

Research Assistant 2004–2006
Peter J. Ortoleva Indiana University
PHP and SQL Development, data visualization and rendering, text parsing, web and text mining, data curation, Transcripional Regulatory Networks, and Gene Ontology at CCVT.

Research Assistant 2003
Rabih Sultan American University of Beirut
Image processing software design, image density calculator, web design chemistry department website at AUB.

Employment

LebSocrates Beirut, Lebanon
Head of Programming and Web Design 2001–2003
Computer sales, maintenance, software development, and web design.

Alaa E. Abi-Haidar

Programming Skills

Expert: R, PHP, Python, SQL, HTML, LaTeX, Linux, Windows, Flash, VI, NetLogo

Intermediate: Matlab, openGL, VTK, JavaScript, Java, C, C++ , CSS

Basic: Perl

Software Skills

Expert: MaYA, GIMP, Microsoft Office and OpenOffice, Adobe Photoshop, Linux and Windows OS, Windows Movie Maker

Intermediate: Poser, AutoCAD, 3DStudio MAX, LightWave, Bryce, Cinema 4D, Corel Draw Studio, Borland CBuilder, Apache, MySQL , Macromedia Flash

Language Proficiency

Fluent[1]: English, Italian, French, Levantine Arabic, Portuguese, Spanish, Modern Standard Arabic

Basic: Modern Greek, Biblical Hebrew, Armenian

Service Activities

Coffee Artist and Photographer with four exhibitions and one live coffee painting demo at the Venue Gallery.

Founder and organizer of the French Table, *L'hebdofrancofolie*, at Indiana University in Bloomington, IN (2004–2008)

Founder of Friends of Artists (2011)

Free-time photographer for UPashion

Prizes Awards

Winner of best poster at the 1st Portuguese Forum on Computational Biology (FPBC 2008)

Awarded internships at the Instituto Gulbenkian de Ciência (IGC) in 2006, 2007, 2008 and 2010.

Awarded Bursaries to attend AlifeX and ICARIS2010.

Winner of the best paper award in the International Conference of Artificial Immune Systems (ICARIS 2010)

[1] Speaking and Writing

Alaa E. Abi-Haidar

Most Recent Publications

A. Abi-Haidar and Luis. M. Rocha [2011]. *Collective Classification of Textual Documents by Guided Self-Organization in T-Cell Cross-regulation Dynamics.* Journal of Evolutionary Intelligence. Evolutionary Intelligence, DOI: 10.1007/s12065-011-0052-5

A. Loureno, M. Conover, A. Wong, F. Pan, Alaa Abi-Haidar, A. Nematzadeh, H. Shatkay, and L.M. Rocha [2010].*Testing Extensive Use of NER tools in Article Classification and a Statistical Approach for Method Interaction Extraction in the Protein-Protein Interaction Literature.* Proceedings of the BioCreative III Workshop 2010, Bethesda, Maryland, September 13-15, 2010.

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). Winner of Best Paper Award. E. Hart et al (Eds.) Lecture Notes in Computer Science. Springer-Verlag, 6209, 237-249.

A. Abi-Haidar and L.M. Rocha [2010]. *Collective Classification of Biomedical Articles using T-Cell Cross-regulation.* In: Artificial Life XII: Twelfth International Conference on the Simulation and Synthesis of Living Systems. H. Fellermann et al et al (Eds.). MIT Press, pp. 706-713.

A. Kolchinsky, A. Abi-Haidar, J. Kaur, A.A. Hamed and L.M. Rocha [2010]. *Classification of protein-protein interaction full-text documents using text and citation network features.* IEEE/ACM Transactions On Computational Biology And Bioinformatics, 7(3):400-411.

A. Kolchinsky, A. Abi-Haidar, J. Kaur, A.A. Hamed and L.M. Rocha [2009].*Classification of protein-protein interaction documents using text and citation network features.* Proceedings of the BioCreative II.5 Workshop 2009: Special Session on Digital Annotations, Madrid, Spain, October 7-9, 2009. pp 34.

A. Abi-Haidar and L.M. Rocha [2008]. *Adaptive Spam Detection Inspired by a Cross-Regulation Model of Immune Dynamics: A Study of Concept Drift.* In: Proceedings of 7th International Conference on Artificial Immune Systems (ICARIS 2008). Doheon Lee, Peter Bentley, Sungwon Jung (Eds.) Lecture Notes in Computer Science. Springer-Verlag, Volume 5132/2008 36-47 doi:10.1007/978-3-540-85072-4.

A. Abi-Haidar and L.M. Rocha [2008]. *Adaptive Spam Detection Inspired by the Immune System.* In: Artificial Life XI: Eleventh International Conference on the Simulation and Synthesis of Living Systems. S. Bullock, J. Noble, R. A. Watson, and M. A. Bedau (Eds.). MIT Press, pp. 1-8.

A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Retchsteiner, K. Verspoor, Z. Wang, and L.M. Rocha [2008]. *Uncovering protein interaction in abstracts and text using a novel linear model and word proximity networks.* Genome Biology 2008, 9(Suppl 2):S11doi:10.1186/gb-2008-9-s2-s11.

A. Abi-Haidar, J. Kaur, A. Maguitman, P. Radivojac, A. Retchsteiner, K. Verspoor, Z. Wang, and L.M. Rocha [2007]. *Uncovering Protein-Protein Interactions in the Bibliome.* Proceedings of the Second BioCreative Challenge Evaluation Workshop (ISBN 84-933255-6-2).

Qu, K., A.E. Abi Haidar, J. Fan, D. Basu, G. Lin, L. Ensman, M. Jolly, P. Ortoleva.

[2007]. *Cancer Onset and Progression: A Genome-Wide, Nonlinear Dynamical Systems Perspective on Onconetworks.* Journal of Theoretical Biology. Volume 246, Issue 2, 21 Pages 234-244

Sun, J., K. Tuncay, A.Abi Haidar, F. Stanley, M. Trelinkski, and P. Ortoleva. [2007]. *Transcriptional Regulatory Network Discovery via Multiple Method Integration: Application to E.coli K12.* Algorithms in Molecular Biology, 2:2 doi:10.1186/1748-7188-2-2

Kagan Tuncay, Lisa Ensman, Jingjun Sun, Alaa Abi Haidar, Frank Stanley, Michael Trelinski and Peter Ortoleva [2006]. *Transcriptional regulatory networks via gene ontology and expression data.* In Silico Biology 7, 0003

R.Sultan, Z. Shreif, Lara, A. Abi-Haydar [2004]. *Taming ring morphology in 2D* $Co(OH)_2$ *Liesegang patterns.* Phys. Chem. Chem. Phys., 6, 3461 - 3466

Selected Press Media Releases

IGC News (17/08/2010) Best paper awarded to IGC scientist at the 9th ICARIS

BBC Digital Planet Interview (2008) .

Telegraph (05/08/2008) Can we make software that comes to life?

Science Daily (06/08/2008) ALife Conference To Reveal Bio-inspired Spam Detection

Silicon Republic (06/08/2008) Fighting spam and thinking robots – bio-inspired AI

Portuguese Press (07/08/2008) Correio Manhã (p20 ext136180), Diário Notícias (p15 ext79040), Jornal Notícias (p5,28 ext150515), Público (p8 ext75000)

CiênciaHoje (11/08/2008) Cientistas portugueses criaram aspirador de lixo electrónico

Tecnomania (12/08/2008) Podemos hacer software que nazca a la vida?

L'atelier (22/08/2008) Spams et bactéries : même combat !

Extracurricular Press Media Releases

Telegraph (09/08/2008) Spore: The science of Spore

Indiana Daily Student (02/03/2010) Local artist uses coffee to make brewed art

Indiana Daily Student (03/03/2010) Student transforms coffee into surrealist art