A Multi-Layered Immune Inspired Approach to Data Mining

Thomas Knight and Jon Timmis
Computing Laboratory
University of Kent at Canterbury
Canterbury, Kent, CT2 7NF, United Kingdom
e-mail: {tpk1, jt6}@ukc.ac.uk

Artificial Immune Systems (AIS) have recently been proposed as an additional soft computing paradigm. This paper presents a new immune inspired algorithm, which has been designed to augment a framework for engineering AIS. Initial results have proven to be promising and are detailed here. The new algorithm has shown that it is capable of data compression, representation and clustering.

Keywords: Artificial Immune Systems, Data Mining, Unsupervised Machine Learning, Clonal Selection, Immune Networks, Cluster Analysis, Kohonen Networks

1 Introduction

Soft Computing has been described as computational systems that exploit tolerance for imprecision, uncertainty, partial truth and approximation [8]. Such systems include *artificial neural networks, fuzzy systems, evolutionary algorithms* and *probabilistic reasoning. Artificial Immune Systems* (AIS) have recently been proposed as an additional soft computing paradigm [5]. It has been argued that AIS exhibit similar characteristics to other soft computing paradigms and therefore can be used to complement and augment existing soft computing techniques. AIS can be defined as *adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models which are applied to problem solving* [4]. This paper presents a new immune inspired algorithm that augments the AIS framework proposed in [4, 5]. Preliminary results on three numerical data sets are given and future directions are discussed. This paper argues that this new algorithm adheres to the soft computing philosophy and therefore proposes this as a novel algorithm with regard to both soft computing and immune inspired algorithms.

2 Context of Work

The motivation for this work was to produce an immune inspired system that is capable of continuous learning, identification of clusters within data, automatic stopping criteria and production of simplified maps of the input data. These goals have been identified from some of the shortfalls of the previous work [11]. An additional aim was to develop an algorithm that would augment the framework described in section 2.2 by expanding on previous work outlined in section 2.1. This is done by re-visiting the immune system and looking for new inspiration that can be abstracted to address shortfalls in the previous work.

2.1 Previous Work

The natural immune system is interesting from a computational computing perspective as it is *distributed, diverse, self-organizing, dynamic* and capable of *recognition learning* and *memory* [2].

An immune inspired data mining algorithm, AINE (Artificial Immune NEtwork) was proposed by work in [11]. The motivation behind this work was to abstract processes and characteristics from the immune system, more specifically the immune network theory [7] to develop a novel *unsupervised machine learning* algorithm. The main mechanism of AINE is the co-stimulating and suppressing interactions between B-cells that are inspired by Jerne's Immune Network theory and the resultant response of these B-cells to antigenic stimulation. AINEs artificial immune networks were applied to unsupervised machine learning benchmark data and was reported to perform well on benchmark tasks. It was anticipated by the authors that this algorithm would be suitable for continuous learning with initial work showing that stable populations within the networks could be achieved [11]. However, more recent work has shown that these networks suffer strong evolutionary pressure and converge to the strongest class represented in the data [9]. Whilst this is an interesting development that could potentially be applied to optimization, however, with regard to data mining it would not be a preferential outcome. From a continuous learning point-of-view it is more desirable if all patterns persist over time rather than the strongest. In parallel to the work in this paper [10] has since developed a form of the original algorithm that is capable of finding stable clusters. Here, a different population control mechanism based on exponential decay of stimulation levels has been used. This work uses a new resource allocation mechanism and stimulation level calculations. However, one drawback of this work it that there is poor data compression in the final networks. These observations prompted a step back from the existing work to re-evaluate the approaches taken. It was noted that a more holistic approach may provide a better solution in the search for an immune inspired data mining algorithm capable of continuous learning.

2.2 A Framework for Artificial Immune Systems

A framework for engineering artificial immune systems has been proposed by [4] as an attempt to formalise the development of artificial immune systems. Here the basic requirements for a AIS or Biologically inspired algorithm algorithms are stated as:

- A *representation* for the components of the system;
- A *set of mechanisms to evaluate the interactions of individuals with the environment and each other*. The environment is usually simulated by a set of input stimuli, one or more fitness function(s), or other mean(s);
- *Procedures of adaptation* that govern the dynamics of the system, i.e. how its behavior varies over time.

When one considers the natural immune system in terms of the framework, the representation can be thought of as B-cells, T-cells, antibodies and antigens; the mechanisms that evaluate the interactions are antigenic binding (affinity measures), shape space, feedback mechanisms and procedures for adaption can be though of as; *Negative Selection* [12], *Clonal Selection* [1] and *The Immune Network Theory* [7].

2.3 Immunology

Here, only parts of the adaptive immune systems are going to be discussed and therefore if the reader wishes to know more about the innate immune system, antigen-presenting cells or T-cells, then they are directed to the description of the immune system in [4]. Our bodies are under constant attack by viruses and bacteria that are foreign to our bodies, these are know as *antigens*. As part of the adaptive immune system a B-cell is capable of recognising free antigens without the assistance of MHC (Major Histocompatibility Complex) molecules. B-cells have

surface receptors known as *antibodies* that are antigen specific (will only respond to a specific antigen). On encountering an antigen the B-cell will become stimulated and will proliferate and differentiate into plasma cells that secrete antibody molecules in high volumes. These neutralize the antigen and lead to the removal of the antigen from the system. Some of these activated B-cells and T-cells differentiate into memory cells which remain in the immune system for long periods of time and help produce a more rapid response in the future. Previous work [11] has concentrated on the interactions of B-cells only as described in the immune network theory, however this paper proposes a new algorithm that draws its inspiration from not only B-cells, but the antibodies and memory cells that they produce. Conceptually it is possible to divide the immune system into layers, with each layer having its own specific interactions. Thus, a multi-layered artificial immune system is proposed in the following section.

3 A Multi-layered Artificial Immune System

Much of the previous work in the field of artificial immune systems for data mining has been in developing systems based on the interactions of B-cells. Both clonal selection and the immune network theory have been tried with some success [3, 6, 11]. As an addition to these existing techniques, this paper proposes a multi-layered algorithm that is inspired by the clonal selection theory and incorporates a feedback mechanism much like the co-stimulation that is seen in the immune network theory model. The proposed algorithm also incorporates the idea of a primary immune response to handle the event of unknown data being presented to the system.

At a more abstract level the immune system can be though of in terms of layers. The algorithm is divided into three layers; A *free-antibody layer*, a *B-cell layer* and a *memory cell layer*. In order to simplify the algorithm T-cell, MHC and APC (Antigen Presenting Cell) interactions are not directly considered within the model, however the signaling action that they employ has been incorporated in part to the feedback mechanism between the free-antibody layer and the B-cell layer. The first level of interactions an antigen has is with the free antibodies generated by the B-cells. This interaction occurs in the free-antibody layer. At the next level we consider the antigen interacting with the B-cells which occurs in the B-cell layer. Finally the memory cells that are produced by the B-cells interact in a memory cell layer. The interactions between the layers are described in Figure 1. A more detailed description of each layer is given in the following sections. Throughout the description of the new algorithm training data will be referred to as *antigens*, and the *free antibodies, B-cells* and *memory cells* all have data items which exist in the same shape space.

3.1 Free-antibody Layer

In the free-antibody layer (Figure 1a) the analogy that, upon entry to the immune system an antigenic substance does not get presented to the entire system, only part of it, has been used. As the antigen travels through the system it comes in contact with numerous free-antibodies, some of which are sufficiently similar to bind to the antigen. Whether a free-antibody binds to the antigen is determined by a user defined affinity threshold (σ_{fab}). Once the antigen pattern has been presented to a percentage of the population (ζ) it is then transferred from the free-antibody layer to the B-cell layer.

3.2 B-cell Layer

On entering the B-cell layer (Figure 1b) the antigen is randomly presented each to cell in the B-cell population until it can bind (where the Euclidean distance between the two cells is less than a predefined threshold) to a B-cell. Binding is determined by a user defined affinity threshold

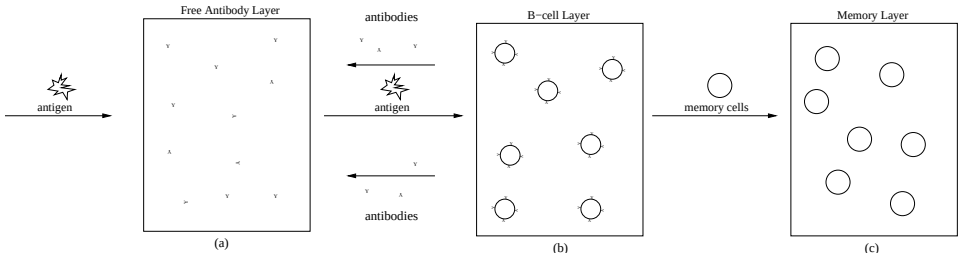

Figure 1: The three layers of the new algorithm showing the flow of cells between each layer

(σ_{bcell}). The antigen stimulates the B-cell that it is attached to, the amount of stimulation is determined by how many of the free antibodies are attached to that antigen. If the stimulation level exceeds a predefined threshold (σ_{stim}) then the B-cell will produce a clone that is then mutated using guided affinity maturation. The B-cell also produces a clone that becomes a memory cell which is then added to the memory layer. After stimulation a B-cell will produce free antibodies according to the equation:

$$n_f = (S_{max} - a_{(ag,bcell)}) * K$$

where n_f is the number of free antibodies produced, S_{max} is the maximum possible distance in the data space between two cells, $a_{(ag,bcell)}$ is the affinity between the antigen (ag) and the B-cell (bcell) and K is a constant. Each free antibody is a slight mutant of the original B-cell. If an antigen does not find a B-cell with sufficient affinity to elicit a response then it undergoes a *primary response* where the antigen data is copied into a new B-cell and added to the B-cell population. This provides new data with an opportunity to be incorporated into the system.

3.3 Memory Layer

On entering the memory layer (Figure 1c) the new memory cell binds to all cells in the layer. This is to ascertain if there are any cells that have an affinity below the memory threshold (σ_{mem}). On encountering the first cell that has an affinity less than σ_{mem} the existing cell is replaced by the new one if the new one has a lower affinity with the antigen that created it in the B-cell layer.

3.4 Population Control

A death threshold (σ_{death}) exists in all layers of the system. If the length of time between the last stimulation of a cell and the current time is greater than σ_{death}, it is removed from the population.

4 Results

This section details the preliminary testing that has been performed on the algorithm. The algorithm has been tested on 3 data sets detailed in Table 1. A plot of the 3Circle data can be seen in Figure 2 and the 3D data set known as the doughnut data originates from [3]. The algorithm was run for 500 iterations using a default set of parameters. The parameters used were empirically defined during the initial development of the algorithm. The experiments were repeated 50 times and the population sizes for each layer were recorded on each iteration. The mean population size for each data set was then calculated and the standard deviation recorded. These results are presented in Table 2. Figure 3 shows the resulting pattern in the Memory

Table 1: Test Data

Name	Type	Dimensions	size	classes
Simulated	numeric	2	30	3
3Circle	numeric	2	600	3
Doughnut	numeric	3	221	2

Table 2: Average Mean ($\bar{x}$) and Standard Deviation (σ) of populations sizes within the three layers.

Layer	Simulated Data		3 Circle Data		Doughnuts Data	
	σ	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{x}$
Free Antibody	14.341	162.315	77.499	1143.492	76.684	1441.08
B Cell	3.503	51.639	43.346	1126.834	18.054	387.755
Memory	1.027	4.86	3.860	50.34	5.185	74.72

Layer after 5 iterations of the algorithm. It can be seen that after a short period of time the pattern is visually representative of the original data. By looking at the mean number of cells in the Memory Layer and the comparing with the original number of data items in the data set we can see that the algorithm achieves good compression ratios (91.6% for the 3 circle data and 66.5% for the doughnut data). The tests have also showed that over 500 iterations the system does not suffer from the loss of information as seen in the AINE system.

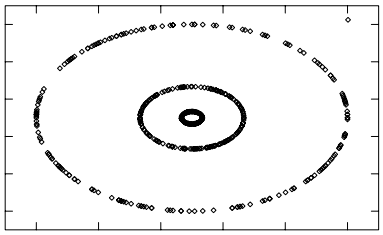

Figure 2: The 3 circle training data set (600 items) in normalised form

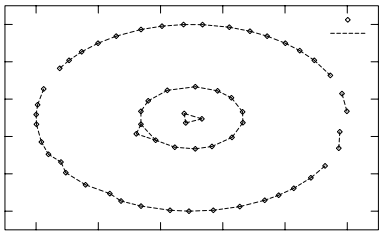

Figure 3: Patterns in the memory layer evolved by the new algorithm after 5 iterations (58 items)

5 Discussion

Previous work has shown that aspects of the human immune system lend themselves to solving computational problems such as data mining. Previously, an artificial immune system for data mining (AINE) [11] was created and was used with some success on benchmark data. However, more recent work has shown that AINE does not perform as the the authors intended it to. It was identified that there were problems concerning the stability of networks produced and how an AIS that produces stable networks is desirable. Parallel to this work [10] has re-designed AINE and succeeded in producing the performance originally intended. The relative youth of the field also has identified that there may well be benefits from trying different solutions to those that already exist. To this end, a new algorithm has been devised that seeks to produce stable results and explores new ideas. The design has also been based on the framework for developing AIS and is proposed as an augmentation for the *Procedures for Adaption*. It can also be said to adhere to the Soft Computing philosophy because it exploits the imprecision,

approximation and uncertainty of the natural immune system.

6 Future Work

Future work will comprise of further evaluation and testing of algorithm on more benchmark data sets. Comparisons with other techniques will be made and a full investigation of the parameters undertaken. Analysis of the clusters generated will also be undertaken to establish how accurate the results are.

7 Acknowledgements

Thomas Knight would like to thank SUN^{TM} Microsystems for their continued financial support for his PhD Studies.

References

[1] F M Burnet. *The Clonal Selection Theory of Immunity.* Vanderbilt University Press, Nashville, TN, 1959.

[2] D Dasgupta. *Artificial Immune Systems and their Applications*, chapter An overview of Artificial Immune Systems and their Applications, pages 3–21. Springer, 1998.

[3] L N de Castro and F J Von Zuben. *Data Mining: A Heuristic Approach*, chapter aiNet: An Artificial Immune Network for Data Analysis, pages 231–259. Idea Group Publishing, London, UK, 2001.

[4] L.N. de Castro and J. Timmis. *Artificial Immune Systems: A New Computational Approach.* Springer-Verlag, London. UK., September 2002.

[5] L.N de Castro and J. Timmis. Artificial Immune Systems as a Novel Soft Computing Paradigm. *Soft Computing*, 7(7), July 2003.

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

[7] N K Jerne. Towards a Network theory of the Immune System. *Annals of Immunology*, 125c:373–389, 1974.

[8] Y Jin. What is soft computing. *http://www.soft-computing.de*, 2002.

[9] T Knight and J Timmis. AINE: An Immunological Approach to Data Mining. In N Cercone, T Lin, and Xindon Wu, editors, *IEEE International Conference on Data Mining*, pages 297–304, San Jose, CA. USA, December 2001. IEEE.

[10] M Neal. An artificial immune system for continuous analysis of time-varying data. In *1st International Conference on Artificial Immune Systems (ICARIS)*, pages 0–2, Canterbury, UK, 2002.

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

[12] I R Tizzard. *Immunolgy An Introduction.* Saunders College Publishing, 4th edition, 1995.