Clonal Selection Based Artificial Immune System for Generalized Pattern Recognition

Terry Huntsberger, Senior Member, IEEE
Autonomous Systems Division
NASA Jet Propulsion Laboratory,
California Institute of Technology
Pasadena, CA, 91109 USA
Terry.Huntsberger@jpl.nasa.gov

Abstract—The last two decades has seen a rapid increase in the application of AIS (Artificial Immune Systems) modeled after the human immune system to a wide range of areas including network intrusion detection, job shop scheduling, classification, pattern recognition, and robot control. JPL (Jet Propulsion Laboratory) has developed an integrated pattern recognition/classification system called AISLE (Artificial Immune System for Learning and Exploration) based on biologically inspired models of B-cell dynamics in the immune system. When used for unsupervised or supervised classification, the method scales linearly with the number of dimensions, has performance that is relatively independent of the total size of the dataset, and has been shown to perform as well as traditional clustering methods. When used for pattern recognition, the method efficiently isolates the appropriate matches in the data set. The paper presents the underlying structure of AISLE and the results from a number of experimental studies.

Keywords—artificial immune system; pattern recognition; classification

I. INTRODUCTION

Artificial Immune Systems (AIS) are loosely modeled after the human immune system. Without going too deeply into the details, the human immune system basically consists of a collection of lymphocyte cells that are produced in the thymus (T-cells) and in the bone marrow (B-cells). The functions of the T-cells include the regulation of other cells’ actions and direct attacks on the host infected cells. The main functions of the B-cells include the production and secretion of antibodies as a response to antigens such as bacteria, viruses and tumor cells. Bindings between the antibody and the antigen occur at epitopic sites on the surface of the antigen where there are “shape” matches. The physical shape and chemical affinity of each antibody makes it specific to a very narrow range of antigens, so they act as efficient pattern recognition engines.

The dynamics of the immune system reaction to an antigen intrusion lends itself well to AIS methods for a wide range of applications including network intrusion detection [1], job shop scheduling [2], pattern recognition [3, 4, 5], and robot control [6, 7]. There have also been a number of studies into use of AIS for image classification [8, 9, 10, 11]. These types of problems are characterized by a large set of data elements (size of images) that the system must be able to handle in an efficient manner. A commonly used short cut is to map the image data elements into a binary representation that leads to fast comparisons between data elements and the existing lymphopathic cells in the system. A recent comprehensive bibliography can be found in [12], and there are also many surveys of the field [13, 14, 15, 16, 17].

AIS approaches can be generally broken into two classes: population based and network based. The population-based algorithms rely on the dynamics of populations of individual immune-like cells. Examples of population-based systems include self/non-self [18, 19], immune libraries [20], clonal selection [21, 22], and hybrids [23]. The network-based algorithms are derived from the Immune Network Theory of Jerne [24], and rely on the interactions between networks of immune-like cells. Examples of network-based systems include case-base reasoning system for classification [25], anomaly detection [26], unsupervised learning [27, 28, 29], and continuous learning [30].

JPL has developed an integrated learning/pattern recognition/classification system called AISLE (Artificial Immune System for Learning and Exploration) based on models of B-cell dynamics found in human immune systems. It is a hybrid of a population-based and a network-based system, where the network interactions are modeled through cross-linking between B-cell populations, and clonal selection is accomplished by solving a differential equation of the B-cell population dynamics.

The next section discusses the B-cells dynamics model and gives some examples of its behavior when “exposed” to multi-
dimensional patterns. The next section describes three experimental studies using a standard machine learning dataset, star tracking, and a satellite hyperspectral sensing dataset. The final section is a summary and description of future directions.

II. B-CELLS DYNAMICS MODEL

Since the B-cell dynamics are used for the system described in this paper, the following discussion will concentrate on them. An abstract model of binding in which receptors and antigens are considered as points in a ‘shape space’ (shown in Figure 1) with affinity measured as a function of the distance between such points [31] is a convenient computational AIS model, although there is still controversy surrounding the biological basis [32]. The shape space is equivalent to the feature space used for pattern recognition. The distance between shape space points is typically measured using Euclidean, Hamming, or Manhattan measures. The stronger the affinity between receptor and antigen, the greater the response of the B-cells. Population growth is directly dependent, and is coupled with somatic hypermutation [33] which is inversely dependent, on the strength of this interaction. Somatic hypermutation is the process whereby poorly matching clonal populations are selectively mutated to provide wider coverage of the shape space. In addition, there is a cross-linking between populations of B-cell receptors.

An expression has been derived for the population dynamics of the B-cells that incorporates all of these characteristics [34]. The population of the B-cells is governed by the differential equation:

\[ \frac{\partial b_i}{\partial t} = s + b_i \left[ \frac{p}{\theta + b_i} f(h_i, h_i) + K_i + d \right] \]  

(1)

which is a modification of the original dynamics equation [34] that restores a term for the direct affinity of the cell for the incoming antigen [35]. In this expression, \( b_i \) is the number of cells of clone \( i \), \( s \) is the rate of influx of newly generated cells, \( p \) is the maximal growth rate, \( \theta \) is the growth clone-size threshold, \( f(h_i, h_i) \) is the cell activation function, \( K_i \) is the affinity that the cell \( i \) has for the incoming pattern, and \( d \) is the death rate. The cell activation function is dependent on the binding between the individual B-cell populations in the network. Cells with higher affinity to the incoming pattern will cling themselves (with or without mutation) faster than those that are not as good a match. The cell activation function is defined as:

\[ f(h_i, h_i) = \frac{(1 + 4w) h_i}{w(1 + h_i)^2 + h_i} \]  

(2)

where the binding field is:

\[ h_i = \sum_j K_{ij} b_j \]  

(3)

and the affinity is:

\[ K_{ij} = e^{-\frac{(x_i - x_j)^2}{2\sigma^2}} \]  

(4)

The exponent \( \eta \) in the cross-linking term (4) controls the width of the cross-linking field. The distance between the shapes \( x_i \) and \( x_j \) in (5) is measured using a representative metric, and \( \sigma \) is the standard deviation. There is some evidence that the affinity (5) is only approximated by a Gaussian in the actual biological systems [36] with the width and height dependent on equilibrium constants, but for the purposes of the AIS the symmetrical Gaussian will serve the purpose.

In addition to the clonal population dynamics given in expression (1) through (5), there is a longer duration process called Memory Cell generation for the most responsive B-cells found during the exposure to the incoming patterns. The Memory Cell population persists for a longer duration and reacts much more quickly to new patterns [37]. This type of
behavior is in some sense equivalent to the training phase of a traditional learning system.

The AISLE system algorithm is shown in Figure 2. For unsupervised learning, a variable number of B-cell populations are randomly initialized and each input pattern is presented to the system. Equation (1) is used to process each data element, with a Memory Cell population being maintained based on the surviving populations after each data presentation. As each generation of B-cells is modified based on (1), cell death very quickly leads to a decreased number of interactions. The final surviving B-cell is put into the Memory Cell population to be included in the next exposure of the system.

III. EXPERIMENTAL STUDIES

A series of three experiments were run in order to test the learning and pattern recognition capabilities of AISLE. The first study uses the Fisher Iris dataset, the second study uses information derived from the Hipparcos star catalog, and the third study uses 18 dimensional spectral data acquired by the MODIS satellite. In all experiments, \( b_i \), the number of cells of clone \( i \), was initially set to 100 for all \( i \), the influx rate \( s \) was set to 0.0001, the maximal growth rate \( p \) was set to 8, the growth clone-size threshold \( \theta \) was set to 75, the activation \( w \) was set to 0.1, weight \( \sigma \) in the affinity function (5) was set to 5.0, the exponent \( \eta \) in the cross-linking term (4) was set to 1.0, and the death rate \( d \) was set to 1. All updates of (1) were done using a Euler rule with a time step of 1.0, the death decision for any population \( i \) was set to 50.0, and a maximal number of generations for each B-cell clone set to 5000.

A. First Study

The first study used one of the most studied classification testbeds, the Fisher Iris dataset [38], a collection of 150 flower samples with the four variables of sepal length, sepal width, petal length, and petal width. The dataset has been broken into the three classes Setosa, Versicolor, and Virginica based on visual analysis of the samples. The three classes are not linearly separable, with mixing between Versicolor and Virginica. This study was used to demonstrate the unsupervised learning capabilities of AISLE using a single exposure of 75 randomly selected members of the dataset to the system. The Memory Cell population in the worst case would have 75 elements corresponding to an element for each member of the data subset. There were 50 trials performed using different starting subsets. Memory Cell population members were labeled with the class of the input sample that triggered the B-cell population in order to determine the classification accuracy of the system. The final Memory Cell population had 11 members after the single exposure of the data subset to AISLE.

The full dataset was presented to AISLE in recall mode using only the Memory Cell population to determine how well the system learned. The overall classification accuracy was 94.67\%, with a 100\% identification of Setosa, eight miss-classified Versicolor as Virginica, and one misclassified Virginica as Versicolor. This classification performance is equivalent to that of a backpropagation network with one hidden layer of 19 nodes or the SVM algorithm, and only slightly less than the 96\% accuracy obtained by a recursive partitioning algorithm [39]. AISLE achieves this level of classification performance with an order of magnitude better than that of the backpropagation network in terms of computational cycles. A plot of the AISLE results is shown in Figure 3. The plot shows petal length versus sepal length, and the clear delineation of Setosa is evident. Members of the Memory Cell population are also shown on the plot, and the source of the misidentification of Versicolor and Virginica lies in the Memory Cell (shown with arrow) within the boundary zone between the two classes. Also shown are members of the somatic hypermutated population that did not persist through the initial presentation of the dataset due to low final populations.

B. Second Study

The second study uses data from a subset of the Hipparcos star catalog. This catalog contains 118,218 stars that were observed by the European Space Agency's Hipparcos Satellite, operational from late 1989 to 1993. The Hipparcos catalog data is purely based on observations performed in space, except for the global orientation of its reference frame that was adjusted to the existing system by a variety of mainly ground-based techniques. By international agreement the catalog is the standard reference for optical astrometry. The subset of the catalog used for this study contained 2854 stars all with an apparent brightness magnitude cutoff of 5.5 found within 33 parsecs of Earth. This type of dataset would be used for star catalogs onboard a satellite to determine overall attitude within a global reference frame based on observed stars. An example of sample images taken at two instants in time is shown in Figure 4 for a simulated yaw maneuver of 1.41 degrees. The sizes of the stars indicate relative magnitudes.
Usually, only the star magnitudes are used as an index into the catalog to determine position. The relative positions of the stars are as important for matching, since the catalog has a large number of members and there will be a lot of mismatches if based on apparent magnitude alone. Characterizing the relative positions of the stars has to be done in such a way that is rotationally invariant, because the satellite can be in any orientation. For this reason, a composite vector was automatically built with sets of star triangles generated from the catalog using a Delaunay decomposition of the data. An example of the Delaunay decomposition is shown in Figure 5, where 13 triangles capture the entire content of the image. The rotationally invariant representation is built using six variables derived from the triangles: the three first order shape moments and the three star magnitudes on the corners of each triangle.

The first image in the sequence is fed into AISLE using a master Memory Cell population built from vectors derived from the Hipparcos catalog. The subset of matched vectors are then used as a new Memory Cell population for analysis of the next image. The number of surviving populations gives an indication of the degree of match, and any other triangle magnitude/moment vectors that are generated (epitopes) are the new stars that have entered the field of view of the satellite sensor due to the rotation. Figure 6 shows a plot of the final populations and the number of generations (iterations) for each of the triangle magnitude/moment vectors. The recognized vectors are characterized by relatively rapid convergence with convergence in all cases in about 300 msec making the algorithm capable of running in real time on a conventional processor or spacecraft. The matched triangle magnitude/moment vectors are shown in Figure 7, where the newly found vectors are labeled as epitopes.

C. Third Study

The final study demonstrates the clustering performance of AISLE. Usually, the object of a cluster analysis of remotely sensed data is to determine the dominant “classes” in the dataset for identification of land/water types. The dataset used in the study is from a hyper-spectral (18 bands) MODIS satellite feed (shown in Figure 8) taken over the Marqueses Islands (140ºW, 10ºS) located in French Polynesia in the Pacific Ocean. There are heavy currents that tend to obscure the boundaries between land and sea areas (especially at the relatively coarse 36 km cell size resolution of the data). Each of the 64,000 18-dimensional vectors corresponding to a location in the images was fed into AISLE much the same way as the first study – any surviving B-cell population is added to the Memory Cell population and used for the exposure of the next data item. The final Memory Cell population had 14 members that would correspond to 14 traditional classes for an unsupervised clustering algorithm. The study zone and output of the system is shown in Figure 9, where color coding is used in the inset to indicate the classes. The Marqueses Islands and Tuamotu Archipelago (circled in red) have been clearly extracted (yellow-tan labeled cells).

IV. SUMMARY AND CONCLUSIONS

This paper has presented AISLE, a clonal B-cell population based AIS that demonstrated generalized learning and pattern recognition capabilities. The Memory Cell populations are automatically generated by AISLE during the presentation of patterns. AISLE scales well for higher dimensional data, and all of the experimental studies were done using an ordinary MacBook Pro laptop (2.66 GHz Intel Core i7 CPU, 4 GB 1067 MHz RAM). The computational complexity of the system was kept under control using shape space culling of the initial large number of B-cell populations. The experimental studies demonstrated the unsupervised learning capabilities with a 94.67% accuracy on recall for the Fisher Iris dataset after only
a single pass, pattern recognition capabilities on a star tracking
dataset with matching of common features in a sequence, and
clustering capabilities on an 18 dimensional feature vector.

Current research directions include investigation of using
GPU (Graphics Processing Units) on the laptop for better
performance, developing a resolving strategy for degenerate
matches (using other matches within the image), parallel
processing, and developing an affinity based method for
somatic hyper-mutations. Also underway are studies
investigating the scalability of AISLE for larger hyperspectral
datasets (256 bands) in an effort to determine what are the
dominant bands that are useful for discrimination.

Figure 7. Matched triangle magnitude/moment vectors from the Memory Cell population derived from the first frame of the
sequence for the second frame of the sequence. A total of 7 new triangle magnitude/moment vectors (epitopes) were found and these
are then matched to the Master Hipparcos Star catalog to determine the movement of the satellite in between frames.

Figure 8. MODIS ocean data from Dec 1, 2000 for 18 bands of a study region centered on Marquesas Islands at
36km spatial resolution: (a) Gelbstoff absorption coefficient at 400 nm; (b) instantaneous absorbed radiation by
phytoplankton for fluorescence; (c) calcite concentration; (d) phytoplankton absorption coefficient at 675 nm; (e)
chlorophyll-a concentration (SeaWiFS analog - OC3M); (f) chlorophyll-a concentration (somatic); (g)
chlorophyll fluorescence efficiency; (h) chlorophyll fluorescence line height; (i) chlorophyll-a concentration
(HPLC, empirical); (j) instantaneous photosynthetically available radiation; (k-q) normalized water-leaving
radiance at 412, 433, 488, 531, 551, 667, 678 nm; (r) sea surface temperature (11/12 micrometer).

ACKNOWLEDGMENT
The author would like to thank JPL for providing the
funding for this research under the Research and Technology
Development Program.

REFERENCES
immune system,” Proc. Genetic and Evolutionary Computation
Immune System for Job Shop Scheduling,” Proc. Second International
Conference on Artificial Immune Systems (ICARIS), September 1-3,
2003, Napier University, Edinburgh, UK.


