AIDEN: A Density Conscious Artificial Immune System for Automatic Discovery of Arbitrary Shape Clusters in Spatial Patterns

Vishwambhar Pathak
Birla Institute of Technology, Mesra, Ranchi, India
pathakvishi@gmail.com

Praveen Dhyani
Banasthali University, Jaipur Campus, India
dhyani_p@yahoo.com

Prabhat Mahanti
University of New Brunswick, New Brunswick, Canada
pmahnti@gmail.com

Abstract

Recent efforts in modeling of dynamics of the natural immune cells leading to artificial immune systems (AIS) have ignited contemporary research interest in finding out its analogies to real world problems. The AIS models have been vastly exploited to develop dependable robust solutions to clustering. Most of the traditional clustering methods bear limitations in their capability to detect clusters of arbitrary shapes in a fully unsupervised manner. In this paper the recognition and communication dynamics of T Cell Receptors, the recognizing elements in innate immune system, has been modeled with a kernel density estimation method. The model has been shown to successfully discover non spherical clusters in spatial patterns. Modeling the cohesion of the antibodies and pathogens with '*local influence*' measure inducts comprehensive extension of the antibody representation ball (ARB), which in turn corresponds to controlled expansion of clusters and prevents *overfitting*.

Keywords: Artificial Immune System; Density Based Clustering; spatial patterns

1. Introduction

The mechanism of artificial immune systems (AIS) has been comprehensively applied in unsupervised learning problems, for example the tasks of anomaly detection and of clustering. Imitation of the recognition process of natural immune cells eventually imposes grouping of 'self' and 'non-self' patterns. The recognition process of the natural immune cells has been modeled using a variety of formulations comprising of mathematical and statistical principles leading to several models of AIS [1][2][3]. Given the heavily complex internal communication structure of immune system, models exploiting holistic analogies of the same have been rare in literature. Selective interpretations of the internal functionality and interaction among the immune system components have generated dependable solutions to a wide range of applications with typical characteristic constraints.

The task of clustering data in $\Re^N$ carries several challenges including arbitrary shapes of inherent clusters, automatic detection of clusters, and robustness against noise besides others. Further, the algorithm must also address the issues of scalability and robustness. Preliminary schemes like the hierarchical and the partitioning based clustering methods have been largely utilized in serving several clustering applications. However these clustering processes essentially depend on user intervention either before start of the learning (prior specification of cluster number K in K-means) or after the completion of clustering process (for interpretation of relevant clustering level in hierarchical clustering). One standard approach addressing such issues includes the generative models, wherein algorithms like EM are used to learn mixture densities. These models have proven capability of automatic detection of clusters inherent in the input data. The initial generative models suffer from limitations like the need for simplifying assumptions of Gaussian

density of clusters, occurrence of local minima of the log likelihood, and their sensitiveness to initialization of model parameters, for example the mean and the std. deviation in case of the Gaussian mixtures [4]. Several stochastic models have been reported to deal with the limitations of the generative models. Few remarkable models applied Markov Chain Monte Carlo (MCMC) method to improve robustness amidst missing values [5]. Solutions employing non Gaussian mixtures from the exponential families have also been shown to be capable of detecting arbitrary shapes of cluster [5] [6]. Majority of such promising schemes with strong theoretical foundations have been found to be expensive in implementation. Recent computing applications have incorporated extensive involvement of a variety of data types e.g. spatial data with inherent complex characteristics namely existence of non linear and overlapping segments/classes. Development of improved data clustering models to aptly handle such underlying issues is a research topic of great interest. Recently spectral clustering method has emerged as a favorable alternative applied in several applications in machine learning [2]. In this method, the top eigenvectors of a matrix derived from the distance matrix of data points are used to derive clusters. Spectral analysis has also been widely used in analysis of patterns in spatial data [3] [4]. However, implementation of this method poses difficulty in decision about which eigenvector to use and to derive clusters from them [2].

This paper presents a clustering method based on artificial immune system (AIS). AIS essentially imitates the recognition process of the natural immune cells to detect pathogen, learn the patterns, and develop antibodies to attack any future occurrence of such pathogens. The mechanism of AIS has found its application in the task of clustering, since the imitation of the recognition process of the immune cells eventually imposes grouping of 'self' and 'non-self' patterns. The present work reports modeling and results of implementation of Artificial Immune system with DENsity sensitiveness (AIDEN). AIDEN in principle exploits the inherent density-closeness of the items in the input dataset to automatically discover the inherent cluster. The experiments show the utility of the model in clustering of spatial patterns. The robustness, scalability, and convergence of the algorithm have been discussed in relevant section.

2. Functional Elements of Artificial Immune System

The immune system involves complex interaction between four major components in fighting off pathogens: i) Antibodies ("immunoglobulins") and the (B) cells that make them, ii) Complement iii) T cells, and iv) Non-specific effector cells [7]. B Cells essentially make the Antibodies which specifically bind to pathogens and pass them to the Complement and phagocytic cells. A Complement is a cascade of small proteins that bind to pathogens and poke holes in their outer surface causing death. Complement proteins can bind to some pathogens directly but the activity of Complement is much amplified by the presence of antibody bound to the pathogen. The T Cells help the B cells become antibody producing cells (APC) and help other cells perform effector functions. Various ways to model an approximate function of the immune systems have been proposed in literature [9]. Non-specific effector cells like macrophages and neutrophils kill pathogens, much effectively if antibody is bound to the pathogen. Other effector cells, like NK cells kill self cells that have been infected with pathogens (such as viruses). The Mast cells secrete factors that create inflammation in the area of the pathogen to allow rapid access of other immune components to the site. The T lymphocyte (T cell) in a mammalian immune system is capable of performing fine grain discrimination of peptide bound Major Histocompatibility Complex (pMHC) molecules on APCs through its T cell Receptor (TCR) and intracellular signaling network. So it discriminates between abundant self-pMHC all pMHC on an APC, and non-self-pMHC [8][10][11].

Biologically, TCRs induce activation signal in a T-cell through a process involving the steps of kinetic proofreading, negative feedback, negative feedback destruction, and tuning. Kinetic proofreading involves energy consumption steps that must be overcome before TCR may generate an activation signal. The negative feedback generated through progression of the kinetic proofreading eventually reverses the process regardless of the TCR-pMHC binding strength.

However on successful completion of kinetic proofreading a TCR generates activation signal which is then amplified and provides protection to all other TCRs from the negative feedback. The coreceptor density can be understood as a tuning parameter, small increases in which increases the probability of activation.

3. Dynamic Density Conscious Clustering with T-Cell Receptor Signaling

A natural correspondence between the biological T Cell receptor signaling and kernel density estimation was established in [10]. We observe selective analogies of the biological features of the signaling and immunization dynamics of TCell-BCell-APC configuration of the immune system in the context of density-conscious clustering. The kinetic proofreading by a single TCell through its receptor and the co-receptor (CD8) tuning can be modeled using the concept of *k*-nearest-neighbors based '*density-reachability*' applied in density based clustering technique DBSCAN [9]. The present work shows that the static stimulation of the TCRs and the dynamic behavior of the immune system resulting from intra-cellular interactions can be modeled in the perspective of density conscious cohesions. The cohesion measurement has been modeled using the concept of *k*-nearest neighborhood and local-density-factor.

We consider a stream of points x1, x2, x3 . . . ∈ x at receptors positions t1, t2, t3 . . . ∈ $\Re^N$.

The stimulation of TCR on presentation of a pathogen, as a result of kinetic proofreading can be represented with a receptor stimulation r_p (x) and negative feedback r_n(x) at all points in $\Re^N$, as given in equation (1) and equation (2). The static receptor stimulation may be generated by a function *f* (.), which is a model parameter, exemplified in the section describing the algorithm.

$$r_p(x) = f(r_p, x) \qquad \text{-- (1)}$$

$$r_n(x) = \begin{cases} r_p(x) - \beta & if \;\; r_p(x) > \beta \\ 0 & otherwise \end{cases} \qquad \text{--(2)}$$

A TCR at position p is stimulated if r_p (x) - r_n(x) > *l*. Figure 1 depicts this process. When a T Cell receives stimulations on more than k receptors, it generates activation signal to a B Cell, as represented in Figure2.

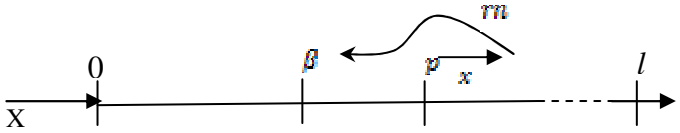

Figure1. Static stimulation of a single TCR- The kinetic proofreading by the receptor on input x Є X forwards the receptor position p toward l. The receptor will generate negative feedback if p > β. The receptor will generate success signal when p== l.

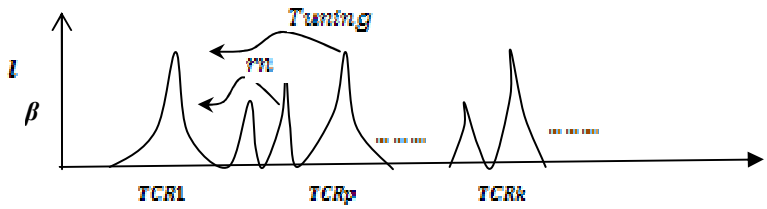

Figure2. Generation of negative feedbacks gets tuned once a TCR completes stimulation beyond the threshold l. A TCell generates activation signal to BCell once it gets stimulation of its k-TCRs.

On receiving activation signal from TCell, a BCell produces *k*-antibodies in the APC associated with it, corresponding to the pathogens stimulating respective T Cell. Intercellular interaction among BCells give rise to antigen recognition balls (ARBs). APCs connect together if

their *inter-cell-cohesion factor (cf)* is beyond a threshold ψ. To formulate *cf* for our model, the concept of local scaling [12][13], which was originally introduced to determine the scale factor for clusters with different densities by scaling the distances around each point in the dataset with a factor proportional to its distance to its *k*th nearest neighbor. Typically ψ is taken as $2^{-\alpha}$, α being another model parameter determined by the data characteristics. Figure3 depicts this process. The model with the above specification then effectively detects self or non-self pathogens. In terms of its application to the task of clustering, this interpretation means making the affinities high within clusters and low across clusters. A pathogen corresponding to an outlier would not stimulate a TCR sufficiently and may not form part of any ARB.

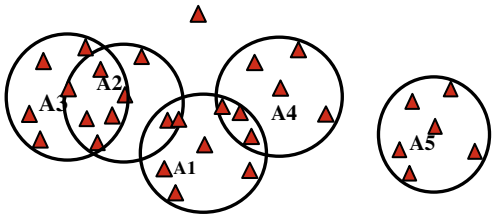

Figure3. APCs A1-A4 connected within range of cohesion-factor-threshold form members of one ARB, while A5 lies there single and is likely to form another ARB with progress of the recognition process. The single point represents a pathogen not sufficiently stimulating any TCR and corresponds to an outlier.

4. Algorithm and Explanation

AIDEN()
Input: D: Ag-Ab Affinity matrix, k: input to *kNN*-dist function, n: number of dimensions, cohesion_factor_threshold (ψ)
Output: Allocation of points to clusters.
begin
    for pgen Є *Pathogens* **do**
        pgen.*ARB* = *UNDETERMINED*;
        [pgen. *BestAffinedist*, pgen.affines] = *kNNDistVal*(D, pgen, k);
    end
    Pathogens.sort(); /* Sort on *BestAffinedist* */
    arbID = 1;
    for pgen Є *Pathogens* **do**
        if pgen.*ARB* == *UNDETERMINED* **and** *localDense*(pgen) **then**
            ARBGrow (pgen, *arbID*, n, ψ);
            arbID = *arbID* + 1;
        end
    end
end

***ARBGrow*()**
Input: pgen, *arbID*, n, ψ
Output: ARB corresponding to current Id Extended with inclusion of APCs with sufficient local-cohesion.
begin
    pgen.ARB = *arbID*;
    APC= pgen.affines
    for Ag Є APC
        if Ag.ARB == *UNDETERMINED*
            Ag.ARB = *arbID*;

```
            else
                APC.remove(Ag)
            end if
        end for
        while Exists (APC)
            Ag = APC.next()
            if  Ag.BestAffinedist ≤ ψ x pgen. BestAffinedist
                newAgs = Ag.affines
                for nAg ∈ newAgs
                    if nAg.ARB == UNDETERMINED
                    APC.append(nAg)
                        nAg.ARB = arbID
                    end if

                end for
            end if
            APC.remove(nAg)
        end while
end
```

5. Explanation of the Implementation and Experimental Results

The algorithm presented above implements the model described in section 3. The input matrix D provides the values for $f(x_i, x_j)$ in Equation 1. All checks related to stimulations at various stages would use this matrix only for the purpose. To keep the translation simple, it has been assumed that the value of β is set much higher, as a system parameter itself, so that its effect is nullified by equation 2. For the same reason, the effects of negative feedback destruction and that of tuning are rendered theoretically running in the background, hence having no direct mention in the above implementation. The function *localDense*(pgen) determines from among the unclassified data-point in *k*-nearest-neighborhood of *pgen* having highest density estimated on the basis of proximity of their own k-NN elements . Further, ARBGrow() implements the ARB formation and expansion process discussed in the section 3.

The program was implemented in Matlab and tested with several patterns. The first, dataset1 consisted of 2 patterns each comprised of 100 points falling on two concentric circles of radii 10 and 20 respectively. The second, dataset2 consisted of 3 patterns each of 100 points falling on three concentric circles of radii 10, 15, and 20 respectively. The model was further tested for its capability to find clusters in patterns of open spatial form using dataset3 and dataset4 consisting of 200 and 300 points falling on 2 and 3 concentric semi circles respectively. As shown in the Figure2.a and Figure2.b, the algorithm is capable of determining spatial association of a data point with other data points belonging to its appropriate circle only. The results successfully demonstrated the capability of our model to automatically detect clean clusters of arbitrary shapes in the input data represented in closed spatial form. The model was found even capable of determining clusters of open spatial forms also, as shown in Figure2.c and Figure2.e. However, the output of the algorithm was found affected by the values of the algorithm parameters *k* and α. In the present experiment, k =8 and α=10 was sufficient for performing correct cluster associations. On the other hand, correct clustering for the dataset2, could be obtained with 10NN estimation i.e. k =10, with α=15. Moreover setting k =15, with α=15 was required for dataset4, as clustering error was observed with k =10, with α=15, as in Figure2.d. Figure2.f and Figure2.g show the correct clustering even in presence of combination of open and closed form of input patterns. In each figure, the first sub-plot shows the original data and the second sub-plot shows the clusters identified by our program.

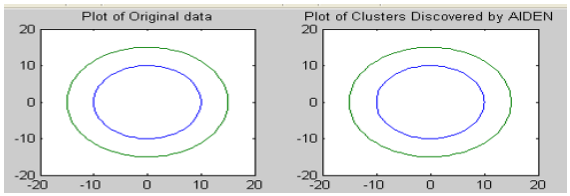

Figure2.a. dataset1. k=8(10), alpha=10(15)

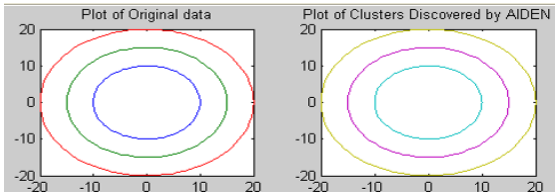

Figure2.b. dataset2. k=10, alpha=15

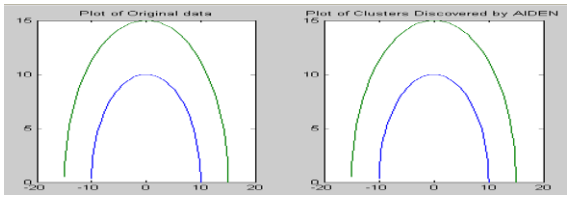

Figure2.c. dataset3. k=8(10), alpha=10(15)

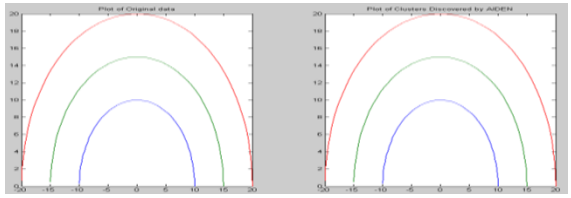

Figure2.d. dataset4. k=10, alpha=15

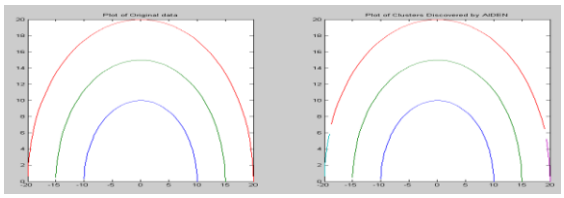

Figure2.e. dataset4. k=14, alpha=15

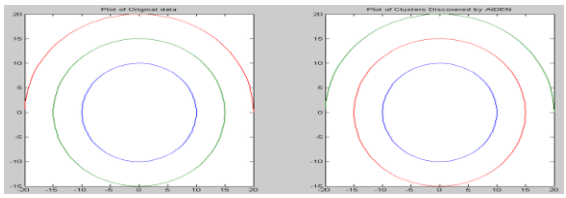

Figure2.f. dataset5. k=10, alpha=24

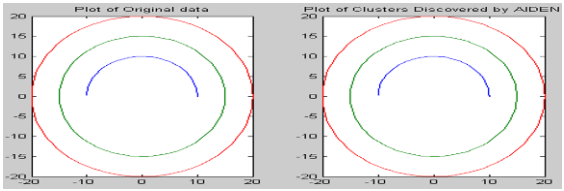

Figure2.g. dataset6. k=8(10), alpha=14(15)

6. Conclusion

The traditional models based on cluster mean as representative of clusters have been sensitive to outliers and parameter initialization. Our model associates a data item (pathogen) with an existing cluster (BCell), only when they are density-attracted to any of the 'dense' elements in the cluster. This property helped in expanding the range of association through recognition by each representative data item (antibody) in non spherical dimension. The strategy of cluster expansion based on *local influence* measure serves the twofold purpose; one, it prevents overfitting; two, it simultaneously assures immunity against outliers. As the algorithm follows a deterministic approach to perform computations for each data point in the input, the convergence of the algorithm is theoretically plausible, and has the same been confirmed with the experimental results. The correct results obtained for different number of patterns convince about scalability of the model. Theoretically the algorithm may be observed to work in a polynomial time. The correct clustering results obtained in presence of both open and closed patterns give a clue towards robustness of the model. A further work is aimed to investigate performance of the model in presence of overlapping patterns. Moreover study toward applicability of the model in relation to clustering and spectral analysis of real world spatial data is planned.

References

[1] Jain Anil K. 2008. Data Clustering: 50 Years Beyond KMeans1, Machine Learning and Knowledge Discovery in Databases, Volume: 31, Issue: 8, Springer.

[2] Nasraoui Gonzalez, Cardona, and Dasgupta. 2002. Scalable Artificial Immune System Based Data Mining, NSF-NGDM, Nov. 1-3, , Baltimore, MD

[3] Greensmith Julie, Whitbrook, Amanda, and Aickelin Uwe. 2010. Artificial Immune Systems, In Handbook of Metaheuristics, Springer.

[4] Pathak V., Dhyani, Praveen, and Mahanti Prabhat. 2011. Data Clustering with Artificial Innate Immune System Adding Probabilistic Behaviour, International Journal of Data Mining and Emerging Technologies Vol. 1 No.2, November, 2011, 77-84 DOI: 10.5958/j.2249-3212.1.2.5

[5] Marlin, Benjamin, M. 2008. Missing Data Problems in Machine Learning, PhD Thesis, Graduate Department of Computer Science University of Toronto.

[6] Orbanz Peter. 2008. Infinite-Dimensional Exponential Families in Cluster Analysis of Structured Data, PhD thesis, Diss. ETH No. 17822, ETH Zurich

[7] Lee, Timothy, and Issekutz Andrew. 2009. Immunology for 1st Year Medical Students,

[8] Timmis, Jon, Hart, Emma, Hone, Andy, Neal Mark, Robins, Adrian, Stepney, Susan, and Tyrrell Andy. 2010, Immuno-engineering.

[9] Han J., Kamber M. 2006. Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufman

[10] Owens, N. D.L., Greensted, Andy, Timmis, Jon, and Tyrrell Andy. 2009. T Cell Receptor Signalling Inspired Kernel Density Estimation and Anomaly Detection, ICARIS 2009, LNCS 5666, pp. 122–135. Springer-Verlag Berlin Heidelberg

[11] Germain, R.N., Stefanov, I. 1999. The dynamics of T cell receptor signaling: complex orchestration and the key roles of tempo and cooperation. A. Rev. Imm., 17

[12] Zelnik-Manor, Lihi, and Perona Pietro. 2004. Self-tuning spectral clustering. In Eighteenth Annual Conference on Neural Information Processing Systems.

[13] Bicici, E. and Yuret, D. 2007. Locally Scaled Density Based Clustering, Lect. Notes Comput. Sc., 4431, 739–748.