Discovery of Emerging Patterns with Immune Network Theory

Bo Chen\textsuperscript{a} and Chuanzhi Zang\textsuperscript{b}

\textsuperscript{a}Dept. of Mechanical Engr.–Engr. Mechanics/Dept. of Electrical & Computer Engineering, Michigan Technological University, 1400 Townsend Drive, Houghton, MI, USA 49931-1295

\textsuperscript{b}Dept. of Mechanical Engr.–Engr. Mechanics, Michigan Technological University, 1400 Townsend Drive, Houghton, MI, USA 49931-1295 and Shenyang Institute of Automation, Chinese Academy of Science, Nanta Street 114, Shenyang, Liaoning, P.R. China, 110016

ABSTRACT

This paper presents an immune network-based emergent pattern recognition method. The artificial immune network provides more flexible learning tools than neural networks and clustering technologies. With a neural network, a network structure has to be defined first. The immune network allows their components to change and learn patterns by changing the strength of connections between individual components. The presented computational model achieves emergent pattern recognition by dynamically constructing a network of feature vectors to represent the internal image of input data patterns. The immune network-based emergent pattern recognition approach has tested using a benchmark civil structure. The test result shows the feasibility of using the presented method for the emergent structural damage pattern recognition.

Keywords: Artificial immune network, emergent pattern recognition, distributed sensing and monitoring systems

1. INTRODUCTION

Due to ever-increasing complexity and unpredictable working conditions, there is a crucial need for the monitoring systems to identify fault/damage at an early stage to avoid potential disaster in power grids, traffic systems, and critical civil infrastructures. The biological immune system is able to handle this challenging problem much more efficient than engineered systems [1]. The immune system has an ability to display emergent behaviors, which is often resilient and robust. Based on this observation, immune-inspired computational approaches have been investigated in the past decade for the anomaly detection and pattern recognition. The Artificial Immune Systems (AIS) is suitable to handle the great complexity of the reality [2]. The reason behind this is that the natural immune system incorporates a variety of artificial intelligence techniques, such as pattern recognition through a network of collaborating agents, adaptive learning through memory, and an advanced selection mechanism of the best B-cells [3].

This paper presents an immune network-based emergent pattern recognition algorithm (INEPR). The presented pattern recognition approach achieves emergent pattern recognition by dynamically constructing a network of antibody memory cells as an internal image of the input data patterns. The memory cells are evolved through a clonal immune response initiated by each input data pattern (antigenic pattern). The memory cells with high affinities to the input data pattern will be recruited into the network, while the memory cells with low affinities to the input data pattern will be eliminated from the network. To classify antibody memory cells into a number of clusters (patterns) which are corresponding to the input data patterns, the agglomerative clustering algorithm is used to generate a hierarchy of memory cell clusters. The number of clusters of the memory cells is automatically determined by the evaluation graphs and the $L$ method proposed in [4].

The rest of the paper is organized as follows. Section 2 introduces the immune network-based emergent pattern recognition algorithm. Section 3 tests the ability of the INEPR algorithm for the recognition of emergent damage patterns using a benchmark civil structure. Section 4 discusses the impact of model parameters on the number of memory cells and the decision of the best number of clusters. Finally, conclusions are made in Section 5.

1 bochen@mtu.edu; phone 1 906 487-3537; fax 1 906 487-2822; www.imes.mtu.edu
2. IMMUNE NETWORK-BASED EMERGENT PATTERN RECOGNITION ALGORITHM

The goal of the presented immune network-based algorithm is to recognize new data patterns in real-time monitoring networks. Unlike neural network, in which the network structure should be defined first [5], the immune networks can learn patterns by changing the strength of connections between individual components. The immune network theory [6] suggests that the immune system is composed of a regulated network of cells and molecules that recognize one another even in the absence of antigens. As shown in Figure 1 (a), immune receptors (antibodies) can be recognized by other receptors through the binding between their idiotopes and paratopes. The behavior of the immune network theory is illustrated in Figure 1 (b) [7]. When the immune system is primed with an antigen, its epitope is recognized by a set of different paratopes, called P1. The set I2 of idiotopes is called the internal image of the epitope because it is recognized by the same set P1 that recognized the antigen.

In pattern recognition, the patterns to be classified are usually the groups of measurements, defining points in an appropriate multidimensional space [8]. The measurements used for the classification are described by features. If \( p \) features are used for \( i = 1, 2, \ldots, p \), these \( p \) features can form a feature vector \( F=(f_1, f_2, \ldots, f_p)^T \), where \( T \) denotes transposition. The affinity of two measurements is defined as a function of the distance between the corresponding feature vectors of the measurements. Let \( \beta=(\beta_1, \beta_2, \ldots, \beta_p)^T \) and \( \gamma=(\gamma_1, \gamma_2, \ldots, \gamma_p)^T \) denote two feature vectors. The affinity between the two feature vectors can be defined as equation (1).

\[
\text{aff}(\beta, \gamma) = 1 / \sqrt{\sum_{i=1}^{p}(\beta_i - \gamma_i)^2}
\]  

The INEPR algorithm consists of four steps as shown in Figure 2. The first step is data preprocessing in which the sensor data from multiple sensors are reduced to one time series using Principal Component Analysis (PCA) method. The autoregressive (AR) model is employed to fit the compressed time series and the coefficients of the AR model are used to form the feature vector of the time series. The details of the data preprocessing are described in [9]. The second step is internal image generation in which an immune network computational model is employed to dynamically construct a network of antibody memory cells to represent the input data patterns. The third step is hierarchical clustering in which a hierarchical clustering algorithm is used to create a hierarchy of nested clusters for the generated memory cells. The number of clusters in the hierarchy is determined in the fourth step using evaluation graphs and the \( L \) method proposed in [4].

![Figure 1. Immune network [7].](image1)

2.1 Antibody Memory Cell Generation Using an Immune Network-Based Computational Model

An immune network computational model is able to generate a set of antibody memory cells for the input data patterns. The antibody memory cells can recognize and represent the data structural organization of the input data set.
memory cells are generated by a series of clonal immune responses that are initiated by antigens (input data) to the antibody set. The antibodies that can successful recognize antigen will be cloned and mutated. The newly generated antibodies with high affinities to the antigen will be recruited into the antibody memory cell set. The antibodies who fail to recognize antigen will be eliminated from the antibody memory cell set to improve the affinity level of the representative antibody memory cells.

A number of computational models based on the immune network theory have been developed and applied for data mining, pattern recognition, and multimodal function optimization [10]. aiNet [11] is one of the artificial immune network models with the goals of clustering and filtering input data set. The output of the aiNet model is a reduced data set of the input data and data structure information, including the spatial distribution of antibody memory cells.

To describe aiNet model in detail, following notation is adopted. Let $ag$ denote an antigen and $ag.f$ denote the feature vector of the antigen; $ab$ denote an antibody and $ab.f$ denote the feature vector of the antibody; $A$ denote a set of antibodies ($A \in R^{N \times P}$), $M$ denote a set of antibody memory cells ($M \in R^{K \times P}$); $G$ denote a set of antigens ($G \in R^{M \times P}$); and $\sigma_d$ and $\sigma_s$ denote natural death threshold and suppression threshold, respectively.

In the aiNet model, the antibody set $A$ and the antibody memory cell set $M$ are firstly initialized. The initial antibody set $A$ is randomly generated, and the initial antibody memory cell set $M$ is an empty set. For each antigen $ag_j$, it stimulates the evolution of the antibody set $A$. The affinities among the stimulating antigen and the antibodies in the antibody set $A$ are calculated. The $n$ number of antibodies in $A$ with highest affinities to the antigen $ag_j$ is selected. The selected antibodies will be cloned and mutated, and saved to a temporary data set $B$. The $\zeta\%$ of the cloned and mutated antibodies in $B$ with highest affinities to the antigen $ag_j$ is saved to the clonal matrix $M_j$. The antibody clones in $M_j$ will suffer a natural death elimination and clonal suppression. In the natural death elimination process, antibodies in $M_j$, whose distances to the antigen $ag_j$ are greater than the natural death threshold $\sigma_d$, are eliminated. The remaining antibody clones in $M_j$ will subject to a clonal suppression process. If the distance between two antibody clones is less than the suppression threshold $\sigma_s$, one of the antibodies will be eliminated from $M_j$. The antibody memory cells after the natural death elimination and clonal suppression process form a resultant clonal memory $M'_j$ for the antigen $ag$. The antibody memory cells in $M'_j$ are then added to the antibody memory cell set $M$. After all the antigens stimulate the antibody set, the newly generated memory cell set $M$ will suffer a network suppression process which is similar to the clonal suppression process. The purpose of the network suppression is to further reduce the number of the antibody memory cells in $M$.

The aiNet model has been used to generate antibody memory cells for the three data patterns shown in Figure 3. The generated antibody memory cells of the input data patterns are shown in Figure 4 in which three distinct clusters of memory cells map those of data patterns in the input data. The number of memory cells in each cluster is less than the number of data points in the original data set. When a new data pattern is input into the INEPR model, a corresponding memory cell cluster will be generated in the antibody memory cell set.
2.2 The Hierarchy of Nested Clusters of the Antibody Memory Cells

To classify the antibody memory cell set $M$ shown in Figure 4, we need to find out how many clusters the $M$ has and the number of memory cells belonging to each data pattern. Hierarchical clustering techniques [8, 12] are robust network interpretation strategies. In this work, hierarchical clustering algorithms are employed to generate a hierarchy of nested clusters for the antibody memory cell set. The memory cell set $M$ as defined is a set of $p$-dimensional vectors $M = \{mc_i, i = 1, ..., K\}$, where $mc_i, i = 1, ..., K$ is the $i$-th memory cell in $M$ and $K$ is the cardinality of the $M$. Let $\Delta$, $\Delta = \{C_i, i = 1, ..., m\}$, denote an $m$-clustering of the set $M$. The subset $C_1, ..., C_m$ in $\Delta$ meets following rules: (1) $C_i \neq \emptyset$, $i = 1, ..., m$; (2) $\bigcup_{i=1}^m C_i = M$; (3) $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, ..., m$. For two clusterings $\Delta_1$ and $\Delta_2$, the clustering $\Delta_i$ is said to be nested in the clustering $\Delta_j$, $\Delta_i$ nested in $\Delta_j$, when each cluster in $\Delta_i$ is a subset of a set in $\Delta_j$ and the cardinality of $\Delta_i$ is larger than the cardinality of $\Delta_j$.

The hierarchical clustering algorithms produce a series of clusterings. If the cardinality of $M$ is $K$, the hierarchical clustering algorithms have $K$ steps. At each step $t$, a new clustering is generated based on the clustering created at the step $t-1$. There are two main types of hierarchical clustering algorithms: the agglomerative and the divisive hierarchical clustering algorithms. For the agglomerative clustering algorithms, there are $K$ clusters in the initial clustering $\Delta_0$ and each cluster contains only one node, one memory cell in $M$. At each step, two clusters are merged into one new cluster. Finally, in clustering $\Delta_{K-1}$, there is only one cluster, $M$. The divisive algorithms follow the inverse path as the agglomerative algorithms. It starts with an initial clustering $\Delta_0$, which contains the set $M$. At each following step, one selected cluster is divided into two non-empty sub-clusters. At the final step $K-1$, there are $K$ clusters. In this work, we use agglomerative algorithms to generate a hierarchy of nested clusterings for the antibody memory cell set. The agglomerative algorithm scheme is stated in Algorithm 1. The clusters $C_i$ and $C_j$ are merged into a single cluster $C_q$ if the distance between them is the smallest one for all the possible pairs of clusters at the level $t$.

![Algorithm 1: agglomerative algorithm scheme](image)

When $C_i$ and $C_j$ are merged into a new cluster $C_q$, the distance between $C_q$ and one of the old clusters $C_s$, $d(C_q, C_s)$, can be calculated by

$$d(C_q, C_s) = \min \{d(C_q, C_i), d(C_q, C_j)\} \quad (2)$$

where $d(C_q, C_s)$ is the cluster distance defined in the single link agglomerative algorithm.

The distance between two clusters (cluster-to-cluster) or two memory cells (point-to-point) can also be considered as the dissimilarity between them. The dissimilarity measures between two memory cells will also affect the performance of clustering and classification. The initial clusters in the hierarchical clustering algorithm contain only one memory cell, and the distance between any two clusters is the distance between the two corresponding memory cells. For the antibody memory cell set $M \in \mathbb{R}^{K \times P}$, there are $K$ number of $P$-dimensional memory cells $mc_1, mc_2, ..., mc_K$, which forms a $K \times P$ matrix $M$. The most common dissimilarity measures between two memory cells $mc_i$ and $mc_j$ are defined as

$$d(mc_i, mc_j) = \min \{d(mc_i, mc_k), d(mc_j, mc_k)\} \quad (2)$$

where $d(mc_i, mc_j)$ is the cluster distance defined in the single link agglomerative algorithm.
follows. These point-to-point dissimilarity measures are used to test the performance of the presented algorithm in section 4.

- **Euclidean distance:**
  
  \[ d^2_{rs} = (mc_r - mc_s)(mc_r - mc_s) \]  

- **Standardized Euclidean (Seuclidean) distance:**
  
  \[ d^2_{rs} = (mc_r - mc_s)D^{-1}(mc_r - mc_s) \]  

where \( D \) is a diagonal matrix with diagonal elements given by \( \sigma_j^2 \), which denotes the variance of the \( j \)-th feature over the \( K \) memory cells.

- **Correlation distance:**
  
  \[ d_{rs} = 1 - \frac{ \left( mc_r - \bar{mc}_r \right) \left( mc_s - \bar{mc}_s \right) }{ \left( \bar{mc}_r - \bar{mc}_s \right) \left( \bar{mc}_r - \bar{mc}_s \right) } \]  

where \( \bar{mc}_r = \frac{1}{p} \sum m_{cr} \), \( \bar{mc}_s = \frac{1}{p} \sum m_{cs} \).

The hierarchy of nested clusterings generated by hierarchical clustering algorithms can be visualized by dendrogram plots. Figure 5 shows the dendrogram of the antibody memory cells shown in Figure 4.

### 2.3 Determining the Number of Clusters of the Antibody Memory Cell Set

The output of the hierarchical clustering algorithm is a hierarchy of nested clusters as shown in Figure 5. The dendrogram can be broken at different levels to yield different number of data patterns. Various methods for the choice of the best number of clusters are discussed in [4, 13]. In this paper, we employ evaluation graphs and the \( L \) method [4] to determine the number of clusters.

The evaluation graph is a two-dimensional plot where the \( x \)-axis values are the possible number of clusters, and the \( y \)-axis values are dissimilarity measures of a clustering consisting of \( x \) number of clusters. The evaluation metric used to compute the \( y \)-axis values is the evaluation of the two emerging clusters in each clustering step as shown in equation (2). The evaluation graph of the antibody memory cells in Figure 4 is shown in Figure 6. The reasonable number of clusters should be in the region of curved area, or the knee of the evaluation graph.

To find the knee of an evaluation graph, the \( L \) method introduced in [4] is employed. As shown in Figure 7, the right and left regions of the evaluation graph presents linear characteristic, as a result, two straight lines can be used to fit to data points in these two regions. The intersection of these two lines is located in the region of the knee, and is an approximation of the knee. The \( x \)-axis value closest to the knee is then be used as the best number of clusters.
3. CASE STUDY

The presented emergent pattern recognition approach has been tested using a benchmark structure proposed by the International Association for Structural Control - American Society of Civil Engineers SHM Task Group [14]. The benchmark structure is a 2x2 bay, four story steel structure. The structural data used in our study are the experimental data. In the experimental setup, a variety of damage cases were simulated by removing braces in the test structure. The details of the damage patterns used in the validation are listed in Table 1.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Fully braced configuration (normal pattern)</td>
</tr>
<tr>
<td>4</td>
<td>Removed braces on 1st and 4th floors in one bay at south east corner</td>
</tr>
<tr>
<td>5</td>
<td>Removed braces on 1st floor in one bay at south east corner</td>
</tr>
<tr>
<td>7</td>
<td>All braces removed on all faces</td>
</tr>
</tbody>
</table>

Table 1. The configurations of the simulated damage patterns

In the experimental benchmark study, a total of 15 accelerometers were used to measure the acceleration data of the structure. The acceleration data for each damage pattern or the normal pattern were recorded in a data file. Three damage patterns (configuration 4, 5, and 7) and the normal pattern (configuration 1) were selected to validate our algorithm. To generate feature vectors for each data pattern, first 3000 data points are dropped, and the following 12,000 data points in each data file are used to form 54 of 4000-point time series by advancing 150 points each time. Time series data from 15 accelerometers were compressed to one time series using the principal component analysis method. The auto-regression models are used to fit to 54 time series of acceleration data for each pattern. The AR order is selected to be 20. The coefficients of the AR models are used to form the feature vectors. Since each pattern has 54 feature vectors, a total number of 54*4=216 feature vectors were generated for three damage patterns and the normal pattern. These feature vectors are high-dimensional vectors. To visualize feature vectors, the 20-dimensional feature vectors are reduced to three dimensions using PCA dimensionality reduction method. Figure 8 shows the feature vectors of the first three data patterns listed in Table 1, and Figure 10 shows all the feature vectors of four data patterns.

Two tests were conducted to verify the INEPR algorithm. In the first test, acceleration data of the normal pattern and the first two damage patterns, shown in Figure 8, were used to find appropriate values of the INEPR parameters. The third damage pattern was introduced in the second test to simulate the emergent damage pattern. In the first test, a set of appropriate parameter values were found, \( \sigma_j = 0.3 \), \( \sigma_d = 5.5 \), \( n = 6 \), and \( \zeta = 0.2 \), which could generate reasonable number of antibody memory cells. Once the memory cells were generated, the single link hierarchical clustering algorithm was applied to produce a hierarchical clustering. The point-to-point dissimilarity measure was the Euclidean distance. To classify antibody memory cells, the evaluation graph for the clustering of three data patterns were calculated, and the \( L \) method was used to determine the best number of clusters as shown in Figure 9. In this case, the best number of clusters is 3. The classified memory cells shown in Figure 10 represent an internal image of the three input data patterns in Figure 8.

Figure 8. Three data patterns input to the INEPR algorithm.
Figure 9. The best number of clusters in the first test.
Figure 10. Generated internal image for three input data patterns.
In the second test, four data patterns shown in Figure 11 were input to the INEPR algorithm. The result of the L method in Figure 12 shows that the best number of clusters in the second test is 4. The generated internal image for the four input data patterns in Figure 13 demonstrates that the INEPR algorithm is able to recognize the new data pattern.

4. PERFORMANCE ANALYSIS

This section discusses the impact of the values of the model parameters on the number of antibody memory cells and the impact of the dissimilarity measures on the decision of the number of clusters.

Firstly, we discuss how aiNet parameters $\sigma_s$ and $n$ impact on the number of the antibody memory cells. The parameter $\sigma_s$ plays a major role in clonal suppression and network suppression. It directly controls the final size of the antibody memory cell set. The parameter $n$ is used to specify the number of antibodies that are selected for clone and mutation. Figure 14 shows the relation of the number of memory cells with the suppression threshold $\sigma_s$. In this figure, $\sigma_s$ changes from 0.05 to 1.25 with a step size of 0.01. The values of rest parameters are chosen as: $n=6$, $\zeta\% = 20\%$, and $\sigma_d = 7.3$. Figure 14 illustrates that the number of memory cells is very sensitive to the value of $\sigma_s$ within the range of $[0.05, 0.3]$. When the value of $\sigma_s$ is greater than 0.3, the number of memory cells is very small. The larger the $\sigma_s$ value, the higher the chance of eliminating antibody clones and memory cells in the clonal suppression and network suppression processes.

Figure 15 shows the impact of the parameter $n$ on the number of memory cells. In Figure 15, the values of the parameters $\sigma_s$, $\sigma_d$, and $\zeta\%$ are chosen as: $\sigma_s = 0.15$, $\sigma_d = 3$, and $\zeta\% = 9\%$. For each value of $n$, the number of memory cells is the average of five runs. Figure 15 shows that the number of memory cells increases as the value of $n$ increases. The larger value of $n$ means more antibodies being selected for clone and mutation. This increases the
diversity of antibody clones and reduces the number of antibody clones and memory cells being eliminated in the clonal suppression and network suppression.

Secondly, we investigated the impact of dissimilarity measures, on the decision of the best number of clusters of the antibody clustering. Different types of distance measures are applied to the clustering generation and the evaluation graph calculation. Figure 16 to Figure 18 show the \( L \) method results when the Correlation distance, Euclidean distance, and Seuclidean distance, are used for the point-to-point dissimilarity measure. The input data used for these three tests is the same data set shown in Figure 3. For the Correlation distance or Euclidean distance, the best number of clusters is decided to be 3, which is consistent with the original data set. For the Seuclidean distance, the best number of clusters is decided to be 4 instead of 3. These test results show that the point-to-point distance measure should be carefully chosen for a specific application.

5. CONCLUSIONS

The INEPR algorithm is based on the immune network theory and hierarchical clustering techniques. The goal of the INEPR is to dynamically generate an internal image mapping to the input data patterns without the need of specifying the number of clusters in advance. This goal is achieved through the construction of a network of antibodies memory cells, generation of a hierarchy of clustering for the antibody memory cell set using hierarchical clustering algorithms, determining the best number of clusters for the memory cell clustering with evaluation graphs and \( L \) method, and classifying the memory cells to form an internal image for the input data patterns. The INEPR model has been tested using damage data of a benchmark civil structure. The test results illustrate that the INEPR model is able to recognize emerging damage patterns. The impact of model parameters and dissimilarity measures on the number of memory cells and the number of clusters is also investigated. The investigation results show that model parameters \( \sigma \) and \( n \) have a significant impact on the number of memory cells.

REFERENCES


