Study on extension negative selection algorithm

Tianzhu Wen*

Department of Scientific Research,
Naval Aeronautical and Astronautical University,
Yantai, Shandong 264001, China
Email: wentianzhu1987@aliyun.com
*Corresponding author

Aiqiang Xu and Jiayu Tang

Department of Scientific Research,
Naval Aeronautical and Astronautical University,
Yantai, Shandong 264001, China
Email: hy_xuaiqiang@163.com
Email: 342118507@qq.com

Abstract: Aimed at the problems of low generation efficiency, serious redundancy and poor matching capability of detectors, the extension negative selection algorithm (ENSA) is proposed by fusing extenics and artificial intelligence system. The basic conceptions of ENSA are described by basic element, and the affinity between detector and antigen or antibody is calculated by dependent function. The algorithms of extension detector generation and optimisation are designed, and the parameters of them are analysed. Furthermore, the performance of ENSA is analysed both in theory and simulation experiment. The results from the Iris dataset show that when generating five detectors, the coverage rate of ENSA is 87.5% which is 70.28% higher than that of RNSA and 76.95% higher than that of V-Detector algorithm; when the expected coverage rate is 90%, three detectors are required in ENSA, which is 14 fewer than that of RNSA and 74 fewer than that of V-Detector algorithm; when the same antibodies are tested, the correct rate of ESNA and RNSA is 100% while the VDetector algorithm’s is 90%.

Keywords: extenics; negative selection algorithm; detector generation; detector optimisation; dependent function.


Biographical notes: Tianzhu Wen received his BSc in 2009 and MSc in 2011 both from Naval Aeronautical and Astronautical University. Currently, he is a Doctoral student in Naval Aeronautical and Astronautical. His main research interests include artificial intelligence and fault diagnosis.

Aiqiang Xu received his PhD in 2000 from Harbin Engineering University. Currently, he is a Professor and PhD supervisor in Naval Aeronautical and Astronautical. His main research interests include automatic test system and fault diagnosis.

Jiayu Tang received her BSc in 2014 from Naval Aeronautical and Astronautical University. Currently, she is a Master student in Naval Aeronautical and Astronautical. Her main research interest is artificial immune system.

This paper is a revised and expanded version of a paper entitled ‘Study on extension negative selection algorithm’ presented at The 2014 Asia-Pacific Services Computing Conference, Fuzhou, China, 4–6 December 2014.

1 Introduction

Artificial immune system (AIS) is a computational system inspired by biological immune phenomena and biological medical theories (Dasguptaa et al., 2011). Different immune algorithms are derived from diverse biological immune theories, such as immune genetic algorithm based on evolution theory, clonal selection algorithm based on clonal selection theory and negative selection algorithm based on self-non-self recognition theory. In the negative selection algorithm, at first the self-tolerance process of T cells in thymus is simulated to generate mature detectors; then the
non-self-identification process of T cells is simulated to detect antigen intrusion. As the negative selection algorithm does not need a priori knowledge of abnormal pattern, it provides a new method for anomaly detection, especially for the detection problems which are lack of abnormal data.

Negative selection algorithm (NSA) was first proposed by Forrest et al. (1994), in which antibody and antigen are encoded by string and the affinity between them is calculated by the r-contiguous-bits matching rule (Jin et al., 2013). Subsequently the real-valued encoding and matrix representation of antibody and antigen are proposed, and matching rules, such as Hamming distance, Minkowski distance and membership function, are developed (Xu et al., 2009). But the inherent limitations of NSA bring on low generation efficiency, serious redundancy and lack dynamic of detectors. Furthermore, the problems of deficiency of fuzziness, robustness and capability to resist noise are occur in the detectors. It is hard to solve these problems by improving the algorithm on deep understanding of related immune mechanism. So an idea of intelligent complementation is introduced to build a new artificial intelligent model which takes advantages of several artificial intelligent technologies. In this way, the practical problems can be solved more effectively.

Extenics is a discipline proposed by Prof. Wen Cai to solve the problems of incompatibility and contradiction, which combines qualitative analysis with quantitative calculation (Chunyan and Wen, 2013). Meanwhile its dependent function provides a new way for designing matching rules, thus make it possible to combine extenics with immune algorithms. Xiang (2009) optimises the classical range of dependent function by introducing the immune network theory into the extension control theory and solves the subjectivity problem resulted from expertise in controlling mobile cranes. Xiang and Huang (2008) presents an extension immune algorithm, in which each kind of fault samples is learning combined with the clonal selection theory, and achieves a more effective classical field in fault diagnosis of turbo generators. The two algorithms can both get a satisfying result, but in essence they just optimise the parameters of superiority evaluation method in extenics by immune algorithm, and do not take full advantages of immune algorithm in solving practical problems. 

So the extension negative selection algorithm (ENSA) is proposed. Firstly antibody and antigen are described by matter-element in extenics, which can effectively solve the fuzziness problem of antibody and antigen expression, and then the dependent function in extenics is used to calculate the affinity between detector and antigen or antibody, which can enhance the robustness and capability to resist noise of algorithm. Secondly, with the development of the detector generation algorithm, the detector’s generation efficiency is improved with no redundancy, further the detector optimisation algorithm is designed to optimise mature detectors, and the total number of detectors is decreased with larger coverage space of each detector. Thirdly the parameters of ENSA are analysed, and the availability of the algorithm is verified by simulation experiment.

2 Fundamental definitions

Basic element, a general designation of matter-element, event-element and relation-element, is a fundamental conception of extenics (Chunyan and Wen, 2013). It can describe matter, event or relation in a formalised way. Fundamental definitions of ENSA, such as problem space, samples and detectors are represented by basic element.

Definition 1: Problem space $U$ indicates a limited closed region where problems lie. Feature vectors $x_1, \ldots, x_n$ correspond to each dimension of the problem space $U$ separately. It can be expressed by basic element as

$$M = \begin{bmatrix} U, & x_1, & V_1 \\ x_2, & V_2 \\ \vdots \\ x_n, & V_3 \end{bmatrix}$$

where for $i=1,2,\ldots,n$, $V_i = [v_i^l, v_i^l]$ is the value range of feature vector $x_i$, $v_i^l, v_i^l$ are the minimum and maximum value of feature vector $S = \prod_{i=1}^{n}(v_i^l - v_i^l)$ is the whole coverage range of problem space $U$.

Definition 2: Self set $U_S \subset U$ is a set of all normal samples and non-self set $U_N \subset U$ is a set of all abnormal samples in the problem space $U$, and $U_S \cup U_N = U$, $U_S \cap U_N = \emptyset$.

Definition 3: Self-training set $U_T = \{t_k, k = 1, 2, \ldots, N_T\} \subset U_S$ is a set full of known normal samples in the self set. A normal sample $t_k$ can be expressed by basic element as

$$T_k = \begin{bmatrix} t_k, & x_1, & v_{t_k} \\ x_2, & v_{t_2} \\ \vdots \\ x_n, & v_{t_n} \end{bmatrix}$$

where $v_{t_k}$ is the value of feature vector $x$, of normal sample $t_k$ and $v_{t_k} \in V_t, i = 1, 2, \ldots, n, k = 1, 2, \ldots, N_T$.

Definition 4: Detector set $U_D$ is composed of detectors generated by self-tolerance process which can identify non-self samples in antigen. The detector $d_j$ can be expressed by basic element as

$$D_j = \begin{bmatrix} d_j, & x_1, & V_{d_j} \\ x_2, & V_{d_2} \\ \vdots \\ x_n, & V_{d_n} \end{bmatrix}$$

where $V_{d_j} = [v_{d_j}^l, v_{d_j}^l]$ is the value range of feature vector $x$, of detector $d_j$, $v_{d_j}^l, v_{d_j}^l$ are the minimum and maximum value of feature vector $x_i$ and $V_{d_j} \subset V_t, i = 1, 2, \ldots, n,$
\[ j = 1, 2, \ldots, N_d. \ S_j = \sum_{i=1}^{n} (v_i^j - v_i^j) \] is the whole coverage range of detector \( d_j. \)

**Definition 5:** Detector coverage \( C \) is the ratio of the coverage range of detector set to the whole range of the problem space \( U \), which is given by

\[
C = \left( \sum_{j=1}^{S_j} S_j - S' \right) / S
\]

where \( S' \) is the overlapping range among detectors, namely the redundant coverage range of detectors.

**Definition 6:** Detection rate \( P_{nt} \) is the ratio of the number of non-self samples detected by detectors to the total number of non-self samples.

\[
P_{nt} = \frac{N_{st}}{N_{nt} + N_{nf}}
\]

where \( N_{st} \) is the number of non-self samples exactly detected by detectors and \( N_{nf} \) is the number of non-self samples which are falsely recognised as self ones.

**Definition 7:** False alarm rate \( P_{fa} \) is the ratio of the number of self samples falsely detected as non-self samples to the total number of self samples.

\[
P_{fa} = \frac{N_{sf}}{N_{sf} + N_{st}}
\]

where \( N_{sf} \) is the number of self samples which are falsely recognised as non-self ones and \( N_{st} \) is the number of self samples exactly detected.

### 3 Extension negative selection algorithm

#### 3.1 Matching rule of extension detectors

Based on the fundamental definitions of ENSA, the affinity between original detectors and training sample or between mature detectors and samples to be detected can be calculated by dependent function. Take the tolerance process of detectors as an example (Gonzalez et al., 2002). For feature vector \( x_i, i = 1, 2, \ldots, n \), the dependent function of attribute \( v_i \) of training sample \( t_k \) about feature vector range \( V_g = [v_i^j, v_i^j] \) and detector \( d_j \) is given by

\[
r_{ijk} = \begin{cases} \rho(v_{ik}, V_g) \rho(v_{ik}, V_i) & 1, \rho(v_{ik}, V_g) = \rho(v_{ik}, V_i) = v_{ik} \notin V_g \\ D(v_{ik}, V_g, V_i) & \rho(v_{ik}, V_i) \end{cases}
\]

where \( \rho(v_{ik}, V_g) \) is extension distance between point \( v_{ik} \) and interval \( V_g \). \( \rho(v_{ik}, V_i) \) is extension distance between point \( v_{ik} \) and interval \( V_i \). \( \rho(v_{ik}, V_g, V_i) \) is the place value of point \( v_{ik} \) about the nest of interval composed of interval \( V_g \) and interval \( V_i \). \( \rho(v_{ik}, V_g) \) and \( \rho(v_{ik}, V_i) \) are given by

\[
\rho(v_{ik}, V_g) = v_{ik} - v_i^j + \frac{v_i^j - v_i^j}{2}
\]

\[
\rho(v_{ik}, V_i) = v_{ik} - v_i^j + \frac{v_i^j - v_i^j}{2}
\]

When

\[
\rho(v_{ik}, V_g) = \rho(v_{ik}, V_i), \quad D(v_{ik}, V_g, V_i) = v_i^j - v_i^j;
\]

when

\[
\rho(v_{ik}, V_g) \neq \rho(v_{ik}, V_i) \quad \text{and} \quad v_{ik} \notin V_g,
\]

\[
D(v_{ik}, V_g, V_i) = \rho(v_{ik}, V_i) - \rho(v_{ik}, V_g);
\]

when

\[
\rho(v_{ik}, V_g) \neq \rho(v_{ik}, V_i) \quad \text{and} \quad v_{ik} \in V_g,
\]

\[
D(v_{ik}, V_g, V_i) = \rho(v_{ik}, V_i) + \rho(v_{ik}, V_g);\]

According to the characteristics of dependent function, when \( v_{ik} \in V_g, k > k_{ik} > 0 \), which means point \( v_{ik} \) lies in the interval \( V_g \); when \( v_{ik} \notin V_g, v_{ik} \in V_i, k > k_{ijk} \geq -1 \), which means point \( v_{ik} \) lies in the interval \( V_i \) but outside the interval \( V_g \). Define the comprehensive affinity between training sample \( T_k \) and detector \( D_j \) as

\[
\text{Aff}_{jk} = \sum_{i=1}^{n} w_i \cdot n_{jk}
\]

where \( \text{Aff}_{jk} \) is the affinity of training sample \( T_k \) and detector \( D_j \) about feature vector \( x_i, k = 1, 2, \ldots, N_r, j = 1, 2, \ldots, N_d, w_i \) is the weight of feature vector \( x_i \) and \( \sum_{i=1}^{n} w_i = 1, w_i \geq 0, i = 1, 2, \ldots, n \). \( \text{Aff}_{jk} \geq \delta \) indicates that training sample \( T_k \) matches with detector \( D_j \) while \( \text{Aff}_{jk} < \delta \) represents the mismatch.

Supposed that the matching threshold \( \delta = 0 \). When \( \text{min}(r_{ijk}) \geq 0 \), for any feature \( x_i \), attribute value \( v_i \) of sample \( t_k \) is within the feature vector range \( V_i \) corresponding to detector \( d_j \), namely sample \( t_k \) is inside the coverage range of detector \( d_j \) and \( \text{Aff}_{jk} > 0 \). When \( \exists r_{ijk} < 0 \), for any feature vector \( x_i \), there exists attribute \( v_i \) of sample \( t_k \) exceeds the feature vector range \( V_i \), namely sample \( t_k \) is outside the coverage range of detector \( d_j \) and \( \text{Aff}_{jk} \) is related to weight \( w_i \) which means sample \( t_k \) and detector \( d_j \) can either match or mismatch. So the match is controlled by the weight of feature vectors instead of the absolute judgment according to whether the sample is inside or outside the range of detector, thus the robustness and the capability to resist noise of algorithm is enhanced.

#### 3.2 Extension detector generation algorithm

According to the matching rule of extension detector, a new extension detector generation algorithm is proposed (Krishnamoorthi and Natarajan, 2013; Wang et al., 2010). It divides the problem space into \( m \) small hyperrectangles which are viewed as \( m \) original detectors and calculates the
affinity between each original detector and self-training sample. When \( \text{Aff}_{jk} < \delta, k = 1, 2, \ldots, N, \) the detector is marked as semi-mature detector; when \( \text{Aff}_{jk} \geq \delta, \) a second division is executed to acquire new detectors for matching with self-training samples. This process is repeated until the coverage of semi-mature detector reaches \( C, \) then the semi-mature detector set \( M_{ID} \) is attained. Detailed procedures are as follows.

Step 1 Initialise problem space \( M \) and self-training set \( M_T = \{ Ti_k, k = 1, 2, \ldots, N_T \} \). Regard problem space as immature detector, then semi-mature detector set \( M_{ID} = \emptyset \) and immature detector set \( M_{ID} = \{ M \} \).

Step 2 For semi-mature detector \( YD_j \in M_{ID}, j = 1, 2, \ldots, \ N_{ID} \), the division process is that randomly choose point \( y_i \in (v^j_{yi}, v^{j'}_{yi}) \), \( i = 1, 2, \ldots, n \) on feature vector \( x_i \), along the perpendicular direction of \( y_i \) on feature vector \( x_i \) divide the semi-mature detector \( YD_j \) into \( m = 2^n \) original detectors which are expressed by basic element as \( D_1, D_2, \ldots, D_m \).

Step 3 For each got original detector \( j = 1, 2, \ldots, m \) calculate its affinity \( \text{Aff}_{jk} \) about self-training samples. When \( \max_{k=1,2,\ldots N_T} \text{Aff}_{jk} < \delta, \) the original detector \( D_j \) is mismatched with all the samples in the self-training set \( M_T \), then add the original detector \( D_j \) to the semi-mature detector set \( M_{ID} = M_{ID} \cup \{ D_j \} \), otherwise add it to the mature detector set \( M_{MD} = M_{ID} \cup \{ D_j \} \).

Step 4 Calculate the coverage \( C \) of semi-mature detector set \( M_{ID} \). If the coverage meets the requirement, terminate the algorithm, else for every semi-mature detector \( YD_j \) in the set \( M_{ID} \), repeat Step 2 and Step 3 till the coverage \( C \) meets the requirement.

3.3 Extension detector optimisation algorithm

Although the semi-mature detectors got by extension detector generation algorithm do not have redundancy, it is possible to merge with the adjacent semi-mature detectors. So the extension detector optimisation algorithm is proposed for acquiring fewer mature detectors (Ayara et al., 2002; Zheng et al., 2013). According to the inverted sequence of semi-mature detector generation, along every direction of feature vector \( x_1, x_2, \ldots, x_n \), the semi-mature detector set \( M_{ID} \) is merged till the mergence cannot go on, then the mature detector set \( M_{MD} \) is attained. Detailed procedures are as follows.

Step 1 Label elements of semi-mature detector set \( M_{ID} \) in the inverted sequence of generation and attain semi-mature detector set \( M_{ID} = \{ HD_j, j = 1, 2, \ldots, N_{ID} \} \).

Step 2 According to the label sequence of semi-mature detectors, along every direction of feature vectors \( x_1, x_2, \ldots, x_n \), semi-mature detectors are merged in turn.

Step 3 For each semi-mature detector \( HD_i, i = 1, 2, \ldots, N_{ID} - 1 \) successively decide whether it can merge with detector \( HD_j, j = i + 1, i + 2, \ldots, N_{ID} \). If possible, mark the new detector as \( HD' \) with the label of all semi-mature detectors after \( HD \) reduced by one, else go on the judgment of \( HD_{i+1} \) till \( i = N_{ID} - 1 \).

Step 4 For the new semi-mature detector set \( M_{MD} = \{ HD_j, j = 1, 2, \ldots, N_{MD} \} \) repeat Step 3 till all semi-mature detectors cannot merge anymore, then the mature detector set \( M_{MD} = \{ MD_j, j = 1, 2, \ldots, N_{MD} \} \) is attained.

3.4 Time complexity analysis

The time complexity of ENSA is

\[
O \left( \frac{N_D \cdot N_T}{(1 - P_m)^{N_T}} + \frac{N_D (N_D - 1)}{2} \right).
\]

Proof: The self-tolerance process contributes to the main time cost of ENSA, RNSA and V-Detector algorithm. Supposed that there are \( N \) original detectors and \( d \) \( N \) detectors are generated after self-tolerance process. According to Lemma \( N \cdot (1 - P_m)^{N_T} = N_N \), so

\[
N = \frac{N_D}{(1 - P_m)^{N_T}}.
\]

namely the time complexity of RNSA and V-detector algorithm is

\[
O(N \cdot N_T) = O \left( \frac{N_D \cdot N_T}{(1 - P_m)^{N_T}} \right).
\]

The detectors need to be optimised after self-tolerance process in the ENSA, and the time complexity of optimisation is

\[
O \left( \frac{N_D (N_D - 1)}{2} \right),
\]

so the total time complexity of ENSA is

\[
O \left( \frac{N_D \cdot N_T}{(1 - P_m)^{N_T}} + \frac{N_D (N_D - 1)}{2} \right).
\]

4 Simulation experiment

The Iris dataset is utilised to compare the performance of ENSA, RNSA and V-Detector algorithm (Lifeng et al., 2012; Gonzalez et al., 2003; Chen et al., 2013). There are three kinds of data in the Iris dataset including setosa, versicolor and virginica. Each kind of data is composed of 50 samples with four attributes which are sepal length, sepal width, petal length and petal width. In order to show the
extension detector generation procedure directly, sepal length and petal length from the preceding 25 samples in setosa are chosen for the self-tolerance process. From a practical perspective, the data is not normalised to preserve the physical characters of original data. The midpoint of feature vectors \( y_i = \frac{v_{ij}^L + v_{ij}^H}{2}, \ i = 1, 2, \ldots, n \) is selected as the division point in the extension detector generation algorithm. The generation process of extension detectors is shown in Figure 1.

Figure 1  Extension detector generation process (see online version for colours)

In the generation process of extension detector shown in Figure 1, suppose that the matching threshold \( \delta = 0 \) and the coverage of detector set \( C = 92\% \). The simulation results show that four times of division are needed to meet the coverage requirement, and nine mature detectors are generated with the coverage \( C = 93.75\% \). The number marked in Figure 1 denotes the generating sequence of semi-mature detectors. From Figure 1, it is known that semi-mature detectors can be optimised by mergence. The optimisation process of extension detectors is shown in Figure 2.

In the optimisation process of extension detector shown in Figure 2, semi-mature detectors are labelled in the inverted sequence of generation. The detectors are merged along every direction of feature vectors till no more detectors can be merged. Three mature detectors are obtained after optimisation. In order to further verify the performance of ENSA, the algorithm presented is compared with fixed-radius RNSA and variable-radius V-Detector algorithm. First the coverage of all algorithms is compared when both generating five detectors; then the number of mature detectors is compared under the same coverage 90%; finally the detection results of each algorithm towards antigen are compared, all of them are shown in Figure 3.

The top three pictures in Figure 3 show the comparison of three algorithms when generating five detectors. The coverage of RNSA, V-Detector algorithm and ENSA is 10.55\%, 17.22\% and 87.50\% respectively where the detector radius of RNSA \( r_d = 5 \). The middle three pictures show the comparison of three algorithms when the coverage \( C \geq 90\% \). There are 77, 17 and three detectors needed in RNSA, V-Detector algorithm and ENSA respectively. The comparison results of antigen detection performance are shown in the bottom three pictures in Figure 3. Five groups of setosa sample, ten groups of versicolor sample and ten groups of virginica sample in the Iris dataset are chosen as antigen data, which are represented by ‘o’, ‘+’ and ‘*’ respectively. The detection rate of all the three algorithm are higher than 90\%, with 100\%, 90\% and 100\% respectively. It is known from comparison results that the generation efficiency of ENSA is much higher than that of RNSA and V-Detector algorithm with fewer detectors, larger coverage range and no redundancy.

Figure 2  Extension detector optimisation process (see online version for colours)
5 Conclusions

ENSA is a new kind of negative selection algorithm which is formed by fusing extenics and artificial intelligence system. Basic element in extenics is adopted to represent antigen and antibody whose attributes are used to restrict the size of problem space and detector set, so the requirements of practical problems can be met more. Dependent function in extenics is adopted to calculate the affinity between detector and antigen or antibody, and the robustness and capability to resist noise of algorithm are improved. Furthermore, the extension detector generation and optimisation algorithms are proposed for the purpose of higher generation efficiency, larger coverage range and less need of detectors. Finally the performance of ENSA is verified by comparing it with fixed-radius RNSA and variable-radius V-Detector algorithm, the performance of algorithm is analysed not only in theory but also in simulation experiment utilising Iris dataset. Both the theoretical analysis and simulation results prove that the performance of ENSA is better than RNSA and V-Detector algorithm.

References


Study on extension negative selection algorithm


