International Journal of Computational Intelligence Systems, Vol. 4, No. 4 (June, 2011), 410-419

A NEGATIVE SELECTION ALGORITHM BASED ON HIERARCHICAL CLUSTERING OF SELF SET AND ITS APPLICATION IN ANOMALY DETECTION

Wen Chen, Xiao-Jie Liu*, Tao Li, Yuan-Quan Shi, Xu-Fei Zheng, Hui Zhao
College of Computer Science, Sichuan University
Chengdu, 610065, China
E-mail: cwcwcwcw2006 @163.com, liuxiaojie8 @126.com

Received 12 September 2010
Accepted 29 April 2011

Abstract

A negative selection algorithm based on the hierarchical clustering of self set HC-RNSA is introduced in this paper. Several strategies are applied to improve the algorithm performance. First, the self data set is replaced by the self cluster centers to compare with the detector candidates in each cluster level. As the number of self clusters is much less than the self set size, the detector generation efficiency is improved. Second, during the detector generation process, the detector candidates are restricted to the lower coverage space to reduce detector redundancy. In the article, the problem that the distances between antigens coverage to a constant value in the high dimensional space is analyzed, accordingly the Principle Component Analysis (PCA) method is used to reduce the data dimension, and the fractional distance function is employed to enhance the distinctiveness between the self and non-self antigens. The detector generation procedure is terminated when the expected non-self coverage is reached. The theory analysis and experimental results demonstrate that the detection rate of HC-RNSA is higher than that of the traditional negative selection algorithms while the false alarm rate and time cost are reduced.

Keywords: Artificial Immune System, Negative Selection, Detector, Cluster.

1. Introduction

Negative selection is a biological process by which the immune system generates non-self detectors that do not match self structures. The biological negative selection process can be mapped to the computational domain as a two class pattern classification problem in the artificial immune system (AIS) [1], in which the normal states correspond to self antigens while the abnormal states correspond to non-self antigens. In AIS, negative selection algorithm (NSA) is an important method for the generation of detectors. NSA is designed by modeling the biological process in which T-cells mature in thymus through being censored against self cells [2]. After negative selection, the left mature (valid) detectors are used for further applications such as anomaly detection [3], machine learning [4], pattern recognition [5], intrusion detection [6], and etc.

The native negative selection algorithm (NNSA) defines the self/non-self discrimination problems using binary representations and calculates the affinities between binary strings by the *r*-contiguous-bits method [2,7,8]. Li and T. Stibor pointed out that the efficiency of NNSA is too low to be applied [9,10,11]: under the given failure rate $P_f \approx e^{-Pm|D|}$, where P_m is the match probability between random detector and antigen, the least number of detector candidates N_0 is $-ln(P_f)/(P_m \cdot (1-P_m)^{Ns})$ which means N_0 is exponentially related to N_s and the time

*Corresponding author. Tel.: +86-28-85405568. E-mail address: liuxiaojie8@1263.com(X.J. Liu).

complexity of NNSA is $O(N_0 \cdot N_s)$ [2]. Thus, the time cost of NNSA cannot be accepted when the self size is large [10].

Gonzalez and Dasgupta introduced the real negative selection algorithm RNSA [12,13], which normalized detectors and antigens into $[0, 1]^d$. And then, Ji and Dasgupta improved RNSA using variable detector radius called V-Detector, which set detector radius by the nearest self distance to enlarge the non-self coverage with little number of detectors [14,15].

T. Stibor indicated that RNSA and V-detector also suffered to the curse of dimensionality [16,17]. On the one hand, the distances between antigens in the high dimensional data space converges to a constant value. Therefore, there is little distinctiveness between self and non-self antigens, resulting in the higher false alarm rate. On the other hand, the algorithms terminate with only a very small number of large radii detectors (hyperspheres) which are covering a limited number of spikes. As a result a large proportion of the volume of the hypercube ($[0, 1]^d$) does not lie within the hyperspheres, it lies in the remaining (high-volume) spikes. Thus the detection rate is lower.

Additionally, for most pattern recognition algorithms, the distance calculation is the main source of time consuming [18,19]. However, NNSA, RNSA and V-detector didn't take any strategy to reduce the cost of distance calculation: the distances from detector candidates to the self set have to be calculated, resulting in the lower efficiency [9]. Furthermore, as there are many overlapped detection regions, the reduction of detector redundancy must also be taken into consideration.

A real negative selection algorithm based on the hierarchical clustering of self set (HC-RNSA) is present in this article. The underline idea is that first, the self data set is preprocessed using Principle Component Analysis (PCA) method to reduce the data dimension, and then the self set is hierarchically clustered. During the detector generation process the detector candidates, restricted in the lower coverage space, are compared with the cluster centers using fractional distance function to eliminate the self reactive detectors. The detector generation process is recursively continued from the higher cluster level to the lower level until the cluster radius is less than the self radius, and in each cluster level, the exit condition is to reach the expected non-self coverage.

2. Basic Definition

In the AIS, antibodies are defined as detectors which are used to recognize non-self elements[2]. Therefore, the accuracy of the detection result is determined by the quality of detectors. As the randomly generated detector candidates may matched self elements, resulting in self reactive [2,12]. The negative selection algorithm, inspired by the censoring process of antibody cells in the biological body, was designed to eliminate the self reactive detectors. The basic conceptions are defined as:

Def 1. All the character strings abstracted from the sample space constitute the antigen set $U = \{g \mid g = (f_1, f_2 \ldots f_n), f_i \in [0, 1]\}$, where n is the data dimension and f_i is the *ith* normalized attribute.

Def 2. The self set $S \subset U$ is the character strings abstracted from the normal samples, $r_s \in R^+$ is the variability threshold of the self points; Non-self set $N = U - S$, which are character strings abstracted from the abnormal samples, and $S \cup N = U$, $S \cap N = \Phi$.

Def 3. Detector $d = \langle c, r \rangle$, where $c \in N$, c is the central vector which represents the location of d in the sample space, $r \in R^+$ is the detector radius. Antigens which are close to d less than r will be identified as non-self elements.

Def 4. The non-self coverage of detectors is defined as the ratio of the volume of the non-self space that can be recognized by any detector to the volume of the entire non-self space [14].

$$p = \frac{V_{covered}}{V_{nonself}} = \frac{\int_{covered} dx}{\int_{nonself} dx}. \quad (1)$$

Def 5. Anomaly detection is to find a functional mapping f:

$$R^N \xrightarrow{f} \{C_0, C_1\}. \quad (2)$$

using training data samples generated according to an unknown probability distribution $P(x, y)$:

$$(x_1, y_1), \ldots (x_n, y_n) \in R^N, Y = \{C_0, C_1\}. \quad (3)$$

where C_0 is the set of normal samples, C_1 is the set of abnormal samples, and such that f will correctly classify unknown examples (x, y). For AIS, the training set only contains normal samples $(x, y \in C_0)$ and the task is to detect abnormal samples $(x, y \in C_1)$ with the function f trained by normal samples. As described in Ref. 17,

abstracting these principles and modeling immune components according to the AIS framework, we obtain a technique for anomaly detection:

Input: S = set of points $\in [0, 1]^n$ gathered from normal behavior of a system.

Output: D = set of hyperspheres, which recognizing a proportion of the total space $[0, 1]^n$, except the normal points.

Detector generation: While non-self coverage of detectors is not reached, generate hyperspheres.

Classification: If unknown point lies within a hypersphere, it does not belong to the normal behavior of the system and is classified as an anomaly.

3. The Description of HC-RNSA

3.1. *The strategies of detector generation*

3.1.1. The estimation of non-self coverage

As Eq. (1) is hard to calculate, we select fixed number of samples in the non-self space and then estimate p using statistical inference method: the probability of a random sample to be recognized by detector set D obeys binomial distribution [15], $P\{x = 1, x \text{ is covered}\} = p$, $P\{x = 0, x \text{ is uncovered}\} = 1-p$. According to the Neyman-Pearson theorem, there exists a most powerful test for the hypothesis testing problem in Eq. (4):

$$H_0 : p < p_{exp},\ H_1 : p \geq p_{exp}. \tag{4}$$

where H_0 is the hypothesis that the expected non-self coverage is not reached and H_1 is on the contrary. The rejection region of Eq. (4) is the same as that of Eq. (5) [20].

$$H_0 : p < p_{\exp},\ H_1 : p > p_1 (> p_{\exp}). \tag{5}$$

where p_1 is a random value bigger than p_{exp}. The likelihood ratio of Eq.(5) is calculated through a random sample set $\{x_1, x_2, \ldots x_n\}$.

$$\frac{L(p_1)}{L(p_{\exp})} = \frac{\prod_{i=1}^{n} P(x_i, p_1)}{\prod_{i=1}^{n} P(x_i, p_{\exp})} = \frac{\prod_{i=1}^{n} [(p_1^{x_i}(1-p_1)^{1-x_i}]}{\prod_{i=1}^{n} [(p_{\exp}^{x_i}(1-p_{\exp})^{1-x_i}]}. \tag{6}$$

Suppose $\zeta = \sum_{i=1}^{n} x_i$,

$$\begin{aligned} \frac{L(p_1)}{L(p_{\exp})} &= (\frac{p_1}{p_{\exp}})^{\zeta} (\frac{1-p_1}{1-p_{\exp}})^{n-\zeta} \\ &= (\frac{1-p_1}{1-p_{\exp}})^{n} [\frac{p_1(1-p_{\exp})}{p_{\exp}(1-p_1)}]^{\zeta} \geq C. \end{aligned} \tag{7}$$

As $p_1 > p_{exp}$, $p_1 (1-p_{exp}) > p_{exp} (1-p_1)$, therefore Eq. (7) equals:

$$\zeta \geq \frac{n \log[\frac{c^{1/n}(1-p_{\exp})}{(1-p_1)}]}{\log[\frac{p_1(1-p_{\exp})}{p_{\exp}(1-p_1)}]}. \tag{8}$$

So the rejection range of H_0 is:

$$\chi\{\zeta \geq b\}. \tag{9}$$

where $b = n \log[\frac{c_\alpha^{1/n}(1-p_{\exp})}{(1-p_1)}] / \log[\frac{p_1(1-p_{\exp})}{p_0(1-p_1)}]$.

When ζ is bigger than b, the initial hypothesis H_0 is rejected, and the detector generation procedure is stopped.

3.1.2. The value ranges of detector candidates

The random value ranges (RVR) are set of d dimensional hypercube: ranges = {hypercube | hypercube = ([low_1, $low_2 \ldots low_d$], [$high_1$, $high_2 \ldots high_d$])}. $High_i$ and low_i represent the upper and lower bounds of the ith attribute of the detectors' central vectors. The RVR of the ith cluster of cluster level l is defined in Eq. (10):

$$\text{Hypercube}_{li} = ([c_{i1}-r_l \ldots c_{id}-r_l], [c_{i1}+r_l \ldots c_{id}+r_l]). \tag{10}$$

where c_i is the cluster center and r_l is the cluster radius. During the detector generation procedure of level l+1, the non-self space outside the RVR of level l has been covered by detectors, so the new generated detector candidates should be located in the RVR of level l to reduce detector redundancy.

As Fig. 1 shows, during the detector generation process of level 2, the non-self space outside the RVR of level 1 has been covered by detectors, therefore the detector candidates in level 2 are generated within the

RVR of level 1 to reduce the detector redundancy. So does the detector generation process in the level 3.

3.1.3. The probability of generating invalid detectors

As the randomness during the detector generation procedure, many invalid detectors (covered self samples) are generated which results in the lower efficiency. Let p represents the probability of generating an invalid detector. For RNSA [12] and V-detector [14], as the central vectors of the detector candidates are randomly sampled from the unit hypercube $[0, 1]^d$, p is the ratio of the self hyperspheres' volume to the volume of the unit hypercube:

$$P_1=\frac{n\cdot V_{self}}{V_{cube}}=\frac{n\cdot r_s^d\cdot\pi^{d/2}}{\Gamma(d/2+1)}. \tag{11}$$

where n is the self number, V_{self} is the volume of single self hypershpere, V_{cube} is the volume of the unit hypercube.

In HC-RNSA, samples are chosen from the random value range which is hypercube with edge length $4r$, r is the cluster radius, so p is the ratio of the cluster hyperspheres' volume to the volume of the $4r$-hypercube:

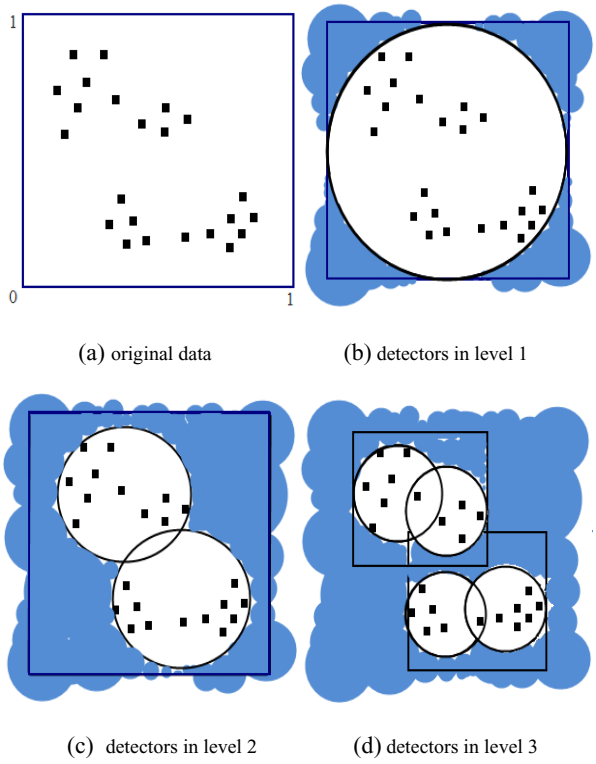
$$P_2=\frac{m\cdot V_{clu}}{V_{random\_cube}}=\frac{m\cdot r^d\cdot\pi^{d/2}}{(4\cdot r)^d\cdot\Gamma(d/2+1)}. \tag{12}$$

where m is the number of clusters, V_{clu} is the volume of a cluster hypersphere, V_{random_cube} is the volume of the $4r$-hypercube (RVR).

To compare the efficiency of RNSA、V-detector and HC-RNSA, we define the coefficient as follows:

$$\rho=\frac{P_1}{P_2}=\frac{\dfrac{n\cdot r_s^d\cdot\pi^{d/2}}{\Gamma(d/2+1)}}{\dfrac{m\cdot r^d\cdot\pi^{d/2}}{(4\cdot r)^d\cdot\Gamma(d/2+1)}}=\frac{n}{m}(4r_s)^d. \tag{13}$$

From Eq. (13), we get coefficient $\rho=(\sqrt[d]{\frac{n}{m}}\cdot 4r_s)^d$,



(a) original data (b) detectors in level 1

(c) detectors in level 2 (d) detectors in level 3

Fig. 1. The rectangles are random value ranges, the cycles are clusters and the shadows are regions covered by detectors. From the higher cluster level to the lower level, the cluster radius is halved.

when ρ is bigger than 1, the efficiency of HC-RNSA is higher than that of the traditional algorithms. As Fig. 2 shows, when the data dimension is lower than 20 and the self radius is bigger than 0.05, ρ >1, otherwise the efficiency of HC-RNSA is lower than that of the traditional NSAs.

Therefore, the data pretreatment process is needed to reduce the data dimension when deals with high dimensional data. In the article, the Principle Component Analysis (PCA) method is employed:

First, n antigen samples $x_1, x_2\ldots x_n$, $x_i = (x_{i1}, x_{i2}\ldots x_{ip})$, $1 \le i \le n$ are selected to calculate the correlation coefficient matrix R:

$$R_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}. \tag{14}$$

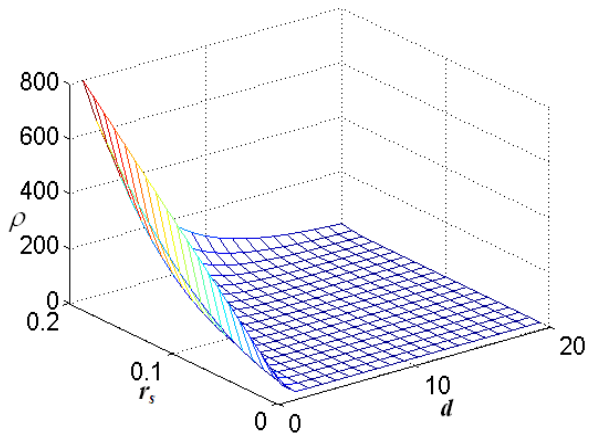


Fig. 2. The relationship between ρ, r_s and d

Then the eigenvector of R is calculated and ordered by values. The first m eigenvector whose accumulative contribution rate is more than threshold are selected as the principle components.

3.1.4. The fractional distance function

Theorem 1. *Let F is a random distribution of two antigens, for the L_k metric,* $\lim_{d\to\infty} E\left[\frac{D_{max}-D_{min}}{d^{1/k-1/2}}\right]=C_k$, *where d is the data dimension, C_k is a constant dependents on norm k, D_{max} and D_{min} are the maximum and minimum distances from antigens to the origin using the L_k metric.*

Proof. Let $A = (P_1 \ldots P_d)$ and $B=(Q_1 \ldots Q_d)$, with P_i and Q_i being drawn from F. Let $d_A = \sqrt[k]{\sum_{i=1}^{d} P_i^k}$ is the distance from A to the origin using the L_k metric and $d_B = \sqrt[k]{\sum_{i=1}^{d} Q_i^k}$ is the distance from B to the origin.

As the d attributes of A and B are drawn from the distribution F with mean μ and standard deviation σ, which means [21]: $\frac{d_A^k}{d} \to_p \mu$, $\frac{d_B^k}{d} \to_p \mu$, therefore:

$$\frac{d_A}{d^{1/k}} \to_p \mu^{1/k}, \frac{d_B}{d^{1/k}} \to_p \mu^{1/k}. \quad (15)$$

The comparison is between two random antigens, so

$$|D_{max} - D_{min}| = |d_A - d_B| = \frac{|d_A^k - d_B^k|}{\sum_{i=0}^{k-1}(d_A)^{k-i-1}(d_B)^i}. \quad (16)$$

From Eq.(16) we get:

$$\frac{|d_A - d_B|}{d^{1/k-1/2}} = \frac{|\sum_{i=1}^{d}(P_i^k - Q_i^k)|/\sqrt{d}}{\sum_{i=0}^{k-1}(d_A d^{-1/k})^{k-i-1}(d_B d^{-1/k})^i}. \quad (17)$$

Since each $R_i = P_i^k - Q_i^k$, $1 \le i \le d$ is a random variable with zero mean and finite variance $\sqrt{2}\sigma'$, where σ' is the standard deviation of P_i^k. The sum of different values of R_i over d dimensions will converge to a normal distribution: $\sum_{i=1}^{d}(P_i^k - Q_i^k) \sim N(0, 2d\sigma'^2)$ because of the central limit theorem, so the expected value of the numerator is a constant C.

$$|\sum_{i=1}^{d}(P_i^k - Q_i^k)|/\sqrt{d} \to_p C. \quad (18)$$

According to Slutsky's Theorem, put Eq. (15) into the denominator of Eq. (17), get

$$\sum_{i=0}^{k-1}(d_A d^{-1/k})^{k-i-1}(d_B d^{-1/k})^i \to_p k\mu^{(k-1)/k}. \quad (19)$$

Combine the results of Eq. (18) and Eq. (19) to obtain

$$\lim_{d\to\infty} E\left[\frac{D_{max}-D_{min}}{d^{1/k-1/2}}\right] = C_k. \quad (20)$$

where C_k is some constant dependents on k. □

From Fig. 3 we can see D_{max} - D_{min} is increasing with $d^{(1/k)-(1/2)}$, which inspires us to use the fractional norm distance function in Eq. (21) to enhance the distinctiveness between self and non-self antigens.

$$dis(x,c) = \sqrt[l]{\sum_{i=1}^{d}|x.f_i - c.f_i|^l}, 0<l<1. \quad (21)$$

3.2. *The negative selection algorithm HC-RNSA*

The pseudo-code of HC-RNSA is:

HC-RNSA(C, D, P_{exp}, N)

C: cluster centers, P_{exp}: expected coverage

D: detector set, N: non-self sample size

Step1 Preprocess the self data set using PCA method to reduce the data dimension.

Step2 Hierarchically cluster the self set, the cluster result is stored in set C.

Step3 Initialize cluster level i =1, non-self number n =0, cover count m =0.

Step4 Sample non-self data x from the random value ranges of the *ith* cluster level.

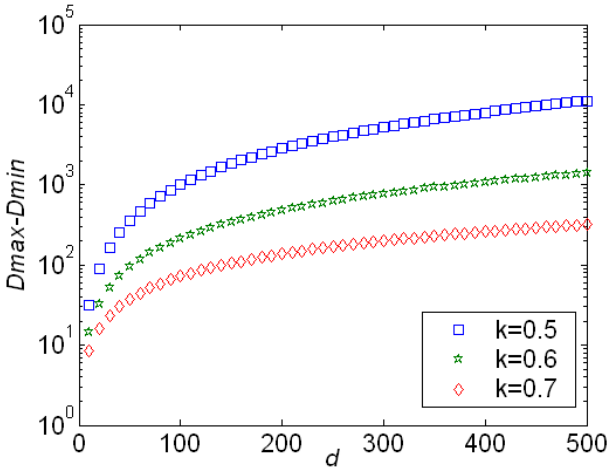
Step5 Calculate the distance $dis(x, c)$ between x and each center c in C_i by Eq. (21), if $dis(x, c)$ is less than the cluster radius r_i, drop x, go to step 4**,** else increase n.

Step6 If n equals N, calculate the rejection range of H_0 by Eq. (9). If H_0 is rejected then increase i, reset m, n, go to step 4; else put the detectors from Td into D.

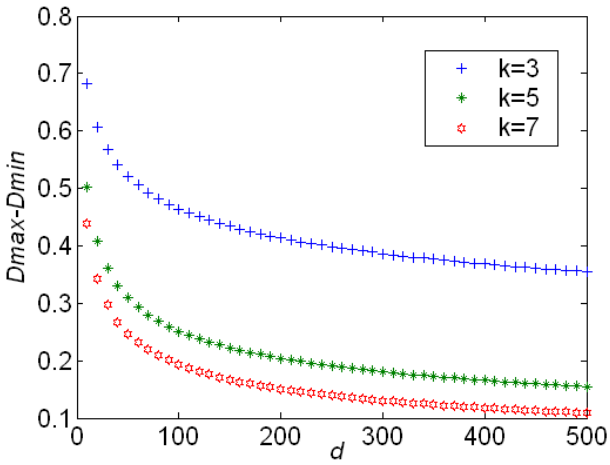
Step7 If x is covered by detectors in D, increase the cover count m

Step8 Generate detector $d<x, r>$. Put $d<x, r>$ into the temporary set Td, go to step4.

Steps 1-3 are the data pretreatment stages, in which the data dimension is reduced and the self set is hierarchically clustered. Steps 4-8 are the detector generation process. In step4 the detector candidates are restricted to the random value ranges to reduce the detector redundancies. In step 5, the self data is replaced by the cluster centers to compare with the detector candidates. As the number of cluster centers is far less than the self set size, the efficiency of the negative selection process is much enhanced. In step 6, the termination criterion is based on the hypothesis test of the non-self coverage. If the hypothesis H_0 in Eq. (4) is rejected, the generation procedure can be terminated. Step 7 test whether the non-self sample x is covered by detectors to accumulate the cover count m. In step 8, the non-self sample x is reused to generate detector $d<x, r>$, where r is the nearest distance between x and the cluster ranges, and r is a byproduct of step 5. As the non-self coverage of detectors is varying when new detectors are put into D, so the detector set is unchanged until the number of non-self samples equals the predefined size N.



(a) Fractional norm distance



(b) Integer norm distance

Fig. 3. The relationship between d and D_{max}-D_{min} under different distance norm

3.3. *The cost of HC-RNSA*

Theorem 2. *The time complexity of the detector generation process of HC-RNSA is irrelevant to the self set size.*

Proof. In HC-RNSA, the time consuming of step 4, 6 and 8 is a constant time t which could be ignored. In step 5, the distances from the sample x to the cluster centers are calculated, the time complexity of this step is O ($|C_l|$), C_l

is the set of cluster centers in level l. Assuming the number of samples in cluster level l is N_l, the number of non-self samples of step 7 is $(1-P_l)\cdot N_l$, P_l calculated by Eq. (12) is the probability of generating self samples. The distances between the non-self sample x and the detector set D are calculated in step 7, and the time complexity of this step is $(1-P_l)\cdot N_l\cdot|D|$. So the time complexity of the detection generation process in level l is $O(N_l\cdot|C_l|+(1-P_l)\cdot Nl\cdot|D|)$. As the level of the hierarchical clusters is $k=\left\lfloor \log_2 \frac{\sqrt{n}}{r_s} \right\rfloor$, the total time complexity of the detector generation process is $O(\sum_{l=1}^{k}(N_l\cdot|C_l|+(1-P_l)\cdot N_l\cdot|D|))$. The total number of samples is $N_0=\sum_{l=1}^{k}N_l\approx\frac{|D|}{1-\bar{P}}$, thus the simplified time complexity is $O(N_0\cdot|\bar{C}|+(1-\bar{P})\cdot N_0\cdot|D|)$, where $\bar{P}$ is the average probability of generating self samples and $|\bar{C}|$ is the average number of clusters in each level. Therefore, the time complexity of the detector generation process of HC-RNSA is irrelevant to the self set size. □

Table 1. The time complexity of the detector generation process.

Algorithm	Time complexity
NNSA	$O(\frac{-\ln(Pf)}{Pm\cdot(1-Pm)^{Ns}}\cdot Ns)^{2}$
RNSA	$O(\frac{\|D\|}{(1-P)^{Ns}}\cdot Ns)^{12}$
V-Detector	$O(\frac{\|D\|}{(1-P)^{Ns}}\cdot Ns)^{14}$
HC-RNSA	$O(N_0\cdot\|\bar{C}\|+(1-\bar{P})\cdot N_0\cdot\|D\|)$

As Table 1 shows the self set size is exponentially related to the time complexity of the traditional NSAs. Therefore, when the number of self data increases, the time consuming increases incredibly. But for HC-RNSA, the time complexity of the detector generation process is irrelevant to the self set size, which means HC-RNSA is suitable for the detector generation under large number of self data.

4. Experiment

To test the anomaly detection performance of HC-RNSA and compare it with the traditional NSAs: NNSA[2], RNSA[12] and V-detector[14], comparison experiments are designed based on different classic UCI (University of California Irvine) data sets [22], which have been widely used in the fields of anomaly detection, disease diagnose, equipment detection, etc [22].

The detection rate (DR), false alarm rate (FA) and time cost are three evaluation criterions of NSAs.

$$DR = TP/(TP+FN). \tag{22}$$

$$FA = FP/(FP+TN). \tag{23}$$

where TP, TN, FP and FN are the counts of true positive, true negative, false positive and false negative respectively.

The experimental data properties are described in Table 2. In the data sets Ball-bearing and Delft pump, all the records with normal equipment state are taken as self data, others are non-self data. In the other data sets, the records collected from healthy people are self data, and records from the unhealthy constitute non-self data set. The data records are first normalized into $[0, 1]^d$, and then NSAs including: HC-RNSA, NNSA[2], RNSA[12], V-detector[14] are employed to generate detectors based on these data sets. The parameters are shown in Table 3.

Table 2. The data prosperities of the UCI data sets.

Data set	Dimension	Record number	Training set	Test set	
				self	non-self
Ball-bearing	32	4150	593	320	3237
Delft pump	64	1500	244	132	1124
Arrhythmia	278	420	154	83	183
B.Cancer	9	699	156	85	458
Biomed	5	194	82	45	67
Diabetes	8	768	174	94	500

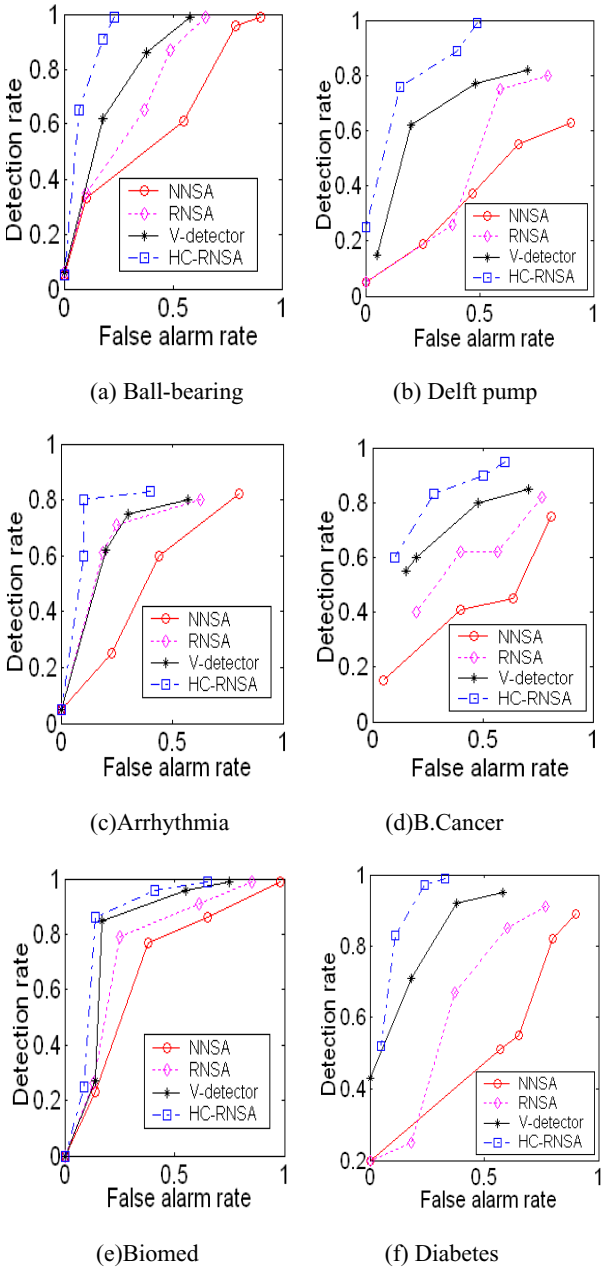
Table 3. Parameter set.

Parameter	Value
expect coverage	20%~100%
self radius	0.05~0.15
initial cluster radius	$\sqrt{d}$
PCA threshold	85%
distance norm	0.5

*d is the data dimension

The receiver operating characteristic curves (ROC) are generated by repeated experiments under different expected coverage. In Fig. 4, the horizontal axis represents the false alarm rate while the vertical axis represents the detection rate. Therefore the ideal curve is the vertical axis, which means the false alarm rate will always be zero with any detection rate.

Fig. 4 shows that the four NSAs get similar results on the data set Biomed; on the other data sets, they get much different results, and however, HC-RNSA always gets better results than others. Apparently, on Ball-bearing, Delft pump and B.Cancer, the detection results of HC-RNSA are much better than that of the traditional NSAs. Combining Table 2 and Fig. 4 we can see that, little self set size results in poor performance of the traditional NSAs. On the one hand, that is because the traditional NSAs rely on the self set to training detector candidates, thus the lack of self elements will result in the generation of self reactive detectors and poor performance. For HC-RNSA, the coverage of detectors is decided by the nearest self cluster margin. Therefore, the absence of some self elements will not affect the training of detectors. On the other hand, the data distribution of self data is not taken into consideration in the traditional NSAs. However, HC-RNSA discriminates the self and non-self regions by the self cluster ranges and generates mature detectors based on the discrimination, which reduced the false alarm rate.



(a) Ball-bearing (b) Delft pump

(c)Arrhythmia (d)B.Cancer

(e)Biomed (f) Diabetes

Fig. 4. The receiver operating characteristic curves

Table 4. The time cost of the detector generation process (*h*).

Dataset	V-detector	NNSA	RNSA	HC-RNSA
Ball-bearing	5.37	8.23	3.57	2.16
Delft pump	2.34	2.63	1.92	1.02
Arrhythmia	1.75	1.96	1.30	0.81
B.Cancer	2.33	2.3	1.93	0.96
Biomed	1.12	0.86	0.71	0.55
Diabetes	2.66	3.21	2.56	1.12

The time cost of the detector generation process is shown in Table 4. From the table we can see that the time cost of HC-RNSA on each data set is less than that of the traditional NSAs. As discussed in Sec. 3.2, during the detector generation process, the self data are replaced by self cluster centers to compare with the detector candidates. Usually, the number of cluster centers is much less than the self set size, so the efficiency of the detector generation process is much improved.

The HC-RNSA algorithm without PCA pretreatment stage is called HC_1 and the HC-RNSA using integer norm distance function is called HC_2. The detection result of HC-RNSA, HC_1 and HC_2 on the same UCI data sets is shown in Table 5.

As Table 5 shows, on the data sets: Ball-bearing, Delft pump and Arrhythmia, the detection rate of HC-RNSA is higher than that of HC_1 and HC_2 while its false alarm rate is lower. The result demonstrates that on the

higher dimensional data sets, the PCA pretreatment and fractional distance function is essential to the improvement of the algorithm performance. However, on the lower dimensional data sets: Biomed and Diabetes, HC_1 and HC_2 have better performance. On the one hand, that is because after the PCA pretreatment process, the data dimension is reduced, on the same time, some useful distinctiveness information is also lost; on the other hand, as discussed in Sec. 3.1.4 that D_{max}-D_{min} does not coverage in the lower dimensional space, so the fractional distance function is not needed. Therefore, when we chose negative selection algorithms to generate immune detectors, the self data distribution, data dimension, self set size, self radius and etc. must be taken into consideration.

Table 5. Detection result (%) of HC-RNSA, HC_1 and HC_2 under expected coverage 95%.

Dataset	HC-RNSA		HC_1		HC_2	
	DR	*FA*	*DR*	*FA*	*DR*	*FA*
Ball-bearing	81.7	15.1	66.1	19.3	70.2	16.7
Delft pump	75.5	17.4	65.4	23.4	52.3	30.3
Arrhythmia	72.6	23.3	59.3	27.9	51.9	26.4
B.Cancer	94.5	9.5	96.5	15.1	89.7	11.6
Biomed	86.9	10.3	97.7	3.6	87.2	5.9
Diabetes	87.1	7.6	89.1	8.5	90.6	6.3

5. Conclusions

Artificial immune theory is an intelligent soft computing technique which has the adaptive learning ability to study from the training data. But the application of the artificial immune system has been confined by the lower efficiency of the detector generation procedure. Therefore, a new negative selection algorithm HC-RNSA based on the hierarchical clustering of self set is proposed in this article. In HC-RNSA, the self data is replaced by the cluster centers to compare with the detector candidates to reduce the cost of distance calculations; the PCA method and fractional distance function are employed to improve the detection performance in the high dimensional space. The non-self coverage is estimated based on statistical inference method to dynamically terminate the detector generation procedure, which is more reasonable than the traditional exit conditions based on a given detector set size. The theory analysis and experimental results demonstrate that HC-RNSA is an effective algorithm to generate artificial immune detectors for anomaly detection.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No.60873246), Important Innovation Project Foundation of Higher Education of China (No. 708075) and Doctoral Fund of Ministry of Education of China (Grant No. 20070610032).

References

1. J. M. Shapiro, G. B. Lamont, and G. L. Peterson, An Evolutionary Algorithm to Generate HyperEllipsoid Detectors for Negative Selection, in: *Proceedings of the International Conference on Genetic and Evolutionary Computation*, (Washington, DC , 2005), pp.337-344.
2. S. Forrest, A. S. Perelson, L. Allen, and R. Cherukuri, Self-nonself discrimination in a computer, in: *Proceedings of the IEEE Symposium on Research in Security and Privac*, (Oakland, CA , 1994) pp.202–212.
3. X.C. Zhao, G.L. Liu, G.H. Zhao, and S.Z. Niu, A new clonal selection immune algorithm with perturbation guiding search and non-uniform hypermutation, *International Journal of Computational Intelligence Systems*, 3(1) (2010) pp.1-17.
4. X.Z. Gao, S.J. Ovaska, and X. Wang, Re-editing and censoring of detectors in negative selection algorithm, *International Journal of Computational Intelligence Systems*, 2(3) (2009) pp. 298-311.
5. K. Cengiz, E. Orhan, and Y.M. Kerim, A new artificial immune system algorithm for multiobjective fuzzy flow shop problems, *International Journal of Computational Intelligence Systems*, 2(3) (2009) pp. 236-247.
6. X. Wang, X.Z. Gao, and S.J. Ovaska, A hybrid artificial immune optimization method, *International Journal of Computational Intelligence Systems*, 2(3) (2009) pp. 249-256.
7. S.A. Hofmeyr, S. Forrest, and P. D'haeseleer, “An immunological approach to distributed network intrusion detection”, in: *First International Workshop on the Recent Advances in Intrusion Detection,* (Louvain-la-Neuve, Belgium, 1998) pp.210-215.
8. J.E. Hunt and D.E. Cooke, Learning using an artificial immune system, *Journal of Network and Computer Applications*, 19(1) (1996) pp.189-212.
9. T. Stibor, P. Mohr, and J. Timmis, Is negative selection appropriate for anomaly detection?, in: *Proceedings of*

the International Conference on Genetic and Evolutionary Computation, (Washington, DC , 2005), pp.498-505.
10. T. Li, An immunity based network security risk estimation, *Sci China Ser F-Inf Sci*, 48(5) (2005) pp.557-578
11. T. Li, *Computer immunology*, (Publishing House of Electronics Industry, Beijing, 2004).
12. F. Gonzalez and D. Dasgupta, Anomaly detection using real-valued negative selection, *Genetic Programming and Evolvable Machine*, 6 (12) (2003) pp.383-403.
13. F. Gonzalez, D. Dasgupta, and L.F. Nino, A randomized real-valued negative selection algorithm, in: *Proceedings of the 2nd International Conference on Artificial Immune Systems*, (Springer-Verlag, Edinburgh, 2003), pp.261-272.
14. J. Zhou and D. Dasgupta, Real-valued negative selection algorithm with variable-sized detectors, in: *Proceedings Genetic and Evolutionary Computation Conference*, (Washington, USA, 2004), pp. 287-29.
15. J. Zhou and D. Dasgupta, V-detector: An efficient negative selection algorithm with "probably adequate" detector coverage, *Information sciences*, 179 (2009) pp.1390-1406.
16. T. Stibor, J. Timmis, and C. Eckert, On the use of hyperspheres in artificial immune systems as antibody recognition regions, *Lecture Notes in Computer Science*, 4163 (2006) pp. 215-228
17. T. Stibor, J. Timmis, and C. Eckert, A comparative study of real-valued negative selection to statistical anomaly detection techniques, *Lecture Notes in Computer Science*, 3627(2005) pp. 262-275
18. E. Chavez and G. Navarro, Measuring the dimensionality of general metric spaces, Technical Report TR/DCC-00-1, (Department of Computer Science, University of Chile, 2000).
19. M. Skala, Measuring the difficulty of distance-based indexing, *Lecture Notes in Computer Science*, 3772 (2005) pp. 103-114.
20. R.V. Hogg and E.A, Tanis, Probability and Statistical Inference, sixth edn, (Prentice Hall, 2001).
21. C. C. Aggarwal, A. Hinneburg, and D. A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: *Proceedings of 8th International Conference on Database Theory*, (London, UK, 2001), pp. 420-434.
22. http://homepage.tudelft.nl/n9d04/occ/index.html