T-detector Maturation Algorithm with Overlap Rate

Jungan Chen, Wenxin Chen, Feng Liang
Electronic Information Department
Zhejiang Wanli University
No.8 South Qian Hu Road
Ningbo, Zhejiang, 315100, China
friendcen21@hotmail.com, yulong@zwu.edu.cn, liangf_hz@hotmail.com

Abstract: A parameter called overlap rate is proposed to control the number of valid detectors generated for a T-detector Maturation Algorithm. The achieved algorithm TMA-OR can reduce the number of detectors for abnormal detection. Experiment results show that TMA-OR is more effective than V-detector algorithms such as naive estimate and hypothesis testing method and can be applied on different data sets.

Key-Words: Artificial immune system, overlap rate, match range

1 Introduction

Nowadays, Artificial Immune System (AIS) is used to construct the algorithms based on negative selection, immune network model, or clonal selection[1][2][3].It is applied in many areas such as anomaly detection, classification, learning and control algorithm[4][5][6]. Negative Selection Algorithm (NSA) is first proposed to generate detectors which are applied to abnormal detection [1]. It has two phases. First, detector set is generated in fig.1. Random string becomes a detector when it does not match any self string. Second, unkown string is taken as nonself(abnormal) when it match any of detectors.

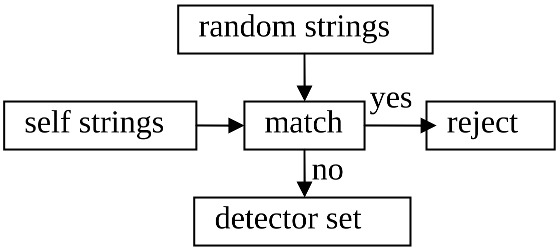


Fig.1 generation of detector set

In NSA, Match rule is one of the most important components and used to decide whether two string is matched. There are several major types [7] [8] [9]. But no matter what kind of match rule, the match threshold (r) is constant. A real-valued negative selection algorithm with variable-sized detectors (V-detector Algorithm) is proposed to generate detectors with variable r. A statistical method (naïve estimate) is used to estimate detect coverage [10]. The number of detectors and match threshold(r) is not required to be set manually. But as reported in Stiboret later work, the performance of V-detector on the KDD Cup 1999 data is unacceptably poor[11]. It is because that the detector coverage to be estimated is changing during the detector generation. So a new statistical approach (hypothesis testing) is used to analyze the detector coverage [12]. But hypothesis testing requires np>5, n(1-p)>5 and n>10. When p is set to 90%, n must be set to at least 50, which is not rational as the number of detectors affect the detect performance.

Actually there is another reason cause that naïve estimate method shows unacceptable result on the KDD data. In naïve estimate method, the candidate detector is added to valid detector set only when it is not detected by any of valid detectors, which means that the distance between candidate detector and any of valid detectors is bigger than the match threshold of the related valid detector. This process can maximize the distance among valid detectors. But it is difficult to find valid detector with the number of valid detectors increasing. At worst, tt is possible that there is no other valid detector generated after one valid detector is generated, which leads that naïve estimate method shows unacceptable result on the KDD data. So the distance among valid detectors chosen in naïve estimate method can affect the number of detectors generated.

To choose the appropriate distance among valid detectors and achieve the appropriate number of detectors generated, a parameter (overlap rate) is proposed in this paper to control the distance among detectors and impact on the number of valid detectors. The candidate detector can be added to valid detector set only when the overlap rate between it and any of valid detectors is smaller than specialized overlap rate *Omin*.

This work describes T-detector Maturation Algorithm with Overlap Rate (TMA-OR). The parameter (overlap rate) is used to achieve less number of detectors which lead to higher detect performance. In the algorithm, match range model is used to cover more detect area [13]. Statistical method naive estimate is used to estimate the coverage of detectors [10].

2 Related Work

2.1 Match Range Model

In human immune system, T-cells maturation goes through two processes, positive and negative selection [8]. Positive selection requires T-cells to recognize self cells with lower affinity, while T-cells must not recognize self cells with higher affinity in negative selection. So there is a range between lower and higher affinity.

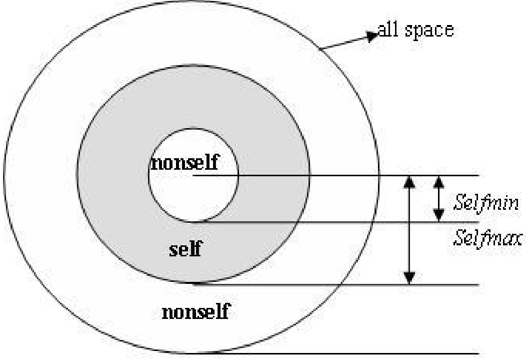


Fig.2 Match Range Model

Inspired from T-cells maturation, a match range model shown in Fig.2 is proposed [13]. *Selfmax* is the maximal distance between detector and selves. *Selfmin* is the minimal distance and must be bigger than 0. The range between *selfmin* and *selfmax* is belonged to the self space. When the distance is bigger than *selfmax* or smaller than *selfmin*, a nonself is detected.

In traditional NSA, the match threshold (r) is needed and must be set at first. To solve the problem, T-detector Maturation Algorithm (TMA) with match range model is proposed TMA can be adapted to the change of self data and cover more detect area [13]. But the parameter, number of detectors that affect the size of detect coverage and detect performance, is difficult to set manually. In this work, naïve estimate method is used to control the number of detectors.

2.2 Statistical method (naïve estimate) used to estimate the coverage

If m points are sampled in considered space and only one point is not covered, the estimated coverage would be 1-1/m. Therefore, when m times are randomly tried without finding an uncovered point, it can be concluded that the estimated coverage is at least α =1-1/m. Thus, the necessary number (m) of tries to ensure estimated coverage α is m=1/(1-α)[10]. The method is called naive estimate, which is used to estimate detect coverage In V-detector Algorithm.

3 Algorithm

3.1 Match Range Model

U={0，1}^{n} ,n is the number of dimensions. The normal set is defined as *selves* and abnormal set is defined as *nonselves*. *selves*∪*nonselves=U*. *selves*∩*nonselves=Φ*.There is two point $x=x_1x_2…x_n$, $y=y_1y_2…y_n$. The Euclidean distance between *x* and *y* is:

$$\mathrm{d(x,y)} = \sum_{i=1}^{n}\left(x_i - y_i\right)^2 \tag{1}$$

The detector is defined as *dct = {<center, selfmin, selfmax > | center ∈ U, selfmin, selfmax ∈N}*. *center* is one point in U. *selfmax* is the maximized distance between *dct.center* and *selves* and *selfmin* is the minimized distance. The detector set is definined as *DCTS*. *Selfmax* and *selfmin* is calculated by *setMatchRange(dct, selves), dct.center∈U, i ∈*[1, |*selves*|], *self$_i$ ∈selfves*：

$$setMatchRange = \begin{cases} \mathrm{selfmin = min(\{d(self_i, dct.center)\})} \\ \mathrm{selfmax = max(\{d(Self_i, dct.center)\})} \end{cases} \tag{2}$$

[*selfmin,selfmax*] is defined as self area. Others is as nonself area. Suppose there is one point *x*∈U and one detector *dct ∈ DCTS*. When *d(x,dct)* ∉ [*dct.selfmin*, *dct.selfmax*], *x* is detected as abnormal. So one rule called Range Match Rule (RMR), *RMRMatch(x,dct)* shown in equation 3, is proposed. In equation 3, value 1 means that x is abnormal.

$$RMRMatch = \begin{cases} \mathrm{0, d(x,dct.center) \in [dct.seflmin, dct.seflmax]} \\ 1, d(x, dct.center) \notin [dct.sefl\,\mathrm{min}, dct.sefl\,\mathrm{max}] \end{cases} \tag{3}$$

Based on RMR, the detect procedure *detect(x,DCTS)* is defined as equation 4, true means that x is abnormal.

$$Detect = \begin{cases} true, RMRMatch(x, dct_k) = 1, \exists dct_k \in DCTS \\ \quad false, \qquad\qquad others \end{cases} \tag{4}$$

3.2 Overlap Rate

With different statistical method, different v-detector algorithm is proposed. Naive estimate is first proposed [10]. But as reported in Stiboret later work, the performance of V-detector on the KDD Cup 1999 data is unacceptably poor[11].

Actually, in naive estimate method, there is one reason which leads to poor performance of naive estimate on the KDD data. In V-detector algorithm, candidate detector is added to valid detector set only when it is not covered by the entire existed valid detector [10] [12]. With the number of detectors increase, it becomes more difficult to find such valid detector on some application area. To overcome this shortage, Overlap Rate (o) is proposed in the paper. Equation 5 can reflect the overlap area between two rings. R and r are especially the value of *selfmin* of candidate detector and valid detector. When d becomes small, the overlap area becomes large in Fig.3.

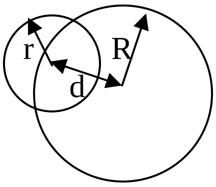


Fig.3 Overlap Rate Model

$$o = 1 - \frac{d}{R + r} \tag{5}$$

In the algorithm, one parameter *Omin* is used to control the overlap rate. When the overlap rate between candidate detector (*dctx*) and any of the detectors is bigger than *Omin*, candidate detector is not added to the valid detector set. The procedure *isvalid (dctx,DCTS)* is shown in equation 6:

$$IsValid = \begin{cases} false, OverlapMatch(dctx, dct_k) = true, \exists dct_k \in DCTS \\ true, \quad others \end{cases} \tag{6}$$

$$OverlapMat\,ch = \begin{cases} true, 1 - \dfrac{d}{dctx.self\ \min + dctk.self\ \min} > O_{\min}, d < dct.self\ \min \\ false, \quad others \end{cases} \tag{7}$$

$$d = d(dctx.center, dctk.center) \tag{8}$$

3.3 The model of algorithm

In this algorithm shown bellow, p_c is defined as the desired coverage in naïve estimate method and used to control the number of detectors. r_s is defined as the radius of self and used to control the false alarm rate.

```
1.  Set the desired coverage $p_c$, Self radius $r_s$, Omin
2.  Generate one candidate detector dctx randomly
/* set the properties of candidate detector including selfmax and selfmin according equation 2*/
3.  setMatchRange(dctx,selves)
/* candidate detector should not covered by self with $r_s$ or it will be dicard.*/
4.  if dctx.selfmin< $r_s$ then Go to 2;
/* judge whether candidate detector is a valid detector according equation 6. */
5.  if isvalid(dctx,DCTS) then
6.          dctx is added to detector set DCTS
/* covered is used to count the number of valid detectors which is covered by the candidate detector.*/
7.                  covered=0
8.  Else
9.                  covered ++
/* estimate the detect coverage according naïve estimate method */
10. If covered <1/(1- $p_c$) then goto 2
```

4 Experiments

The objective of the experiments is to:

1. Compare between TMA-OR and V-detector algorithm including both naive estimate and hypothesis testing.
2. Investigate the effect of *Omin*, the match range model and r_s.
3. Investigate the performance's of TMA-OR

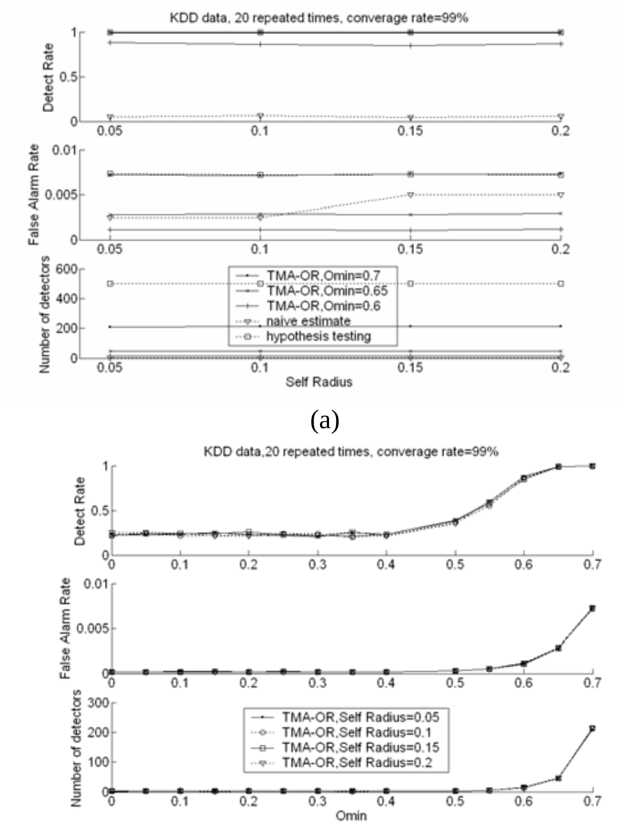
For the purpose of comparison, experiments are carried out using every data set list in table 1.

In table 1, 2-dimensional synthetic data(shown in appendix) is described in Zhou's paper[14]. Over the unit square $[0,1]^2$,various shapes are used as the self region. In every shape, there are training data (self data) of 1000 points and test data of 1000 points including both self points and nonself points. In the famous benchmark Fisher's Iris Data, one of the three types of iris is considered as normal data, while the other two are considered abnormal [10]. As for KDD data, 20 subsets were extracted from the enormous KDD data using a process described in [11]. Self radius from 0.01 up to 0.2 and *Omin* used in TMA-OR from 0 up to 0.7is conducted in these experiment. All the results shown in these figures are average of 100 or 20 (see table 1) repeated experiment with coverage rate 99%.

Table.1 data set and parameters used in experiments

Data set		Parameters		
		r_s	*Omin*	Repeated times
2-dimensional synthetic data	Comb	0.01 ~ 0.2	0 ~ 0.7	100
	Cross			
	Ring			
	Triangle			
	Stripe			
	Intersection			
	Pentagram			
Iris data	Setosa as self data			
	Versicolor as self data			
	Virginica as self data			
KDD data		0.05~0.2		20

4.1 Comparison



(a)

(b)

Fig.4 average results on 20 subsets of KDD data

Fig.4 shows the results of V-detector algorithm including naïve estimate and hypothesis testing method, TMA-OR with different *Omin*. Naïve estimate method has low detect rate in the first figure of Fig.4(a) because it generates less detectors in the third figure of Fig.4(a). It can be concluded that naïve estimate method is difficult to find valid detectors which can not be detected by any of other detectors. Also it is proved that the performance of V-detector on the KDD Cup 1999 data is unacceptably poor[11].Hypothesis testing method and TMA-OR can achieve high detect rate in the first figure of Fig.4(a). To achieve the same detect rate in the third figure of Fig.4(a), TMA-OR generates about 200 detectors. But hypothesis testing method generates about 500 detectors as it requires np>5, n(1-p)>5 and n>10. When p is set to 99%, the number of detectors n must be set to at least 500. Fig.4 (b) shows that the parameter (overlap rate) *Omin* can control the number of valid detector in the third figure and affect the detect rate in the first figure. When *Omin* increased, the overlap area among detectors is permitted to become larger. So valid detector is easy to find and the number of valid detectors increases with *Omin*.

Table.2 results on Iris data set (rs=0.03)

Data Set	Algorithm	Detect Rate	False Alarm Rate	Number of Detectors
Setosa as self data	Hypothesis Testing	1.000	0.000	500.840
	TMA-OR(Omin=0.7)	1.000	0.000	77.350
	TMA-OR(Omin=0)	1.000	0.000	14.350
	Naïve Estimate	1.000	0.000	14.390
Versicolor as self data	Hypothesis Testing	0.998	0.000	500.000
	TMA-OR(Omin=0.7)	0.998	0.000	153.930
	TMA-OR(Omin=0)	0.975	0.000	26.410
	Naïve Estimate	0.961	0.000	26.270
Virginica as self data	Hypothesis Testing	0.996	0.000	501.120
	TMA-OR(Omin=0.7)	0.997	0.000	195.340
	TMA-OR(Omin=0)	0.986	0.000	33.770
	Naïve Estimate	0.985	0.000	34.520

Table 2,3 shows that TMA-OR (*Omin*=0). has the same performance with naïve estimate method. So does between hypothesis testing method and TMA-OR(*Omin*=0.7). It is concluded that TMA-OR with different *Omin* can achieve the same performance with both naïve estimate and hypothesis testing method. Through comparison between TMA-OR(*Omin*=0) and naïve estimate method(in bold font), it is illustrated that TMA-OR achieve bigger detect rate with less number of detectors. This demonstrates the effect of match range mode. i.e., V-detector detects nonself only when the distance between detect and nonself is smaller than match threshold. However, detector with match range model detects nonself when the distance is smaller than *selfmin* or bigger than *selfmax*.Other result is given in appendix.

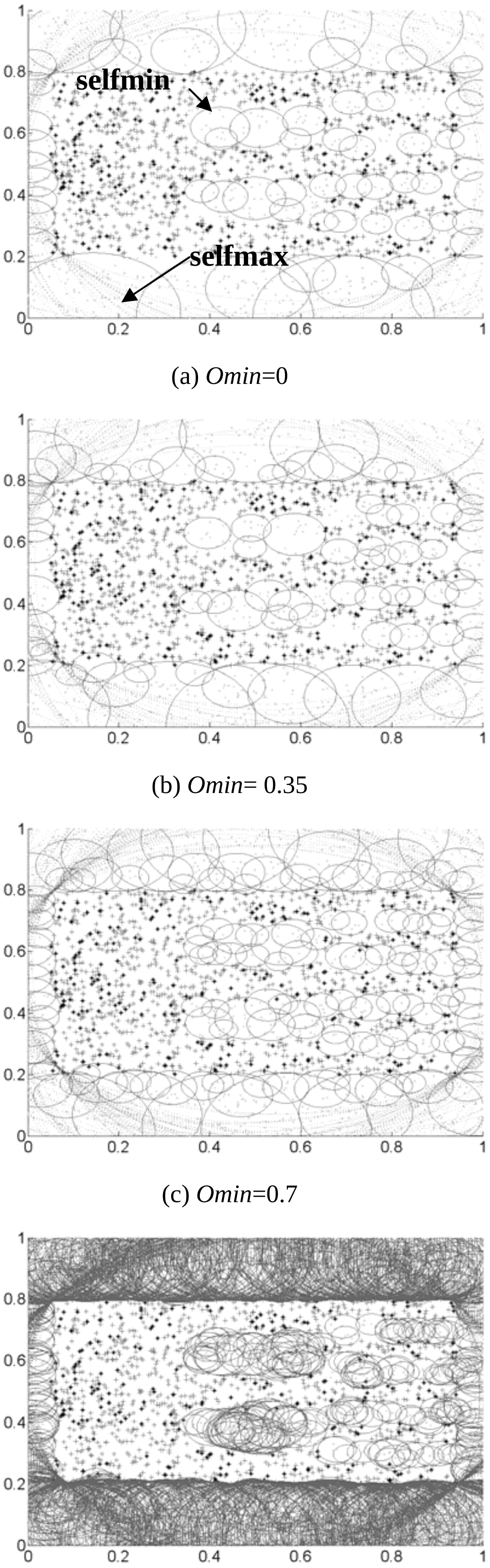
Table.3 results on 2-dimensional synthetic data set (rs=0.03)

Data Set	Algorithm	Detect Rate	False Alarm Rate	Number of Detectors
Comb	Hypothesis Testing	0.972	0.124	509.730
	TMA-OR(Omin=0.7)	0.971	0.119	143.560
	TMA-OR(Omin=0)	**0.925**	**0.088**	**71.650**
	Naïve Estimate	**0.921**	**0.082**	**71.900**
Cross	Hypothesis Testing	0.998	0.083	503.730
	TMA-OR(Omin=0.7)	0.995	0.076	73.800
	TMA-OR(Omin=0)	**0.967**	**0.046**	**27.480**
	Naïve Estimate	**0.966**	**0.043**	**29.140**
Ring	Hypothesis Testing	1.000	0.130	500.000
	TMA-OR(Omin=0.7)	0.998	0.107	82.530
	TMA-OR(Omin=0)	**0.983**	**0.064**	**26.930**
	Naïve Estimate	**0.962**	**0.045**	**27.480**
Triangle	Hypothesis Testing	1.000	0.031	500.000
	TMA-OR(Omin=0.7)	0.999	0.022	52.330
	TMA-OR(Omin=0)	0.981	0.008	14.510
	Naïve Estimate	0.977	0.008	14.220
Stripe	Hypothesis Testing	1.000	0.048	500.000
	TMA-OR(Omin=0.7)	0.999	0.039	55.300
	TMA-OR(Omin=0)	**0.973**	**0.017**	**14.560**
	Naïve Estimate	**0.970**	**0.013**	**14.940**
Intersection	Hypothesis Testing	0.999	0.117	500.000
	TMA-OR(Omin=0.7)	0.996	0.091	98.990
	TMA-OR(Omin=0)	0.976	0.054	36.940
	Naïve Estimate	0.964	0.044	36.660
Pentagram	Hypothesis Testing	0.998	0.023	500.000
	TMA-OR(Omin=0.7)	0.996	0.017	68.230
	TMA-OR(Omin=0)	**0.969**	**0.009**	**22.890**
	Naïve Estimate	**0.967**	**0.009**	**23.760**

4.2 The effect of Omin, Match Range Model and rs

In Fig.5, dot point is as nonself test data, * point is as self test data and plus point is self training samples. The inside of solid ring is the coverage of detectors with *selfmin*. The outside of dotted ring is the coverage of detectors with *selfmax*. It is concluded that detectors can detect the main outline of outside of self area with *selfmax* and the nonself area embed in self area with *selfmin*.

Fig.5 shows that the number of detectors increases with *Omin* because the distance among detectors reduced. So the parameter (overlap rate) *Omin* can control the distance among detectors and impact on the number of valid detectors.



(a) *Omin*=0

(b) *Omin*= 0.35

(c) *Omin*=0.7

(d) *Omin*=1.0

Fig.5 results on Comb data set with different Omin (rs=0.03)

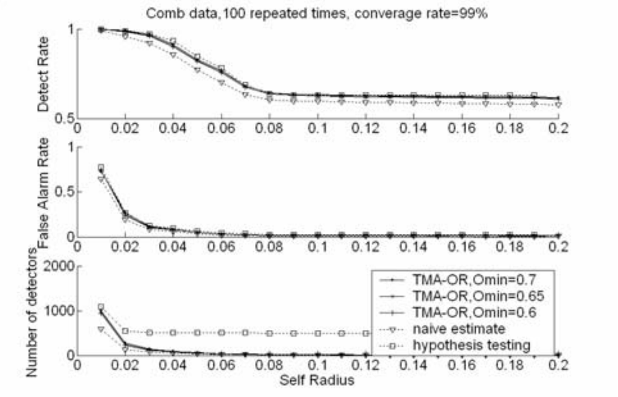


Fig.6 results on Comb data set with different rs

The second figure of Fig.6 shows that self radius rs is used to control the false alarm rate. When rs becomes smaller, self is easier to be detected as nonself by valid detector generated Step.4 in the model of algorithm.

4.3 Performance of TMA-OR

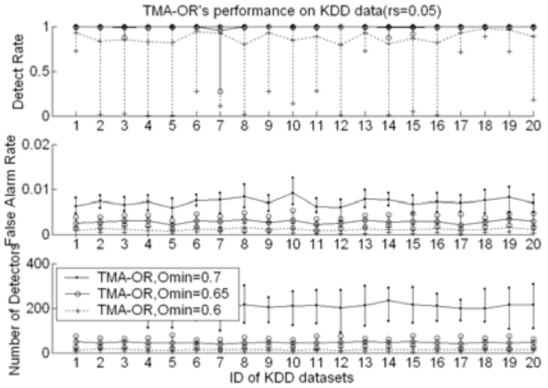


Fig.7 TMA-OR's performance on KDD data sets

Fig.7 shows that TMA-OR achieves the best result with Omin=0.7. Detect rate changes dramatically when Omin is set to 0.6 and 0.65. So Omin is a key parameter.

5 Conclusion

This work proposed T-detector Maturation Algorithm with Overlap Rate (TMA-OR). The parameter (overlap rate) *Omin* can control the distance among detectors and impact on the number of valid detectors generated. So TMA-OR with different *Omin* can achieve less number of detectors which lead to higher detect performance. Furthermore, match range model is used to cover more detect area and detectors can detect the main outline of outside of self area with *selfmax* and the nonself area embed in self area with *selfmin*. Experiment results conclude that TMA-OR is more effective than V-detector algorithm such as naïve estimate and hypothesis testing method and can be applied on different data set.In TMA-OR, Omin is an important parameter and affect the detect rate. Rs is used to control the False Alarm Rate.

Of course, there are many aspects that are worth further research. The overlap rate model is required to be improved in order to reflect the distance among detectors well and truly.

Acknowledgement

This work is supported by Ningbo Nature Science Foundation 200701A6301043 and Scientific Research Fund of Zhejiang Provincial Education Department 20070731. Thanks for the assistance received by using KDD Cup 1999 data set [http://kdd.ics.uci.edu/databases / kddcup99/ kddcup99.html] , the 2-dimensional synthetic data set [https:// umdrive.memphis.edu/ zhouji/ www/ vdetector.html].

References:

[1]Forrest S., Perelson A. S., Allen L. and Cherukuri, R., Self-nonself Discrimination in a Computer, Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy, 1994.

[2]de Castro L. N. and Von Zuben F. J., aiNet: An Artificial Immune Network for Data Analysis, Book Chapter in Data Mining: A Heuristic Approach, H. A. Abbass, R. A. Sarker, and C. S. Newton (eds.), Idea Group Publishing, USA, Chapter XII, pp. 231-259, 2001

[3]de Castro L. N. and Von Zuben F. J., Learning and Optimization Using the Clonal Selection Principle, IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems, 6(3), pp. 239-251. 2002

[4]T.-C. Chen, C.-Y Chen, Satellite-Derived Land-Cover Classification Using Immune Based Mining Approach, 5th WSEAS International Conference on APPLIED COMPUTER SCIENCE, 2006.

[5]Lee, V. C. S. and Yang, XingJian, An Artificial Immune System Based Learning Algorithm for Abnormal or Fraudulence Detection in Data Stream, in Proceedings of WSEAS International Conference on Applied Computing, ACOS'2006

[6]T.L. Huang, K.T. Lee , C.H. Chang and T.Y. Hwang, Two-Level Sliding Mode Controller Using Artificial Immue Algorithm, WSEAS Transcations on Power Systems, 2006

[7]Hofmeyr S. A., An Immunological Model of Distributed Detection and its Application to Computer Security, PhD Dissertation, University of New Mexico, 1999.

[8]Gonzalez F., A Study of Artificial Immune Systems applied to Anomaly Detection, PhD Dissertation, The University of Memphis, May 2003.

[9]Zhou Ji, Dipankar Dasgupta, Revisiting Negative Selection Algorithms, Evolutionary Computation, 2007

[10]Zhou Ji, Dipankar Dasgupta, Real-valued Negative Selection Algorithm with Variable-Sized Detectors, Genetic and Evolutionary Computation Conference (GECCO), 2004

[11]Stibor, T., Timmis, J. and Eckert, C. A Comparative Study of Real-Valued Negative Selection to Statistical Anomaly Detection Techniques, ICARIS 2005,2005

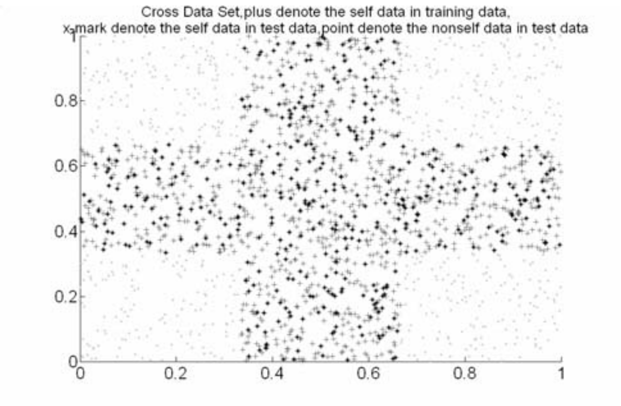
[12]Zhou Ji, Dipankar Dasgupta, Estimating the Detector Coverage in a Negative Selection Algorithm, Genetic and Evolutionary Computation Conference (GECCO), 2005

[13]Jungan Chen, T-detectors Maturation Algorithm with Min- Match Range Model,the 3rd IEEE International Conference on Intelligent System, 2006

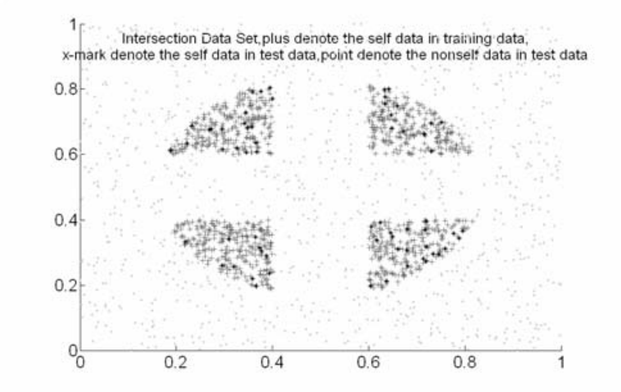
[14]Zhou Ji, Negative Selection Algorithms: from the Thymus to V-detector. PhD Dissertation, University of Memphis, 2006.

Appendix

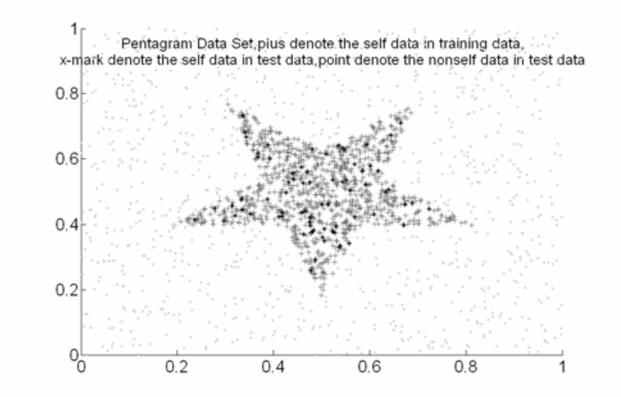
1 2-dimensional synthetic data



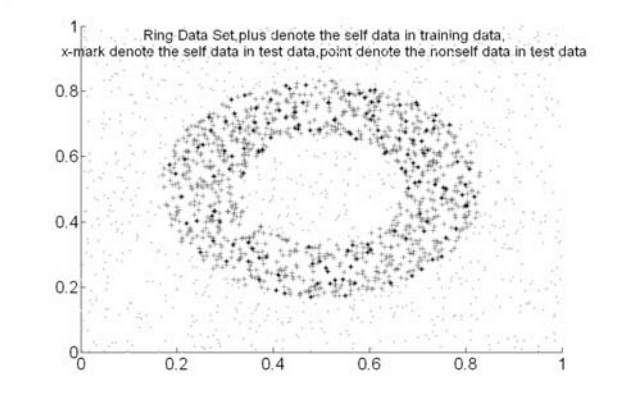
(a) Cross



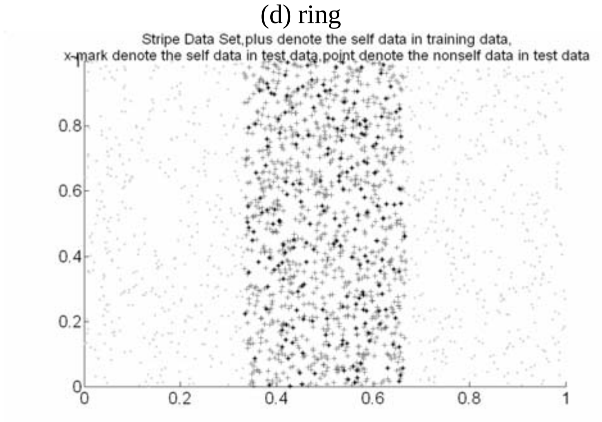
(b) Intersection



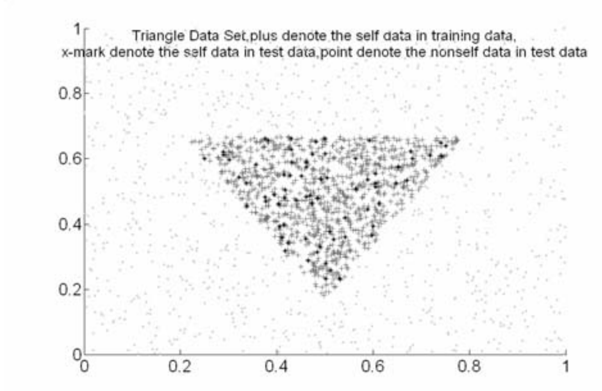
(c) Pentagram



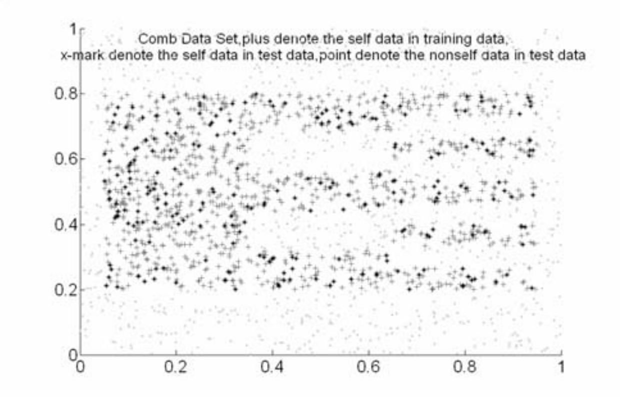
(d) ring



(e) stripe



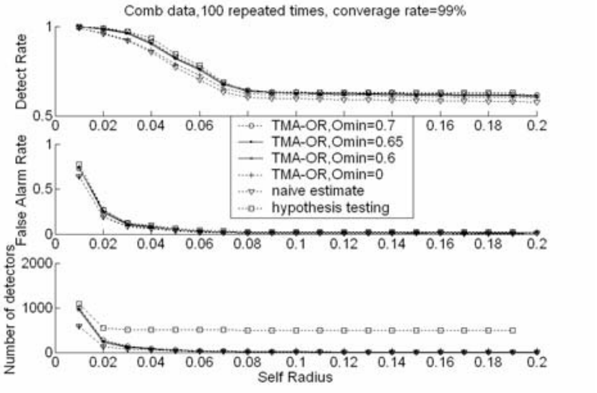
(f) Triangle



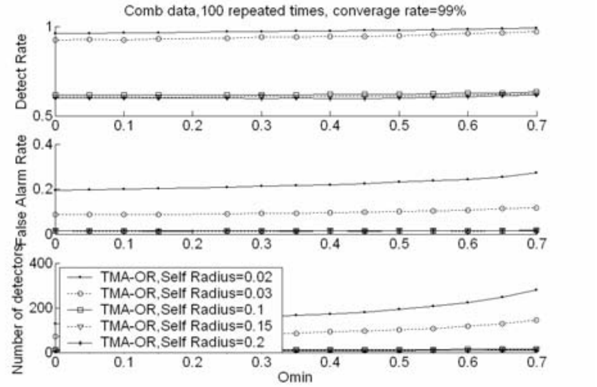
(g) Comb

Fig.8 2-dimensional synthetic data

2 results on 2-dimensional synthetic data set

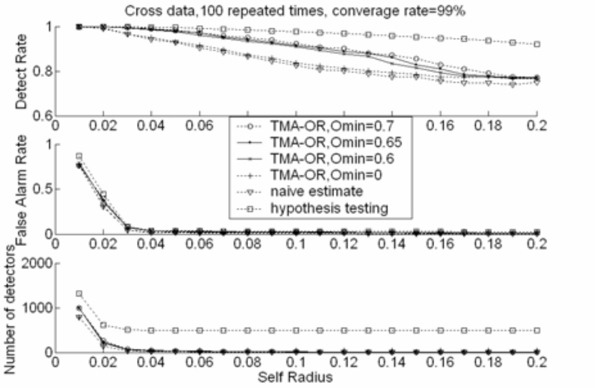


(a)

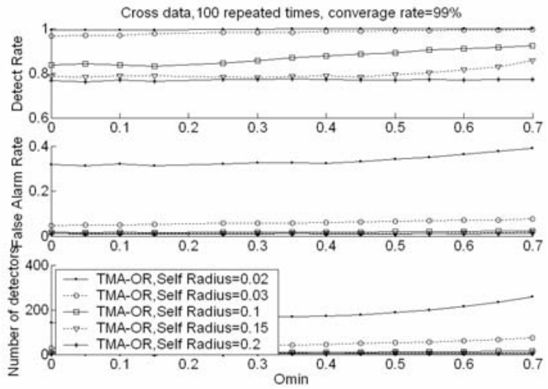


(b)

Fig.9 Results on Comb

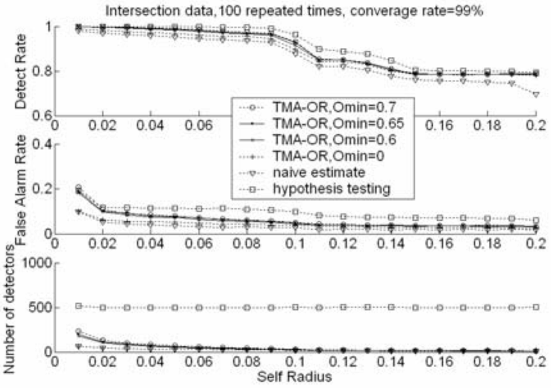


(a)

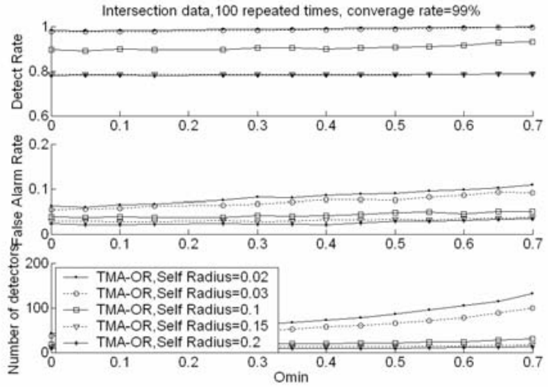


(b)

Fig.10 Results on Cross

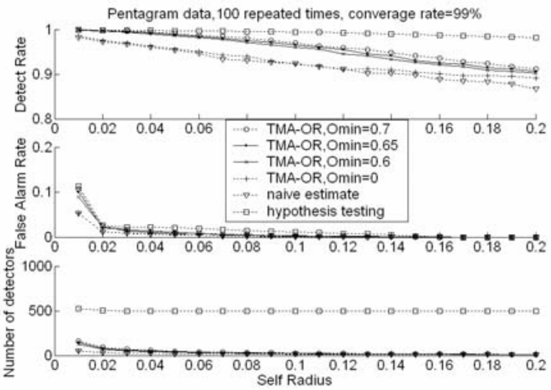


(a)

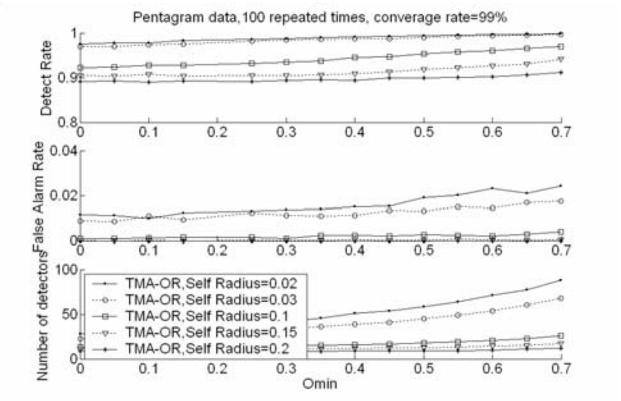


(b)

Fig.11 Results on Intersection

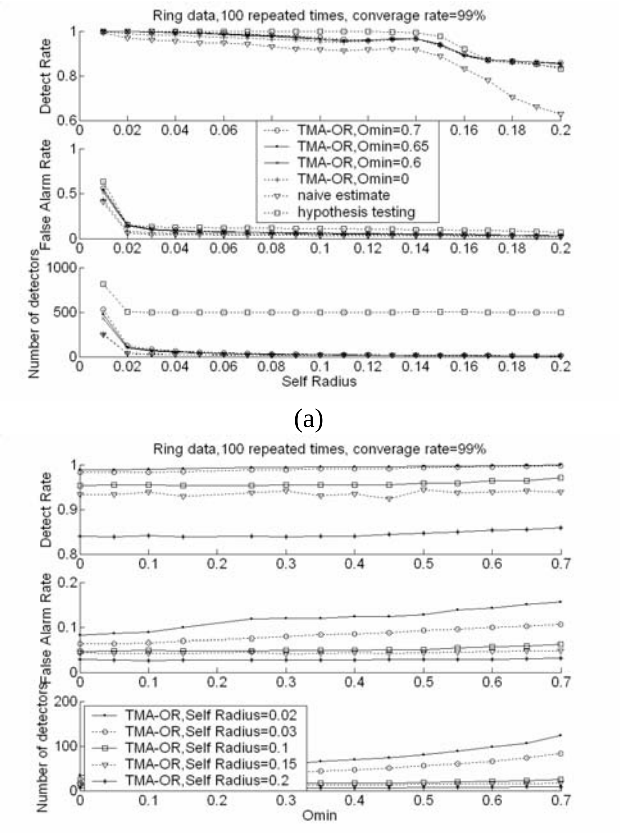


(a)



(b)

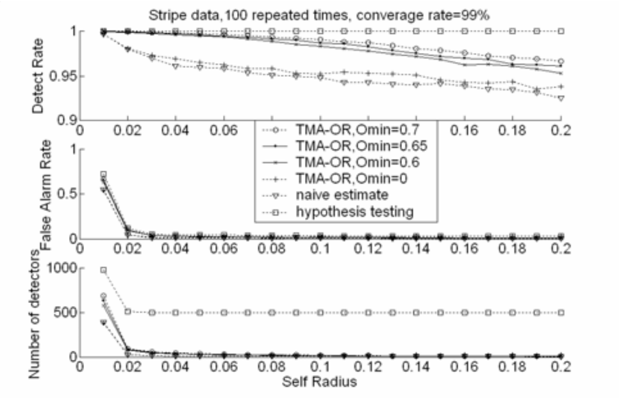
Fig.12 Results on Pentagram



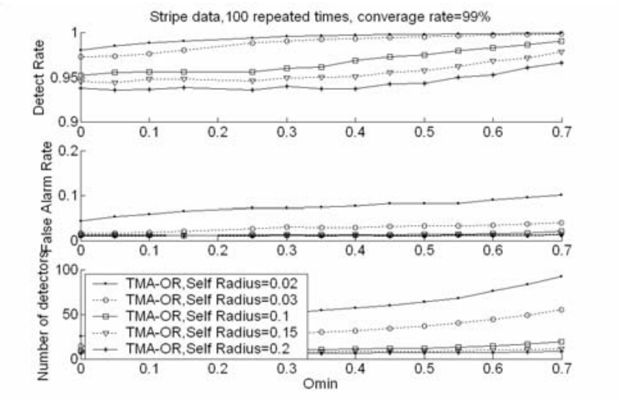
(a)

(b)

Fig.13 Results on Ring

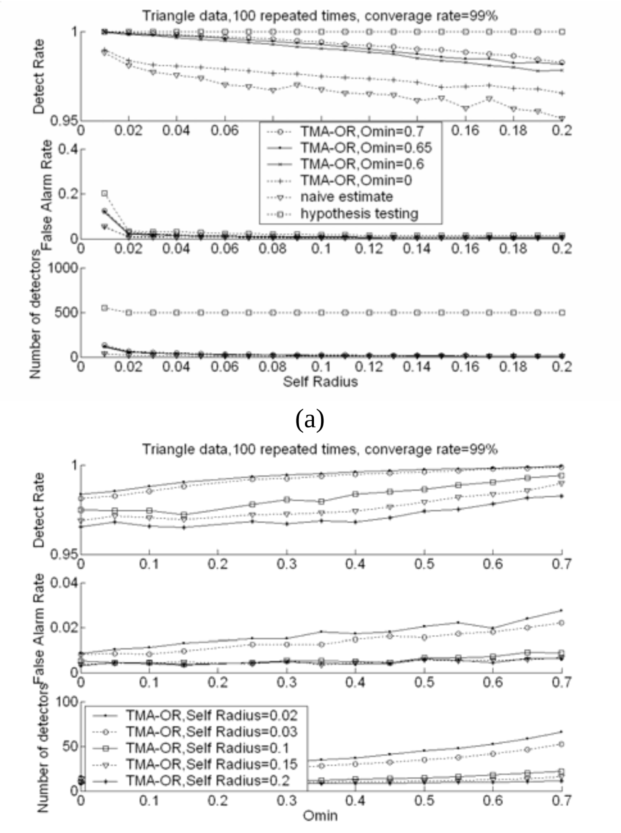


(a)



(b)

Fig.14 Results on Stripe



(a)

(b)

Fig.15 Results on Triangle