

Augmented Negative Selection Algorithm with Variable-Coverage Detectors

Zhou Ji

St. Jude Children's Research Hospital
Memphis, TN 38105
Email: zhou.ji@stjude.org

Dipankar Dasgupta

The University of Memphis
Memphis, TN 38152
Email: ddasgupt@memphis.edu

Abstract - An augmentation of negative selection algorithm is developed featuring detectors that have variable coverage. While the detectors can have different kinds of variable properties in the light of this concept, the paper mainly describes the experiments of variable-sized detectors in real-valued space. Effects of the two main control parameters, self radius and expected coverage, are discussed and experimented with both synthesized and real-word datasets. The new approach improves efficiency and reliability without compromising the order of magnitude of complexity.

I. INTRODUCTION

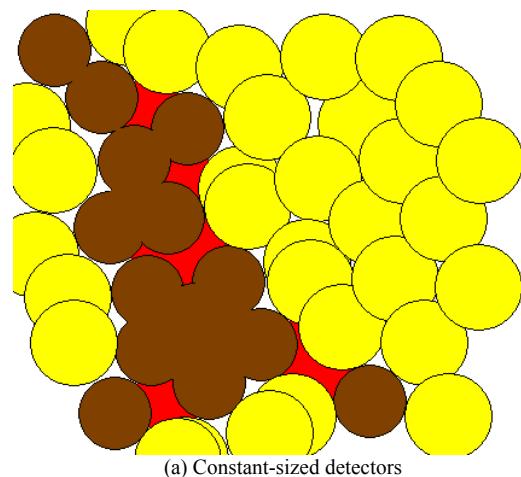
As one of the relatively new areas of soft computing, Artificial Immune Systems (AIS) generally construct the algorithms based on negative selection, immune network model, or clonal selection [1][2][12][14]. The inspiration of negative selection, or negative detection, came from the T cell development process in the thymus. If a T cell recognizes self cells, it is eliminated before deployment for immune functionality [16]. In an analogous manner, the negative selection algorithm generates the detector set by eliminating any detector candidates that match self samples. It is thus used as an anomaly detection mechanism with the advantage that only the negative (or 'normal') training data are needed [3].

Researches in negative selection usually discuss the problems in binary representation [6][15]. There are at least two good reasons to support this choice: first, binary representation provides a finite problem space that is easier to analyze mathematically; second, binary presentation is convenient to use for categorized data. However, many applications are natural to be described in real-valued space. Furthermore, these problems can hardly be processed properly using negative selection algorithm in binary representation [4]. On the other hand, this paper and some other works [5][10] demonstrated that despite the intrinsic difficulty of real-valued problem space, it also provides unique opportunity.

Matching rule is one of the most important components in a negative or positive selection algorithm [4][6][11][13][15]. For binary representation, there are several major types of matching rules like rcb (r-contiguous bits), r-chunks, and Hamming distance [6][4]. For real-valued representation, matching rules are generally one way or another based on Euclidean distance between the data to be tested and the detectors [4][5][7][10]. Matching is usually defined as a distance that is within a certain threshold. In some cases, it may be

a variation of Euclidean distance. For example, [7] used a Euclidean distance defined in a lower dimensional space, which was projected from the original problem space by contiguous- or random-chosen dimensions.

No matter what kind of matching rule is used, the detectors's basic characteristics are usually constant, e.g., the number of bits r in binary representation, or the distance threshold to decide a matching in real-valued representation. In the latter case, the detectors are in fact hyper-sphere-shaped although it is adequate to represent them as points. The threshold is actually the radius of the detectors. However, the radius or other basic properties used in matching rule doesn't have to be constant for all the detectors. The algorithm introduced in this paper is an attempt to demonstrate that allowing the detectors to have some variable properties will enhance the negative selection algorithm from several aspects. We call this idea and the algorithms based on it V-detector. In real-valued application using Euclidean distance matching rule, the radius of detectors is an obvious choice to make variable considering that the non-self regions to be covered by detectors are very likely to be in different scales. The flexibility brought by variable radius is easy to see. Furthermore, variable radius is not the only possibility provided by V-detector. The shapes of detectors or even the types of matching rules can be extended to be variable too to augment negative selection algorithm.



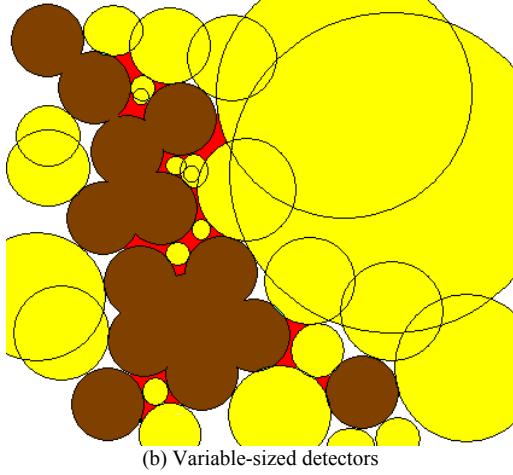


Figure 1. Comparison of constant-sized detector and variable-sized detectors (V-detector) in negative selection algorithm

The diagrams in figure 1 illustrate the core idea of variable detectors in 2-dimensional space. The dark area shows the assumed self regions, which usually are represented by the training data (self samples). The light-shaded circles are the possible detectors. Figure 1(a) demonstrates the situation when the detectors are constant-sized in real-value space. On the right side of the diagram, a large number of detectors are needed to cover the large area of non-self space. The well-known issues of “holes” are illustrated in medium shade. In figure 1(b), variable-sized detectors deal with both issues. The large area of non-self space is now covered by much fewer detectors. At the same time, smaller detectors work much better to cover the holes. Since the number of detectors is saved on one hand, it becomes more feasible to use smaller detectors when necessary.

Another advantage of this new method is that estimated coverage instead of the number of detectors is used as a control parameter. It estimates automatically when the detector set is generated. Comparatively, we need to set the number of detectors in advance when constant detectors are used. This will be discussed in more details in the following sections.

II. ALGORITHM AND ANALYSIS

To illustrate the new feature, let us first describe a negative selection algorithm of “constant detectors” in real-valued space. Using Euclidean distance matching rule, generation algorithm of “constant detectors” is shown in figure 2. The time complexity of this algorithm is $O(m|S|)$, where m is the preset number of detectors and $|S|$ is the size of training set (the set of self samples). Self radius r_s in this case is the same as detector radius, which is the allowed variability of the self space [10].

Detector - Set(S, m, r_s)

S : set of self samples

m : number of detectors

r_s : self radius

1: $D \leftarrow \emptyset$

2: Repeat

3: $x \leftarrow$ random sample from $[1, 0]^n$

4: Repeat for every s_i in $S = \{s_i, i = 1, 2, \dots\}$

5: $d \leftarrow$ Euclidean distance between s_i and x

6: if $d \leq r_s$, go to 2

7: $D \leftarrow D \cup \{x\}$

8: Until $|D| = m$

9: return D

Figure 2 Algorithm of detector generation using constant detectors

V-detector algorithm’s detector generation phase is shown in figure 3. Comparing with the constant-sized version, the most important difference lies in steps 13 through 15. Now that we let each detector have its own radius in addition to location, the radius is basically decided by the closest self sample. Self radius still specifies the variability represented by the training data, but it is not used as detector radius anymore.

The algorithms of detection phase are similar for constant and variable detectors except that matching threshold for each variable-sized detector is unique. In the experiments of this paper, matching is decided by any of generated detectors.

The control parameters of V-detector are mainly self radius r_s and expected coverage c_0 . Maximum number of detectors, shown as T_{max} in figure 3, is preset to the largest number that we are willing to tolerate in practice, which does not need much further discussion. Self radius is an important mechanism to balance between detection rate and false alarm rate, in the other words, the sensitivity and accuracy of the system.

Expected coverage is a by-product of variable detectors. If we sample m points in the considered space and only one point is not covered, the expected coverage would be $1-1/m$. Therefore, when we randomly try m times without finding an uncovered point, we can conclude that the expected coverage is at least $\alpha=1-1/m$. Thus, the necessary number of tries to ensure expected coverage α is

$$m = 1/(1-\alpha) \quad (1)$$

```

V - Detector - Set(S, Tmax, rs, co)
S : set of self samples
Tmax : maximum number of detector
rs : self radius
c0 : expected coverage

1: D  $\leftarrow \emptyset$ 
2: Repeat
3:   t  $\leftarrow 0$ 
4:   T  $\leftarrow 0$ 
5:   r  $\leftarrow$  infinite
6:   x  $\leftarrow$  random sample from  $[1, 0]^n$ 
7:   Repeat for every  $d_i$  in D = { $d_i$ .i = 1, 2, ...}
8:      $d_d \leftarrow$  Euclidean distance between
      x( $d_i$ ) and x, where x( $d_i$ ) is the location of  $d_i$ 
9:     if  $d_d \leq r(d_i)$  then, where r( $d_i$ ) is the radius
       of detector  $d_i$ 
10:    t  $\leftarrow t + 1$ 
11:    if  $t \geq 1/(1 - c_0)$  then return D
12:    go to 4:
13:   Repeat for every  $s_i$  in S
14:     d  $\leftarrow$  Euclidean distance between  $s_i$  and x
15:     if  $d - r_s \leq r$  then r  $\leftarrow d - r_s$  :
16:     if r  $> r_s$  then D  $\leftarrow D \cup \{<x, r>\}$ , where <x, r>
       is a detector with location x and radius r
17:   else T  $\leftarrow T + 1$ 
18:   if T  $> 1/(1 - \text{maximum self coverage})$  exit
19: Until |D|  $\leq T_{max}$ 
20: return D

```

Figure 3 Algorithm of Variable-sized Detector Generation (V-detector).

Despite the enhancement, complexity of detector generation in V-detector algorithm is not increased comparing with the algorithm using constant-sized detectors. The computation of radius has linear complexity with respect to the number of the training samples. In figure 3, steps 13 through 15 has complexity $O(|S|)$, where $|S|$ is the size of training set. That is the same as steps 4 through 6 in the constant detector version (figure 2). Furthermore, not only are the orders of

magnitude of complexity in the two algorithms the same, but the times to actually compute the distance, which is potentially a costly step, is the same as well. If the final number of detectors is m , the complexity to generate the entire detector set is $O(m|S|)$. If m is the same order of magnitude as the preset number in constant version, the complexity doesn't change; if m is reduced significantly, the complexity is in fact improved.

After the detector sets are generated, the detection phase of the algorithms using constant- or variable-sized detectors is similar to each other. The complexity of detection phase is $O(m)$ although m has different interpretation in the two methods. The difference in the size of needed memory also only lies in the possible different m .

The V-detector algorithm normally converges in one of the two ways. Type 1 convergence happens when the expected coverage is reached (step 11 in figure 3). This is the scenario that V-detector shows more of its strength in controlling detector number. Type 2 convergence occurs when the pre-defined maximum number of detectors is reached (step 19). Even in this case, the algorithm still has the potential to cover the holes better than algorithm using constant detectors. As an exception, the algorithm may also terminate when it fails to sample any non-self point after many repetitions (step 18). That implies that the self region covers almost the entire space. It may happen when the self samples are randomly distributed over the space, or the chosen self-radius is too big.

In V-detector, the small "holes" are easier to be covered not by just using smaller detectors, rather by using the automatic decision of how small the detectors need to be. The total numbers of detectors, on the other hand, are regulated by using larger detectors whenever possible.

Described above is the framework of the V-detector algorithm. However, there is still room for improvement in this model. For example, we will discuss the issue of "boundary dilemma" in the following section of experiments.

III. EXPERIMENTS AND RESULTS

To demonstrate the basic behavior of V-detector algorithm, synthesized datasets are used. Then, the benchmark Fisher's Iris data and more real-world datasets are used to further examine its performance and compare with other methods.

Figure 4 shows a 2-dimensional data where a cross-shaped area on the unit square $[0, 1]^2$ is assumed to be the normal or self region. Figure 4(a) is the shape of the self region. The training set is 100 randomly picked points in the self region and the test data are 1000 randomly distributed points over the entire square. Figure 4(b) and 4(c) show the area actually covered by the generated detector set. The dark area is where the data will be claimed abnormal. Comparing the coverage in (b) and (c),

it is easy to see the effect of self radius on the results. The smaller self radius would result in high detection rate but high false alarm too, so it is suitable for the scenario when detecting all or most abnormalities is very important. On the other hand, larger self radius would result in low detection

rate and low false alarm, thus suitable when we need to try the best to avoid false alarm. Figure 5 demonstrates similar situation when the self region is ring-shaped.

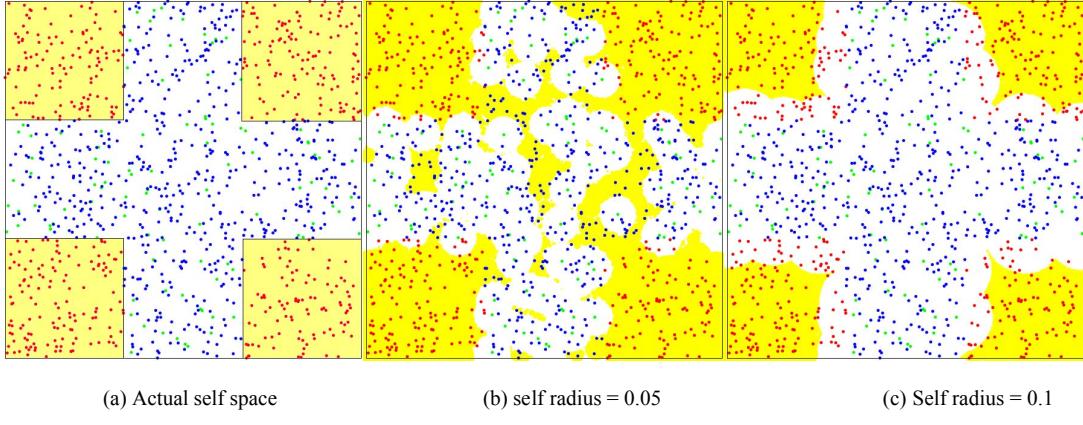


Figure 4 Cross-shaped self space

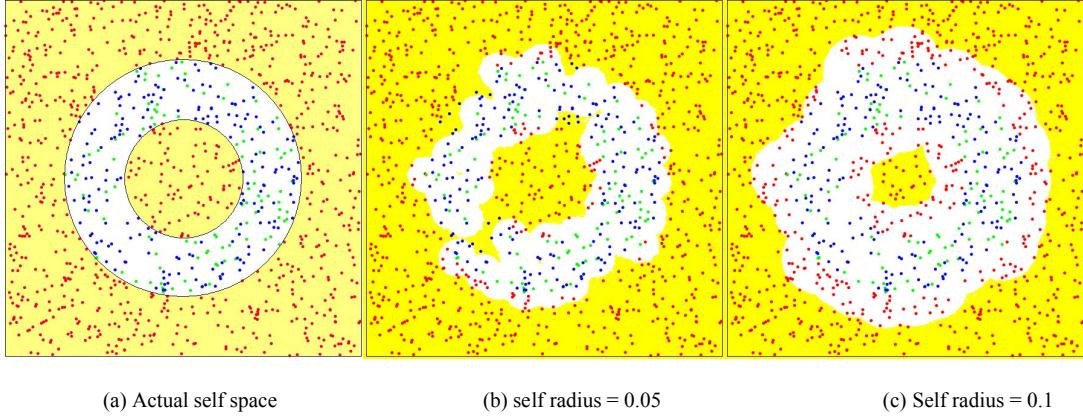


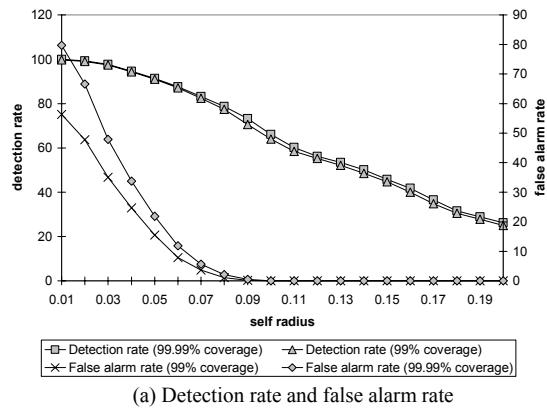
Figure 5 Ring-shaped self region

Figures 6 and 7, for the cross-shaped self region and ring-shaped self region, respectively, show the complete trend of self radius's effect on the results for self radius from 0.01 up to 0.2. All the results shown in these figures are averages over 100 repeated experiments. Detection rate and false alarm rate are defined as

$$DR = TP/(TP+FN). \quad (2)$$

$$FA = FP/(FP+TN). \quad (3)$$

respectively, where TP, FN, FP, TN are the counts of true positive, false negative, false positive, and true negative. As shown in these results, high detection rate and low false alarm rate are the two goals between which we need to balance according to specific application.



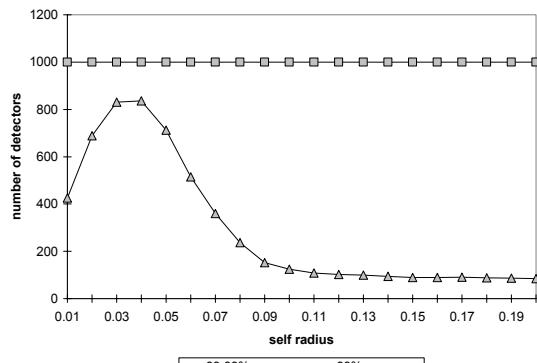
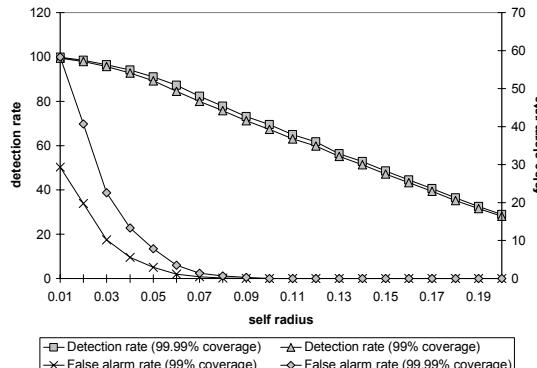


Figure 6 Cross-shaped self region
(b) Number of detectors



(a) Detection rate and false alarm rate

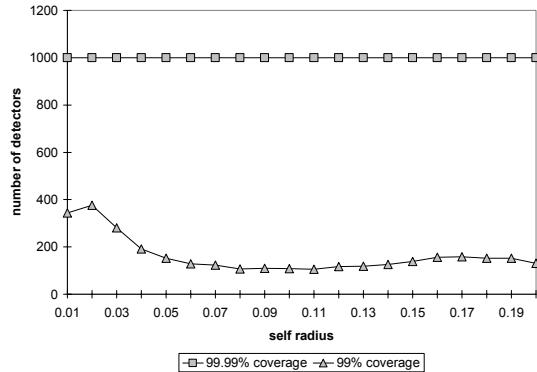


Figure 7 Ring-shaped self region

Effect of expected coverage c_0 was also demonstrated by these results. Two different values of “expected coverage”, 99.99% and 99% were used. The results in figures 6 and 7 display no significant difference in detection rate, while the numbers of detectors are much lower and false alarm rate is lower too for 99% coverage. It is clear that 99% expected coverage not only reduces the number of detectors, but also controls the false alarm rate better. In the case of cross-shaped self region, more

detectors are needed when the self radius is chosen between 0.02 and 0.05. This happens when the regions not covered by self “circles” becomes separated from one another. Comparing figure 4 and figure 5, we can see that this trend is not very similar for the two different shapes of self regions.

When the self radius is larger, the detection rate decreases. As we can see from figures 4 and 5, a larger portion of the non-self region is covered by the circular regions around the self points with larger self radius. This causes what we called the “boundary dilemma” – if a self sample point is very close to the boundary of self region, it is likely that only one of the two sides of this point is really in self region. In fact, the algorithm using constant detectors as presented in figure 2 does not have this problem, since the detectors in that case reach the self sample points instead of stay r_s away from them. The self region is safe from being detected because the intervals between self samples are supposed to be small enough to prevent any detectors to be generated in between. The dilemma exists because it is usually unknown whether a sample point is at the boundary or not. As a compromising solution, we can modified the V-detector algorith by combining the ideas of the two algorithms in figures 2 and 3, in the other words, allowing the varaiable-sized detectors to reach the sample points but excluding those whose radius is smaller than self radius r_s . This way, V-detector’s ability to cover the holes is largely limited by the possibility that any sample point may be a bounary point.

To explore the property and possible advantages of V-detector, experiments are also carried out to compare with the results obtained by AIS methods reported in [7], namely NSA (negative selection algorithm) and MILA (multilevel learning algorithm). NSA used there is a real-valued version of rcb (r-contiguous bits) – r contiguous dimensions out of all the dimensions are used to calculate Euclidean distance [7][8]. MILA is a multilevel version, which combines negative selection and positive selection [7][8]. Table I shows the comparison using the benchmark Fisher’s Iris Data. The results shown are the averages of 100 runs for each method with different parameter setting. Standard deviation of these results are all within 3%. One of the three types of iris data is considered as normal data, while the other two are considered abnormal. The available normal data are either completely or partially used to train the system. Although the partial training set may seem small in this case, it is necessary to demonstrate the system’s capability to recognize unknown normal data. V-detector has similar or better detection rates but lower false alarm rates in most cases, especially when fewer training data are used.

The results in table I are obtained using self radius $r_s = 0.1$ considering that the NSA and MILA results used threshold 0.1. It is to be noted that the threshold used in NSA or MILA is not strictly comparable to the self radius for V-detector. NSA and MILA used sliding windows of

size 2 in the results cited here. In the other words, the distance is defined in 2-dimensional space instead of in original 4-dimensional space. So self radius 0.1 used in table I is to some extend arbitrary.

TABLE I. COMPARISON USING FISHER'S IRIS DATA

		Detection rate	False alarm rate
Setosa 100%	MILA	95.16	0
	NSA	100	0
	V-detector	99.98	0
Setosa 50%	MILA	94.02	8.42
	NSA	100	11.18
	V-detector	99.97	1.32
Versicolor 100%	MILA	84.37	0
	NSA	95.67	0
	V-detector	85.95	0
Versicolor 50%	MILA	84.46	19.6
	NSA	96	22.2
	V-detector	88.3	8.42
Virginica 100%	MILA	75.75	0
	NSA	92.51	0
	V-detector	81.87	0
Virginica 50%	MILA	88.96	24.98
	NSA	97.18	33.26
	V-detector	93.58	13.18

Besides better detection results, V-detector's another important advantage is the potentially smaller number of detectors. Table II shows that V-detector used fewer detectors in all the cases reported in table I. NSA used 1000 detectors; MILA used 1000 T-detectors and 1000 B-detector groups. The maximum detector set size of V-detector is set to 1000 for the reason of comparison. Table II shows that less number of detectors are actually used.

TABLE II. NUMBER OF DETECTORS USED V-DETECTOR

	mean	max	Min	SD
Setosa 100%	20	42	5	7.87
Setosa 50%	16.44	33	5	5.63
Veriscolor100%	153.24	255	72	38.8
Veriscolor 50%	110.08	184	60	22.61
Virginica 100%	218.36	443	78	66.11
Virginica 50%	108.12	203	46	30.74

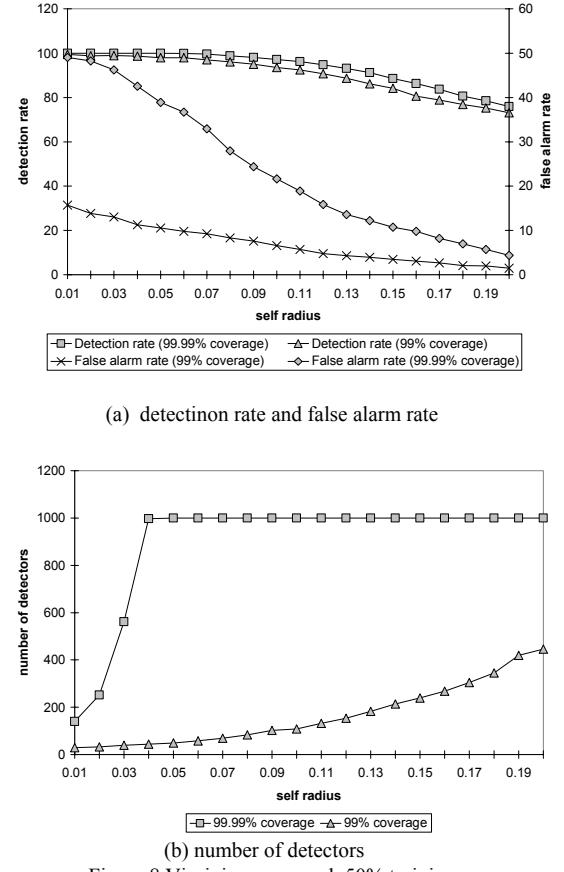


Figure 8 Virginica as normal, 50% training

Figure 8 shows the effect of the two control parameters, self radius and expected coverage, on the results when virginica is considered “normal” and half the available data are used as training set. As we have seen for the synthesized data, using 99% coverage didn't degrade detection rate very much comparing with 99.99% expected coverage, but largely saved the number of detectors and lowered false alarm rate.

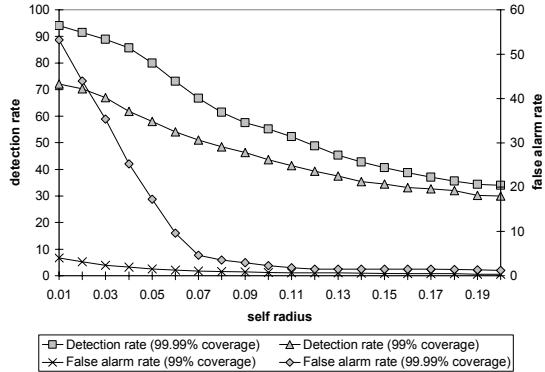
Influence of self radius is also similar to the results of synthesized data. As we have seen, self radius is an important control parameter of V-detector to balance between high detection rate and low false alarm rate in more general cases. False alarm does not really become a problem when all available training data are used, so the issue is more readily illustrated when only partial data are used to train. The results in figure 8 show that V-detector has advantage over other methods in terms of balancing detection rate and false alarm rate.

The results when setosa or versicolor is considered as “normal” are similar too, except that in the case of setosa the detection rate is almost always 100% due to the fact the setosa type is more clearly separated from the other two in the data space.

Similar comparison is done for another dataset, referred to as “biomedical data” from blood measurement of a group of 209 patients [9]. Each patient has four different types of blood measurements. These blood measures are used to screen a rare genetic disorder. 75 of those patients are carrier of the disease. 134 patients are normal. In this case, the carrier patients are considered “abnormal” data.

Biomedical data set is also 4-dimensional as iris data. Different percentage of normal data are used in the experiments as training data. Figure 9 shows the results when 25% of the available normal data are used. All the trends are similar to those of above results. Also as expected, if all available normal data are used to train, false alarm didn’t appear to be a problem.

What is different from the previous results is that the detection rate is lower, and influence of lower expected coverage is more obvious in this case. Comparing between different percentage of normal data used to train, the detection rate is actually lower when more data are used as training data. The same trend can be seen in the results obtained with MILA or NSA. It is thus attributed to the distribution of the original data set. Considering the balance between detection rate and false alarm, V-detector’s results are either better or comparable

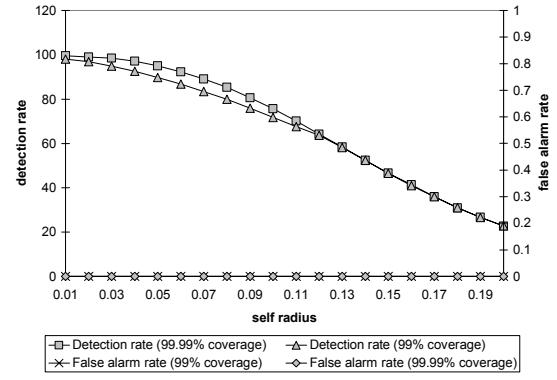


(a) detection rate and false alarm rate

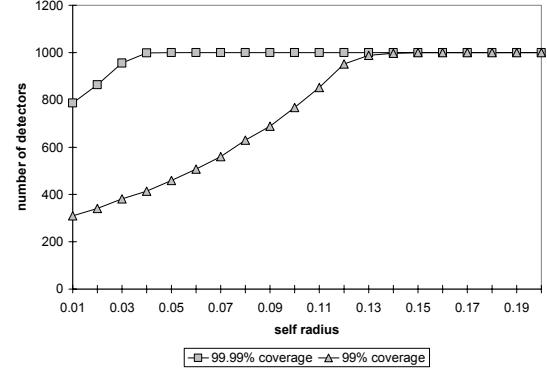
(b) number of detectors

Figure 9 Result of biomedical data, using 25% training data

V-detector was also tested using a pollution data set [9]. This data set is different in that it is high-dimensional. Each record consists of 16 measurements concerning air pollution. All the real data are “normal data”. The testing data were made by changing some components out of the range of known normal data. Figure 10 shows the results, which confirm the same conclusion about the influence of self radius and expected coverage.

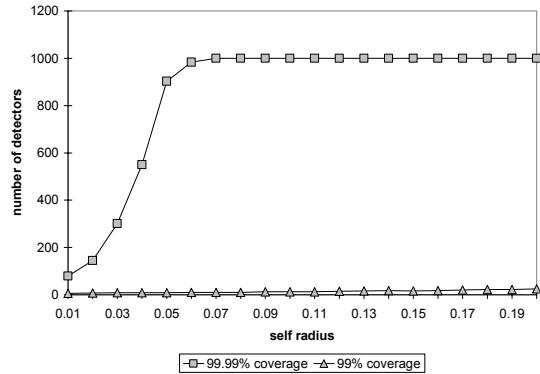


(a) detection rate and false alarm



(b) number of detectors

Figure 10 Results on pollution data



IV. DISCUSSION AND CONCLUSION

In comparison with real-valued negative selection algorithm of constant-sized detectors, V-detector is more effective by using fewer detectors due to its detectors’ variable size. Essentially, it provides a more concise representation of detection rules learned from training data. Detector set generated by V-detector is more reliable because the expected coverage instead of arbitrary detector number can be achieved by the algorithm.

A small number of detectors reduces space requirement and saves time to generate the detector set and to detect new cases.

Holes are covered better in the new method because smaller detectors are more acceptable when fewer larger detectors are used to cover the large non-self region.

Coverage estimate is very useful when evaluation of detection rate is not enough. For specific experiments, detection rate may not accurately reflect how the algorithm works because the training data is either too noisy or not representative enough. In those cases, confident estimate of coverage is more useful.

Influence of self radius and expected coverage as control parameters was preliminarily analyzed, where 0.1 is found to be a typical value of self radius. Expected coverage of 99% appears to be necessary. Optimal parameters are expected to depend on the specific application and available training data. One of the goals for further research is to decide these parameters automatically. Interpretation of training data, especially how each self sample represents the self region, is also an important topic that needs to be further explored.

Boundary dilemma will be an important issue to improve the performance of negative selection algorithm, where the problem is more obvious in real-valued space. It depends on how the training normal data are obtained and interpreted.

Future works along the line can be extended to variable shape of detectors, variable number of dimensions, etc. It also has the potential for certain problems that are hard to deal with otherwise, e.g. those in very high dimensional space where only a small number of dimensions affect the classification. For binary representation, it is easy to extend to variable dimension, which has possible advantage similar to what we discussed in this paper. Limited number of detector dimensions has additional benefit of extracting and representing the knowledge or rules in a more comprehensible form.

ACKNOWLEDGMENTS

This work was supported in part by NIH Cancer Center Support Core Grant CA-21765 and the American Lebanese Syrian Associated Charities (ALSAC).

REFERENCES

- [1] de Castro, L. N., et al, *Artificial Immune System: A New Computational Intelligence Approach*, Springer-Verlag, 2002
- [2] Dasgupta, D., et al, *Artificial Immune System (AIS) Research in the Last Five Years*, CEC-03, 2003
- [3] Dasgupta, D., et al, An Anomaly Detection Algorithm Inspired by the Immune System, in *Artificial Immune System and Their Application*, ed. by D. Dasgupta et al, 1999
- [4] Gonzalez, F., D. Dasgupta, J. Gomez, The Effect of Binary Matching Rules in Negative Selection, GECCO-03, 2003
- [5] Gonzalez, F., D. Dasgupta, L. F. Nino, A Randomized Real-Valued Negative Selection Algorithm, ICARIS-03, 2003
- [6] Esponda, F., S. Forrest, P. Helman, A Formal Framework for Positive and Negative Detection Scheme, IEEE Transaction on Systems, Man, and Cybernetics, 2003
- [7] Dasgupta, D., et al, MILA – Multilevel Immune Learning Algorithm, GECCO-03, 2003
- [8] Ji, Z., Multilevel Negative/Positive Selection in Real-Valued Space, Research Report, UM, 12/21/2003
- [9] StatLib – Datasets Archive, <http://lib.stat.cmu.edu/dataset/>
- [10] Gonzalez, F., D. Dasgupta, Anomaly Detection Using Real-Valued Negative Selection, Genetic Programming and Evolvable Machine, 4, 383-403, 2003
- [11] Ceong, H. T., et al, Complementary Dual Detectors for Effective Classification, ICARIS-03, 2003
- [12] Hofmeyr, S., and S. Forrest, Architecture for an artificial immune system, Evolutional Computation Journal, vol. 8, no.4, 2000
- [13] Kim, J., et al, An evaluation of negative selection in an artificial immune system for network intrusion detection, in Proceedings Genetic Evolutionary Computation Conference, San Francisco, 2001
- [14] de Castro, L., N., J. I. Timmis, Artificial Immune Systems as a Novel Soft Computing Paradigm, Soft Computing Journal, 2003
- [15] Ayara, M., J. Timmis, R. de Lemos, L. de Castro, and R. Duncan, Negative Selection: How to Generate Detectors, 1st ICARIS, 2002
- [16] Janis Kuby, *Immunology*, W. H. Freeman and Company, 1997