1. INTRODUCTION

Negative selection algorithm is one of the main algorithms of artificial immune system, which has been widely used in different areas \[1\]. In essence, the negative selection algorithm is a self/non self recognition technique that is inspired by the maturation process of T cells in the thymus: if T cells recognize autologous elements, it will be cleared. The typical negative selection algorithm is composed of two stages. First of all, it is the generation of the detector set, and then uses the detector set to check the new samples. So it can be seen that the quality of the detector directly affects the result of the test. The detector expression method is a binary representation and real valued representation, due to the real valued said is more suitable to describe practical problems, the researchers attention \[2\]. literature \[3\] first proposed the real valued negative selection algorithm and the fixed length detector, due to the difficulty in determining the appropriate detector radius. Therefore, detection performance is poor; literature \[4\] proposed variable length detector generating algorithm, improves the detection rate, and has been widely used in the different of the anomaly detection domain. On the basis of literature \[4\], different researchers have made different improvement work \[5-9\]. However, almost all of the algorithms are considered as candidate detectors for each individual point as a candidate for the individual and excluded from the detector. This approach may lead to some deviations, especially in the regions of the self and non self regions of the boundary. The traditional detection rate and false alarm rate are mainly caused by the boundary of self and non self, so the algorithm must consider the performance of the self / non self boundary. This paper calls this problem for the border dilemma.

Based on this, this paper proposes a new algorithm of edge detection. The basic idea of this algorithm is that training data is considered as a whole, and not independent of the point; independent self point cannot be declared a regional self region, only a set of points together to show that a self region. Based on the VRNSA \[4\] algorithm, using real valued, using variable length detector. The experimental results show that this algorithm has certain advantages, especially for the case of self / non self boundary element.

2. BOUNDARY PROBLEM OF NEGATIVE SELECTION ALGORITHM

In terms of real valued representation, according to the meaning of self region, a self elements or self elements set specific said what kind of data depends on matching rule \[8\]. Usually, it can be considered that a region of a circle is normal (self), which is a generalization of the self, which is similar to that of the binary representation. At the same time, it can be considered that only the normal state of the body is normal, and the other is abnormal state. In this
paper, the former is called "radical (aggressive)"; the latter is "conservative". It is clear that each of the points can be used as a proof of the self region near the region. On the other hand, from the perspective of fairness, it is assumed that the self points can be obtained from any part of the whole body. Therefore, no matter which matching rules are used, the points that are close to the boundary of the self and non self region cannot be ignored. That is, how to generalizing the self and non self boundary of the autologous sample is a difficult problem. Therefore, it is difficult to make a reasonable explanation of these elements. Figure 1 shows the boundary problem. The shaded part said actual self region, the white area said non self region and dots indicate self data, circle is autologous generalization. As can be seen from Figure 1, the self sample points close to the boundary inevitably extend the range of the actual self region algorithm. If the threshold is too small, the space between the samples will not be expressed, in other words, more samples will be needed to train the system. If the threshold is too large, the error of the self region is too large to be accepted. In particular, when the nonself region is in two regions of autologous thin strips, as shown in Figure 2, there may simply cannot be said (yellow bar indicates the non self region). In this case, the boundary problem is more serious. Therefore, it is necessary to design a new algorithm to solve the boundary problem.

3. BASIC STEPS OF ALGORITHM AND ANALYSIS

3.1 Basic steps of algorithm

In this paper, the design of the detector generation algorithm, the basic steps are outlined as follows. Where $S$ represents a collection of self samples; $N_{\text{max}}$ represents the maximum number of preset detectors; $r_s$ represents the self radius; $a$ is the desired coverage. $D$ Representing the generated detector set, $d_i(i=1,2,...n)$ is a detector in the detector set; $t$ is a random sample points covered by detector.

Step 1: generate candidate detector $x$ (using a combination of \[9\] and random generation from the gene pool);

Step 2: calculate the Euclidean distance between the $x$ and the detector $d_i(i=1,2,...n)$ in the detector set $L_{di}=\text{Euclidean}(d_i,x)$;

Step 3: if the distance $L_{di}$ is less than the radius $r(d_i)$ of the detector $d_i$, the point $x$ has been covered by the detector, then the counter $t=t+1$; otherwise the turn step 5;

Step 4: if $t \geq 1/(1-a)$ , then the coverage is sufficient, the end of the algorithm; otherwise, turn step 1;

Step 5: computing point $x$ and self point $s$, from Autologous sample set $S$,it’s Euclidean distance is $L_i=\text{Euclidean}(s_i,x)$ , $x$ radius $r$ is determined by the recent self elements, remembering distance from $x$ recent self point distance $d$ detector $(x,r)$ which $r = d - r_b (0 \leq r_b < r_s)$; $r_b$ said as the threshold boundary control parameters, is used to control the detector "aggressive";

Step 6: add the $(x,r)$ as a new detector to the detector set $D$;

Step 7: if the detector set $|D|$ to the maximum number of detectors $N_{\text{max}}$, then the end; otherwise; turn step 1;

Algorithm basic process as shown in figure 3.
3.2 Algorithm characteristics and advantages analysis

(1) The biggest difference of this algorithm lies in: This paper sets a boundary threshold $r_b$ to measure the aggressiveness of boundary detector in the self region boundary. We can use the $r_b$ size of the different, to control the performance of the algorithm.

(2) Compared with the fixed length detector [3], the radius of each detector in this algorithm is variable. Due to the variable radius, the algorithm has better coverage ability to detect "black hole".

(3) And VRNSA compared [4], the algorithm in selecting the candidate detectors are proposed by combination of randomly generated and the gene pool, both to ensure the diversity of the detectors, and improve the production efficiency of the detectors, and detection rate can be increased.

(4) The detector generates an integrated coverage estimation method, which is used as the termination condition of the algorithm, rather than simply depends on the maximum number of preset detectors. Therefore, this algorithm provides the following two ways: the first case is: when the estimated coverage of the algorithm, which is the end of the algorithm, this algorithm is unique, you can avoid the production of redundant detectors. The second case is: the number of detectors to reach the preset value. Even in this case, because the detector is longer, the algorithm still has the ability to better cover the black hole. The correctness of the point estimate (step 4) is shown as follows:

1) Generating random points as candidate detectors;
2) If it is a non self point and is not covered, on top of it produced a new detector; if it is covered by the self point is not as the candidate detector, but attempt was recorded in a counter $t$ to estimate coverage;
3) If sampling $m$ points in the non self space, there is only one point that has not been covered, by point estimation, the non covered area is: $1/m$, the coverage rate is estimated as: $a = 1 - 1/m$;
4) Therefore, when a random attempt is made to $t$ times without finding an uncovered point (point all covered), the actual coverage rate has been estimated to be $a$. Therefore, $t$ does not need to default, it is determined by the target coverage rate:

$$t = \frac{1}{1-a}$$

4. EXPERIMENT AND RESULT ANALYSIS

In order to verify the performance of the algorithm, experiments on both synthetic data and real data are carried out to verify the performance of [10-11], and compared with the related algorithms.

4.1 Results on synthetic data

The proposed algorithm is validated by using a synthetic 2D data set [10]. The whole space is 2 dimensional region $[0,1]^2$. This data set includes a variety of different shapes of the self distribution. As the main boundary detection problem, the band of self region is the most simple and can explain the difference of shape. Therefore, this paper chooses the band to verify the performance of the algorithm (Figure 2). Assuming that the training data points are randomly distributed over the entire space, using 100 self points. After the detector is generated, the performance is evaluated using random distribution across the space of the point. Test data are 1000 randomly distributed data points, including self and non self. According to the existing literature research results [8,9], the choice of parameters in the experiment: the maximum number of detectors...
is $N_{\text{max}} = 1000$, $r_s = 0.1$, $a = 99\%$, the threshold of 0.01-0.1. The result of the experiment is to use the same parameter to run 100 times. Detection performance using the detection rate (rate detection) and false alarm rate (alarm rate false) to measure. Contrast algorithm for the classic [4] algorithm VRNSA.

The experimental results are shown in figure 4.

![Figure 4. Experimental results.](image1)

In the figure, the horizontal coordinates represent the boundary threshold, and the longitudinal coordinates respectively represent the detection rate and false alarm rate. As can be seen from the figure, with the increase of the boundary threshold, the detection rate of the algorithm is reduced, while the false alarm rate is also reduced. Mainly due to: with the change of the boundary threshold, the algorithm detects the small radius of the detector which is close to the edge of the non self space. From the figure can be seen, the detection rate of the proposed algorithm is much higher than the VRNSA [4] algorithm, the false alarm rate is slightly increased. The main reason is that: in the VRNSA algorithm, a large part of the non self will be used as a body, therefore, reduce the detection rate. Considering synthetically, the algorithm has certain advantages. At the same time, the boundary threshold is adjustable, which can be used to balance the detection rate and false alarm rate.

### 4.2 Experimental results on real data

In order to verify the performance of the algorithm, the well-known Iris Fisher's data set is used to compare with other related algorithms [11]. The data set has 3 classes, each class of 50 samples, each sample is a 4 dimensional feature vector. Are the mountain (Iris setosa), iris iris versicolor (Iris versicolor) and Virginia (Iris virginica). When the experiment, a kind of data is used as the normal data, the other two kinds of abnormal data. Normal data is used to train the system. Among them, 50% of the Setosa said setosa is considered normal data, the other two types (versicolor and virginica) was as abnormal data, 50% setosa data is used as training data, all three types of data (including has been used as training data) as the test data and other (Setosa 25 percent) said that the meaning is similar. The results of the experiment are the results of the 100 repeated experiments for each method and the parameter settings. And set the same experimental parameters with other algorithms in order to compare. The results of the experiment are shown in table 1.

<table>
<thead>
<tr>
<th>Training data</th>
<th>Relevant algorithm</th>
<th>Testing rate</th>
<th>False alarmrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setosa100%</td>
<td>This algorithm</td>
<td>99.98</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>95.99</td>
<td>0</td>
</tr>
<tr>
<td>Setosa50%</td>
<td>This algorithm</td>
<td>99.96</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>95.97</td>
<td>0.34</td>
</tr>
<tr>
<td>Versicolor100%</td>
<td>This algorithm</td>
<td>86.37</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>76.95</td>
<td>0</td>
</tr>
<tr>
<td>Versicolor50%</td>
<td>This algorithm</td>
<td>89.65</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>80.31</td>
<td>0.42</td>
</tr>
<tr>
<td>Virginica100%</td>
<td>This algorithm</td>
<td>83.80</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>76.87</td>
<td>1.2</td>
</tr>
<tr>
<td>Virginica50%</td>
<td>This algorithm</td>
<td>94.18</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>89.58</td>
<td>2.18</td>
</tr>
</tbody>
</table>

As can be seen from the table, the algorithm compared with [3] VRNSA, the detection rate has been greatly improved, false alarm rate increased slightly, can effectively improve the detection efficiency.

At the same time, a further experiment was made on the Biomedical data set [11]. This data set is a blood pressure measurement of 209 patients, each patient has four different types of blood pressure measurements (used to determine a rare genetic disease). 134 patients are normal, 75 patients are disease carriers, that is, the abnormal data to be detected. The experimental results show that the algorithm has some advantages, as shown in Table 2. The lower detection rate of the reason is that the radius of $r_s$ is set to 0.1, if you want to get a higher detection rate, you can reduce the radius of appropriately. Obviously, the self radius is still an...
important parameter for the balanced detection rate and the false alarm rate.

Table 2. Comparison of correlation algorithm detection performance (Biomedical data set).

<table>
<thead>
<tr>
<th>Train data</th>
<th>Relevant algorithm</th>
<th>Testing rate</th>
<th>False alarm rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>100% training</td>
<td>This algorithm</td>
<td>69.36</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>60.61</td>
<td>0</td>
</tr>
<tr>
<td>50% training</td>
<td>This algorithm</td>
<td>72.29</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>62.92</td>
<td>6.1</td>
</tr>
<tr>
<td>25% training</td>
<td>This algorithm</td>
<td>86.96</td>
<td>8.5</td>
</tr>
<tr>
<td></td>
<td>VRNSA</td>
<td>78.68</td>
<td>6.2</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS

In this paper, we propose a new algorithm of edge detection based on the shortcomings of the existing detector algorithm in dealing with the boundary element, and the advantage of the algorithm is verified by the synthetic data and the actual data. In practical use, the specific parameter values depend on the specific application problems. The next research work is: how to automatically determine these parameters, such as the radius of $r_s$, the boundary threshold $r_b$, etc.

REFERENCES