Original Research Paper

Computer Science

A NEGATIVE SELECTION ALGORITHM BASED ON HIERARCHICAL CLUSTERING OF DETECTOR SETS

Jun He

School of Computer Science and Software Engineering, Tianjin Polytechnic University, China

ABSTRACT

BACKGROUND- For the anomaly detection problem, the negative selection algorithm (NSA) has a significant detection effect. However, the traditional negative selection process requires a lot of time in the detection phase to calculate the distance from the sample point to the detector.

METHODS- This paper proposes a new negative selection algorithm based on hierarchical clustering of detector sets. Hierarchical clustering of detector sets to reduce the number of detector sets is the primary goal, thereby reducing the time spent in the test phase.

RESULT- Compared with RNSA and V-detector, the results show that in most cases, the algorithm can improve the detector rate and reduce the false detection rate.

KEYWORDS : anomaly detection; negative selection algorithm; hierarchical clustering

1. INTRODUCTION

Negative Selection Algorithm (NSA) is one of the main algorithms of artificial immune system. The application in the field of anomaly detection includes bearing detection[1], credit card fraud detection[2], and small sample detection[3].

The early negative selection algorithm used binary strings to represent self samples (antigens) and detectors (antibodies), but the algorithm has problems with matching computational inefficiencies and binary representations limit its practical application[4,5]. In 2003, Gonzalez et al. proposed the Real-valued Negative Selection Algorithm (RNSA) [6,7]. There are many black hole regions in the RNSA that cannot be detected by the detector. ZhouJi proposed a variable-valued real-valued negative selection algorithm V-detector[8].

These algorithms greatly increase the efficiency of detector generation, but in the sample test phase, the distance between the sample to be tested and all detectors is still calculated. This paper proposes a real-valued negative selection algorithm based on hierarchical clustering of detector set (CV-RNSA). The negative selection algorithm also improves the detection efficiency while improving the test efficiency.

2. CV-RNSAALGORITHM

The algorithm is based on the v-detector generation detector, optimize the test phase to improve detection efficiency.

The test phase of the algorithm is performed in two steps:

(a) Perform hierarchical clustering on the detector set, calculate the distance between the test sample and each cluster center, and compare it with the hierarchical cluster center detector radius to determine whether the test sample is an abnormal sample.

The basic idea of hierarchical clustering is to calculate the similarity between nodes by some similarity measure, and sort the similarities from high to low, and gradually reconnect each node[8]. The purpose of hierarchical clustering of detector sets is to use each clustering center to represent the detector to match the data to be detected, and to collect similar detectors into the same cluster to reduce the computational cost of the testing phase. The clustering process in this paper is hierarchical clustering from top to bottom. The clustering radius of each layer is gradually reduced by half, and the clustering range of the bottom layer will be more convergent, which makes the detection of the data to be tested more accurate.

For a two-dimensional real-valued feature space, the first layer of clustering radius is $r_1=\sqrt{2}$ to contain all of the detector sets in the unit feature space. To obtain the n+1th cluster, the cluster set is performed on the detector set in each cluster of the nth layer with a radius of $r_{n+1}= r_n/2$. The selection criteria of the hierarchical clustering center are as follows: the sum of the distances from the detector set that is not divided into clusters to the centers of the remaining detectors is the smallest until there is no cluster center that can be selected.

(b) For a small amount of data to be tested in the hole area that is not covered by the detector and the self, the V-detector determines all of it as normal data, which also makes the false detection rate higher. According to the detection of the hole area[9], if a test sample is not covered by any one of the detectors or the self-area, its anomaly can be determined by the data to the nearest detector and the nearest self-Euclidean distance (*Dis1* and *Dis2*). If *Dis1>20Dis2* , then the data is marked as normal, otherwise it is abnormal.

The test algorithm is shown in Figure 1. Where S is the self set, D is the detector set, C is the detector cluster center set, and T is the test set.

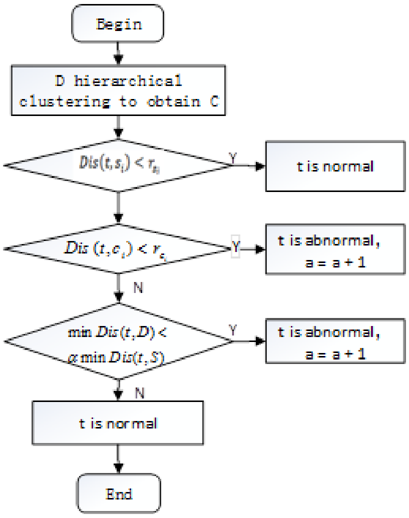


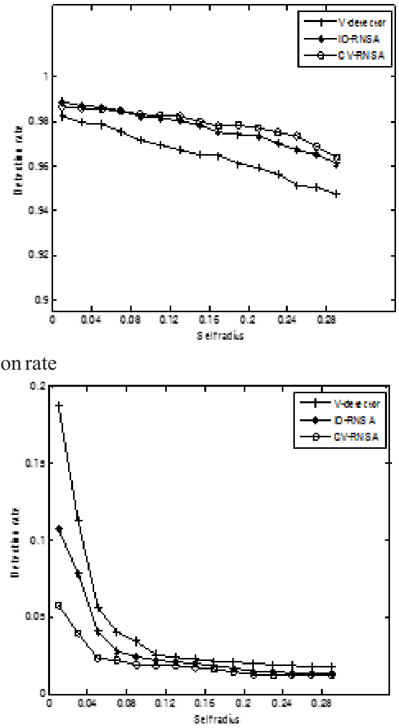
Figure 1: Test phase algorithm for CV-RNSA

3. Experiment and Result

In this section, the CV-RNSA and V-detector[8] and IO-RNSA[10] algorithm methods are compared on the four data sets. The experimental data is derived from the artificial two-dimensional data set used by Zhouji[6], including strips, crosses, pentagrams, and ring data sets selected from the two-dimensional synthetic data set. In this paper, the experimental results of the pentagonal star self-body set are taken as an example. In the two-dimensional real-value space, the training data set size is 1000, and the test data set size is 1000. The experiment uses the detection rate DR (detection rate) and the false alarm rate FA (false alarm rate) to compare the experimental results.

The detector is generated with a target coverage of 95%, and the self-radius is taken as 0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25, 0.27, 0.29, respectively. 50 experiments were performed for each self-radius, and the experimental results were averaged to obtain the experimental results of the coverage and false detection rates shown in Figure 2.

It can be seen from the experimental results that when the auto-radius increases, the detection rate of the CV-RNSA algorithm is slightly higher than that of the V-detector. When the target coverage rate of the detector generation phase is 95%, the CV-RNSA detection rate is more prominent than the V-detector detection result as the auto-collection radius increases, because the hole area increases as the auto-radius increases, the CV-RNSA improves the detection of the hole area by further determining the hole area, and further improves the detection rate of the entire area.



a) detection rate

b) false alarm rate

Figure 2: Target coverage rate of 95%, influence of self-radius

The detectors were generated with target coverage of 90% and 99%, respectively, and the self-radius was taken as 0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.13, 0.15, 0.17, 0.19, 0.21, 0.23, 0.25, 0.27, 0.29, respectively. 50 experiments were performed for each self-radius, and the results of the detection rate were averaged to obtain the results shown in Fig. 3 and Fig. 4.

Figure 3 shows that when the target coverage is 90%, the detection rate of CV-RNSA and IO-RNSA is slightly higher than that of V-detector.

Figure 4 shows that when the target coverage is 99%, the experimental results of CV-RNSA, IO-RNSA and V-detector are almost the same. Because when the target coverage increases, the number of detectors generated is significantly increased, at which time the coverage of the non-self area is significantly increased, and the area of the hole area not covered by the detector is reduced. The advantages of CV-RNSA in the hole area test cannot be exerted, so the experimental results (detection rate) of the three algorithms are almost the same at this time.

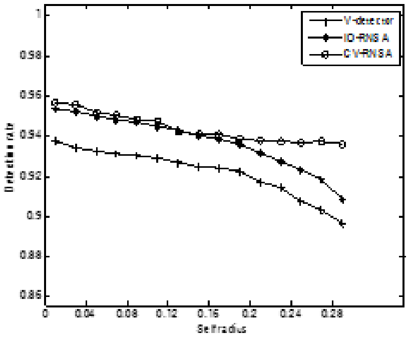


Figure 3: Detection rate when the target coverage is 90%

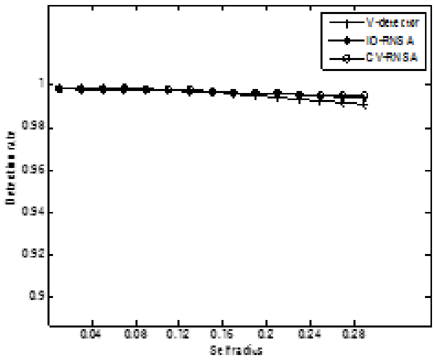


Figure 4: Detection rate when the target coverage is 95%

4. CONCLUSION

This paper proposes a real-valued negative selection algorithm based on hierarchical clustering of detector set (CV-RNSA). The cluster center set is obtained by hierarchically clustering the detector, and the distance between the data to be tested and the cluster center is calculated, thereby performing abnormal or normal classification determination on each data to be tested. For the hole area that is not covered by the detector, the attribute is further judged. The experimental results show that when the target detection rate is lower, the detection effect of CV-RNSA is slightly improved, and the false detection rate is significantly reduced. When the target detection rate is too high (99%), the advantage of CV-RNSA is not obvious.

ACKNOWLEDGMENTS

Supported by the National Natural Science Foundation of China (U1504613).

REFERENCES

1. Li Dong, Liu Shulin, Liu Yinghui, Zhang Hongli. (2014), "A Method for Equipment Abnormality Detection Based on Adaptive Super-Circle Detector." Journal of Mechanical Engineering ,50(12):17-24.
2. Mao Jiali, Jin Cheqing, Zhang Zhigang, Zhou Aoying. (2017) "Trajectory Big Data Anomaly Detection: Research Progress and System Framework." Journal of Software ,28(01):17-34.
3. Ying Wang, Yongjun Shen, Guidong Zhang. "Research on Intrusion Detection System Using Ensemble Learning Methods." 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS2016), Beijing, China.
4. Lin Weining, Chen Mingzhi, Zhan Yunqing, Liu Chuanwei. (2017) "Research on Intrusion Detection Algorithm Based on PCA and Random Forest Classification." Information Network Security, 11, 50-54.
5. Xia Yuming, Hu Shaoyong, Zhu Shaomin, Liu Lili. (2017) "Research on Network Attack Detection Method Based on Convolutional Neural Network." Information Network Security, 11,32-36.
6. González F A, Dasgupta D. (2003) "Anomaly Detection Using Real-Valued Negative Selection." Genetic Programming & Evolvable Machines, 4(4):383-403.
7. Gonzalez F A. (2003) "A study of artificial immune systems applied to anomaly detection." The University of Memphis.
8. Zhou J, Dasgupta D. (2009) "V-detector: an efficient negative selection algorithm with probably adequate detector coverage." Inform sciences, 19: 1390–1406
9. Gong M, Zhang J, Ma J, et al. (2012) "Short Communication: An efficient negative selection algorithm with further training for anomaly detection." Knowledge-Based Systems , 30(2):185-191.
10. Xiao X, Li T, Zhang R.. (2015) "An immune optimization based real-valued negative selection algorithm." Kluwer Academic Publishers.