Information Sciences 420 (2017) 218–234

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Parameter analysis of negative selection algorithm

Chen Wen*, Li Tao

College of cybersecurity, Sichuan University, Chengdu, 610065 China

ARTICLE INFO

Article history:
Received 21 January 2015
Revised 16 August 2017
Accepted 19 August 2017
Available online 23 August 2017

Keywords:
Negative selection algorithm
Parameter estimation
Artificial immune system
Detector
Coverage

ABSTRACT

The performance of negative selection algorithm (NSA) is affected by several parameters such as its self radius and expected coverage. Traditionally, most of the parameters are selected based on experience, resulting in varying performances of NSAs. In the present study, the NSA parameters were analyzed based on a new set of evaluation criteria. The criteria were used to calculate the self radius by an iterative algorithm using some non-self samples as reference points for the nonself boundary. The difficult problem of estimating the overlap volume between immune hyperspheres was solved by the Monte Carlo method. It was found that the error of the estimated nonself coverage was the primary cause of the poor performance of some existing artificial immune systems. A confidence estimation method was thus developed for improving the estimation precision. Experiments were performed in which both fixed- and variable-radius detectors were generated using different parameter value combinations. The results revealed that a significantly higher NSA performance could be achieved by the proposed parameter calculation method. Statistical analysis of the experimental results further confirmed the effectiveness and practicability of the proposed method.

1. Introduction

Negative selection algorithm (NSA) is designed to train immune detectors in the artificial immune system (AIS). NSA simulates the training process of T-cells in the biological marrow to generate mature detectors [8], which are trained to only recognize nonself elements. The mature detectors can be used in the fields of network intrusion detection, mechanical damage detection, disease diagnosis, and other applications [1,22,23,31,36]

In 1994, Forrest firstly designed the original negative selection algorithm [8], which defined antibodies (detectors) and antigens (sample characters) using binary strings and calculated the similarities between them through *r*-contiguous matching rule. As many applications are feasible to be studied in the real-value space, Gonzalez and Dasgupta[10]. proposed real negative selection algorithm (RNSA), in which the antigens and detectors are defined as immune recognition balls (hyperspheres) in the shape space [20], while the similarity was evaluated by the Minkowski distance function.

Many improved NSAs have since been proposed, with the focus on the use of alternative detector generation schemes to optimize the efficiency or detection accuracy. Idris et al. [12] proposed a differential evolution-based NSA referred to as NSA-DE, in which the distribution of the detectors is optimized by differential evolution to reduce the amount of detection holes. Luo et al. [16] developed an NSA based on a genetic algorithm, with the purpose of optimizing the non-overlapping detectors to obtain maximal coverage of the nonself space. Poggiolini and Engelbrecht [21] developed a set of

* Corresponding author.
E-mail address: wenchen@scu.edu.cn (C. Wen).

http://dx.doi.org/10.1016/j.ins.2017.08.062
0020-0255/

feature-detection rules for improving the detection accuracy, and analyzed the relationships between adjacent and non-adjacent antigens to determine whether the antigens should be activated by the artificial lymphocytes. Ma et al. [17] proposed an antigen feedback mechanism for the efficient generation of mature detectors using feedback on unmatched antigens in the maturing process. Ostaszewski et al. [18,19,24] introduced different hypershape detectors such as hyper-ellipsoids and hyper-rectangles, as an improvement on hypersphere detectors. In previous works of the present authors [31,33], tree-based structures and grid files were used to organize the training data and hence reduce the time complexity of the search for the nearest self neighbors. In another work [32], the present authors demonstrated that, in a high-dimensional space, the Euclid distance is adversely affected by the dimensionality, resulting in a more difficult discrimination between self and nonself samples. The fractional order distance is thus used to enhance the discrimination level of the antigens.

NSAs can be categorized into two main classes, namely, those that generate fixed-radius detectors such as the greedy-NSA [6] and RNSA [10]; and those that represent an improvement on the RNSA through the use of variable detector radii, which are dynamically resized to the nearest self margin. Examples of the latter class of NSAs are the GF-NSA [33], V-detector [14], and Ft-NSA [9]. Despite the improvements that have been achieved in NSAs, there performance is still significantly affected by several parameters such as the self radius, detector radius, detector volume, and expected nonself coverage. Forrest [8] established a relationship between the missing detection rate P, the detector volume V_ε and nonself volume $V_{nonself}$, namely, $P \approx e^{-V_\varepsilon / V_{nonself}}$. This implies that P decreases with increasing nonself coverage,given by $V_\varepsilon / V_{nonself}$. Stibor et al. [29] analyzed the relationship between the data dimension and the detector volume and found that the latter rapidly decreased with increase of the former. Ji [14] evaluated the nonself coverage of detectors and used expected coverage as the exit condition for detector generation. Wen et al. demonstrated that large self radii were suitable for situations that were sensitive to self false-alarm, while small self radii were suitable for cases that were not sensitive to such [31]. In another work, the relationship between NSA time cost and self set size was analyzed, and all of the self training samples were preprocessed to reduce the time cost of the negative selection process [33].

However, the parameters of an NSA are traditionally selected based on user experience. This causes the performances of different AISs to vary, with the consequence that AIS researchers can be classified into two groups based on their perception of the system performances. On one hand are those who obtained positive results from an AIS and consequently advocated it as promising [1,4,7,8,20,22,23,31,36]; and on the other hand are those who obtained poor results, and concluded on the inapplicability of the AIS [25–29]. This situation has significantly limited the application of AISs. Some NSA parameter adjustment methods have, however, been proposed in recent years. In ANSA [34], for example, the detector radius is periodically updated based on external confirmation of the detection results. In BIORV-NSA [3], the radii of the boundary self elements are randomly inhibited to reduce the amount of detection holes. Furthermore, in VSRNSA [35] and CHNSA [2], the radii of the self elements and detectors are automatically adapted based on the sample density to reduce overlapping of the self or nonself areas. The objective of all these methods is independent optimization of the immune parameters. However, the NSA parameters are closely related and thus require conjoint analysis and evaluation to achieve global optimization. In this study, we analyzed the relationships between the NSA parameters, as well as the properties of the parameter data. We also developed a group of algorithms for calculating the parameters to avoid the performance degradation caused by their improper setting.

2. Related works

The basic execution flow of the NSAs can be divided into two phases: detector training and antigen detection. In the detector training phase, the candidate detectors are randomly generated and compared by a self set. The candidates that cover any self area are replaced by new ones. After the training phase, the detectors become mature and only cover nonself spaces. Thus, in the detection phase, the antigens covered by the mature detectors are classified as nonself samples.

There are several parameters that substantially affect the detector performance, such as the self radius, detector radius, and exit condition of the detector training phase. The self radius is indeed important but often ignored. In addition, all the self samples may not be collectable, and the self radius is thus used to generalize the self profile. In the present paper, the space profiled by the self samples and self radius is defined as a self region. As an example, the whole self set in Fig. 1(a) is a rectangle defined by $[0, 1]^2$. If only a few self samples are available, as in Fig. 1(b), it would not be possible to use the samples to contour the self profile. In Fig. 1(c), the self region is expanded by the self radius, enabling coverage of most of the self elements. However, as shown in Fig. 1(d), if the self radius is too large, some nonself areas would be wrongly incorporated into the self region, resulting in the creation of detection holes that cannot be covered by the detectors. Choosing an appropriate self radius is thus important in the execution of NSAs.

The detector radius controls the detection range of a detector [14]. As illustrated in Fig. 1(e) and (f), there are two types of detectors, namely, fixed-radius and variable-radius detectors. For the fixed-radius detectors, the detector radius and detector number are two vital factors that affect the nonself coverage. Usually, a smaller radius requires the use of more detectors, and vice versa. Unfortunately, because of the many overlapping regions among the detectors when fixed-radius detectors are used, it is difficult to determine the number of required detectors. Furthermore, researchers [26–28] have found that the volume of a single detector tends to zero in high-dimensional space, and it is thus not clear how the detector radius can be set to avoid these problems.

The exit condition for the generation of detectors may be an expected number of detectors or an expected nonself coverage. The expected number of detectors is usually used in early NSAs [8], especially for the generation of fixed-radius