Knowledge-Based Systems 56 (2014) 26–35

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Negative selection algorithm based on grid file of the feature space

Chen Wen [a,*], Ding Xiaoming [b], Li Tao [a], Yang Tao [a]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] School of Computer and Information Science, Southwest University, Chongqing 400715, China

ARTICLE INFO

Article history:
Received 30 March 2013
Received in revised form 16 October 2013
Accepted 16 October 2013
Available online 1 November 2013

Keywords:
Artificial immune system
Negative selection
Detector
Grid file
Coverage

ABSTRACT

Negative selection algorithm (NSA) is an important algorithm for the generation of artificial immune detectors. However, the randomly generated candidate detectors have to be compared with the whole self set to exclude self reactive detectors. The inefficiency of the comparing process seriously limited the application of immune algorithms. Therefore, a new negative selection algorithm GF-RNSA is proposed in the paper. Firstly, the feature space is divided into a number of grid cells, and then detectors are separately generated in each cell. As candidate detectors just need to compare with the self antigens located in the same cell rather than with the whole self set, the detector training can be more efficient. The theoretical analysis demonstrated that the time complexity of GF-RNSA is effectively reduced that the exponential relationships between self size and time complexity in traditional NSAs is eliminated. The experimental results showed that: not only the time cost of negative selection, but also the time cost of data preprocess and detection are reduced, while the detection accuracy is not much declined.

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 [1], which are trained to only recognize non-self elements. The mature detectors can be used for further applications including machine fault diagnosis, network intrusion detection, etc. [2–6].

The early NSA [4] (native negative selection algorithm, NNSA) defined antibodies (detector) 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, González and Dasgupta et al. [7,8] proposed real negative selection algorithm (RNSA), in which the attributes of detectors and antigens were normalized into *d*-dimensional real valued space $[0,1]^d$, while the similarity was evaluated by the Minkowski distance function. Ji and Dasgupta [9,10] improved RNSA using variable detector radius that the detector radius was dynamically resized to the nearest self margin.

Chavez and Navarro [11] and Skala [12] proposed that for distances based pattern recognition algorithms, the distance calculation is the primary cause of time consuming. However, NNSA, RNSA and V-detector have no strategy to reduce the distance calculation cost: the distances between each candidate detector and the whole self set have to be calculated during the negative selection process, resulting in enormous time cost. Tao [2,3] and Forrest et al. [4] analyzed the efficient of NSA: the probability of a candidate detector to pass the negative selection is $(1-P_m)^{|S|}$, where P_m is the matching probability between random detector and antigen, $|S|$ is the self size, as the increasing of $|S|$, the pass probability will decrease to 0 ultimately; Furthermore, $\frac{-ln(P_f)}{P_m\cdot(1-P_m)^{|S|}}$ candidate detectors have to be generated under the expected failure rate P_f, therefore the time complexity of negative selection is exponentially related to the self size $|S|$, which seriously limited the application of NSA [13–17].

Furthermore, in the traditional NSAs, the candidate detectors are randomly distributed in the feature space, resulting in the unbalanced distribution of detectors [16,17]: part of the non-self space is re-covered by redundant detectors, while some other non-self regions are uncovered. NNSA [1] and RNSA [7] terminate the detector generation process when the expected detector number is reached, but there still exist many detection holes even if the expected number is reached. On the other side, V-detector [10], HC-RNSA [18], CB-RNSA [19] which set expected coverage as the terminate condition, also suffered from the redundant coverage problems as they do not control the distribution of detectors.

2. Related works

In 1990s, inspired by the self tolerance process of T-cells in the biological body, Forrest proposed the negative selection algorithm

* Corresponding author. Tel.: +86 28 85404936; fax: +86 28 85405568.
E-mail address: cwcwcwcw2006@163.com (C. Wen).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.10.018

NSA to generate artificial immune detectors [4]. NSA comprises two phases: First in the generation stage, candidate detectors are compared with the whole self set, the candidates that matched any self antigen will be dropped and replaced by new ones, while others will be incorporated into the mature detector set; then during the detection stage, the antigens that matched mature detectors will be classified as non-self.

By now, many improvements of NSA have been proposed [7–10,18–21], and most of them focused on the mechanisms of detector representation, generation, matching, etc. The detector representation is the foundation of NSA, which decides the detector generation and matching rules. There are mainly two representation mechanisms: binary strings and real vectors. Balthrop et al. [20] demonstrated that the limited discrimination ability of binary detectors is not suitable for the applications in the real environment. Therefore, González and Dasgupta et al. [7,8] proposed real valued negative selection algorithm RNSA. RNSA defines antigens and detectors using real vectors, which expands the discrimination domain into real space. After that, Ji and Dasgupta [9,10] proposed V-detector, which dynamically calculates the detector radius to the nearest self margin to increase the detection region of detectors.

The detector generation mechanisms include: random, mutation and model. The random method is commonly used in NSAs [1,7–10], which randomly generates candidate detectors from a given range. Detectors also can be generated by mutations, which first randomly generate detectors, and then the self reactive detectors are hypermutated along guided directions that far from self elements. Wenjian et al. [21] proposed model based generation mechanism, in which detectors are generated based on some predefined models with partial matching rule.

The matching rules includes: r-continuous bit match rule, r-chunk and distance based rule, etc. r-continuous is suitable for binary strings, which classifies two strings with r continuous same bits as matched strings [4]. As r-continuous has high false detection rate, Balthrop et al. [20] improved r-continuous and proposed r-chunk mechanism, which divides the binary string into some blocks, and two strings matched each other if the number of matched blocks is bigger than the matching threshold. Gonz¢lez et al. [7] proposed Euclid distance based matching rules for real valued NSA, if the Euclid distance between two real vectors is less than threshold (self radius or detector radius), then the two vectors are matched. Wen et al. [18] demonstrated that in high dimensional space, the Euclid distance suffers from the dimension disaster, and the discrimination between self and non-self becomes harder. Therefore, they proposed fractional order distance to evaluate the matching degree between real vectors.

Although many improvements of NSAs have been proposed, the detector generation is still in low efficiency, the fundamental problem is that the randomly generated detectors have to be compared with the whole self set, resulting in exponential time complexity[2–4]. Someone even suspected that this selection problem might be NP-complete [22], although no completeness proof had been shown. These ongoing difficulties led some in the field to conclude that negative selection is computationally too expensive for real-world data sets.

In Ref. [23], Elberfeld and Textor constructed an automaton whose acceptance behavior was equivalent to the NSA's classification outcome, and proved that NSA is not an NP hard problem and detectors can be generated for a given self set in none exponential time. Hofmeyr and Forrest [24] designed liner NSA and greedy NSA. First the binary strings that unmatched self strings were prestored in a dataset, and then candidate detectors were picked up from the set, thus the string comparing between candidate detectors and self set was avoided, and the time complexity was linearly related to the self set size. However, both the algorithms are restricted to the case where self and detectors are strings with r-continuous or r-chunk matching rule. For real valued NSA, in our previous work [18,19], we proposed hierarchical cluster based NSA, HC-RNSA and CB-RNSA to reduce the detector training cost. Both algorithms first preprocessed the self data set into a cluster tree structure before the negative selection process. During the training procedure, in each cluster level the self cluster centers replaced the cluster members to compare with candidate detectors. As the number of clusters is far less than the self size, the detector training is more efficient. In some other works, tree based structures were also employed to organize self data [25], such as using $k-d$ tree structure to store self data to reduce the complexity of searching the nearest self samples. Although tree-based NSAs are efficient in the detector training process, the time cost of preprocessing of self data is increasing with the self size, especially under huge self set and high dimensional space. Therefore, we need a more efficient preprocessing method to organize the self set to support the detector training process.

This paper aimed to improve the exponential worst-case complexity of existing NSA algorithms, and thus removes one major obstacle for applying negative selection to real-world problems. In the paper, a real negative selection algorithm based on the grid file of feature space (GF-RNSA) is proposed. First the n-dimensional feature space is divided by n-1 dimensional hyperplanes to generate orthogonal grid cells; and then candidate detectors are independently generated in each cell. The cells restrict the location of detectors to reduce the redundant distribution of detectors. Most importantly, the candidate detectors only have to be compared with the self-antigens located in the same cell, thus the cost of distance calculation can be much reduced.

3. Basic definition

Immune is the state maintain process of physical body which relies on antibodies to discriminate self and non-self antigens [4]. In the artificial immune theory, antibodies are defined as detectors which are used to recognize non-self elements, thus the detection performance depends on the quality of detectors [16]. The basic definitions of GF-RNSA are defined as:

Definition 1. All the sample character strings abstracted from the feature space constitute the antigen set $U=\{g|g=(f_1,f_2,\ldots,f_n), f_i\in[0,1]\}$, where n is the data dimension and f_i represents the ith normalized attribute.

Definition 2. Self set $Self\subseteq U$ represents the character strings abstracted from the normal samples, $r_s\in R^+$ is the variability threshold of the self points; Non-self set $Nonself\subseteq U$ represents the character strings abstracted from the abnormal samples, and $Self\cup Nonself=U, Self\cap Nonself=\varnothing$.

Definition 3. Detector $d=\langle c,r\rangle$, where $c\in Nonself$, c is the central vector (location) of d, $r\in R^+$ is the detector radius. Antigens which are close to d less than r will be identified as non-self elements.

Definition 4. The coverage of non-self space P is the volume of covered non-self space $V_{covered}$ to the volume of the whole non-self space $V_{nonself}$, and $0\le P\le 1$.

Definition 5. If any self element located in the detection region of detector d, then d is a self reactive detector.

4. The detail of GF-RNSA

In GF-RNSA, the antigen features $(f_1,f_2,\ldots,f_n)$ are normalized into the unit feature space $[0,1]^n$, and then $[0,1]^n$ is divided into a