The Effect of Antibody Morphology on Non-self Detection

Johan Kaers[1], Richard Wheeler[2], and Herman Verrelst[1]

[1] Data4s Future Technologies,
Ambachtenlaan 13G, 3001 Heverlee, Belgium
{johan.kaers,herman.verrelst}@data4s.com
[2] Edinburgh Research and Innovation Ltd.
1-7 Roxburgh Street, Edinburgh EH8 9TA, Scotland
richard.wheeler@ed.ac.uk

Abstract. Anomaly detection algorithms inspired by the natural immune system often use the negative selection metaphor to implement non-self detection. Much research has gone into ways of generating good sets of non-self detectors or antibodies and these methods' time and space complexities. In this paper, the antibody morphology is defined as the collection of properties defining the shape, data-representation and data-ordering of an antibody. The effect these properties can have on self/non-self classification capabilities is investigated. First, a data-representation using fuzzy set theory is introduced. A comparison is made between the classification performance using fuzzy and m-ary data-representations using some benchmark machine learning data-sets from the UCI archive. The effects of an antigen data reordering mechanism based on Major Histocompatibility Complex (MHC) molecules is investigated. The population level effect this mechanism can have by reducing the number of holes in the antigen space is discussed and the importance of data order in the r-contiguous symbol match-rule is highlighted. Both are analysed quantitatively using some UCI data-sets.

1 Introduction

Anomaly detection algorithms based on the biological immune system have been applied to a variety of problems, including virus detection [7], network intrusion detection [12] [11] and hardware fault tolerance [3]. The T-cell maturation process is used as an inspiration for algorithms that produce a set of *change detectors* or *antibodies*. A *censoring* process removes a *detector* when it *matches* with a cell or *data-string* of the *self* from a large space of possible detectors. The remaining ones are then used to determine if an incoming data-string or *antigen* is part of the self or not. The time and space complexities of these algorithms have been extensively analysed and variations inspired on other immunological phenomena [14] and evolutionary computing [16] [10] have been proposed [12].

One common property is that they all use a binary or m-ary symbol string data-representation. In the biological immune system however, recognition between receptors and antigens is based on their three-dimensional shapes and

J. Timmis et al. (Eds.): ICARIS 2003, LNCS 2787, pp. 285–295, 2003.

other physical properties [15]. Following the biological structure of the antibody-antigen binding process more closely could enable us to transpose the performance and adaptability of the natural immune system onto these computer algorithms. Therefore we investigate in this paper some modifications of the shape, data-representation and data-ordering of artificial antibodies. We will refer to these properties as the *antibody morphology*. They are analysed from the machine learning viewpoint and evaluated according to their usefulness as tools to enhance the classification capabilities.

The Artificial Life hypothesis is used because we assume that it is possible to model the biological immune system as a complex system of many interacting components. Similar to work in Artificial Chemistries [6] we abstract from natural molecular processes to investigate the fundamental dynamics of the complex system. An antibody morphology that uses Fuzzy Set Theory [20] is introduced. It captures the graded nature of the physical antibody/antigen match process and allows the non-self detection algorithms to handle data-sets with complex symbolic structures.

The paper is organized as follows. Section 2 looks at a data-representation inspired by fuzzy set theory and shows how the r-contiguous symbol matching rule can be modified accordingly. This fuzzy morphology is applied to some data-sets and statistically compared with the m-ary one. In section 3, the effect of using a data reordering method inspired by the *Major Histocompatibility Complex* (MHC) molecules is analysed. The importance of the data-order for the matching process is highlighted and the problem of *holes* in the non-self space is addressed by taking advantage of the population level effects resulting from using the MHC method. Section 4 contains the results and details of the various experiments using standard machine learning data-sets from the UCI repostiry [2].

1.1 Antibody Generation

Algorithms that generate antibodies make assumptions about their internal data-representation and therefore are tied to the antibody morphology. From the ones known in the literature [1], the *linear*, *greedy* and *binary template* algorithms build most heavily on the assumption of the binary (or m-ary [19]) string morphology. The ones based on generating random antibodies and/or genetic mutation operators (e.g. *exhaustive*, *NSMutation*) are more easily extended to arbitrary morphologies. The only requirement they have is that the morphology should allow for the generation of random antibodies and that a self to antibody matching scheme is present. Because of this independence of morphology, the *exhaustive* algorithm introduced by Forrest, Perelson et al. in [7] is used in the experiments included in section 4.

1.2 Machine Learning

Throughout the paper, non-self detection is considered as a 2-class classification problem, discriminating between self and non-self classes. The data-sets used