ANOMALY DETECTION IN SINGLE AND MULTIDIMENSIONAL DATASETS USING ARTIFICIAL IMMUNE SYSTEMS

A Thesis

Presented for the

Master of Science

Degree

The University of Memphis

Nivedita Sumi Majumdar

May 2002

This thesis is dedicated to the memory of my parents:

Late Cdr. Subhendu Majumdar and Late Dr. Mina Majumdar.

Acknowledgements

To Dr Dipankar Dasgupta: who introduced me to this field of research in the first place. Thank you for everything, especially for the funding for my masters degree.

To my thesis committee members, Dr Robert Kozma and Dr King-Ip Lin: Thank you for all your encouragement, support, excellent suggestions and feedback.

To Senhua Yu: For your warmly enthusiastic, unrelenting scrutiny of my work and clarification of all my queries.

To Fabio Gonzalez, Jonatan Gomez and Fernando Nino: For your warm enthusiasm in answering all my questions and giving me your excellent suggestions and feedback.

To Madhavi, Prasad, Ravi, Shahid and all others of the Intelligent Security Systems Research Lab: Thanks for your help.

To the department of Computer Science at the University of Memphis, TN: Am extremely grateful to you for your support in so many ways.

A very special thanks to my family and friends: particularly to Sabita, Sukhendu, Gypsy, Krishnendu, Suman, Sunny, Sarmistha, Suchandra and Swapnonil: thank you for everything.

Abstract

Majumdar, Nivedita Sumi. M.S. The University of Memphis. May 2002. Anomaly. Detection in Single and Multidimensional Datasets Using Artificial Immune Algorithms. Major Professor: Dipankar Dasgupta, Ph. D.

The natural immune system is an extremely efficient, complex, adaptive, security system that defends the body from foreign pathogens. It is able to categorize all cells (or molecules) within the body as self-cells or non-self cells. It does this with the help of a totally distributed task force that has the intelligence to take action from a local and also a global perspective using its superior network of chemical messengers for communication. This remarkable information processing bio system has caught the attention of computer science in recent years. The Artificial Immune System (AIS) community is a growing body of researchers working on importing ideas from biological immune systems to computer science and applying those ideas to solve real world science and engineering problems. My thesis discusses the problem of Anomaly Detection in datasets and outlines how the immunity based negative selection algorithm has been applied to anomaly detection in single and multidimensional datasets. It also discusses the development of a new immunology algorithm: MILA (Multi-level Immune Learning Algorithm). The new algorithm combines many of the features attractive in the immune system. The thesis

outlines how this algorithm has been applied to the problem of anomaly detection especially for high dimensional data.

Preface

The natural immune system is an extremely efficient, complex, adaptive, security system that defends the body from foreign pathogens. It is able to categorize all cells (or molecules) within the body as either belonging to its own kind (self-cells) or those that have a foreign origin (non-self cells). Rather than rely on any central control, it has a totally distributed task force that has the intelligence to take action from a local and also global perspective using its superior network of chemical messengers for communication. This remarkable information processing bio system has caught the attention of computer science in recent years. We discuss some of its capabilities of major interest to computer science.

- *Self / Non-Self discrimination:*

  The immune system is able to categorize all cells (or molecules) within the body as either belonging to its own kind (self-cells) or those that have a foreign origin (non-self cells).

- *Feature extraction:*

  Any suspicious candidate invader is recognized by its surface features by the B Cells and macrophages. Then the cell is processed by breaking the cell up into its

constituent peptides. This is similar to feature extraction. Subsequently, helper T Cells make a positive identification of the cell as a non-body cell, following which appropriate actions are taken for its elimination.

- *Learning:*

  The body is not born with defense against all possible forms of attacks. It is able to learn and evolve most suited receptors for each new type of attack that it is exposed to. This is why the baby is most susceptible to diseases. But as its immune system has more and more interactions with foreign invaders, it builds up its repertoire of defense cells suitable to the invaders signature pattern and thus develops a stronger line of defense.

- *Recognition:*

  The vertebrate immune system recognizes particular antigens (viruses and other undesirable foreign substances) by means of antibodies and immune cell receptors that bind to epitopes (small portions of the antigen, consisting of at least 4 to 6 amino acids). It is interesting to note that an exact match to the entire antigen is not attempted; in fact, it is almost certainly a physical impossibility. The immune system makes a first recognition of a foreign peptide, not by an exact match with an immune cell receptor but rather an approximate match. This is a very clever device in order to make the system robust and enable it to recognize variants.

- *Distributed pattern recognizer:*

There is no central controller of the immune system. The response and defense mechanism is totally distributed, which is a major strength for a robust fault tolerant security system.

- *Memory:*

  Immune systems do not forget an attacker fingerprint that it has seen. Some cells are circulated as memory cells so that the system retains the perfect receptor surface evolved for an attacker. When the immune system sees an attacker pattern for the first time, it takes a few days to evolve a suitable receptor for the unknown attacker. But at any subsequent time, this attacker signature is known to the system and it is able to recall that. That is why the immune response for an infection occurring a second time (secondary response) is much faster and stronger than the first one.

- *Elimination/neutralization of intruders:*

  The immune system is able to neutralize the site of reactivity of an attacker it has identified and thus neutralize its harmful effect. Also there are special purpose killer cells to destroy the foreign invaders once they are identified.

- *Use of selective proliferation and self-replication:*

  Once an immune cell identifies a particular attacker, that cell is selected to be cloned and mutated rapidly so that it can mature to create a more perfect fit for the intruder that activated it and thus take efficient steps to inactivate or kill the foreign cell.

All these ideas from immunology have relevance in computer security systems. Various immune system metaphors have already been implemented in software and have been applied to widely different application areas. This is a rapidly developing field and promises to be one of the focal points of computer science research in the coming decade. We call this the field of Artificial Immune systems.

This thesis is presented in the following way. The first chapter has a discussion of the natural immune system with focus on the metaphors used for real-world problem solutions. The second chapter discusses the research that is already underway in this area from a computational perspective and includes a survey of existing applications of artificial immune systems. The third chapter introduces the anomaly detection problem and possible approaches for its solution. In subsequent sections we discuss Forrest's Negative Selection algorithm inspired by immune systems, our application design in details outlining how the Negative Selection Algorithm has been applied to anomaly detection in single and multidimensional datasets. Finally we have the experimental results and some conclusions. The fourth chapter discusses the new immune algorithm developed namely MILA: Multi-level Immune Learning Algorithm. We start with the global characteristics of the algorithm. In subsequent sections of it, the different stages of the algorithm are discussed in details along with some analytic results on the complexity of the proposed algorithm. Application Design details and experimental results are given in the end. The final chapter makes a conclusion of the entire thesis, highlighting the new ideas presented and ends with a discussion on the directions for future research on this work.