CHAPTER (1)

Introduction

Chapter 1 Introduction

1.1 Research Motivation

Security is a major issue for all networks in commercial companies, the enterprise environment and organized business activities aim specifically at growth and profit. Security in general means the state or feeling of being safe and protected. Hackers and intruders have made many successful attempts to bring down the high-level company networks and its Web services. Unfortunately, these companies cannot stop using Internet.

Due to the expansion of high-speed Internet access, the need for secure and reliable networks is more critical. Sophistication of network attacks as well as their severity has also increased recently, so more and more organizations are becoming vulnerable to potential attacks. The geographic spread of Sapphire/Slammer Worm 30 minutes after release (2003) is shown in Figure 1-1.

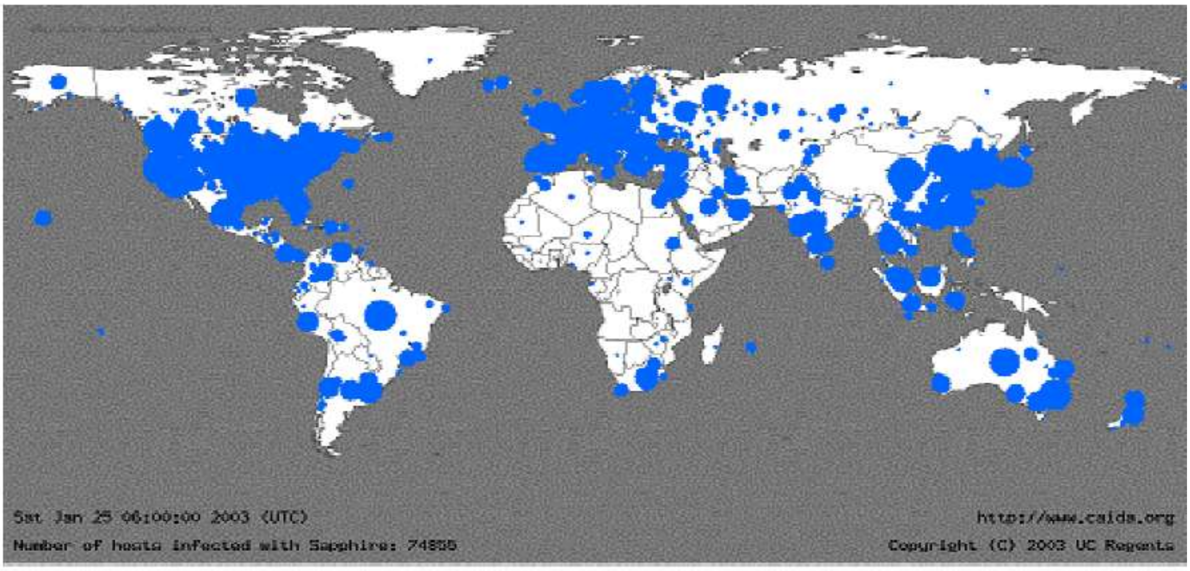
Sat Jan 25 06:00:00 2003 (UTC)
Number of hosts infected with Sapphire: 74855
http://www.caida.org
Copyright (C) 2003 UC Regents

Figure 1-1 The spread of Sapphire/Slammer Worm 30 minutes after release.

Intrusion detection is a very important issue that has to be addressed in order to protect our data and our privacy, etc. It is a program and / or a device designed to detect unwanted attempts to access the computer system or try to disable the system in general and manipulate it, through network, such as the Internet. These attempts can use several forms of attacks, for example, by breaking the protection on the rights of copying or printing, or the use of malicious software and / or use of disaffected staff on their company or similar piece.

1.2 Literature Review

Many researches have been performed by using artificial neural networks (ANNs). In [2], artificial neural networks and support vector machine (SVM) algorithms were applied for intrusion detection (ID) with frequency-based encoding method; in the chosen DARPA data set, the authors used 250 attacks and 41,426 normal sessions and the percentage of detection rate (DR) varied from 100 % to 43.6 % with the percentage of false positive rate (FPR) from 8.53 % to 0.27 % using different settings. In [3], the author describes and concludes that the combination of radial basis function (RBF) and self-organizing map (SOM) is convenient to use as an intrusion detection model. He concludes that the ‘‘evaluation of human integration’’ is necessary to reduce the classification error. The reported experimental results show that RBF–SOM achieves, similar or even better results, compared to RBF. In [4], the authors use hierarchical (SOM) and conclude that the best performance is achieved using a two-layer SOM hierarchy, based on all 41-features from the KDD data set and the ‘Protocol’ feature provides the basis for a switching parameter. The detector provides FPR and DR of 1.38 % and 90.4 % respectively. In [5], the author uses a hierarchical ID model using principal component analysis (PCA) neural networks and results 97.1 % DR and somewhat higher 2.8 % FPR. In [6], a critical study about the use of some neural networks (NNs) is used to detect and classify intrusions, the percentage of DR is 93.83 % on PCA, and the percentage of FPR is 6.16 % on PCA. In [7], the authors present a biologically-inspired computational approach to dynamically and adaptively learn signatures for network intrusion detection using a supervised learning classifier system. The classifier is an online and incremental parallel production rule-based system.

It was noted that most of the previous systems concentrate on either detecting two categories (normal or attack) or detecting a certain category of attack. Also most of the previous works ignore the symbolic features of the KDD cup 1999 data set which adversely affected the accuracy of the detection. In this work, this study suggested a back-propagation neural networks intrusion detection system (BPNNIDS) and radial basis function neural networks intrusion detection system (RBFNNIDS), both can work on either two categories or multi categories detection and at the same time does not ignore the symbolic features.

1.3 Aim of the work

The aim of our research is to classify the network attacks using neural networks (NN) which leads to a higher detection rate in less time. This study focuses on two classification types of records: a single class (normal, or attack), and a multi class (normal, DoS, PRB, R2L, U2R), where the category of attack is also detected by the NN. Extensive analysis is conducted in order to assess the translation of the symbolic data, the partitioning of training data and the complexity of the architecture. The back-propagation neural network intrusion detection system (BPNNIDS) and the radial basis function neural network intrusion detection system (RBFNNIDS) proposed in this thesis are tested against traditional and other machine learning algorithms using common data set: the DARPA 98 KDD99 benchmark data set from the International Knowledge Discovery and Data Mining Tools [8]. BPNNIDS shows superior response compared to other techniques reported in literature especially regarding time, false positive rate (FPR) and human interaction. .

1.4 Contributions

In this thesis, supervised learning approach to the intrusion detection problem is investigated and demonstrated on the International Knowledge Discovery and Data Mining Tools Competition intrusion detection benchmark (the KDDCUP99 data set). To do so back-propagation neural networks intrusion detection system (PBNNIDS) architecture and radial basis function are investigated [1] under two basic data sets; one is limited to 19361 connections (records), which is our selected version whereas the other contains 494021 connections (records), which is the 10% version. The significance of our reduction algorism is to reduce the training time from 40 minutes 18 seconds to 3 minutes 27 seconds in the five categories system, from 48 minutes to 3 minutes and 4 seconds in the two categories system.

The proposed iterative reduction algorithm, encoding of the symbolic features and the complexity architecture of the proposed back-propagation neural network have the great effect on ensuring a high detection rate with a low false positive rate.

This thesis introduced two ways of machine learning in the supervised environment. The first one depends on the back-propagation neural networks and the second one depends on the radial basis function. By comparing the training time, the detection rate and the false positive rate it is concluded that the engine with the back-propagation neural networks produced better results than the one using the radial basis function.

1.5 Thesis Outline

The rest of this thesis follows the structure listed below:

Intrusion detection systems

Chapter 2 is divided into two parts. In the first part, the thesis discusses the basic Intrusion detection terminology, shows the difference between the fire wall and the intrusion detection system and also the locations for the intrusion detection system and explains shortly the structure and the elements of the typical intrusion detection system, in the second part, the thesis explains the different taxonomy of intrusion and also the different taxonomy of intrusion detection system and the pros and cons of each IDS using these categories.

Machine learning and neural networks

Chapter 3 is divided into three parts. In the first part, the thesis introduces the meaning of the machine learning concept and the different classes of the machine learning. In the second part, the thesis introduces the error-propagation method used in the back-propagation algorithm and also the pros and cons of the back- propagation algorithm. In the third part, the thesis introduces the radial basis function as an alternative approach to the back-propagation.

The DARPA 98 KDD99 benchmark data set

Chapter 4, the thesis shortly explains the structure of the network used in to evaluate the performance of the system introduces in that thesis. It also shows the parameter of the DARPA 98 KDD99 benchmark data set from the International Knowledge Discovery and Data Mining Tools and the distributions of the data files.

The proposed system architecture

Chapter 5, the thesis describes the proposed system architecture. Starting with the system design philosophy, and go step by step in the architecture stages. This chapter also explains the proposed iterative reduction algorithm and encoding of the symbolic features.

Experimental results and analysis

Chapter 6, the thesis shows the complexity architecture of the proposed back-propagation neural network which has the great effect on ensuring a high detection rate with a low false positive rate. The thesis works under two datasets for the training process. The thesis works under two engines one uses back-propagation neural network and the other uses the radial basis function neural network. We compared the two engines working under the two datasets with each other and with other machine learning approaches reported in literature.

Conclusion and future works

Chapter 7, the thesis presents the conclusions of the work. It also discusses some future directions.