University of Science and Technology

College for Graduate Studies & Academic Advancement

COMPARTAIVE STDUY BETWEEN MACHINES LEARNING ALGORITHMS SUPPORT VECTOR MACHINE, MULTIPLE LINEAR REGRESSIONS FOR WORM DETECTION

Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Computer Science

By

Maha Adam Gumaa osman

Supervisor

Dr.Altyeb Altaher

September 2012

الخلاصة :

الديدان عبارة عن برمجيات خبيثة تنتشر عبر الانترنت من دون تدخل الانسان ، بما ان الديدان تنتشر بسرعة اكبر من استجابة الانسان ، فأن الحماية الحيوية من الديدان تكون بجعل اكتشافها يتم تلقائيا .

انظمة اكتشاف التسلل تعمل علي اكتشاف الديدان باختبار الحزم او البيانات المنسابة لمعرفة مدى تطابقها مع التوقيعات المعروفة .هذا يشير الي ان انظمة اكتشاف التسلل لا تستطيع اكتشاف الديدان الجديدة الغير معروفة .طفرات الديدان المعروفة تظل غير مكتشفة وذلك لان كل ظفرة يكون لها في العادة توقيع مختلف .الحل البديهي والفعال علي ما يبدو هو كتابة توقيعات عامة لكن هذا ادى الى زيادة معدلات الانذارات الكاذبة ، لذلك الحل غير فعال.

هذا البحث يختبر دقة وامكانية استخدام تقنيات تعلم الالة لاكتشاف ديدان الانترنت بمقارنة خوارزمية (الة الدعم الموجة) و خوارزمية (الانحراف الخطي المتعدد) لمعرفة ايهما اكثر دقة في اكتشاف ديدان الانترنت ، وجدنا ان خوارزمية (الة الدعم الموجة) اكتشفت ديدان الانترنت بدقة اعلى من خوازمية (الانحراف الخطي المتعدد).

Abstract:

Worms are malicious programs that propagate over the Internet without human intervention. Since worms generally spread faster than humans can response, the only viable defense is to automate their detection.

Network intrusion detection systems typically detect worms by examining packet or flow logs for known signatures. This indicates that it cannot detect the new unknown worms. Those mutations of known worms will remain undetected because each mutation will usually have a different signature. The intuitive and seemingly most effective solution is to write more generic signatures, but this has been found to increase false alarm rates, so this solution.

This research examines the accuracy and feasibility of using machine learning techniques for detecting Internet worms. By comparing support vector machine algorithm (SVM) and multiple linear regression algorithms (MLR) to see whichever is higher accuracy in detecting the Internet worms. We found that the SVM algorithm detected the Internet worms with higher accuracy than the MLR algorithms on the tested dataset.

1 Introduction:

Computers play an important role in our daily live, as well as it has become an integral part of the basic elements in the work environment. We do not use computers in the workplace only, but also we use it at home. Although this machine that has made our live easier and more exciting effective. But it has become unsafe as a result of the misuse. Where computers have become vulnerable to malicious software such as viruses, worms, Trojan horses and other, especially worms because worms are small programs that stand alone and not depend on the other applications to propagate on computer network, worms disadvantage include quickly spread and difficult to get rid of them because of its unique ability to coloration and reincarnation and Shuffle [1, 2]. And this makes it spread wider and faster than the virus, thus causing great damage to the computers on the network. Where disruption of system resources that would lead to slower performance, and therefore had to be discovered so that we can try to get rid of them and the protection of computers connected with each other through the network, including, and ensure the best performance.

The initial Internet worm was released in 1988 and brought down hundreds of devices across the USA, [3] at the time **an** important portion of the early Internet. Worms proliferated as the Internet matured into a global network, wreaking havoc and causing ample financial damage. None of the devastation, anyhow, came close to the $2.6

billion caused by Code Red. Code Red used vulnerability in Microsoft's Internet Information Services (IIS) web server to infect its sufferers. The first, rather failed version of Code Red [4, 5] attempted to spread itself by generating a set of random IP addresses that it then tried to infect. However, there was a final flaw in this version it used a static kernel to generate the IP addresses , which meant that all defiled hosts generated the same set of IP addresses.

This flaw stopped the worm from extending far**.** Some days later Code Reds coming a variation in its behavior was watched it started to probe new hosts. The change in behavior was due to an improved version of Code Red, self-same in all aspects except for the random.

This mutation enabled it to infect 359, 000 hosts in less than 14 hours**.** A worm that spread even faster was the Slammer worm, which infected most of its 75, 000 victims within 10 minutes [6]. This worm was the first Warhol worm observed in the wild, a name coined from Andy Warhol's renowned quote that "in the future; everybody, will have 15 minutes of fame, and based on the worm's ability to extend to most in danger machines within15 minutes.

In the face of their prominence, Code Red and Slammer are just two of the more disgraceful worms drawn from the large pool of lethal worms that have swamped the Internet over the last decade. Study of those worms' lead to the following observations about worm behavior [6]:

- The initial release of each worm is typically followed by one or more mutations.
- Each mutation tends to be more lethal than its predecessors, by refining the attack or infection strategy.
- They spread significantly faster than humans can respond.

These observations suggest that an attractive defense strategy against worms. Is to automate the detection of their mutations, and this strategy is the focus of this thesis.

1.2 Background:

Before Internet access became widespread, viruses spread on personal computers by infecting the executable boot sectors of floppy disks. By inserting a copy of it into the machine code instructions in these executables, a virus causes itself to be run whenever a program is run or the disk is booted. Early computer viruses were written for the Apple II and Macintosh, but they became more widespread with the dominance of the IBM PC and MS-DOS system. Executable-infecting viruses are dependent on users exchanging software or boot-able floppies, so they spread rapidly in computer hobbyist circles**.** The first worms, network-borne infectious programs, originated not on personal computers, but on multitasking UNIX systems. The first well-known worm was the Internet Worm of 1988, which

infected SunOS and VAX BSD systems. Unlike a virus, this worm did not insert itself into other programs. Instead, it exploited security holes (vulnerabilities) in network server programs and started itself running as a separate process. This same behavior is used by today's worms as well. With the rise of the Microsoft Windows platform in the 1990s, and the flexible macros of its applications, it became possible to write infectious code in the macro language of Microsoft Word and similar programs. These macro viruses infect documents and templates rather than applications (executables), but rely on the fact that macros in a Word document are a form of executable code [4].

Today, worms are most commonly written for the Windows OS, although a few like Mare-DS and the Lion worm are also written for Linux and UNIX systems. Worms today work in the same basic way as 1988's Internet Worm: they scan the network and leverage vulnerable computers to replicate. Because they need no human intervention, worms can spread with incredible speed. The SQL Slammer infected thousands of computers in a few minutes [4].

1.3 Intrusion Detection:

With the development of networking technologies and applications used, Network attacks are increasing significantly in both number and how., intrusion detection system (IDS) play an important role for the detection of different types of

malicious software, and networks safe .The main purpose of IDS is to find out penetrations between the data which enables to detect, prevent and possibly a reaction to the attack [6].

Intrusion detection is a set of techniques and methods that are used to detect suspicious activity both at the network and host level. Intrusion detection systems fall in to two basic categories: signature-based intrusion detection systems and anomaly detection systems. Intruders have signatures, like computer viruses, that can be detected using software.

Try to find data packets that contain any known intrusion-related signatures or anomalies related to Internet protocols. Based upon a set of signatures and rules, the detection system is able to find and log suspicious activity and generate alerts. Anomaly-based intrusion detection usually depends on packet anomalies present in protocol header parts. In some cases these methods produce better results compared to signature-based IDS. Usually an intrusion detection system captures data from the network and applies its rules to that data or detects anomalies in it

1.4 Research Problem:

Worms are the most dangerous types of malware currently circulating, and that because it does not need human intervention to spread, despite the development of techniques for the detection of worms, but the early detection of new worms is still

a problem. In this research we used apply machine learning techniques for worm's detection by using Support Vector Machines, multiple linear regressions.

1.5 Research Objective:

The contribution of the thesis is the following:

- Comparing how effectively the Support Vector Machines, multiple linear regressions, Machine learning techniques detect worm mutations.
- Show the advantages and disadvantages of each algorithm.