UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADEMIC ADVANCEMENT

Measure of Decision Tree Algorithms for Intrusion Detection System using NSL-KDD Dataset

By

Mustafa Abd Elaziz Elhaj Ahmed

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements
in Partial Fulfillment of the Requirement for the Degree of Master of Science in
Computer Science

Supervisor

Prof. Dr. Noureldein Abd elrahman Noureldein

December 2016

Abstract

In the current era of Internet, network security technology has become crucial in protecting the computing infrastructure on the network, Intrusion detection is an essential mechanism to protect computer systems from many attacks.

Computer network attacks have been increased exponentially due to internet growth.

Intrusion Detection system (IDS) have been reported by many researchers as major security technology for attacks detection.

In this research the problem is to detecting anomalies throw network traffic by measure the performance of Decision Tree machine learning algorithm in detecting various types of attacks using NSL_KDD dataset.

Testing and training is done based on two validation strategies,10 fold cross validation and Percentage Split.

The performance of algorithms is measured using time taken to build model, correctly classified instances, incorrectly classified instances, root relative squared error and accuracy.

The result show that the Decision Tree algorithm Random Forest is the best classifier model for IDS detection attack using NSL_KDD data set.

Finally, avenues for future researches are presented.

المستخلص

في العصر الحالي من الإنترنت أصبحت تكنولوجيا أمن الشبكات أمراً حاسماً في حماية البنية التحتية لشبكات الحاسوب. أن كشف التسلل هو آلية أساسية لحماية أنظمة الكمبيوتر من العديد من الهجمات التي ذادت على شبكات الكمبيوتر بشكل كبير نتيجة للنمو المطرد للإنترنت.

تم التطرق لنظام كشف التسلل من قبل العديد من الباحثين .في هذا البحث المشكلة هى الكشف عن مختلف انواع الهجمات فى حركه البيانات فى الشبكة ذلك من خلال قياس ومقارنة اداء خوارزميات شجرة القرار باستخدام قاعدة البيانات NSL_KDD.

تم الاختبار والتدريب باستخدام اثنين من استراتيجيات التحقق: ١٠ fold cross validation و Percentage split.

تم قياس أداء الخوارزميات باستخدام الوقت المستغرق لبناء النموذج، الحالات المصنفه بشكل صحيح، الحالات المصنفه بشكل غير صحيح، الجذر التربيعي النسبي للخطاء والدقة.

وقد أظهرت نتائج تحليل المقارنات أن (Random Forest) هو أفضل خوارزمية تصنيف للكشف عن الهجمات بإستخدام قاعدة البيانات (NSL_KDD). ختاماً تم وضع أطر لما يمكن تقديمه من بحوث مستقبلية.

1.1 Introduction

Internet is largely used in government, military and commercial institutions. The new emerging protocols and new network architectures permit to share, consult, exchange and transfer information from any place all over the world to any other one situated in different country. Despite the above progress, the actual networks are becoming more complex and are designed with functionality while security is not considered as a main goal.

The concept of Intrusion Detection System (IDS) proposed by Denning (1987) is useful to detect, identify and track the intruders. An intrusion detection system (IDS) is a device or software application that monitors network or system activities for malicious activities or policy violations and produces reports to a management station. The intrusion detection systems are classified as Network based or Host based,the network based attack may be either misuse or anomaly based. The network based are detected from the interconnection of computer systems. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection [1]. It helps in to classify the attacks to measure the effectiveness of the system.

Classification is the process of finding the hidden pattern in data. With the use of classification technique it is easy to estimate the accuracy of the resulting predictive model.

In this research compared five decision tree algorithms in WEKA tool by classifying the data set NSL_KDD with two validation strategies:10 fold cross validation and percentage split 66 % to find the best method from decision tree by using the result analysis of them, and also see the performance detection accuracy.

1.2 Problem Statement

Technology now allows us to capture and store big data. Finding patterns and anomalies in these datasets, and detect attack, now a day When you want to do any things we use network, for that IDS are become very necessary for detecting attack in network . The problem of this research is detecting anomalies throw network traffic by find the best algorithms in decision trees algorithms on detecting various types of attacks using NSL_KDD dataset by measuring and camper them using Time taken to build model ,Correctly Classified Instances, Incorrectly Classified Instances ,Root relative squared error and accuracy.

1.3. Research 0bjective

algorithms decision tree machine learning algorithms are commonly used to detect anomalies in network traffic. Recently, many research studies are focus on determining the optimistic performance of an algorithm. The objectives of the research is To measured and compare different data mining decision tree classifier models in detecting anomalies.

- Review the main concepts, definitions, and terms of Intrusion Detection Systems (IDSs) and main types of Intrusion Detection Systems (IDSs).
- Explore the main types of attacks that threat network.
- To carry out dataset processing and evaluate classifiers accuracy on varying proportions of the dataset.
- Determine the most optimal method.

1.4Research Questions

The questions of this research are:

1. How can we detected the different attack in network traffic by using decision trees algorithms ?
2. What are the performance matrix can be used to measure the performance of machine learning algorithms in detecting attack using decision trees algorithms ?
3. What are the best classifiers that provide higher accuracy and reduced runtime?.
4. -How can we validate the detecting performance results of machine learning decision trees algorithms ?

1.5 Motivation

computer security is the heart of today’s technological in world , intrusion detection is one of core areas of network security that needs to be highly effective.

Any unauthorized access to the networks means lot of problems, we need to think about securing network that begins with detecting of any intrusion in the network.

Many IDS based on machine learning and data mining have been proposed, but which machine learning algorithms is more efficient in detecting anomalies is still a challenging question.

1.6 Research methodology

The methodology of this research using data mining and machine learning language WEKA, test and train NSL_KDD with the decision tree class classification algorithms using two validation strategies 10 fold validation and percentage split 66 %, measure the performance of the 5 decision algorithms through Time taken to build model ,Correctly Classified Instances, Incorrectly Classified Instances ,Root relative squared error and accuracy, camper the result, more details in chapter 3.

1.7 Thesis Structure

This thesis consists of five chapters as follows: Chapter 2 explains what Intrusion Detection Systems is, Anomaly detection systems, Misuse detection systems, specification-based detection, type of IDS, data mining technique, machine learning technique and focus on decision trees classifier and network attack. Chapter 3 shows research methodology that have been used in this research, Chapter 4 shows the experimental analysis and results of two validation applied on Decision Tree algorithms in detecting attacks, Chapter 5 contains the conclusion and the future work that should be done in thesis.