Republic of Iraq

Ministry of Higher Education and
Scientific Research

University of Diyala

College of Science

Department of Computer Science

Enhancement of Intrusion Detection in NSL-KDD Dataset using some machine learning techniques

A Thesis Submitted to Council of College of Science, University of Diyala in Partial Fulfillment of the Requirements for the Degree of Master in computer science

By

Amar Ahmed Othman

Supervisor by

Asst. Prof. Dr. Taha M. Hasan

Asst. Prof. Dr. Safwan O. Hasoon

Aug 2020 IRAQ Dhul Hijjah,1441

بسم الله الرحمن الرحيم

إِنَّ فِي خَلْقِ السَّمَاوَاتِ وَالْأَرْضِ وَاخْتِلَافِ اللَّيْلِ وَالنَّهَارِ لَآيَاتٍ لِأُولِي الْأَلْبَابِ
﴿١٩٠﴾ الَّذِينَ يَذْكُرُونَ اللَّهَ قِيَامًا وَقُعُودًا وَعَلَى جُنُوبِهِمْ وَيَتَفَكَّرُونَ فِي خَلْقِ
السَّمَاوَاتِ وَالْأَرْضِ رَبَّنَا مَا خَلَقْتَ هَذَا بَاطِلًا سُبْحَانَكَ فَقِنَا عَذَابَ النَّارِ
﴿١٩١﴾

آل عمران (١٩٠–١٩١)

صدق الله العظيم

Acknowledgement

Praise be to Allah, Lord of the worlds, and prayers and peace be upon the master of the messengers, our Prophet Mohammed and his family and companions. I would like to express my thanks and appreciation to my supervisors, Dr. Taha M. Hasan and Dr. Safwan O. Hasoon for their faithful guidance, valuable instructions, and constructive comments which have made the completion of this work possible.

Also, I would like to express my gratitude and my thanks to all the Lecturers staff who have taught me. Who provided us with all the care and assistance. Special thanks to the members of the evaluation committee for discussing my thesis, Special thanks are extended to all my friends for their help. I also extend my sincere thanks and gratitude to my friend Dr. Saad Al-Jumaili for his good efforts and continuous support. And thank for Mr. Khalil Al-Karkhi, who provided us with all the care and assistance. My final words go to my family. I want to thank my wife and my children for their love and bear the difficult circumstances that accompanied the completion of this research and helping me in achieving my goals.

AMAR

Dedication

I would like to dedicate this Work To:

The soul of my father

my dear mother, whom I ask Allah to protect

my brothers and sisters, whom I cherish most

my life partner and my road companion my dear wife

the reason for my joy in this life, my children, Rafef and Yazen

AMAR

Abstract

Machine learning today is widespread and probably can be used many time a day, without even realizing. So several scholars believe that machine learning is the best way to advance Artificial intelligence(AI), and machine learning systems must-have qualities of proficiency, especially with the huge amount of data that imposes limitations on the functioning of machine learning systems for achieving the purpose of using it.

This thesis Enhancement of Intrusion Detection in NSL-KDD Dataset using some machine learning techniques investigates intrusion detection implementation on the NSL-KDD dataset, Which is considered as the most important type of intrusion detection data. The proposed system includes three main steps, pre-process (encoded data, normalization), the dimension reduction using two technique Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and classification using Naive Bayesian (NB), Decision Tree (J48), The classification step divides the data into two sections (the training section include 70% of all the data and test section include 30% of all the data), then compare results of classification accuracy and effect of the proposed system steps.

The results of the comparison between the related techniques and algorithms based on the accuracy rate shown the proposed system is succeeded in achieving high accuracy of 99.72% of using normalization step, dimensional reduction technology PCA with 25 eigenvectors, and a j48 classifier. Classification accuracy is achieved 99.291% using the normalization step, dimensional reduction technology PCA with 5 eigenvectors, and a j48 classifier. This confirms a large data shortcut that guarantees efficient operation, non-degradation of the system, and high classification accuracy.

List Of Contents

List of Figures

List of Tables

List Algorithms

List of Symbols

Symbol	**Meaning**
*	Multiplication operation
+	Addition operation
/	Division operation
-	Subtraction operation
C#	C sharp
λ	eigenvalue
σ	Standard Deviation
$\forall j$	for all
$P(A \vert B)$	conditional probability function
μ	mean
&	AND
$\vert$	OR
V	Eigenvector
$\prod$	capital pi
P(A)	probability function
A^T	transpose
$P(A \cap B)$	probability of events intersection
cov(X,Y)	covariance
$\bar{x}$	sample mean
Σ	Summation - the sum of all values in the range of series

$\lvert X \rvert$	The absolute value bars
Log_2	Logarithm **base 2**
%	modulus
$\oplus$	Circled plus / oplus - xor
S	sample standard deviation
$\overrightarrow{x}$	Vector
det(A)	Determinant
A^{-1}	inverse matrix
I	Identity matrix

Chapter One

Introduction

Chapter One

Introduction

1.1 Overview

Computers and networks that link to it have become the target of computer crime, which has increased greatly due to the spread of uses and their connection to all areas of life, money, and business. Some techniques have been used to protect important data such as firewalls and encryption. Any new security technology contains some defects, in design which makes it the target of attacks and penetrations [1]. Accordingly, it became necessary to monitor and protect network security infrastructure to detect intruders or any kinds of intrusions with an intrusion detection system (IDS). The main components of IDS are event generator, analyzer, and response module [2].

There are two major categories of IDS: signature-based IDS and anomaly-based IDS. Signature-based IDS looks for characterized designs inside the examined network traffic. On the opposite side, the anomaly-based IDS" is able to evaluate and predict system behaviour. The signature-based IDS clarifies a very good performance only for specified well-known attacks. On the contrary, the anomaly-based IDS. Shows the ability to detect invisible offside events, which is an important feature for detecting attacks of zero-day. Anomaly-based IDS can be gathered into three major kinds [3]: statistical-based approaches, knowledge-based approaches, and machine learning-based approaches [5].

Machine learning (ML) strategies can predict and recognize threats before they outcome main security accidents [4]. Data mining is the way of extracting interesting data from huge data sets utilizing techniques of ML machine learning [3]. There are different methods available for data mining like association rule learning, clustering, classification, regression, and summarization [6].

Classification can be termed as the process of assigning items in a collection to predefined instances or classes. Classification represents supervised learning because classes are determined before data is examined. Classification of instances into two classes is called binary classification and multi-class classification means a classification of instances into three or more classes [4]. Therefore, researchers have directed to use of smart concepts to solve intrusion detection problems, The designers of intrusion detection systems faced on reducing features in a dataset where they are some of the features of attacks are not important and distort the classification and prediction process for future attacks, In addition to the fact that these dimensions require time to process them and increase the speed of response [3].

To reduce the complexity of computation time and obtain better classification accuracy, researchers use various methods of dimensions reduction. Principal Component Analysis (PCA) and Linear Discriminant Analysis(LDA) are considered as the most common and efficient dimensions reductions[4]. PCA is an orthogonal transformation statistical technique. PCA transforms a group of associated variables into an uncorrelated set. PCA is used for the study of exploratory data, for example, .it can also be used to analyze the relationships between several variables. Furthermore, PCA may also be used to reduce dimensionality. LDA is another common dimensional reduction technique for projecting a dataset with high features into a smaller space with a good quality-separability. That will decrease the computational expenses [7,8]. On basis of Standards such as KDD99 and NSL-KDD, malevolent activities (attacks) in networks are divided into four groups [9]:

- **DoS:** Server denial, an intruder attempts to block authorized users from accessing the server. (e.g. flood SYN).
- **Probe:** An intruder attempts to acquire information about the destination host such as ports scan.

- **R2L:** Remote to Local, intruders are trying to get remote access to target computers like password guessing for brute force.

- **U2R:** User to Root, intruders have local access to the target computer and seek to obtain superuser rights such as privilege increases.

Some studies aim to achieve the average detection rate (classification accuracy) without taking the value of the attacks into account. Because U2R and R2L attacks are established, they can be dangerous compared with other forms because they are relatively uncommon in sampling and analysis, and can also cause serious harms. [9].

1.2 Related Works

Researchers have proposed a variety of related works about enhancing the robustness of machine learning systems via dimensionality reduction. The following are some studies and researchers that can so far associate works to the proposed model of this thesis:

1. **Lakhina, S., et.al. (2010)[8]:** They suggested PCANNA Hybrid Algorithm (primary neural network component analysis algorithm) have been used to reduce the amount of computer resources, memory, and processor time required for detecting attacks. The PCA transformation is used to reduce the functionality and a qualified neural network to recognize new attacks. The experiments with NSL-KDD data show that a proposal model improves and provides robust data representation, since features have reduced data by 80.4 percent, a training time decrease of approximately 40 percent, and a time reduction of 70 percent have been achieved, also increases the classification accuracy.

2. **Revathi, S., & Malathi, A. (2013) [10]**: They introduced an NSL-KDD dataset analysis that solves some issues related to the KD99 cup99 data. The analysis shows that NSL-KDD dataset is very ideal for comparing different intrusion detection models. Using all the 41 features in the dataset to evaluate the intrusive patterns may lead to time-consuming and it also reduces performance degradation of the system. Some of the features in the dataset are redundant and irrelevant for the process. In this work, Correlation-based Feature Selection (CFS) is used to reduce the dimensionality of the dataset and compared between the different types of classifier algorithm (Random Forest, J48, SVM, CART, and Naïve Bayes) based on results of accuracy ratio for two cases (41 Features and 15 Features). The results shown Random Forest has highest accuracy ratio (Normal =99.1,Dos=98.7,Probe= 97.6,U2R=97.5,R2L=96.8)for 41 Features and (Normal =99.8 ,Dos =99.1 , Probe = 98.9 , U2R =98.7, R2L= 97.9) 15 Features.
3. **Dhanabal, L., & Shantharajah, S. P. (2015) [3]:** In this research, the researchers use the CFS approach for reducing dimensionality decreases detection time and improves precision in the identification of anomalies in network traffic patterns, presented NSL-KDD data package experiments used to test the efficacy of diverse classification algorithms (J48, SVM, and Naïve Bayes). Also, it analyzes the relationship between a protocol accessible in the typically used network protocol stack and anomalous network traffic attacked used by intruders. Algorithm J48 for better accuracy classification of NSL-KDD with six features "Normal 99.8, DoS 99.1, Probe 98.9, U2R 98.7, R2L 97.9".
4. **Assi, J. H., & Sadiq, A. T. (2017) [11]:** They suggested to identify network attacks using the data collection NSL-KDD, it provides five key classification methods with three feature selection policy. These methods are "J48 Decision Tree, Vector Support Machine (SVM), Decision Table

(DT), and Bayesian Network". The policies include the collection of the "correlation apps (CFS), information gain (IG), and the Decision Table". Several tests have been introduced to obtain good results with the general attack training and testing of NSL-KDD (Normal and Anomaly). Four types of attacks were performed: "Denial of service attack (DOS), User to root attack (U2R), Remote to Local attack (R2L) and Probing attack". The best result (80.3 percent) with test dataset and (93.9 percent) as a precision training dataset using the J48 classification approach.

5. **Bhagoji, A. N., et.al (2018)[12]:** They Proposing the use of data transformations in protection against Multiple Classifiers (ML) evasion attacks and investigating techniques for including dimensional reduction through PCA, to boost machine learning's resilience and emphasis both on classification and training. Experimental experiments and demonstrations demonstrate the feasibility of linear data transformations as a method of defence against evasion attacks using various data sets in the field. As a consequence, (i) security is efficient against various attacks, such as (Misconformity of classifications, white-box (optimal), white-box (FG), white-box (FGS), white-box (Opt.)), arch. The FC100-100-10 and Arch. Arch. Mismatch (FC100-100-10), (ii) "applicable across" a variety of ML classifiers, including SVM and Deep Neural Networks, and (iii) generalizable to various applications, including image classification and the classification of human behaviour.

6. **Abdulhammed, R., et.al.(2019) [4]:** In this research, use Auto-Encoder (AE) as well as PCA for dimension reduction and well-validated classifiers such as Random Forest (RF), "Bayesian Network (BN), Linear Discriminant Analysis (LDA) and Quadrant Discriminant Analysis (QDA)" are the main factors behind the creation of an anomaly-based IDS (intrusion detection system) machinery learning platform. The resulting small-dimensional characteristics of the two techniques are

used to construct various classification systems for designing intrusion detection systems, such as "Random Forest, Bayesian Network, Linear Discriminant Analytics (LDA) and Quadratic Discriminant Analytics (QDA)". This research effort will which the characteristic dimensions of the CICIDS2017 dataset from 81 to 10 while retaining a high precision of 99.6% in multiple and binary classifications.

7. **Sapre, S., Ahmadi, P., & Islam, K. (2019) [13**]: In this work, the KDDCup99 comparison and NSL-KDD intrusion detection network data sets were implemented using different machine learning methods. The comparison of both data sets by comparing the output of different forms of machine learning has been equipped with a broader variety of classification methods than former researchers. The results showed that the NSL-KDD dataset is higher than the KDDCup99, as its classifiers were 20.18 percent less accurate on average. This is because the classifiers trained in the KDDCup99 data set displayed a bias to their redundancies and allowed them to achieve better accuracy.
8. **Aljawarneh, S. et. al. (2019) [14]:** Researchers thought about this work to increase detection accuracy and efficiency of the new IDS technique, an improved J48 algorithm was developed. The improved J48 algorithm can help detect possible attacks that could threaten the confidentiality of the network. By separating it into two formats, an NSL KDD intrusion dataset was applied. A process of feature selection, based on the program WEKA, was then used to determine the effectiveness of all apps. The results show that this algorithm has provided a higher, precise, and more effective performance without using the above features compared with the function selection process. The implementation of this algorithm ensured the data set classification, based on a detection accuracy of 99.88% in all the functions, 90.01% for the supplied test set following the use of the full test data sets and all features, and 76.23% for delivery

of test sets after the use of the test-21 data set along with all of the features.

9. **Reddy, G. T., et.al (2020) [7]:** They Investigated the effect on ML algorithms of two groundbreaking techniques for reduction of dimension, namely Primary Component Analysis and Linear Discriminant Analysis. These techniques of dimensional reduction are used for Cardiotocography datasets, which are available in the UCI machine learning repository. This dataset consists of 36 dependent attributes. When 95 percent of components were retained by using PCA, the number of dependent attributes was reduced to 26. This reduced data set is trained by the use of four popular classification systems, a Decision Tree classification, a Naive Bayes classification, a Random Forest classification, and an SVM. The results demonstrate that the performance of PCA classifiers is better than that of LDA. Decision Tree and Random Forest classifiers also outdo the other two algorithms without reducing dimensionality and with both PCA and LDA.
10. **Bhattacharya, S., et.al.(2020)[15]:** Proposed a "Hybrid Principal Component Analysis (PCA)-firefly-based" machine learning model to classify IDS. Kaggle gathers the datasets used in the analysis. The system starts with the One Hot encoding method for the IDS dataset transformation. The transformed data is then exposed for the purpose of dimensional reduction to a hybrid PCA firefighter algorithm. The XGBoost algorithm is used to detect unexpected cyber-attacks on this that dataset. The findings show that in contrast with conventional machine learning methods, the method suggested is more accurate. The shortcomings in this work are growing in terms of time complexity as the reduction of dimension and classification training phases take longer.

In the proposed system, the data was pre-processed to solve the problem of the large divergence between the values resulting from the difference in the measurements of the features, and the dimensions were reduced using two different methods(PCA, LDA), which are the most prevalent in the field of dimensional reduction, and for achieving a real evaluation of the proposed techniques and algorithms and their method in Extraction of valid results the process of dimensional reduction was Conducting 27 stages to reduce dimensions to clarify the results of the repeated implementation of dimensional reduction with the change of the value of this reduction and to give a prediction of the behaviour of those techniques if they were used with the same data or with new data.

1.3 Problems statement

The thesis problem is deterioration and slowdown in the performance of some machine learning systems to detect intrusion as the NSL-KDD data set used in this work has huge dimensions, which may delay the rapid response of the request process. Therefore an effective mechanism must be used to reduce dimensions and to enhance intrusion detection systems.

1.4 Aims of the thesis

This thesis proposes intrusion detection models that detect whether network traffic is an attack or a normal network packet based on enhancement machine learning. The objectives are illustrated in what follows:

1- Construct a set of fast and robust classifiers (NB and J48) that has the ability fast and robust to distinguish between normal network packet and attack network packet by using appropriate data pre-processing and dimensional reduction.

2- Categorize the attack network packets according to their types.

3- Conduct a comprehensive investigation to study the effects of dimensional reduction techniques on the results of the classifier accuracy rate.

1.5 Layouts of Thesis

The Thesis is organized into five chapters.

Chapter One Includes the basic introduction, aim of the thesis, related works, and the layout of the thesis.

Chapter Two This chapter includes theoretical background and discusses the algorithms that be use

Chapter three This chapter presents the details of the proposed Enhancement of Intrusion Detection in NSL-KDD Dataset using some machine learning techniques in the implementation of each one.

Chapter four This chapter presents the experimental results and evaluation.

Chapter Five: This chapter offers conclusions and suggestions for future work.