Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 125 (2018) 709–716

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on Smart Computing and Communications, ICSCC 2017, 7-8 December 2017, Kurukshetra, India

Statistical analysis of CIDDS-001 dataset for Network Intrusion Detection Systems using Distance-based Machine Learning

Abhishek Verma[a,*], Virender Ranga[a]

[a]*Department of Computer Engineering, NIT Kurukshetra, India*

Abstract

A lot of research is being done on the development of effective Network Intrusion Detection Systems. Anomaly based Network Intrusion Detection Systems are preferred over Signature based Network Intrusion Detection Systems because of their better significance in detecting novel attacks. The research on the datasets being used for training and testing purpose in the detection model is equally concerned as better dataset quality can advance offline Intrusion Detection. Benchmark datasets like KDD99 and NSL-KDD cup 99 are outdated and face some major issues, which make them unsuitable for evaluating Anomaly based Network Intrusion Detection Systems. This paper presents the statistical analysis of labelled flow based CIDDS-001 dataset using *k*-nearest neighbour classification and *k*-means clustering algorithms. The analysis is done with respect to some prominent evaluation metrics used for evaluating Network Intrusion Detection Systems including Detection Rate,Accuracy and False Positive Rate.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications

Keywords: Anomaly, Signature, Datasets, Labelled flow, *k*-nearest neighbour classification, *k*-means clustering, Analysis, Metrics

1. Introduction

Network security has become one of the most concerning problems for internet users and service providers with drastic increase in the internet usage [1]. A secure network can be defined in terms of its hardware and software protection against various intrusions. A network can be secured by implementing a resilient monitoring, analysis and defence mechanisms. Network Intrusion detection systems (NIDS) [2] forms a class of systems which implement these mechanisms in order to defend a network from insider and outsider intrusions. These systems monitor the incoming and outgoing traffic in a network, perform time to time analysis and report when some intrusion is detected. NIDS can be broadly categorised into Misuse detection (MD) [3], Anomaly Detection (AD) [4]. MD based NIDS use signatures or patterns of already existing attacks to detect intrusions. While AD based NIDS check for strict deviations from the normal profiles of the network traffic and report it as attack. MD based NIDS have a fast Detection Rate (DR) with less False Positive Rate (FPR) as compared to AD. However AD based NIDS are able to detect novel attacks over networks and this property outperforms them over MD based NIDS. MD works on the offline data whereas AD works better on the online data. Machine Learning (ML) [5] is playing a major role in the development of better NIDS.

* Corresponding author
E-mail addresses: abhishek_6170034@nitkkr.ac.in (Abhishek Verma)., virender.ranga@nitkkr.ac.in (Virender Ranga).

1877-0509
Peer-review under responsibility of the scientific committee of the 6th International Conference on Smart Computing and Communications
10.1016/j.procs.2017.12.091

It involves a system to learn from the recorded traffic patterns or signatures and then act accordingly for upcoming traffic patterns. Training and testing are the two major task involved in the ML. ML requires a large and complex datasets comprising of different types of normal and abnormal traffic patterns. There is also a need of applying ML algorithms to NIDS which have a less computational time and space complexity for a better learning. In this work we have analysed CIDDS-001 dataset using some prominent NIDS evaluation metrics like DR, FPR, Accuracy, Precision and F-measure [6]. We have used distance-based ML models [7] like *k*NN classification algorithm [8] due to its better DR and *k*-means clustering [9] due to its fast execution time.

1.1. CIDDS-001 Dataset

CIDDS-001 (Coburg Network Intrusion Detection Dataset) [10] is a labelled flow [11] based dataset developed for the evaluation of Anomaly based NIDS. This dataset contains unidirectional NetFlow data. It consists of traffic data from two server's i.e. OpenStack and External server. The dataset is generated by emulating small business environment which consist of OpenStack environment having internal servers (web, file, backup and mail) and an External Server (file synchronization and web server) which is deployed on the internet to capture real and up-to-date traffic from the internet. The dataset consists of three logs files (attack logs, client configurations and client logs) and traffic data from two servers where each server traffic comprises of 4 four week captured traffic data. The CIDDS-001 has 14 attributes out of which 12 have been used in this empirical study. This dataset consists large number of traffic instances out of which 153026 instances from External Server and 172839 instances from OpenStack Server been used for the analysis. Table 1 provides the description of CIDDS-001 dataset attributes .

Table 1: Classwise detail of CIDDS-001 dataset attributes

Sl. no.	Attribute Name	Attribute Description
1	Src IP	Source IP Address
2	Src Port	Source Port
3	Dest IP	Destination IP Address
4	Dest Port	Destination Port
5	Proto	Transport Protocol (e.g. ICMP, TCP, or UDP)
6	Date first seen	Start time flow first seen
7	Duration	Duration of the flow
8	Bytes	Number of transmitted bytes
9	Packets	Number of transmitted packets
10	Flags	OR concatenation of all TCP Flags
11	Class	Class label (Normal, Attacker, Victim, Suspicious and Unknown)
12	AttackType	Type of Attack (PortScan, DoS, Bruteforce, PingScan)
13	AttackID	Unique Attack id. Allows attacks which belong to the same class carry the same attack id
14	AttackDescription	Provides additional information about the set attack parameters (e.g. the number of attempted password guesses for SSH-Brute-Force attacks)

1.2. Objective

The objective of this research work is to perform analysis of the CIDDS-001 dataset from the ML point of view. In this work we study how the classification and clustering algorithms perform on the dataset considering 12 important attributes. Our objectives include classifying and clustering Network traffic of OpenStack and External Servers into $Normal, Attacker, Victim, Suspicious$ and $Unknown$ classes. We have used some prominent metrics for evaluating *k*NN classifier and shown Confusion Matrix generated from *k*-means clustering.