Intrusion Detection System in Computer Networks Using Machine Learning Techniques

Tejas Srinivas

W1385218

Table of Contents

Table of Figures

Table of Tables

Target Audience

Engineers and analysts in the field of communication networks would discover this report valuable. This report can fill in as an establishment for the individuals who wish to look into in the up and coming field of Computer Networks and Machine Learning. There is no pre-imperative to peruse this report however fundamental comprehension of Computer Networks and Machine Learning would be useful.

Organization

This article begins with the fundamental ideas of Machine Learning and networks took after by presentation where significance of Machine Learning in the field of networks administration is clarified. We at that point talk about in detail unique utilizations of machine learning in the field of systems administration. Chapter 2 talks about the idea of characterizing the intrusions in the system. Chapter 3 examines the use of machine learning in foreseeing the TCP throughput. Chapter 4 at that point closes the report with conclusion and future extent of work.

Abstract

Nowadays, the security of computer networks plays important role in computer system. Effective and responsive activities and preventive strategies ought to be implemented against these dangers. For this propose different programming devices are as of now grew more assets ought to be spent on this region to keep the interruptions from happening. Intrusion Detection System go for discovery the intrusion and taking vital measures from keeping it from occurring later on. The need of different potential classifiers to recognize the sort of assault in the system is exceptionally important.

The most slanting subject is the utilization of Machine Learning systems to computerize the recognition of interruptions in the computer networks. This paper presents data on the grouping of interruptions and it recognition, different ordering calculations that are utilized for characterization, Machine learning calculations used to arrange the interruptions, utilizing Appropriated Framework to enhance the execution of interruption discovery and how Artificial Intelligence can be utilized as a part of identifying interruption.

Chapter 1: Introduction

1.1 What is machine learning?

Machine learning is a utilization of Artificial Intelligence (AI) that enables programming applications to wind up more precise in foreseeing results without being expressly modified. The essential focal point of machine learning is to fabricate calculations that acknowledge input information and utilize factual examination to anticipate an important yield inside a satisfactory range.

1.2 Widely used Machine Learning Algorithms

1.2.1 Naive Bayes Classification: Naive Bayes classifiers are a group of straightforward probabilistic classifiers in light of applying Bayes' hypothesis with independent between the features. The included picture is the condition — with P(A|B) is posterior probability, P(B|A) is likelihood, P(A) is class prior probability, and P(B) is predictor prior probability.

1.2.2 Logistic Regression: Logistic regression is an effective measurable method for demonstrating a binomial result with at least one logical factors. It quantifies the connection between the straight out ward variable and at least one free factors by assessing probabilities utilizing a calculated capacity, which is the total strategic circulation.

1.2.3 Support Vector Machines: SVM is binary classification algorithm. Given an arrangement of purposes of 2 writes in N dimensional place, SVM creates a (N — 1) dimensional hyperplane to isolate those focuses into 2 gatherings. Let's assume you have a few purposes of 2 writes in a paper which are straightly distinguishable. SVM will locate a straight line which isolates those focuses into 2 writes and arranged beyond what many would consider possible from every one of those focuses.

1.2.4 Clustering Algorithms: Clustering is the undertaking of collection an arrangement of items with the end goal that articles in a similar gathering (cluster) are more like each other than to those in different gatherings. Each grouping calculation is extraordinary, and here are a few them:

• Centroid-based calculations

- Connectivity-based calculations
- Density-based calculations
- Probabilistic
- Dimensionality Reduction
- Neural systems/Deep Learning

1.2.5 Decision Trees: A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility.

1.2.6 Ensemble Methods: Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a weighted vote of their predictions.

The original ensemble method is Bayesian averaging, but more recent algorithms include error-correcting output coding, bagging, and boosting.

1.3 OSI network architecture

The essential establishment of Computer Networks is to know the Layers of network architecture. The present network architecture is for the most part in light of the OSI convention layer stack. It is a reasonable structure so we can better comprehend complex co-operations that are going on. Thus, it is essential to get a thought regarding it to comprehend this report. In the OSI show, control is passed starting with one layer then onto the next, beginning at the application (Layer 7) in one station, and continuing to the base layer, over the channel to the following station and go down the chain of command. The OSI demonstrate takes the assignment of between systems administration and partitions that up into what is alluded to as a vertical stack that comprises of the accompanying 7 layers.

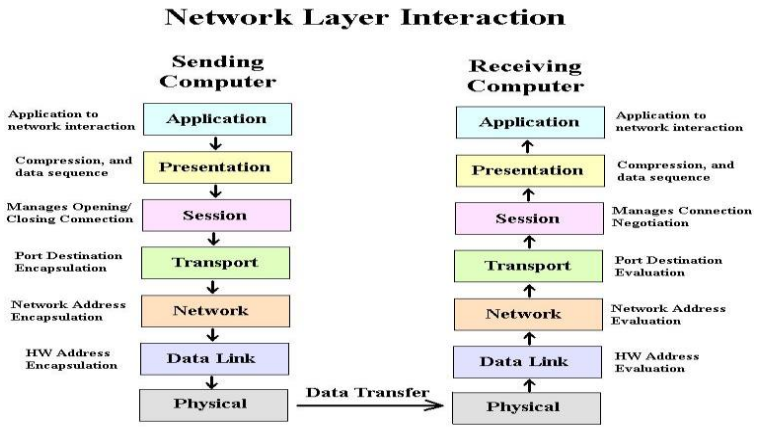


Figure 1.1. OSI Network Layers

1.3.1 Physical (Layer 1): OSI Model, Layer 1 passes on the bit stream - electrical impulse, light or radio flag — through the system at the electrical and mechanical level. It gives the equipment methods for sending and accepting information on a transporter, including characterizing links, cards and physical viewpoints. Fast Ethernet, RS232, and ATM are protocols with physical layer components.

1.3.2 Data Link (Layer 2): At OSI model, Layer 2, information bundles are encoded and decoded into bits. It outfits transmission convention information and administration and handles mistakes in the physical layer, stream control and casing synchronization. The information connect layer is separated into two sub layers: The Media Access Control (Macintosh) layer and the Logical Link Control (LLC) layer. The Macintosh sub layer controls how a PC on the system accesses the information and consent to transmit it. The LLC layer controls outline synchronization, stream control and mistake checking.

1.3.3 Network (Layer 3): Layer 3 gives exchanging and steering advancements, making sensible ways, known as virtual circuits, for transmitting information from hub to hub. Steering and sending are elements of this layer, and in addition tending to, internetworking, blunder taking care of, clog control and parcel sequencing.

1.3.4 Transport (Layer 4): OSI model, Layer 4, gives straightforward exchange of information between end frameworks, or hosts, and is in charge of end-to-end blunder recuperation and stream control. It guarantees finish information exchange.

1.3.5 Session (Layer 5): This layer sets up, oversees and ends associations between applications. The session layer sets up, organizes, and ends discussions, trades, and exchanges between the applications at each end. It manages session and association coordination.

1.3.6 Presentation (Layer 6): This layer gives autonomy from contrasts in information portrayal (e.g., encryption) by making an interpretation of from application to organize arrangement, and the other way around. The 9-introduction layer attempts to change information into the shape that the application layer can acknowledge. This layer arranges and encodes information to be sent over a system, giving opportunity from similarity issues. It is here and there called the linguistic structure layer.

1.3.7 Application (Layer 7): OSI model, Layer 7, bolsters application and end-client forms. Correspondence accomplices are recognized, nature of administration is distinguished, client verification and protection are considered, and any imperatives on information sentence structure are distinguished. Everything at this layer is application-particular. This layer gives application administrations to document exchanges, email, and other system programming administrations. Telnet and FTP are applications that exist altogether in the application level. Layered application structures are a piece of this layer.

1.4 Introduction to Intrusion Detection System (IDS):

Nowadays, the use of new networking paradigms like Internet of Things (IoT) and Cloud Computing has led to new methods to handle security challenges which ensure confidentiality, integrity, scalability and availability of services and information to the users. The most trending topic is the use of hybrid Machine Learning (ML) techniques to automate the detection of intrusions in the computer networks. This paper presents information on the classification of intrusions and it detection, various classifying algorithms that are used for classification, Machine learning algorithms used to classify the intrusions, using Distributed system to improve the performance of intrusion detection and how Artificial Intelligence can be used in detecting intrusion.

In this paper, ML techniques have been used in labeling the data. Because, manual labeling is an expensive task. It takes several attempts to label the data as a potential threat. But

with Optimum Path Forest Clustering we can do it efficiently. Main disadvantage about this is that, the system is strongly dependent on the database, which is not effective against unknown attacks. The use of Intrusion Detection System (IDS) based on Artificial Neural Network (ANN), which has been experimentally proved to have an accuracy of almost 99% based on the classification of pre-defined classes of attacks.

Chapter 2: Classification

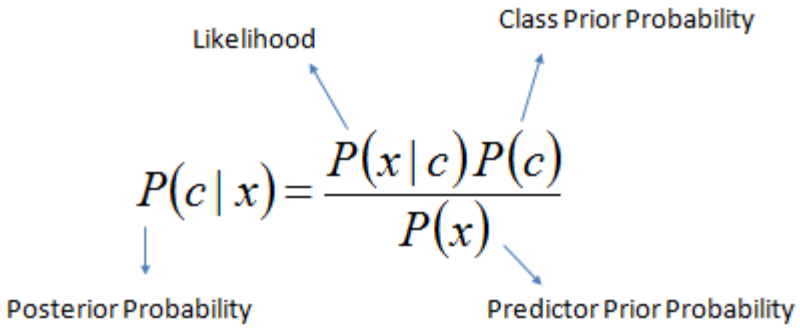
Intrusion Detection System is like a burglar alarm and classifying it is like classifying the potential threat that the intrusion may cause. Generally, IDS are based on 2 models, one that use signature based detection and the other is anomaly based detection. Working of the signature based detection is limited to pattern of attack that has been recorded in the database, based on that required action is taken. Whereas, in Anomaly based IDS, intrusion is detected based on the unusual network activity under normal conditions, but it results in a lot of false positives.

The IDS problems can also be approached with the help of new trending and effective techniques like Artificial Intelligence. Algorithms such as Naïve Bayes, decision trees, Support Vector Machines (SVM), Artificial Neural Network (ANN) and other algorithms are recommended for classification. Handling huge amount of data like text, is way too far than the limited capacity of human processing. So, to process the data we need to use data mining techniques coupled with ML techniques instead of inefficient human processing. This technique automatically adapts to the data and extracts useful patterns needed as reference to the normal behavior or attacks from existing data for the classification of network traffic.

In this paper the classification of IDS will be done using Naïve Bayes method and some concepts of data mining to process data. In-order to classify the attacks, they are grouped into 4 attacks viz, Denial of Service (DoS) attacks, Probe, Remote to Local (R2L) and User to Root (U2R) attacks. The data set used in the reference paper is NSL-KDD99, which is the mostly used data set in IDS though it is old. It is because of lack of availability of datasets which are freely accessible to researchers. On the same lines, the reason for using Naïve Bayes algorithm for classification is based on the experimental results which have proved it efficiency in classification [1] [2] [3].

2.1 Naïve Bayes algorithm:

Naïve Bayes algorithm is based on the Bayes' Theorem. Bayes' theorem is stated mathematically as the following equation:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Figure 2.1. Bayes' Probability Formula

Naïve Bayes algorithm calculates the probability to an event to occur, this technique will result in a very small result set with a huge value in its attribute. This probability value refers to the possibility of the same value, but on the other side it's range is very large that the probability to occur is very small, this usually happens when the data is continuous. This results in poor performance of Naïve Bayes algorithm. In-order to group the continuous type of data we use K-means Clustering [4] [5] [6] method to improve the efficiency of IDS using Naïve Bayes algorithm.

Naïve Bayes is based on normalizing the assumption that the attribute is independent on an input value. The advantage of using this is that it requires only fewer dataset to train the model required for classification process.

2.2 Dataset–NSL-KDD99:

The most common dataset used for research is NSL-KDD, it is the proposed solution in KDD Cup 1999, hence the name KDD-99 [7]. Though it is over a decade old [8], it is still a commonly used dataset in research on the field of IDS due to the lack of datasets that are available to the public freely. Yet KDD-99 has some problems which have overcome in NSL-KDD which also includes detection of redundant and re-proportion of datasets. This shortcoming will play a vital role in performance of the learning algorithms and in evaluating other learning algorithms. Table 2.1 shows Intrusion classes and Attack type for each class.

Intrusion Class	Attack types
DoS	back, land, neptune, pod, smurf, teardrop, apache2, udpstorm, processtable, worm (10)
Probe	satan, ipsweep, nmap, portsweep, mscan, saint (6)
R2L	guess_password, ftp_write, imap, phf, multihop, warezmaster, warezclient, spy, xlock, xsnoop, snmpguess, snmpgetattack, httptunnel, sendmail, named (16)
U2R	buffer_overflow, loadmodule, rootkit, perl, sqlattack, xterm, ps (7)

Table 2.1: Attack Class

2.3 K-Means Clustering:

K-Means Clustering is a method used to group the data into one or more clusters unlike the normal hierarchical clustering methods. The clusters are grouped in such a way that, the data points in the same cluster have same characteristics and those with different are in different clusters. The general K-Means Clustering algorithm is shown below (Figure 2.2). The algorithm is as follows,

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

Figure 2.2. K-Means Clustering Algorithm

Where $x^{(1)}, \ldots, x^{(m)}$ are training set, our goal is to predict k centroids and a label $c^{(i)}$ for each data point. Below is the flow diagram of K-Means clustering algorithm (*Figure 2.3*).

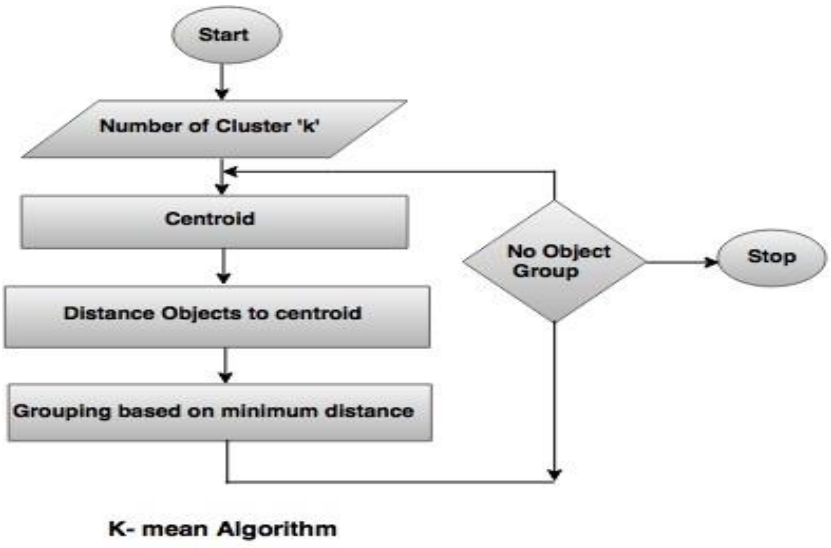


Figure 2.3. Flow Chart of K-Means algorithm

Chapter 3: Spam Detection

3.1 What is Spam and how to detect it?

Abnormalities in computer networks has increased enormously over the last few decades and awareness to create useful techniques to identify the unusual traffic patterns is important. The use of data mining collectively with machine learning techniques to properly identify these abnormalities, in particular spam detection and spams which cause potential threat to the user's privacy. Using the dataset called SPAMBASE we can categorize or classify the type of abnormalities in an efficient way which further will be useful for IDS.

Numerous systems classifiers were utilized as a part of this work, for example, Ada Boost, Radial Base Function (RBF) and Naïve Bayes. In any case, in this area, we quickly survey of the hypothetical foundation in the best systems classifiers, being them, Rotation Forest, Random Forest, Bagging. Likewise, additionally we portrayed the Weka Classifier tool environment.

3.2 Rotation Forest:

The primary motivation behind this procedure is a technique for producing classifier gatherings in view of highlight extraction. To make the preparation information for a base classifier, the list of capabilities is arbitrarily part into K subsets where K is a parameter of the calculation and Principal Component Analysis (PCA) is connected to every subset. All main segments are held with a specific end goal to safeguard the inconstancy data in the information. Subsequently, K pivot turns happen to frame the new highlights for a base classifier. The possibility of the revolution approach is to support at the same time singular exactness and assorted variety inside the troupe. Assorted variety is advanced through the element extraction for each base classifier. Decision trees were picked here on the grounds that they are delicate to turn of the element tomahawks, in light of this data comes the name Forest. Rotation Forest is a proposed technique for building classifier gatherings utilizing freely prepared decision trees and your troupe comprises of decision trees prepared on bootstrap tests from the informational index. The trademark called Additional Diversity is presented by randomizing the element decision at every hub. Amid tree development, the

best element at every hub is chosen among M haphazardly picked highlights, where M is a parameter of the calculation. Rotation Forest utilize the ideas of the technique Random Forest. The base classifiers are likewise freely fabricated choice trees, however in Rotation Forest each tree is prepared all in all informational index in a pivoted include space. As the tree learning calculation assembles the arrangement areas utilizing hyperplanes parallel to the element tomahawks, a little pivot of the tomahawks can prompt different trees, this clever system goes for building precise and various classifiers. Bootstrap tests are taken as the preparation set for the individual classifiers, as in packing. The heuristic for this strategy is to apply highlight extraction and to in this manner remake a full list of capabilities for every classifier in the group.

3.3 Random forest:

Random Forest is a troupe learning calculation strategy creates numerous individual students and totals the outcomes, the methods utilizes an expansion to the bagging approach. In a customary decision tree classifier, a choice at a node split is made in light of all the element properties, yet in Random Forest, the best parameter at every node in a decision tree is produced using a haphazardly chose number of highlights. This random choice of highlights causes Random Forest to not just scale well when there exist numerous highlights per include vector, yet in addition encourages it in decreasing the reliance (relationship) between the component traits and is along these lines less defenseless against characteristic clamor in the information.

3.4Methodology:

In this area, we display the technique utilized to approve the outcomes among all classifiers depicted in Chapter 2. Keeping in mind the end goal to break down and measure the inconsistencies, we have utilized the SPAMBASE dataset, which contains 57 information properties related with the recurrence of a few words in the email's substance. This informational collection was made keeping in mind the end goal to enhance security programming in network systems, as assaults utilizing spam messages can cause misfortunes, for example, superfluous time spending, profitability misfortune, shameful or hostile substance and money related misfortune caused by misrepresentation. The

informational index can be stacked into Weka Classifier tool utilizing the previously mentioned ARFF document organize. In this record, every segment contains a characteristic of the information, which speaks to a word and its recurrence in a given email.

The framework we have arranged with a specific end goal to characterize the information is appeared in *Figure 3.1*. Crude information are fundamentally separated into clear cut classes. At that point crude information were given by classifier with 5-overlap cross approval technique. Thusly, the information were partitioned into 5 sections and chose for testing one time at any given moment, and the rest of the parts were chosen for preparing. Along these lines the classifier is run 5 times and the qualities are arrived at the midpoint of.

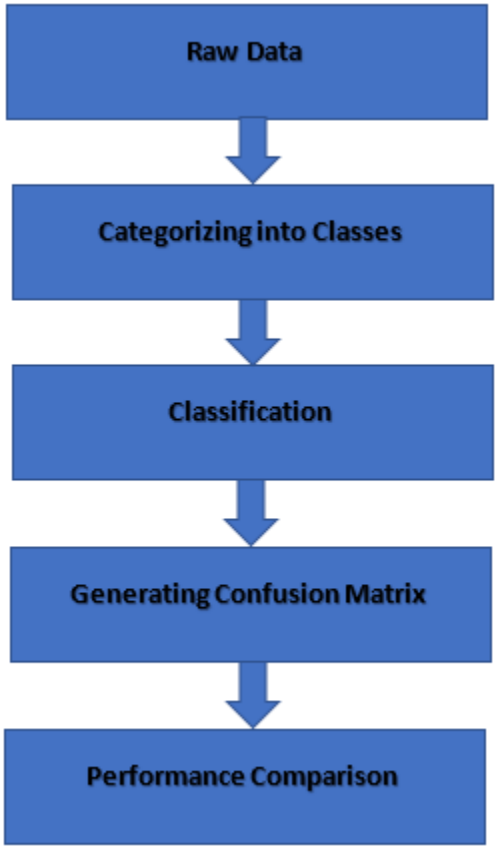


Figure 3.1 Process Flow

3.5 Experimental Outcomes:

The outcomes acquired 2 better methods: Random Forest also, Rotation Forest. The Random Forest calculation accomplished a 99.42% of mean acknowledgment rate, where 99.50% of the spam tests were accurately grouped, and 99.34% of the non-spam tests have been accurately perceived *(Table 3.1, Table 3.2).*

Methods	Recognition Rate %
Random Forest	99.42%
Rotation Forest	98.03%
Nbtree	96.96%
J48	96.51%
Bagging	96.19%
MLP	93.18%
Logit Boost	92.264%
Ada Boost	90.29%
RBF	88.03%
Naïve Bayes	86.63%
OneR	80.79%
ZeroR	49.60%

Table 3.1. Experimental results of detection rates

Spam	Non-Spam	Recognition Rate %
451	2	99.50%
3	450	99.34%
	Mean Rate	99.42%

Table 3.2 Confusion matrix of Random Forest

The Rotation Forest calculation accomplished a 98.03% of mean acknowledgment rate, where 98.70% of the spam tests were accurately ordered, and 97.36% of the non-spam tests have been effectively perceived *(Table 3.3).*

Spam	Non-Spam	Recognition Rate %
447	6	98.70%
12	441	97.36%
	Mean Rate	98.03%

Table 3.2 Confusion matrix of Rotation Forest

Conclusion:

Recently, numerous looks into on Computer Networks have utilized machine learning procedures to investigate the conduct and the capacity to distinguish any conceivable inconsistency in a given system. In this work, we are centered around the issue of intrusion identification utilizing the Weka Classifier so as to look at the viability of some machine learning calculations to this errand. We have utilized a dataset known as SPAMBASE, which includes a gathering of messages marked as spam and non-spam for grouping purposes. Moreover, the accompanying strategies have been tended to: Random Forest, NBTree, Rotation Forest (RF), Decision Tree-based Classifier, Naivë Bayes, Sacking, Multilayer Perceptron (MLP)

For future work we propose the mix of a criticism framework to enable the models to build their execution in future forms. Furthermore, we propose the usage of various hybrid models, making utilization of other calculation blends keeping in mind the end goal to contrast comes about and the models made for this examination.

Acronyms

Qos = Quality of service

TCP = Transmission Control Protocol

SVM = Support Vector Machines

ML = Machine Learning

IP = Internet Protocol

FTP = file transfer protocol

AI = Artificial Intelligence

OSI = Open Systems Interconnection

MAC = Media Access Control

LLC = Logical Link Control

FTP = File Transfer Protocol

PCA = Principal Component Analysis

IDS = Intrusion Detection System

References:

[1] I Nyoman Trisna Wirawan, I. E. (2015). Penerapan Naive Bayes Pada Intrusion Detection System Dengan Diskritisasi Variabel. *Jurnal Ilmiah Teknologi Informasi* , 2

[2] Jacobus, Jacobus, and Edi Winarko. "Penerapan Metode Support Vector Machine pada Sistem Deteksi Intrusi secara Real-time." *Berkala Ilmiah MIPA* 23.2 (2014).

[3] Susanto, Bekti Maryuni. "Naive Bayes Untuk Mendeteksi Gangguan Jaringan Komputer Dengan Seleksi Atribut Berbasis Korelasi." *Bianglala Informatika* 1.1 (2013).

[4] Kusrini, K. Grouping of Retail Items by Using K-Means Clustering. The Third Information Systems International Conference. Surabaya. 2015; Vol. 72, Pages 495–502.

[5] Kusrini, K, Iskandar, M.D., Wibowo F. W. Multi Features Content- Based Image Retrieval Using Clustering And Decision Tree Algorithm. TELKOMNIKA Telecommunication, Computing, Electronics and Control. Yogyakarta. 2016; Vol 14 No 4; Halaman : 1480-1492.

[6] Kusrini. Pendiskritan Kelas Kontinyu dengan Algoritma K-Mean Cluster. Jurnal Dasi. 2010; Vol 11 No 4.

[7] Tavallaee, M. A Detailed Analysis of the KDD CUP 99 Data Set. *IEEE Sysposium on Computational Intelegence in Security and Defense Applications.* 2009, CISDA.

[8] Kusrini, E. T. *Algoritma Data Mining*. Yogyakarta: Andi Publisher. 2009.

[9] H.-J. Liao, C.-H. Richard Lin, Y.-C. Lin, and K.-Y. Tung, “Intrusion detection system: A comprehensive review,” *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 16–24, 2013.

[10] Akamai, “Akamai Q4 2016 State of the Internet / Security Report.” [Online]. Available:

https://www.akamai.com/us/en/about/our-thinking/state-of-theinternet- report/global-state-of-the-internet-security-ddos-attackreports. jsp.