Classification of Intrusion Detection System (IDS) Based on Computer Network

David Ahmad Effendy
Master Program of Informatics Engineering
AMIKOM Yogyakarta University
Jl. Ringroad Utara Condong Catur
Depok Sleman
Yogyakarta 55283 Indonesia
bangdavid07@gmail.com

Kusrini Kusrini
Master Program of Informatics Engineering
AMIKOM Yogyakarta University
Jl. Ringroad Utara Condong Catur
Depok Sleman
Yogyakarta 55283 Indonesia
kusrini@amikom.ac.id

Sudarmawan Sudarmawan
Computer Science Department
AMIKOM Yogyakarta University
Jl. Ringroad Utara Condong Catur
Depok Sleman
Yogyakarta 55283 Indonesia
sudarmawan@amikom.ac.id

Abstract* – Intrusion Detection System (IDS) is made as one of the solutions to handle security issues on the network in order to remain assured free of attack. IDS's work is developed by 2 models that using signature-based detection, how it works is limited to the pattern of attack behavior that has been defined in the database. The next is the Anomaly-based IDS model. It works by detects unusual activity of network in the normal conditions, but this model gives a lot of false positiv messages. Several previous studies have shown that the IDS approach with machine learning techniques can provide high accuracy results. The first step that must be done in the application of mechine learning technique is preprocessing the selection of features / attributes to optimize the performance of learning algorithms. In this study, intrusion detection system with mechine learning classification technique is proposed by using naivebayes algorithm with NSL-KDD dataset. The processes in this reseach start from normalization of data, discretization features continuous variables with k-means method and the selection of features using Information Gain algorithm. The result of this reseach shows that the application of k-means clustering method for continuous variabe discretization and feature selection can optimize the performance of naivebayes algorithm in classifying intrusion types.

***Keywords*— ids; k-means clustering; fitur selection; naivebayes**

I. Introduction

Intrusion Detection System (IDS) is an alarm system that can be used to detect any unusual activity in the network. In general, IDS is developed with 2 models that use signature-based detection, how it works is limited to the pattern of attack behavior that has been defined in the database. Another model is Anomaly-based IDS, this model will detect any unusual activity of network activity under normal conditions, but this model gives a lot of false positive messages [1].

Recently some media reported that, according to the latest information from the Security Incident Response Team on Internet Infrastructure (ID-SIRTII). By mid-May in 2017, some countries become the target of ransomware virus or commonly called Wannacry that is included in Indonesia [2]. there are Harapan Kita and Dharmais Hospital as submitted by the Director General of Informatics Semuel A. Pangerapan, in a release received by CNNIndonesia.com, Saturday (13/5). Although the security system continues to be improved but there are still have many loopholes for hackers finding access to the system. So it needs to be done with new techniques for getting better result.

IDS problems have been approached with some artificial intelligence algorithms such as decision trees, naivebayes, support vector machines, artificial neural networks and other algorithms. Managing large amounts of data such as text, far beyond the limited capacity of human processing capacity [3]. The use of artificial intelligence techniques known as data mining or machine learning as an alternative to expensive and heavy human capabilities. This technique automatically learns the data or extracts useful patterns from the data as references to normal behavior profiles or attacks from existing data for the classification of subsequent network traffic [4].

In this research will be done classification IDS testing using naivebayes method with a combination of other methods to improve the performance of the method. Classification of types of attacks on IDS are grouped into 4 attacks including DoS, Probe, R2L, and U2R. The dataset used in this study is NSL-KDD99. Although the data is old enough, it is still commonly used in research on intrusion detection systems because of the lack of complete datasets available to the public and accessible freely [5]. The underlying reasoning of the researchers using the naivebayes method is some of the research done by previous researchers by [6] [7] [8], and has proven the ability of excellent naivebayes classification.

The naivebayes algorithm is based on the probability value of an attribute to its class, this technique will produce very small probability values if there are very many different values in an attribute. The probability value points to the possibility of the same value coming out in a class but on the other side the range of values on the attribute is so large that the probability value of that value reappears in a class will be very small, This can be caused by the use of attributes of a continuous type. This problem will cause weak performance of naive bayes. To group or extract data with continuous type can be done by using K-Means Clustering [9] [10] [11]. The focus of this research is to improve the capability of intrusion detection system (IDS) by applying K-Means Clustering method to improve the performance of naivebaye algorithm.

978-1-5386-0658-2/17/$31.00 ©2017 IEEE

II. MATERIAL AND METHODS

A. NSL-KDD Dataset

The data used in this study is the NSL-KDD Cup 1999 dataset. The NSL-KDD is the proposed dataset in several studies as a solution to the problem in the KDD Cup 1999 dataset (KDD-99) [12]. The KDD-99 dataset is over 15 years old [5], but is still commonly used in research on the area of intrusion detection systems due to the lack of datasets available to the public and accessible freely. Some of the problems that exist in the KDD-99 dataset have now been addressed in NSL-KDD including deletion of redundant data and re-proportion of datasets [12]. NSL-KDD does not include redundant data on KDD-99 which may affect the performance of learning algorithms while the re-proportion of KDD-99 enables NSL-KDD in the process of evaluating various learning algorithms.

The NSL-KDD dataset has 42 attributes, which consists of 41 input attributes and 1 target attribute. Furthermore the types of attacks are grouped into four categories of intrusion classes namely DoS, Probe, U2R and R2L. Table I shows classes and members (attack type) for each class.

Table I. Attack 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)

B. Normalization

Normalization of data is the preprocessing stage in this study. Normalization is scaling the attribute value of the data so it is in a certain range. Normalization is done with the aim to reduce any errors in the process of classification. The normalized method applied is a useful Siqmoid when the existing data involves outlier data by means of all data presented in the range of 0 to 1, using the formula [13], presented at (1) and (2).

$$y = \frac{x_{ik} - \bar{x}_k}{r\,\sigma_k} \qquad (1)$$

$$\bar{x}_{ik} = \frac{1}{1 + e^{-y}} \qquad (2)$$

Where
x_{ik} : dataset tuple
$\bar{x}_k$: mean value of all dataset record
σ : standard deviation
r : value factor, r value is defined by user, i.e. set to 1
$\bar{x}_{ik}$: result of normalization
e : Euler constant, i.e 2.71..

C. K-Means Clustering

K-means is one of the non-hierarchical clustering methods used to partition data into one or more forms of clusters. This method partitions the data into clusters so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into other clusters [14]. In general the basic K-Means Clustering algorithm is as follows:

- Determine the number of clusters
- Allocate data into the cluster randomly
- Calculate the centroid/average of the data in each cluster
- Allocate each data to the nearest centroid / average
- Return to Step 3, if there is still data that moves the cluster or if the centroid value changes, some are above the specified threshold value or if the value changes on the objective function used above the specified threshold value.

The determination of centroid was randomly determined by the equation formula (4) [15] as follows:

$$c_i = \min + \frac{(i-1) * (\max - \min)}{n} + \frac{(\max - \min)}{2 * n} \qquad (4)$$

Where
c_i : centroid of the i class
min : the smallest value of continuous class data
max : the largest value of continuous class data
n : number of discrete classes

Next calculate the distance between data and centroid using Euclidean Distance Space method. This needs to be done because the results obtained are the shortest distance between two points to be taken into account. Mathematically the Ecluidean Distance method can be formulated: [16], as in (5):

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left\{x_{ik} - x_{jk}\right\}^2} \qquad (5)$$

Where
d_{ij} : distance of object between object i and j
P : dimensions of data
X_{ik} : coordinate of object i in dimension k
X_{jk} : coordinate of object j in dimension k

To get a new centroid can be done by calculating the average value of data that has a closeness to a particular cluster, the iteration process of forming a new cluster will continue to be done and will stop when the amount of difference between the average and the centroid is not more than the fault tolerance. the error tolerance is obtained from the difference between the highest data value and the lowest data value multiplied by the 0.01 delta value.

D. Selection Feature

Feature selection is one of the most important and often uses techniques in pre-processing. This technique is done after the clustering process to be prepared for classification stage by selecting features that have value proximity to the target class. Reducing the number of features involved in determining a target class value, reduce irrelevant, exaggeration and data that cause misunderstanding in target classes that create immediate effects for applications [17]. The main purpose of feature selection is selecting the best feature of a set of data features [18]. Information Gain is the feature selection method in this step.

Information Gain is a feature selection technique that uses scoring methods for nominal or weighting of continuous attributes that are discredited by using the entropy maximum. An entropy is used to define the value of the Information Gain. In pattern, it is written in point 6 and 7[15].

$$Entropy\ (S) = -\sum_{i=1}^{n} \frac{|s_i|}{|s|} \log \frac{|s_i|}{|s|} \qquad (6)$$

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \qquad (7)$$

Where

S : case set

A : attribute

n : number of attribute partition A

$|s_i|$: the number of cases on the i-th partition

|S| : the number of cases in S

The next process after selecting a feature is to classify the dataset into the target class.

E. Naive Bayes Classifier

Naivebayes is a simple probabilistic classifier, this algorithm uses Bayes's theorem and assumes all independent or non-dependent attributes given by the class variable value [19]. Another definition says naivebayes is a classification with probability and statistical methods brought by British scientist Thomas Bayes, predicting future opportunities based on past experience.

Naive Bayes is based on the simplifying assumption that the attribute value is naturally independent if given an output value. In other words, given the value of output, the probability of observing collectively is the product of individual probabilities [20].

The advantage of using Naive Bayes is that it requires only a small amount of training data to determine the approximate paremeter required in the classification process. Naive Bayes often works much better in real-world situations than ever before. The equation of the naivbayes method [21], is defined as in (8)

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \qquad (8)$$

Where

X : data with unknown class

H : the hypothesis that X belongs to a specific class

P(H|X) : the probability of hypothesis H is based on condition X (posteriori probabilitas)

P(H) : probability of hypothesis H (previous probability)

P(X|H) : the probability of X is based on the conditions in hypothesis H

P(X) : probability X

Then calculate the optimum accuracy and the optimum f-measure value after the confusian matrix value is obtained from the classification process. The equation for calculating the accuracy value [22], is defined as in (9) and f-measure is defined as in (10).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (9)$$

Where

TP : False positive, positive tuple labeled true

TN : True negative, negatively labeled negative tuples

FP : False is positive, wrong negative tuple is labeled as positive

FN : False negative, positive tuples misconstrued as negative

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \qquad (10)$$

Where

Precision : The number of positive categorically classified samples divided by the total samples classified as positive samples

Recall : The number of samples classified positively divided by the total sample in the testing set of positive categorized

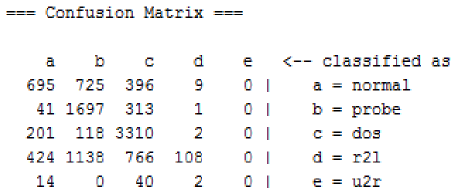
III. RESULT AND IMPLEMENTATION

This research is a type of experimental research. In this study, data collection includes literature studies in the form of books, journals, and relevant scientific papers. In this research, stages are :

- Clear data, redundant and irrelevant data removed.
- Data integration, attack data will be integrated with the attack group.
- Transformation of data, continuous numerical data clustered and changed in discrete form.
- Feature selection, please choose best feature.
- Classification, the algorithm used by naivebayes
- Evaluation, classification results with naivebayes, each group of attacks submitted to the network administrator.

The test dataset uses 10,000 data. In the process using weka applications to process data from the process of selecting features to the process of classification. The model development process is called training data, while the data will be used to validate the accuracy of a model built on a particular dataset using Cross Validation. Model making usually aims to make predictions and classifications of new data that may never have appeared in the dataset. The data used in the model development process is called training data, while the data to be used to validate the model is referred to as testing. In this process the dataset is divided into random number of k-pieces of partition. Then a number of experimental k-times were performed, each experiment using k-partition data as data testing and using the rest of the partition as training data.

The testing stage is done by 2 stages, the first is done without the k - means clustering method. At this stage feature with data type and standard deviation. secondly, clusters are done with a combination of k-means clustering. In the test to be conducted clusters is formed with by 3 kind of clusters, there are 3 cluster, 5 cluster and 8, as the largest cluster. In the test used weka application to get the value of confusin matrix, the results obtained are shown as in Figure 1.



```
=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
 695  725  396    9    0 |    a = normal
  41 1697  313    1    0 |    b = probe
 201  118 3310    2    0 |    c = dos
 424 1138  766  108    0 |    d = r2l
  14    0   40    2    0 |    e = u2r
```

Fig 1. Confusion matrik hasil klasifikasi tanpa menerapkan metode k-means clustering

From result of confusion matrix hence obtained optimum accuracy and optimum f-measure value. The accuracy and f-measure values of the classification results without combination with k-means are presented in Table II. and Table III.

Table II. Value of optimum accuracy

Optimum accuracy value (without k-means)				
Normal	Probe	DOS	R2L	U2R
0,941 with 1 atribute	0,895 with 5 atribute	0.831 with 2 atribute	0.817 with 5 atribute	0.994 with 2 atribute

Table III. Optimum f-measure value

Optimum f-measure value (without k-means)				
Normal	Probe	DOS	R2L	U2R
0.601 with 18 atribute	0.777 with 6 atribute	0.783 with 3 atribute	0.561 with 9 atribute	0.156 with 4 atribute

Next step is classification, is done by continued fitur discretitation with k-means method. And matrix confusion result is showed on table IV. and table V.

Table IV. Optimum accuration value

Optimum accuration value (with k-means)				
Normal	Probe	DOS	R2L	U2R
0,953 with 5 atribute and 8 cluster	0,975 with 5 atribute and 8 cluster	0.96 with 5 atribute and 8 cluster	0.964 with 5 atribute and 8 cluster	0.998 with 5 atribute and 8 cluster

Table v. Optimum f-measure value

Optimum f-measure value (with k-means)				
Normal	Probe	DOS	R2L	U2R
0.658 with 6 atribute and 5 cluster	0.879 with 31 atribute and 8 cluster	0.893 with 30 atribute and 8 cluster	0.822 with 28 atribute and 3 cluster	0.487 with 27 atribute and 5 cluster

Based on previous research conducted by I Nyoman Trisna Wirawan and his colleague Ivan Eksistyanto in the Year 2015 entitled Implementation of Naivebayes on IDS with variable discretization. In the research described discretization process is done using binning technique, by dividing the data into 3 intervals and 5 intervals then produced an accuracy of 89.6% the result is better than testing using 3 intervals. Furthermore, in the research that researchers do this is the discretization of variables with clustering process with k-means clustering medare which is divided into 3 clusters, 5 clusters and 8 clusters. From the results it was found quite clearly the difference from previous research.

The application of k-means to handle the problem of continuous numerical features can provide better classification results than using the discretization technique of the mean and standard deviation. If seen from table 1 and table 2. The value of accuracy and value of f-measure get higher value.

IV. Conclusion

Classification of Intrusion Detection System (IDS) with NSL-KDD99 dataset applies discretization technique of k-means clustering method proved to give better result than using discretization technique of mean and standard deviation. After testing the results of different classification results from the total of 3 clusters, 5 clusters and 8 clusters is the best 8 clusters. And it can be concluded that the number of cluster formation for IDS classification by applying k-means clustering to handle attributes affects the final classification results. The more clusters formed can produce higher accuracy values. In contrast to the result of the f-measure the number of clusters does not always give a better value.

References

[1] Neethu, B. Classification of Intrusion Detection Dataset Using Machine Learning Approaches. International Journal of Electronics and Computer Science Engineering, (2012) pp. 1044-51

[2] Kertopati, L.. *internet news*. access October 12, 2017, from cnnindonesia.com: https://www.cnnindonesia.com/teknologi/20170513191519-192-214642/dua-rumah-sakit-di-jakarta-kena-serangan-ransomware-wannacry/

[3] Toyota, T., Nobuhara, H. Visualization of the Internet News Based on Efficient Self-Organizing Map Using Restricted Region Search and Dimensionality Reduction. *JACIII.* 16: 222. 2012.

[4] Turban. *Decision Support System and Intelligent System, edisi Bahasa Indonesia jilid 1.* Yogjakarta: ANDI. 2005.

[5] Hettich. *The UCI KDD Archive.* California: Department of Information and Computer Science. 1999

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

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

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

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

[10] 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.

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

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

[13] Prasetyo, E. *Mengolah Data Menjadi Informasi Menggunakan Mathlab.* Yogyakarta: Andi Yogyakarta. 2014.

[14] Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, (First Edition). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005.

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

[16] Kohonen, T., An introduction to neural network, Neural Network., 1988, 1:3-16

[17] Djatna, Taufik & Morimoto, Yasuhiko. *Pembandingan Stabilitas Algoritma Seleksi Fitur Menggunakan Transformasi Ranking Normal*. Jurnal Ilmiah Ilmu Komputer, Institut Pertanian Bogor. 2008; Vol. 6 No. 2, ISSN : 1693 -1629

[18] Abadi, Delki. *Perbandingan Algoritme Feature Selection Information Gain Dan Symmetrical Uncertainty Pada Data Ketahanan Pangan*. Skripsi, Institut Pertanian Bogor, Bogor. 2013.

[19] Patil, T. R., Sherekar, M. S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, *International Journal of Computer Scienceand Applications*. 2013; Vol. 6, No. 2, Hal 256-261

[20] Ridwan, M., Suyono, H., Sarosa, M. *Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier*, *Jurnal EECCIS.* 2013; Vol 1, No. 7, Hal. 59-6.

[21] Schneider, M.K, Techniques for Improving the Performance of Naïve Bayes for Text Classification, Proceedings of CICLing, page 682-693, 2005

[22] Han, J. K. *Data Mining : Concepts and Techniques Second Edition.* Morgan Kaufmann. 2006