Jurnal Elektronik Ilmu Komputer Udayana
Volume 9 No. 3, February 2021

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Implementation of Feature Selection using Information Gain Algorithm and Discretization with NSL-KDD Intrusion Detection System

x

I Gusti Bagus Dharma Putra[a1], I Gusti Agung Gede Arya Kadyanan[a2]

[a]Informatics Department, Faculty of Math and Science, Udayana University
South Kuta, Badung, Bali, Indonesia
[1]dharmatkjone@gmail.com
[2]gungde@unud.ac.id

Abstract

Feature selection is one of the research on data mining for datasets that have relatively many attributes. Eliminating some attributes that are irrelevant to the label class will be able to improve the performance of the classification algorithm. The Information Gain algorithm is one of the algorithms for searching for features that are irrelevant to the label class. This algorithm uses wrapper techniques to eliminate irrelevant attributes. This research aims to implement feature selection using the Information Gain algorithm against the NSL KDD intrusion detection dataset which has a large number of relative attributes. The dataset of the selected attribute will be performed by a classification algorithm so that an attribute reduction can improve the compute process and improve the accuracy of the algorithm model used.

Keywords: *Feature Selection, Information Gain, Intrusion Detection System (IDS), NSL-KDD, Discretization*

1. Introduction

Feature selection is one of the most important data mining techniques in preprocessing for the selection of relatively many features on the dataset[1]. It aims to reduce data thereby speeding up computing processes and producing accurate models of the algorithms used. Feature selection is usually used to select optimal features, reduce dimensions, improve algorithm accuracy, and remove irrelevant features[2].

Intrusion detection systems (IDS) have been introduced as security techniques for detecting various attacks. IDS can be identified by two techniques, namely abuse detection, and anomaly detection. Abuse detection techniques can detect known attacks by checking attack patterns, such as virus-detection by antivirus applications. However, they cannot detect unknown attacks and need to update their attack patterns whenever there is a new attack. On the other hand, anomaly detection identifies unusual patterns of activity that deviate from normal use as interference[3].

Intrusion Detection System (IDS) is one of the important research in the field of computer networking or computer security. Several studies have conducted an intrusion detection system with calcification-based data mining to detect attacks on computer networks by analyzing data packets in the network. In the process of doing machine learning must have good data (complete, true, consistent, and integrated). Before data mining is done, the data needs to be processed in advance to ensure its quality. Moreover, many features in the data may reduce classification performance. So it takes a feature selection technique to select the relevant features for the data[4].

The NSL-KDD dataset is a dataset used to benchmark various classification methods for intrusion detection. This dataset has quite a lot of features namely 41 features that are continuous and discrete with normal or anomaly labels (Dos, Probe, R2L, U2R). Feature selection is one of the most important processes for eliminating uns needed features in nsl-kdd datasets. Not all

attributes can have an effect in a label class, therefore eliminating non-essential attributes with label classes is critical to improving classification performance[5].

This study aims to select features that are important or relevant to the label class and reduce computing time. The selection of features of the dataset to be used is the Information Gain (IG) technique. To group data with continuous type can be done by the binning method with the number of bins that is 12.

2. Research Methods

This research is an experimental study with the discrete of continuous numerical value variables using binning methods while using attribute selection. The dataset used in this study is NSL-KDD'99 obtained in http://nsl.cs.unb.ca/NSL-KDD/ which is grouped into 4 categories of attacks namely DoS, R2L, U2R, and Probe.

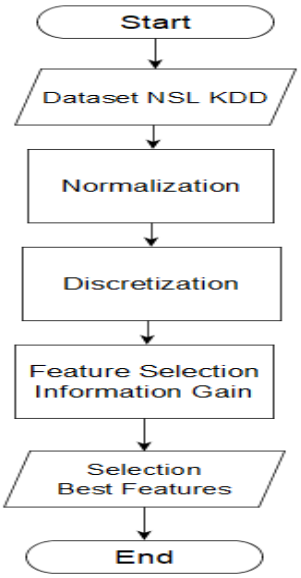

Figure 1. Feature selection process scheme

2.1. NSL-KDD'99 Dataset

The data used in this study is the 1999 NSL-KDD Cup dataset. NSL-KDD is the solution to the problem in the 1999 KDD Cup dataset (KDD-99). Intrusion detection systems because there are not many alternative datasets available and publicly accessible [6]. This dataset consists of a normal class and 39 types of attacks. In this study, the types of attacks contained in the dataset were grouped into 4 categories namely DoS, R2L, U2R, and Probe. As found in Table 1. and the NSL-KDD dataset attribute type is contained in Table 2. below.

Table 1. IDS Attack Category
Source: http://nsl.cs.unb.ca/NSL-KDD/

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Normal (1825)	Dos (3631)	Probe (3631)	R2L (2436)	U2R (56)
Normal (1825)	Back (297) Land (4) Pod (35) Smurf (528) Teardrop (9) apache2 (615) Udpstorm (2) Processtable (572) Mailbomb (245) Neptune (1324)	Satan (614) Ipsweep (121) Nmap (58) Portsweep (133) Mscan (869) Saint (257)	guess_passwd (1031) ftp_write (2) imap (1) phf (2) multihop (13) warezmaster (506) xlock (8) xsnoop (3) snmpguess (282) snmpgetattack (152) httptunnel (110) sendmail (10) named (16)	buffer_overflow (16) loadmodule (2) rootkit (11) perl (2) sqlattack (2) xterm (10) ps (13)

Table 2. Dataset Attribute Type
Source: http://nsl.cs.unb.ca/NSL-KDD/

Nominal	Biner	Numerik
protocol_type(2) service(3) flag(4)	land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21), is_guest_login(22)	duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8), urgent(9), hot(10), num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20), count(23), srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29), diff_srv_rate(30), srv_diff_host_rate(31),

2.2. Normalization

In this preprocessing process, an attribute separation will be performed which is then performed normally. Normalization is a transformation process in which a numeric attribute is scaled in a smaller range such as -1.0 to 1.0, or 0.0 to 1.0. In this study the methods/techniques applied to data normalization are:

$$v' = \frac{v - min_a}{max_a - min_a}(new\_max_a - new\_min_a) + new\_min_a \quad (1)$$

Description :
v : value in numeric data column
$\boldsymbol{v'}$: value result on normalization calculation
$\boldsymbol{min_a}$: minimum value in numeric data column
$\boldsymbol{max_a}$: maximum value in numeric data column
$\boldsymbol{new\_max_a}$: new maximum value or range limit
$\boldsymbol{new\_min_a}$: new minimum value or range limit

2.3. Discrete

In the process of discrete where attributes that have continuous values are then changed in discrete form. This process is done aimed at minimizing the condition of the appearance of small continuous values, as it can affect in the selection process of features. The Binning method used to discrete variables in this study. For the amount of binning in this study use a minimum of 3 and a maximum of 12.

w = (max – min) / (no of bins)
(2)
Description :
w : interval limit
max = maximum value in numeric data
min = minimum value in numeric data
bins = interval size

2.4. Feature Selection

Feature selection is done to reduce less relevant features in the classification process. From the next processing result to the feature selection stage with Information Gain. Calculation of Information Gain. After calculating the 41 attributes with the gain value next select the attribute that has the gain value with the highest weight.

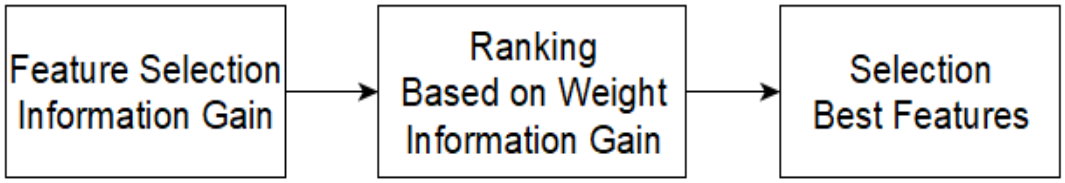

Figure 2. Information gain feature selection process

$Info\ (D) = -\sum_{i}^{c} P_i log_2$
(3)
Description :
c = Number of values in the target attribute (number of classification classes)
pi = Number of samples for class i

$InfoA\ (D) = \sum_{j=1}^{v} \frac{|Dj|}{D}\ x\ Info(Dj)$
(4)
Description :
A = Attribute
|Dj|= Total number of data samples
|D| = Number of samples for j value
v = A possible value for attribute A
Then the information gain value used to measure the effectiveness of an attribute in data claiming can be calculated with the formula below:

$Gain\ (A) = |Info\ (D) - InfoA(D)|$
(5)

3. Result and Discussion

The dataset in this study is an NSL-KDD'99 dataset that has gone through several processes before. The data used was 125,973 data consisting of 39 types of attacks. Of the 39 types of attacks grouped into categories of attacks. Here are the results of the feature workings obtained.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

Number of Classes	Number of features	Feature number
2	41	3,4,29,33,34,12,38,39,25,26,23,32,31,41,2,27,28,40, 36,24,30,35,37,8,1,19,10,22,15,17,14,18,16,11,13,5,7,6,9,21,20

4. Conclusion

From the results of the study can be concluded that information gain can be used to determine the effect of dataset attributes on classification. In the process of selection of dataset, features are carried out binary classification process (normal and attack). The selection method of features proposed in this study can help in the process of improving performance development in the Classification Of Intrusion Detection System (IDS).

References

[1] D. A. Effendy, K. Kusrini, and S. Sudarmawan, “Classification of intrusion detection system (IDS) based on computer network,” *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, pp. 90–94, 2018, doi: 10.1109/ICITISEE.2017.8285566.

[2] H. Malhotra and P. Sharma, “Intrusion Detection using Machine Learning and Feature Selection,” *Int. J. Comput. Netw. Inf. Secur.*, vol. 11, no. 4, pp. 43–52, 2019, doi: 10.5815/ijcnis.2019.04.06.

[3] J. Malviya, V. Patel, and A. Srivastava, “New Naïve Bayes Classifier for Improve Intrusion Detection System Accuracy,” pp. 7–11, 2017, [Online]. Available: www.irjeas.org,.

[4] T. Ahmad and M. N. Aziz, “Data preprocessing and feature selection for machine learning intrusion detection systems,” *ICIC Express Lett.*, vol. 13, no. 2, pp. 93–101, 2019, doi: 10.24507/icicel.13.02.93.

[5] A. Prof and S. H. Hashem, “Denial of Service Intrusion Detection System (IDS) Based on Naïve Bayes Classifier using NSL KDD and KDD Cup 99 Datasets University of Technology - Computer Science Department Hafsa Adil University of Technology - Computer Science Department,” no. 40, 2017.

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

This page is intentionally left blank