Feature Selection for efficient Intrusion Detection using Attribute Ratio

Hee-su Chae, Sang Hyun Choi

***Abstract*—**Network traffic is increasing due to the growing use of smart devices and the Internet. Most intrusion detection studies have focused on feature selection or reduction because some features are irrelevant or redundant which results in a lengthy detection process and degrades the performance of an intrusion identify important selected input features for building an Intrusion Detection System (IDS) that is computationally efficient and effective. To this end, we investigated the performance of standard feature selection methods; CFS(Correlation-based Feature Selection), IG(Information Gain) and GR(Gain Ratio). In this paper, we propose a new feature selection method using feature average of total and each class and applied efficient classifier decision tree algorithm for evaluating feature reduction method. Moreover, we compared the proposed method and other methods.

***Keywords*—**Data Mining, Preprocessing, Feature selection, Intrusion detection system.

I. Introduction

In recent year, due to the increasing use of smart devices and the Internet, the network traffic is rapidly increasing. In Cisco report, "Globall IP traffic in 2012 stands at 43.6 exabytes per month and will grow threefold by 2017, to reach 120.6 exabytes per month" [1].

Intrusions are defined as attempts or action to compromise the confidentiality, integrity or availability of computer or network . Intrusion detection systems (IDSs) are software or hardware systems that automate the process of monitoring the events occurring in a computer system or network, analyzing them for signs of security problems [2].

Feature selection is the process of removing features from the original data set that are irrelevant with respect to the task that is to be performed [3]. It can reduce both the data and the computational complexity, and can also get more efficient and find out the useful feature subsets [4]. So not only the execution time of the classifier that processes the data reduces but also accuracy increases because irrelevant or redundant features can include noisy data affecting the classification accuracy negatively [5]

In this paper, we suggest a new feature selection method that use attribute average of total and each class data. The decision tree classifier will be evaluated on the NSL-KDD dataset to detect attacks on the four attack categories: Dos, Probe, R2L, U2R. The feature reduction is applied using three standard feature selection methods Correlation-based Feature Selection (CFS), Information Gain (IG), Gain Ratio (GR) and proposed method. The decision Tree classifier's results are computed for comparison of feature reduction methods to show that our proposed model is more efficient for network intrusion detection. Rest of the paper is organized as follows: Section 2 give overview of IDS, Feature selection methods, and NSL-KDD. The experimental study discussed in section 3. The section 4 presents the result. Finally the paper is concluded with their future work in section 5.

II. Intrusion Detection System (IDS)

Intrusion is a type of attack that attempts to bypass the security mechanism of a computer system. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems [6].

There are three main strategies of IDS. First, misuse detection attempts to match patterns and signatures of already known attacks in the network traffic. A constantly updated database is usually used to store the signatures of known attacks. It cannot detect new attack until trained for them. Second, anomaly detection attempts to identify behavior that does not conform to normal behavior. This technique is based on the detection of traffic anomalies. The anomaly detection systems are adaptive in nature, they can deal with new attack but they cannot identify the specific type of attack [7]. Third, Specification-based detection depends on program specifications that describe the intended behavior of security-critical programs. The monitoring of executing programs involves detecting deviations of their behavior from these specifications, rather than detecting the occurrence of

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the "Employment Contract based Master's Degree Program for Information Security" supervised by the KISA (Korea Internet Security Agency) (H2101-13-1001). This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center)) support program (NIPA-2013-H0301-13-4009) supervised by the NIPA(National IT Industry Promotion Agency).

Hee-su Chae is with the Department of Information Security Management of Chungbuk National University, Chungbuk, South Korea (e-mail: enigma0724@gmail.com).

Sang Hyun Choi is with the Department of Management Information Systems, Chungbuk National University, 12 Gaeshin-dong, Heungduk-gu, Cheongju, Chungbuk 361-763, South Korea (corresponding author. BK21 Plus, Big Data Service Model Optimization team. phone: +82-43-261-3742; fax: +82-43-273-2355; e-mail: kimts@cbnu.ac.kr).

specific attack patterns. Thus, attacks can be detected even though they may not previously have been encountered[8].

Many researchers have proposed and implemented various models for IDS but they often generate too many false alerts due to their simplistic analysis. An attack generally falls into one of four categories[9]:

1) Denial-of-Service(DoS) : Attackers tries to prevent legitimate users from using a service. For example, there are smurf, neptune, back, teardrop, pod and land.

2) Probe: Attackers tries to gain information about the target host. Port Scans or sweeping of a given IP-address range typically fall in this category (e.g. saint, ipsweep, portsweep and nmap).

3) User-to-Root(U2R) : Attackers has local access to the victim machine and tries to gain super user privileges.cc For example, these are buffer_overflow, rootkit, landmodule and perl.

4) Remote-to-Local(R2L) : Attackers does not have an account on the victim machine, hence tries to gain access. For example, these are guess_passwd, ftp_write, multihop, phf, spy, imap, warezclient and warezmaster.

III. Feature Selection

Feature selection is important to improve the efficiency of data mining algorithms. Most of the data include irrelevant, redundant, or noisy features. Feature selection is process of selecting a subset of original features according to certain criteria, and an important and frequently used technique in data mining for dimension reduction. It reduces the number of features, removes irrelevant, redundant, or noisy features, and brings about palpable effects for applications: speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility [10].

There are two common approaches for feature reduction. A Wrapper uses the intended learning algorithm itself to evaluate the usefulness of features, while filter evaluates features according to heuristics based on general characteristics of the data. The wrapper approach is generally considered to produce better feature subsets but runs much more slowly than a filter [11].

In this paper we are using three feature subset selection methods Correlation-based Feature Selection (CFS), Information Gain (IG) and Gain Ratio(GR) to compare our proposed method.

A. Correlation-based Feature Selection (CFS)

CFS has two concepts. One is the feature-classification (r_{cfi}) correlation and another is the feature-feature (r_{fifj}) correlation. These two concepts are based on following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". The feature-classification correlation indicates how much a feature is correlated to a specific class. the feature-feature correlation is the correlation between two features[12]. CFS can be calculated as (Ghiselli 1964) :

$$M_s(k) = \frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (1)$$

In equation 1, $\overline{r_{cf}}$ is the average feature-classification correlation, and $\overline{r_{ff}}$ is the average feature-feature correlation.

B. Information Gain (IG)

The IG evaluates attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using MDL based discretization method[13]. Information gain for F can be calculated as [14]:

$$\text{Gain}(F) = I(c_1, \ldots, c_m) - E(F) \quad (2)$$

Expacted information ($I(c_1,\ldots,c_m)$) needed to classify a given sample is calculated by

$$I(c_1, \ldots, c_m) = -\sum_{i=1}^{m} \frac{c_i}{c} \log_2 \frac{c_i}{c} \quad (3)$$

C be set consisting of c data samples with m distinct classes. The training dataset ci contains sample of class I. $\frac{c_i}{c}$ is the probability that an arbitrary sample belongs to class Ci. Feature F has v distinct values $\{f_1, f_2, \ldots, f_v\}$ which can divide the training set into v subsets $\{C_1, C_2, \ldots, C_v\}$ where Ci is the subset which has the value f_i for feature F. Let C_j contain C_{ij} samples of class i.

The entropy of the feature F (E(F)) is given by

$$E(F) = \sum_{j=1}^{v} \frac{c_{ij}+\cdots+c_{mj}}{c} \times I(c_{ij}, \ldots, c_{mj}) \quad (4)$$

A. Gain Ratio (GR)

The gain ratio an extension of info gain, attempts to overcome information gain which prefers to select features having a large number of values[15]. Gain ratio applies normalization to info gain using a value defined as

$$\text{NorInfor}_F(C) = -\sum_{i=1}^{v} (|C_i|/|C|) \quad (5)$$

The above value represents the information generated splitting the training data set C into v partitions corresponding to v outcomes of a test on the feature F [13][16]. The gain ratio can be calculated as :

$$\text{Gain Ratio}(F) = \text{Gain}(F) / \text{NormInfor}_F(S) \quad (6)$$

IV. NSL-KDD Data Set

NSL-KDD data set suggested to solve some of the inherent problems of the KDDCUP'99 data set. KDDCUP'99 is the mostly widely used data set for the anomaly detection. But Tavallaee et al conducted a statistical Analysis on this data set and found two important issues which highly affects the performance of evaluated systems, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, they have proposed a new data set, NSL-KDD, which consists of selected records of the complete KDD data set[16] . The following are the advantages of NSL-KDD over the original KDD data set[5] :

First, it does not include redundant records in the train set, so the classifiers will not be biased towards more frequent records. Second, the number of selected records from each difficulty level group is inversely proportional to the percentage of records in the original KDD data set. As a result, the classification rates of distinct machine learning methods vary in a wider range, which makes it more efficient to have an accurate evaluation of different learning techniques. Third, the numbers of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable.

NSL-KDD data includes 41 features, 125,973 instances, and has 4 attacks type and normal data. Attacks can be categorized as the following :

Table. 1 Attack type and their related attack

Category	Attacks
dos (Denial of Service)	back, Neptune,pod, smurf, teardrop, process table, warezmaster, apache2, mail bomb.
Probe	http tunnel, ftp_write, multihop, buffer overflow, root kit, xterm, ps.
R2L (Root to Local)	guess_passwd, named, snmpgetattack, xlock, send mail
U2R (Unauthorized to Root)	ipsweep, nmap, port sweep, satan, mscan, saint

. Table. 2 shows number of instances for each class.

V. Experimental Study

We explained above that network traffic data is increasing rapidly. In order to detect intrusion from large traffic data, detection algorithm, and feature selection method have to more efficient. The above three feature selection methods use a complex calculation. For this reason, these methods is inefficient for large scale data. In this paper, we propose a simple and efficient feature selection method.

A. Descriptive Statistics of NSL-KDD

NSL-KDD data has three features types : Numeric, Nominal, and Binary. Features 2, 3, and 4 are nominal, features 7, 12, 14, 15, 21, and 22 are binary, and the rest of the features are numeric type. Table 4 show the average of feature 23 which is numeric type. The total average is bigger than normal, R2L, and U2R average value and less than the Dos and Probe average value.

Table. 2 Number of NSL-KDD

class	number
normal	67,343
dos	45,927
probe	11,656
R2L	995
U2R	52
total	125,973

Table 3. Type of features in NSL-KDD

Type	Features
Nominal	Protocol_type(2), Service(3), Flag(4)
Binary	Land(7), logged_in(12), root_shell(14), su_attempted(15), is_host_login(21),, is_guest_login(22)
Numeric	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), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38), dst_host_srv_serror_rate(39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

Table 3 show the average of feature 23 which is numeric type. The total average is bigger than normal, R2L, and U2R average value and less than the Dos and Probe average value. Therefore, We used average value to calculate AR of numeric data type.

Table. 4 Average value of feature 23

Class	Mean
Total	0.16459408
Dos	0.348512787
Normal	0.044066495
Probe	0.150787218
R2L	0.002539183
U2R	0.011365423

Table 5 shows frequency of feature 12 for each class and total. Feature 12 is binary type consisting of 0 and 1. There is

Table. 5 Frequency of feature 12

	Dos	Normal	Probe	R2L	U2R	Total
0	44970	19486	11573	86	6	76121
1	957	47857	83	909	46	49852

We propose feature selection method to using the Attribute Ratio(AR). AR is calculated by average value and frequency of features.

B. Proposed Method

In section 4, we explain NSL-KDD data which has three attribute types. We use attribute average and frequency for each class calculate the AR from numeric and binary type. AR can be calculated as :

$$\mathbf{AR(i) = MAX\left(CR(j)\right)} \quad (7)$$

Class Ratio (CR) is attribute is ratio of each class for Attribute i. CR is calculated by two methods according to the type of attributes. CR can be calculated as for numeric :

$$\mathbf{CR(j)} = \frac{\mathbf{AVG(C(j))}}{\mathbf{AVG(total)}} \quad (8)$$

CR can be calculated as for binary :

$$\mathbf{CR(j)} = \frac{\mathbf{Frequency(1)}}{\mathbf{Frequency(0)}} \quad (9)$$

After calculating AR(i), Features rank ordering larger AR. Table 4 shows the rank of features with a calculated AR. We did not use nominal type features to calculate AR.

C. Experimental Setup

We used WEKA 3.7 a machine learning tool [17], to compute the feature selection subsets for CFS, IG, and GR, and to evaluate the classification performance on each of these feature sets. We chose the J48 decision tree classifier [18] with full training set and 10-fold cross validation for the testing purposes. In 10-fold cross-validation, the available data is randomly divided into 10 disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining nine sets are used for building the classifier. The test set is then used to estimate the accuracy, and the accuracy estimate is the mean of the estimates for each of the classifiers. Cross-validation has been tested extensively and has generally been found to work well when sufficient data is available [14].

Table. 6 Calculated AR value of NSL-KDD

Rank	Feature	AR	Rank	Feature	AR
1	2	nominal	22	41	3.668
2	3	nominal	23	28	3.668
3	4	nominal	24	27	3.646
4	18	326.114	25	29	3.444
5	9	173.040	26	40	3.280
6	17	62.234	27	31	3.082
7	11	46.039	28	8	2.743
8	10	40.775	29	39	2.673
9	33	11.700	30	26	2.643
10	12	10.600	31	25	2.631
11	6	9.155	32	38	2.629
12	5	8.464	33	16	2.609
13	1	7.226	34	23	2.117
14	37	5.757	35	24	1.177
15	35	4.837	36	14	1.000
16	19	4.695	37	22	0.459
17	36	4.393	38	7	0.000
18	13	4.339	39	15	0.000
19	34	4.223	40	21	0.000
20	30	4.069	41	20	0.000
21	32	3.813			

VI. Results

We used three standards and on proposed method for feature selection. The feature selection performed on 41 features. We used selected features and all nominal features.

To evaluate the results of classifier, we used accuracy.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FN+FP+TN)} \times 100 \quad (10)$$

Table. 7 Accuracy and AR value

Features #	Accruncy(%)	AR
13	99.152	7.226
14	99.312	5.757
15	99.660	4.837
16	99.663	4.695
17	99.728	4.393
18	99.729	4.339
19	99.753	4.223
20	99.788	4.069
21	99.792	3.813
22	99.794	3.668
23	99.792	3.668
24	99.787	3.646
25	99.786	3.444
26	99.778	3.280
27	99.779	3.082
28	99.785	2.743

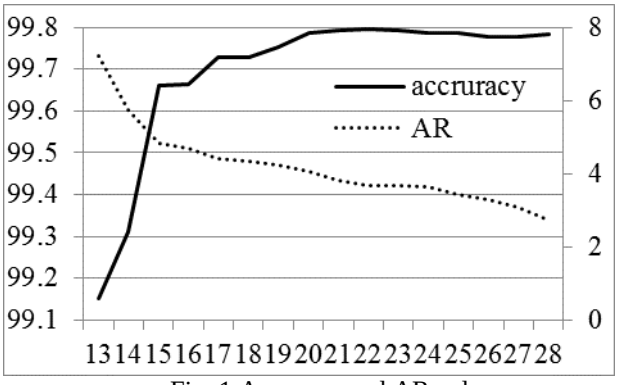

Fig. 1 Accuracy and AR value

Table 6 and Figure 1 show accuracy for the accumulation of the number of features using AR ranker. These show inverse correlation between accuracy and AR until 22 features. It is clear that the highest accuracy is 99.794% at 22 features.

Table. 8 Accuracy(%) of AR,CFS, Info.Gain, GR, Full data

Feature #	19	20	21	22	23	24	25
AR	99.75	99.78	99.79	**99.79**	99.79	99.78	99.78
CFS	99.76	99.76	99.77	99.77	99.77	99.76	**99.78**
Info. Gain	99.77	99.76	99.76	99.77	**99.78**	99.77	99.76
GR	**99.79**	99.79	99.78	99.78	99.77	99.77	99.78
Full data	99.76						

Table 8 and Figure 2 show accuracy of AR, CFS, Info.Gain, and GR for the accumulation of the number of features and Full data. Accuracy of full data is 99.763. CFS' highest accuracy is 99.781 with 25 features, IG is 99.781% with 23 features, and GR is 99.794 with 19 features

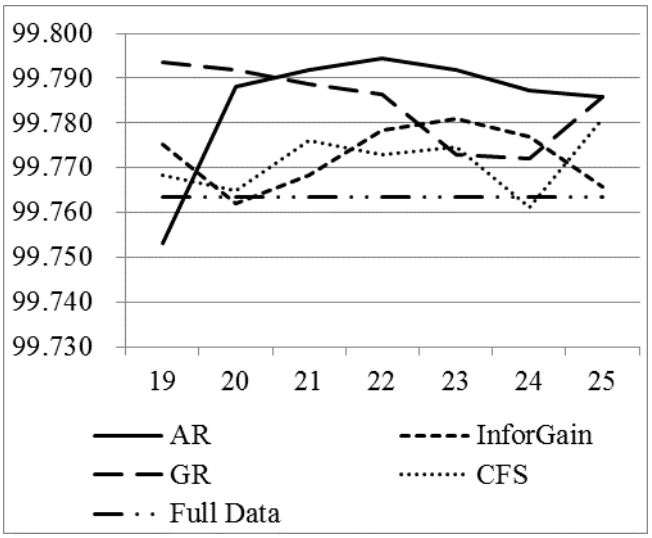

Fig. 2 Accuracy and AR value

VII. Conclusion and Future Work

In this paper we have proposed feature selection methos using AR and compared it with three feature selectors CFS, IG, and GR.

The experiment shows that between accuracy and AR value is inverse correlation in our feature selection method and the highest accuracy is 99.794% using 22 features. The accuracy of our method is higher than the accuracy of full data and is also as highly as accuracy of other methods. Future work will include a comparison of calculation time for our method and other methods. Also. we will calculate the True Positive Rate(TPR), False Positive Rate(FPR), and accuracy for each attack type.

Acknowledgment

This research was supported by the MSIP (Ministry of Science, ICT & Future Planning), Korea, under the "Employment Contract based Master's Degree Program for

Information Security" supervised by the KISA (Korea Internet Security Agency) (H2101-13-1001). This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center)) support program (NIPA-2013-H0301-13-4009) supervised by the NIPA(National IT Industry Promotion Agency)

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology", 2012-2017, Cisco, 2013.
[2] R.Bace and P. Mell, "NIST Special Publication on Intrusion Detection Systems", 2001.
[3] Yang, Yiming; Pedersen, Jan O., "A comparative study on feature selection in text categorization" *In: ICML*, 1997, pp. 412-420.
[4] Kun-Ming Yu, Ming-Feng Wu, and Wai-Tak Wong, "Protocol-based classification for intrusion detection", *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering. World Scientific and Engineering Academy and Society*, 2008, pp.29-34.
[5] Lakhina, Shilpa; Joseph, Sini; Verma, Bhupendra, "Feature reduction using principal component analysis for effective anomaly–based intrusion detection on NSL-KDD", *International Journal of Engineering Science and Technology Vol.2, No.6*, 2010, pp.1790-1799.
[6] Bace, R., "Intrusion Detection", Macmillan Technical Publishing, 2000.
[7] Anuar, Nor Badrul, Sallehudin, Hasimi, "Identifying false alarm for network intrusion detection system using data mining and decision tree", *Proceedings of the 7th WSEAS International Conference on DATA NETWORKS, COMMUNICATIONS, COMPUTERS*, 2008. pp.22-28.
[8] Rouached, M., Sallay, H., Ben Fredj, O., Ammar, A., Al-Shalfan, K., & Ben Saad, M., "Formal analysis of intrusion detection systems for high speed networks." *In Proceedings of the 9th Conference in Advances in E-Activities, Information Security and Privacy*. 2010. pp. 109-115.
[9] Srinoy, S., Chimphlee, W., Chimphlee, S., & Poopaibool, Y., "A fusion of ICA and SVM for detection computer attacks." *In: Proceedings of the 5th WSEAS international conference on Applied computer science. World Scientific and Engineering Academy and Society (WSEAS)*, 2006, pp. 986-990.
[10] Liu, H., Motoda, H., Setiono, R., & Zhao, Z., "Feature Selection: An Ever Evolving Frontier in Data Mining." *Journal of Machine Learning Research-Proceedings Track 10*, 2010, pp.4-13.
[11] Kim, Yong, W. Nick Street, and Filippo Menczer., "Feature selection in data mining." *Data mining: opportunities and challenges Vol.3, No.9*, 2003, pp.80-105.
[12] S. Doraisami, S. Golzari, "A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music", *Content-Based Retrieval, Categorization and Similarity 1*, 2008, pp.331-336.
[13] j.Han ,M Kamber, "Data mining : Concepts and Techniques", Morgan Kauffmann Publishers, 2001.
[14] Saurabh Mukherjeea, Neelam Sharmaa, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", *Procedia Technology 4*, 2012, pp.119-128.
[15] I.H.Witten, E.Frank, M.A. Hall, "Data Mining Practical Machine Leanrning Tools & Techniques Third edition", Morgan kouffman. 2011.
[16] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set", *2009 IEEE Int. Conf. Comput. Intell. Security Defense Appl.*, 2009, pp.53–58.
[17] http://www.cs.waikato.ac.nz/~ml/weka/
[18] G. Kalyani, A. Jaya Lakshmi, "Performance Assessment of Different Classification Techniques for Intrusion Detection", *IOSR Journal of Computer Engineering (IOSRJCE) Vol.7, No.5*, 2012, pp.25-29.

Hee-su Chae is a master course student at the Department of Management Information Systems at Chungbuk National University. He received his bachelor degree in Science in Business Administration from Chungbuk National University. His research areas include data analytics, data mining, big analytics, information security, and business process analytics.

Sang Hyun Choi received the B.S. degree in industrial engineering from Hanyang University, Seoul, Korea, in 1991, the M.S. degree in industrial engineering and the Ph.D. degree in management information systems from Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 1993 and 1998, respectively.
From 1998 to 2002, he was a Senior Consultant at the Entrue Consulting, LG CNS. He is currently an Professor in the Department of management Information Systems, Chungbuk National University, Chungbuk, Korea. He has authored or coauthored more than 20 papers in international conference proceedings and journals. His research interests include bigdata analytics, recommendation systems, data mining, electronic commerce, and information systems planning.