Denial of Service Intrusion Detection System (IDS) Based on Naïve Bayes Classifier using NSL KDD and KDD Cup 99 Datasets

Asst. Prof. Dr. Soukaena H. Hashem
soukaena.hassan@yahoo.com
University of Technology - Computer Science Department
Baghdad, Iraq
Hafsa Adil
h_adel_89@yahoo.com
University of Technology - Computer Science Department
Baghdad, Iraq

Abstract: *Intrusion Detection Systems (IDS) become necessary to protect data from intruders and reduce the damage of the information system and networks especially in cloud environment which is next generation Internet based computing system that supplies customizable services to the end user to work or access to the various cloud applications. This paper concentrates the views to be noted that; the attacks in cloud environment have high rates of Denial of service (DoS) attacks compared with the usual network environment. This paper will introduce Naïve Bayes (NB) Classifier supported by discrete the continuous feature and feature selection methods to classify network events as an attack (DoS, Probe, R2L and U2R) or normal. The influence of use all features and use set of features by applying two methods of feature selection methods has been studied in this paper. The performance of the proposed system was evaluated by using KDD 99 CUP and NSL KDD Datasets, and*

from experimental works the results are; proposal improves the performance of NIDS in term of accuracy and detecting DOS attack, where it detected 94%, 97% and 98% of DoS attacks for three experimental test datasets in KDD Cup 99 dataset when used twelve features selected by gain ratio, while in NSL KDD Dataset the accuracy of detecting DoS is 86%, 87% and 88% for three experimental test datasets when select only ten features by applied gain ratio.

Keywords: IDS, Data mining, Multiclass classification, and Naïve Bayes (NB)

1. Introduction

Now in these days Cloud environment growing rapidly and is an ingenious model that make the users access Internet based applications and data storage services in easy way. The data stored in remote data center can be accessed or managed through the cloud services provided by the cloud service providers. So the data store should be done with utmost care for data preprocessing [1]. As a result of the growing popularity of cloud, it is important to provide security to cloud environments, but the traditional security methods are not enough to ensure security for cloud environments, so it is important to provide IDS which becomes an essential component in terms of cloud security [2].

IDS is a security tool used to strengthen the security of information system and communication by detecting unauthorized intrusions into networks and computer systems [3]. Intrusion detection (ID) methods can be classified into misuse detection and anomaly detection, where in misuse detection the system collect the information, analyze and compare with huge databases of attack signatures while in anomaly detection, the system manager defines the baseline or criterion such as protocol, network traffic load, packet size and breakdown [4].

Data mining (DM) is the process of extracting relevant information from huge database, ID is a data analysis process where DM techniques are used to automatically discover and detect normal and intrusive patterns. DM commonly involves four classes of task. Clustering, Classification, Regression and Association rule learning [5]. Classification is the process of taking every instance in the dataset and determining which class is belong to, that means known structure will be used for new instances [6].

2. Naïve Bayes Classifier

Naïve Bayes classifier (NB) is a popular DM classification method that has been applied to several fields, including ID, which depend on applying Bay's theorem with strong independence assumption. That means the probability of a feature doesn't influence the probability of the other features [7]. NB classifier has two types of variables: the class C variable and a set of features X =$\{X_1; X_2;\ldots;\ X_n\}$, on a dataset D which consists of $\{E_1, E_2,\ldots, E_t\}$ instances and can be defined as the Eq. (1), Then, with the consideration of Naïve assumption of independence of the attributes given the class as in Eq. (2) [8].

$$c(E) = \arg\max_{c\in C} P(c)P(a_1, a_2, \ldots, a_n|c) \qquad \text{Eq. (1)}$$

$$P(E|c) = P(a_1, a_2, .., a_n|c) = \prod_{i=1}^{n} P(a_i|c) \qquad \text{Eq. (2)}$$

The conditional independence assumption leads to posterior probabilities. NB classifier is easily constructed because of the simplicity of computing P (C) and $P(a_i|c)$. NB classifiers simplify the computations and give high accuracy and speed when applied to large databases [9].

The feature selection method is still an important preprocessing step in IDS; one of the most common methods in feature selection is information gain (IG) which measure the information gain of each attribute by evaluating the worth of an attribute based on

entropy with respect to the class; the attribute which have higher entropy is the more information content. Entropy can be defined as a measure of uncertainty of the system. The Expected information (Entropy) of a feature Y is defined as Eq. (3), Information needed (after using A to split D into v partitions) to classify D is mention in Eq.(4). Information gained by branching an attribute A as in Eq. (5) [10].

$$\text{Info D} = H(Y) = -\sum_{y \in Y} P(y) \log_2 P(y) \qquad \text{Eq. (3)}$$

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * I(D_j) \qquad \text{Eq. (4)}$$

$$\text{Gain (A)} = \text{Info D} - Info_A(D) \qquad \text{Eq. (5)}$$

Gain Ratio (GR) is a feature selection method that adjusts the information gain for each attribute to allow for breadth and uniformity of the attribute values [11]. GR takes the size and number of branches into consideration when selecting an attribute as It corrects the information gain by taking the substantial information of a split into consideration (i.e. the amount of information needed to determine which branch an instance related to) where substantial information is the entropy of a distribution of instances into branches as Eq.(6). This value is the potential information caused as a result of splitting the training dataset as in Eq.(7) [12].

$$\text{Split Info}_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \qquad \text{Eq. (6)}$$

$$\text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{Split Info (A)}} \qquad \text{Eq. (7)}$$

3. Related Work

In [7], Mukherjeea S. et al., 2012, discussed the importance of reduce features to build efficient and effective IDS. They investigated the performance of feature selection methods using three methods (Information Gain, Gain Ratio and Correlation based Feature Selection); they proposed method for Feature Vitality Based Reduction Method to identify the importance of reduce feature. They applied NB classifier on NSL KDD dataset for ID. Experimental results showed that selected some Features give better performance to design effective and efficient NIDS.

In [11], Ghosh P. et al., 2014, proposed a Hybrid Multilevel IDS to Cloud Environment by applied KNN as a binary classifier for anomaly detection. Neural Network is applied for detecting abnormal classes after KNN classification. Before classification, feature selection has been used to select relevant features. They used NSL-KDD dataset where all samples of "KDDTrain+" used as training dataset and "KDDTest+" samples are used as testing dataset. They used Rough Set Theory and Information Gain to select relevant features. Experimental results show that they get better accuracy with their proposed hybrid KNN_NN classifier model for Intrusion Detection.

In [2], Padmakumari P. et al., 2014, presented IDS to Cloud Environment, by use k-means clustering and combine it with a frequent attacks generation module for anomaly detection by applied Apriori algorithm to detect attacks that are frequently occurring in various network environments. Experimental result showed that applying clustering algorithm separately for different attributes enhance the accuracy of detection. The frequent attack detection module achieved increasing reliability and reduce false alarm rate. They used KDD 99 CUP dataset to evaluate the performance of their system.

4. Datasets and Attacks in Cloud Environment

The KDD Cup 99 10 % dataset is knowledge discovery competition for network intrusion detection composed of 10% of the original dataset that is about 494,020 records each of which contains 41 features (continuous and discrete see Table1) and is labeled either normal or attack (DoS, Probe, R2L, U2R). The dataset has 80.31% attack and 19.69% normal connections. It has been most widely used in attacks on network [13]. NSL KDD dataset is a reduced version of KDD 99 dataset, which composed of the same features as KDD cup 99. The class feature has 21 types that fall under four kinds of attacks: DOS attacks, Probe attacks, R2L attacks and U2R attacks see Table 2 [14].

Table 1: Attributes of KDD Cup 99 and NSL KDD Datasets

Continuous	Duration, src_bytes, dst_bytes, worong_fragment, urgent, hot, num_failed_login, num_compromised, num_root, num_file_creations, num_shell, snum_access_files, num_outbound_cmds, count, serror_rate, error_rate, same_srv_rate, diff_srv_rate, srv_count, srv_serror_rate, sre_rerror_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate.
Discrete	Protocol_type, service, flag, land, logged_in, root_shell, su_attempted, is_hot_login, is_guest_login.

Table 2: Description of attacks in KDD Cup 99 and NSL KDD Datasets

Attack type	Description	Types
DOS	Denial of Service attacks	Pod ,Land , smurf , back etc.
Probe	Surveillance and probing	Satan, ipssweep, nmap etc.
R2L	Unauthorized access from remote machine to local machine	Guess_passwd, ftp_write, imap, phf etc.
U2R	Unauthorized access to local superuser priviledges by a local unpiviledge user	Rootkit, buffer overflow, loadmodule etc.

The NSL KDD dataset is different from the KDD 99 dataset in the following points:[15]

(1) It doesn't have redundant records in training dataset.
(2) It doesn't have duplicate records in testing dataset.
(3) From every difficulty level set, the number of records that selected is inversely commensurate to the percentage of records in KDD 99 dataset.

Cloud computing is used to overcome overhead on information technology (IT) of systems and users by increase system flexibility and provide high security level and decrease total cost of ownership. Cloud environment compromise the availability, confidentiality and integrity of resources or computer systems. But one of the most important requirements in security of cloud environment is availability. This Challenge appears as DoS attack, a major threat to availability which makes resources or services unavailable for indeterminate period of time by flooding it with useless traffic [16].

5. Proposal NIDS for Cloud Environment

The proposed system is multiclass NIDS based on NB classifier applied to detect DoS attacks that consider the most dangers attacks especially in cloud environment, the reason of using network intrusion detection is that the host intrusion detection can be easily detected by antivirus so the important is to detect the network intrusion, to evaluate the system we used the well-known dataset KDD Cup 99 and NSL KDD Dataset, Figure (1) Depicts the general structure of the proposed NIDS.

The proposed system consists of the following steps

1. Normalization.
2. Discretization.
3. Feature Selection.
4. Training and Testing.

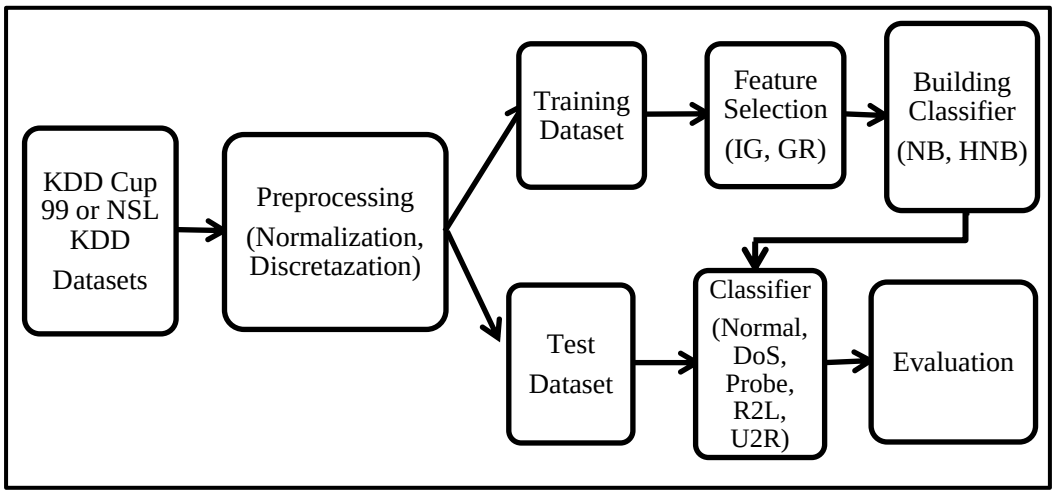

Figure 1: Block Diagram of the proposed NIDS

5.1 Normalization dataset

The first step after gain the dataset from Internet is conversion to access database then applied the normalization process to the continues feature which is so important to improve the performance and effectiveness of the system by make the values

of attribute within specific range from 0 to 1, in our system we used the Min-Max normalization process (see Algorithm 1).

Algorithm (1) Normalization Dataset

Input: Continuous feature of Datasets(training and testing).

Output: Normalization dataset with values between 0 and 1

Begin

Steps:

For each Feature in Dataset

  Find the Maximum value (Max)

  Find the Minimum value (Min)

    For each value X in Feature

      $Value\ X = \frac{ValueX - Min}{Max - Min}$

    End For

End For

End

Example of Normalization

For feature duration

Maximum value= 15168

Minimum value=0

Suppose X= 5051 then

$Xnew = \frac{5051 - 0}{15168 - 0} = 0.33$

5.2 Discretization Dataset

Since the KDD Cup 99 and NSL KDD Datasets contains continues and discrete feature it is so important to convert the continuous attribute to discrete to ensure the efficiency of the system and to solve the problem of appearing new value when test dataset which it's not appeared in training dataset.

5.3 Feature Selection

One of the most important preprocessing in data mining techniques is feature selection methods that used to remove the redundant and unrelated features in large dataset like KDD Cup 99 and NSL KDD, and to enhance the performance of the system by using the correct features and reduce the consuming time. In this study, we used information gain and gain ratio feature selection methods (see Algorithm 2).

Algorithm 2: Feature selection based on info gain and gain ratio

Input: Training dataset after normalization and discrete processes
Output: Set of feature which have highest gain values

Begin
Steps:

1) compute the size of training dataset D
2) For each class in dataset
   find the Pro(c) by dividing the frequency of class on D
   compute the entropy of five class to find info D by use Eq.3

$$\text{Info D} = -\sum P\text{ro(normal)}\,\text{Log}_2\,\text{P(normal)} + \text{pro(DOS)}\,\log_2\,\text{p(DOS)} + \text{pro(Probe)}\,\text{log2}\,\text{pro(probe)}\ldots..\text{etc.} \quad \text{Eq.(3)}$$

   End for

3) For each Feature F in training dataset
   For each value j in Feature F
   compute the frequency of value in all training dataset Ft
   compute the frequency of value with each class F1,F2,F3,F4,F5
   compute the entropy for each value with the five class by using Eq.(3)

$$\text{I}(D_j) = \sum (\frac{F1}{Ft}\text{Log}_2\ \frac{F1}{Ft}) - (\frac{F2}{Ft}\log_2\frac{F2}{Ft}).\ldots\ldots\text{etc.}$$

   End For
End For

4) For Each Feature in training dataset
   Compute info A by used Eq.(4), where v is number of values

$$Info_A(D) = \sum_{j=1}^{v} \frac{|Ft|}{|D|} * I(D_j)$$

   Compute gain for each Feature as in Eq.(5)

Gain (A)= Info D - $Info_A$(D)

End For

5) For each Feature in training dataset
   Compute Split Info by use Eq.6

$$Split\ Info_A(D) = -\sum_{j=1}^{v} \frac{|Ft_j|}{|D|} \log_2 \frac{|Ft_j|}{|D|}$$

   Compute the Gain ratio by use Eq.(7)

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ Info\ (A)}$$

End For

6) IF information gain is used as feature selection Then
   Select the set of features that have the highest gain
   Else if gain ratio is used for feature selection Then
   Select the set of features that have the highest gain ratio
   End IF

5.4 Training and Testing

In learning step the system applied NB classifier (see Algorithm 3) on 4000 records in the training process by select 2169 DOS, 388 probes, 173 R2L, 35 U2R and 1235 normal in both datasets (KDD cup 99 and NSL KDD).

In test phase 1200 samples are used to evaluate the work in KDD Cup 99 Dataset and two other datasets (600,900) samples to validate the performance of the system, where the numbers of samples selected for each class demonstrate in (Table 3). While in

NSL KDD Dataset the test samples that have been used is 1028 and two other dataset to validate the performance with (795 and 566), see Table 4 to demonstrate the numbers of samples selected for each class. The reason of selecting test samples of NSL KDD Dataset with different size from the original KDD Cup 99 Dataset is that, in NSL KDD Dataset the samples of attack is less than the KDD Cup 99 Dataset as a result of removing the redundant records in NSL KDD Dataset, it is important to note that the selection process have been chosen randomly.

Table3: Test KDD Cup 99 Dataset selected

Dataset	DOS	Probe	R2L	U2R	Normal
600	342	74	23	4	157
900	515	111	36	5	233
1200	680	133	53	8	326

Table 4: Test NSL KDD Dataset selected

Dataset	DOS	Probe	R2L	U2R	Normal
566	326	68	10	6	156
795	434	100	17	11	233
1028	539	122	24	13	330

Algorithm 3: Naïve Bayes Classifier

Input: Training and Testing dataset after normalization and discrete processes

Output: Classification the test dataset

Begin

Step1:**Training phase**

1) For each class c in training dataset
   Compute P(c) from training dataset
   End for
2) For each feature F_i in training dataset
   For each value v_j and c in training dataset
   find the frequency of v_j with c1, c2,c3,c4,c5
   if v_j with c1=0 or v_j with c2=0 or v_j with c3=0 or v_j with c4=0 or v_j with c5 =0
   then probability of v_j =(freq. of v_j +1)/(freq. of c+ No. of value in F_i)
   else probability of v_j =Freq. of v_j /freq. of class
   End for
   End for

Step 2: Testing phase

3) For each record in test dataset
   For each value in test dataset
   find probability of v_j with c in training dataset
   End for
   multiply the probability of each record as Eq.(2)

$$\mathrm{P(E|c)} = P(a_1, a_2, \dots, a_n|c) = \prod_{i=1}^{n} P(a_i|c)$$

   Classify the record by Multiply the result of Eq. 2 with probability of class and choose the maximum value to classify the record as Eq.(1)

$$\mathrm{c\ (E)} = \arg\max_{c \in C} P(c) * P(a_1, a_2, .., a_n|c)$$

   End For

End

Example of Naïve Bayes Classifier

To demonstrate how NB applied in NIDS, see the following example

1- Find the probability of each class where the probability of class = (Frequency of class in the training dataset) / (total size of training dataset)
Pro(normal)=1235 / 4000 = 0.30875

Pro(Dos)=2169 / 4000 = 0.54225

Pro(probe)= 388 /4000 = 0.097

Pro(R2L)= 173/4000 = 0.04325

Pro(U2R)= 35/ 4000= 0.00875

2- Finding the probability of each values in each features for all class where the probability of value within class = (Frequency of value in feature that appear within class) / (Frequency of class in training dataset).

One of the problems that occurred in NB classifier is zero probability that happens when frequency of value in one class or more is zero that leads to make the result of test is zero since the posterior probabilities used multiplication operand. The zero probability is solved by using Laplace estimator where adding one to the frequency of values and adding a number of values in feature to the frequency of class in training dataset. Probability of value within class = (Frequency of value in the feature that appear within class +1) / (Frequency of class in training dataset + number of values in feature), Table 5 show the Frequency of each value in attribute flag for all class, while Table 6 shows the probability for each value**.**

Example:

In attribute flag when the value is =0 the frequency of value in U2R class is 0

Probability (= 0, U2R) = $\frac{0+1}{35+3}$ = 0.0263

Table 5: Frequency of each value in attribute flag for all class

Attribute name	Values range	Normal	DoS	Probe	R2L	U2R
Flag	= 0	37	1969	376	53	0
	0_0.1	1198	200	12	109	35
	0.1_1	0	0	0	11	0

Table 6: Probability of each value in attribute flag for all class

Attribute name	Values range	Normal	DoS	Probe	R2L	U2R
Flag	=0	0.03069	0.9069	0.964	0.3068	0.0263
	0_0.1	0.97004	0.0922	0.0309	0.63005	1
	0.1_1	0.0008	0.00046	0.00255	0.0681	0.0263

Test phase will use Eq.(2) to find the posterior probability for each record in test dataset, Table 7 describes the test phase by selecting one record and find the probability for each value in each class. Then use Eq.(1) will be use to classify the record as follows:

Pro(normal)* pro(record| normal)= 0.30875* 0.0017529= 0.0005412

Pro(DoS)* pro(record|DoS)= 0.54225* 0.00006022= 0.00003265

Pro(probe)* pro(record| Probe)= 0.097* 0.00000539= 0.0000000523

Pro(R2L)* pro(record| R2L)= 0.04325* 0.007008= 0.00303

Pro(U2R)* pro(record| U2R)= 0.00875* 0.000738= 0.000000646

By taking the maximum value, the record will be classified as normal event

Table 7: Test for one record

No.	Attribute name	Record value	Normal	DoS	Probe	R2L	U2R
1	Flag	=0	0.03069	0.9069	0.964	0.3068	0.0263
2	Num_failed_login	=0	1	1	1	0.791907	0.971428
3	Logged_in	1	0.952226	0.0922	0.0154	0.658959	0.942857
4	Root_shell	0	0.999190	1	1	0.982658	0.514285
5	Num_file_ creation	0	0.999190	1	1	0.959537	0.514285
6	Num_shells	0	1	1	1	0.994219	0.885714
7	Is_guest_ login	0	0.989473	1	1	0.641618	1
8	Diff_srv_ rate	=0	1	0.977408	0.221649	0.803468	1
9	Srv_count	=0	1	0.964038	0.280927	0.763005	1
10	Dst_host_srv_diff _host_ rate	0.01	0.06072	0.000922	0.0103	0.190751	0.142857
11	Dst_host_rerror_r ate	=0	1	0.910557	0.203608	0.774566	0.971428
12	Dst_host_srv_rerr or_rate	=0	1	0.910557	0.278350	0.803468	0.942857
Pro_class			0.30875	0.54225	0.097	0.04325	0.00875
Result of multiplication			0.0017529	6.022E-05	5.39E-07	0.007008	0.000738
Result			0.0005412	3.265E-05	5.23E-08	0.000303	6.46E-06

6. Experimental Work and Result

The proposed network intrusion detection system used KDD cup 99 and NSL KDD Dataset to evaluate the system where the records selected randomly; the system tests three classifiers (NB classifier, NB classifier with info Gain and NB classifier with Gain Ratio).

6.1 Performance Measure

To evaluate the proposed NIDS effectiveness, it was used confusion matrix, accuracy, detection rate and error rate, the confusion matrix is a quality measurement of the classifier.

$$DR = \frac{TP}{TP+FN}$$

$$ER = \frac{FP+FN}{TP+TN+FP+FN}$$

$$\text{Accuracy binary} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Accuracy of class} = \frac{\text{number of samples of a class correctly classified}}{\text{number of samples of the class}}$$

$$\text{Accuracy of multiclass} = \frac{\sum \text{number of samples of a class correctly classified}}{\text{total number of test samples}}$$

Where

TP: number of attack events correctly classified as attack.

TN: number of normal events correctly classified as normal.

FP: number of normal events are incorrectly classified as attacks.

FN: number of attack events are incorrectly classified as normal.

6.2 KDD CUP 99 Dataset Evaluations

Table 8 shows the evaluation of classification in three KDD cup 99 test datasets with feature selection methods, (accuracy binary) mean the accuracy of detecting normal and attack while the (accuracy of multiclass) is the accuracy of detecting normal, DoS, Probe, R2L and U2R.The table also shows the detection rate (DR) and error rate (ER).

As shown in Table 8, the results are close to each other, but the importance of the proposed system is the ratio of detecting DoS attacks that viewed in Table 9 which mentioning the accuracy of each class that demonstrates the accuracy of detecting DoS attack is best when selecting 12 features by used gain ratio.

Table 10, Table 11, and Table 12 demonstrate the confusion matrix for best result in Test1, Test2, and Test3 of KDD Cup 99 dataset based on selection 12 features by applied gain ratio, Table 5 shows the rate of detecting R2L as R2L attack and U2R as U2R attack is low but when you look at Table10, Table11, and Table12 you can observe that, the proposed system is detected it but as another kind of attacks.

Table 8: performance measure of KDD cup 99 Dataset

DS	No. Feature & FS	Accuracy multiclass	Accuracy binary	DR	ER
Test1	41	0.92	0.98	100	0.01
Test1	20 IG	0.92	0.99	0.98	0.01
Test1	12 IG	0.91	0.98	0.98	0.02
Test1	20 GR	0.91	0.97	0.98	0.25
Test1	12GR	0.91	0.97	0.97	0.03
Test2	41	0.91	0.98	100	0.01
Test2	20 IG	0.92	0.96	0.96	0.03
Test2	12 IG	0.93	0.97	0.99	0.02
Test2	20 GR	0.91	0.97	0.99	0.02
Test2	12GR	0.92	0.96	0.97	0.03
Test3	41	0.96	0.98	100	0.01
Test3	20 IG	0.96	0.98	0.97	0.01
Test3	12 IG	0.95	0.98	0.99	0.01
Test3	20 GR	0.96	0.98	100	0.01
Test3	12GR	0.93	0.97	0.97	0.02

Table 9: Accuracy for each class in KDD Cup 99 Dataset

DS	No. Feature & FS	DOS	Probe	R2L	U2R	Normal
Test1	41	0.90	100	0.95	0.25	0.93
Test1	20 IG	0.90	0.98	0.73	0.25	0.99
Test1	12 IG	0.89	100	0.73	0	0.96
Test1	20 GR	0.90	0.98	0.78	0.25	0.93
Test1	12GR	0.94	0.89	0.47	0.50	0.94
Test2	41	0.92	0.90	0.77	0	0.93
Test2	20 IG	0.93	0.91	0.80	0	0.96
Test2	12 IG	0.95	0.91	0.72	0	0.92
Test2	20 GR	0.92	0.91	0.75	0	0.93
Test2	12GR	0.97	0.83	0.38	0	0.93
Test3	41	0.97	0.96	0.98	0	0.95
Test3	20 IG	0.95	100	0.98	0	0.99
Test3	12 IG	0.96	100	0.92	0	0.95
Test3	20 GR	0.97	100	0.96	0	0.95
Test3	12GR	0.98	0.87	0.49	0.62	0.95

Table 10: confusion matrix for test1

	Normal	DOS	probe	R2L	U2R
Normal	**148**	**0**	**0**	**9**	**0**
DOS	**0**	**322**	**0**	**20**	**0**
probe	**0**	**8**	**66**	**0**	**0**
R2L	**9**	**3**	**0**	**11**	**0**
U2R	**0**	**1**	**0**	**1**	**2**

Table 11: confusion matrix for test2

	Normal	DOS	probe	R2L	U2R
Normal	218	6	0	9	0
DOS	0	504	0	11	0
probe	6	12	93	0	0
R2L	5	8	6	14	3
U2R	5	0	0	0	0

Table 12: confusion matrix for test3

	Normal	DOS	probe	R2L	U2R
Normal	310	1	0	15	0
DOS	0	669	0	11	0
probe	0	16	117	0	0
R2L	18	7	0	26	2
U2R	0	1	0	2	5

6.3 NSL KDD Dataset Evaluations

The evaluation of classification in three NSL KDD test datasets that viewed in Table 13 (accuracy binary, accuracy of multiclass, detection rate (DR) and error rate (ER)) proved that, the best result of detect attack and normal when select 10 features by applied gain ratio, while the accuracy for each class that shown in Table 14 demonstrate the accuracy of detecting DoS attack in Test1 is high when select 12 and 10 fetaures by applied IG but the remain attacks cannot detect it , so it's not efficient that means the best result when select 10 or 12 features by use GR, while in Test2 and Test3 the results of select 10 and 12 feature is close to each other**.**

Table 13: Performance measure of NSL KDD Dataset

DS	No. Feature & FS	Accuracy multiclass	Accuracy binary	DR	ER
Test1	41	0.68	0.88	0.97	0.1
Test1	20 IG	0.68	0.80	0.87	0.2
Test1	12 IG	0.80	0.82	0.88	0.18
Test1	10 IG	0.80	0.82	0.88	0.18
Test1	20 GR	0.79	0.84	0.90	0.16
Test1	12GR	0.76	0.83	0.90	0.17
Test1	10GR	0.74	0.89	0.99	0.11
Test2	41	0.72	0.86	100	0.1
Test2	20 IG	0.76	0.85	0.98	0.1
Test2	12 IG	0.65	0.80	100	0.15
Test2	10 IG	0.64	0.81	100	0.15
Test2	20 GR	0.69	0.84	0.97	0.12
Test2	12GR	0.78	0.84	0.89	0.12
Test2	10GR	0.72	0.89	0.98	0.08
Test3	41	0.76	0.91	0.97	0.08
Test3	20 IG	0.81	0.85	0.88	0.14
Test3	12 IG	0.82	0.86	0.89	0.13
Test3	10 IG	0.82	0.84	0.87	0.16
Test3	20 GR	0.83	0.86	0.89	0.14
Test3	12GR	0.72	0.89	0.97	0.08
Test3	10GR	0.76	0.92	0.98	0.08

Table 14: the accuracy for each class in NSL KDD

DS	No.Feature & FS	DOS	Probe	R2L	U2R	Normal
Test1	41	0.67	100	60	16	0.63
Test1	20 IG	0.67	100	0	0	0.63
Test1	12 IG	0.86	100	0.20	0	0.64
Test1	10 IG	0.86	100	0.10	0	0.66
Test1	20 GR	0.85	0.98	0.80	0	0.66
Test1	12GR	0.88	0.63	0.70	0.50	0.63
Test1	10GR	0.88	0.35	0.70	0.33	0.61
Test2	41	0.51	0.93	0.57	0.05	0.54

Test2	20 IG	0.98	0.61	0	0	0.52
Test2	12 IG	100	0	0	0	0.37
Test2	10 IG	100	0	0	0	0.32
Test2	20 GR	0.50	0.95	0.25	0.23	0.54
Test2	12GR	0.89	0.64	0.70	0.54	0.70
Test2	10GR	0.87	0.35	0.29	0.36	0.66
Test3	41	0.64	100	0.62	0.15	0.78
Test3	20 IG	0.85	100	0	0	0.78
Test3	12 IG	0.86	98	0.20	0	0.79
Test3	10 IG	0.85	100	0.20	0	0.77
Test3	20 GR	0.86	98	0.87	0	0.78
Test3	12 GR	0.87	0.35	0.22	0.36	0.66
Test3	10 GR	0.86	0.35	0.62	0.38	0.76

Table 15, Table 16, and Table 17 demonstrate the confusion matrix for Test 1, Test2, and Test3 of NSL KDD dataset based on the ten selected features by using a gain ratio method.

Table 15: confusion matrix for test1

	Normal	**DOS**	**probe**	**R2L**	**U2R**
Normal	97	58	0	1	0
DOS	0	290	6	30	0
probe	0	44	24	0	0
R2L	0	3	0	7	0
U2R	1	2	0	1	2

Table 16: confusion matrix for test2

	Normal	**DOS**	**probe**	**R2L**	**U2R**
Normal	156	69	0	8	0
DOS	1	378	16	39	0
probe	0	65	35	0	0
R2L	7	5	0	5	0
U2R	2	2	1	2	4

Table 17: confusion matrix for test3

	Normal	DOS	probe	R2L	U2R
Normal	252	70	0	8	0
DOS	0	468	21	50	0
probe	0	79	43	0	0
R2L	1	8	0	15	0
U2R	2	1	2	3	5

7. Conclusions

This paper indicates the importance of using NIDS to detect the most harmful attack in network which is a DoS attack that effect the availability of the resource, The experimental results proved that, when use Naïve Bayes classifier supported by discretization and feature selection by applied gain ratio and selecting only 12 features from 41 features in KDD Cup 99 dataset, the proposed system achieves high accuracy rate, reduce the computation time and reduce the error rate as mention in Table 5, while in NSL KDD it is best to select only 10 features by applied gain ratio method as shown in Table 9 . The proposed system proved that. The use KDD Cup 99 dataset to detect DoS attack is better than NSL KDD, and use gain ratio as feature selection method is better than information gain.

Reference

[1] Mell P. and Grance T., "The NIST Definition of Cloud Computing", National Institute of Standards and Technology, 2011.

[2] Padmakumari P., Surendra K., Sowmya M. and Sravya M.," Effective Intrusion Detection System for Cloud Architecture", ARPN Journal of Engineering and Applied Sciences VOL. 9, NO. 11, 2014.

[3] Agrawal S. and Agrawal J., "Survey on Anomaly Detection using Data Mining Techniques", Elsevier B.V., 2015.

[4] Lodhi M. B., Richhariya V. and Parmar M., "Survey on Data Mining based Intrusion Detection Systems", International Journal of Computer Networks and Communications Security Vol. 2, No. 12, 2014.

[5] Tewatia R. and Mishra A., "Introduction To Intrusion Detection System: Review", International Journal Of Scientific & Technology Research, Vol. 4, Issue 05, 2015.

[6] Abhaya, Kumar K., Jha3 R. and Afroz S., "Data Mining Techniques for Intrusion Detection: A Review", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 6, 2014.

[7] Mukherjeea S. , Sharmaa N., "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", Elsevier Ltd., 2012.

[8] Ibáñez A., Bielza C. and Larrañaga P., "Cost sensitive selective naïve Bayes classifiers for predicting the increase of the h-index for scientific journals", Elsevier B.V., 2014.

[9] Yassin W., Udzir N. I, Muda Z., and Sulaiman N., "Anomaly-Based Intrusion Detection Through Kmeans Clustering And Naives Bayes Classification",International Conference on Computing and Informatics, ICOCI Universiti Utara Malaysia, 2013.

[10] Choudhary M., Prity and Choudhary V., "Performance Analysis Of Data Reduction Algorithms Using Attribute Selection In NSL-KDD Dataset ", International Journal of Engineering Science & Advanced Technology Vol. 4, Issue-2, 2014.

[11] Ghosh P. , Debnath C. , Metia D. and Dr. Dutta R., " An Efficient Hybrid Multilevel Intrusion Detection System in Cloud Environment", IOSR Journal of Computer Engineering Vol. 6, Issue 4, 2014.

[12] Choudhary M., Prity and Choudhary V., "Performance Analysis Of Data Reduction Algorithms Using Attribute Selection In NSL-KDD Dataset", International Journal of Engineering Science & Advanced Technology, Vol. 4, Issue-2, 2014.

[13] Balogun A. O. and Jimoh R. G., "Anomaly Intrusion Detection Using an Hybrid Of Decision Tree And K-Nearest Neighbor", A Multidisciplinary Journal Publication of the Faculty of Science, Adeleke University, Ede, Nigeria, 2015.

[14] Dhanabal L. and. Shantharajah S. P., "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 6, 2015.

[15] Ibrahim L. M., Basheer D. T. and Mahmod M. S., "A Comparison Study For Intrusion Database (KDD99, NSL-KDD) Based On Self Organization Map (SOM) Artificial Neural Network", Journal of Engineering Science and Technology Vol. 8, No. 1, 2013.

[16] Alsharafat W. S. and Prince Hussein Bin Abdullah, "Classifier System in Cloud Environment to Detect Denial of Service Attack ", International Journal of Computer Applications (0975 – 8887) Vol 85 – No 14, 2014.

نظام لكشف تطفل هجوم حجب الخدمة بالاعتماد على مصنف النظرية الافتراضية بأستخدام KDD Cup 99 وNSL KDD

أ.م.د. سكينة حسن هاشم
soukaena.hassan@yahoo.com
الجامعة التكنولوجية - قسم علوم الحاسبات
حفصه عادل محمود
h_adel_89@yahoo.com
الجامعة التكنولوجية - قسم علوم الحاسبات

المستخلص:

أن نظام كشف التطفل اصبح ضروري لحماية البيانات من المتطفلين ولتقليل الاضرار في نظام المعلومات والشبكات خاصة في بيئة السحابة التي تعتبر الجيل الجديد للانترنت بالاعتماد على نظام الحوسبة الذي يجهز المستخدمين مختلف انواع الخدمات للعمل والوصول الى تطبيقات السحابة المختلفة. يركز هذا البحث على ملاحظة ان المتطفلين في بيئة السحابة يكونون بنسبة كبيرة من نوع هجوم حجب الخدمة بالمقارنة مع الشبكات الاعتيادية وسوف يتم تقديم النظرية الافتراضية مع تجزئة قاعدة البيانات و اختيار الصفات الملائمة لتحسين اداء النظام وسوف يتم دراسة تاثير استخدام كل الصفات او تحديد مجموعة من الصفات في قاعدة البيانات بأستخدام طريقتين من اختيار الصفات. حيث اظهرت النتائج ان النظام المقترح حسن نسبة اكتشاف هجوم حجب الخدمة حيث تم اكتشاف 94% ، 97% و 98% بأستخدام قاعدة البيانات KDD Cup 99 بتطبيقها على 12 صفة اختيرت بواسطة GR بينما تم اكتشاف 86 %، 87 % و 88 % بأستخدام قاعدة البينات NSL KDD بتطبيقها على 10 صفات تم اختيارها بواسطة GR ايضا.

الكلمات الرئيسية: نظام كشف التطفل، تعدين البيانات، التصنيف المتعدد، النظرية الافتراضية.