Propose NB/HNB Classifiers to Build NIDS

Soukaena H. hashem, Ph.D (Asst. Prof.)* **Hafsa Adil***

Abstract: This paper indicates that the potential attack to traditional/cloud network is Denial of Service (DoS) attack that effect on the availability of the resource, to solve this problem; this paper propose hidden naïve bays(HNB) classifier to enhance the accuracy of detect DoS attack in cloud network with taking into consideration the traditional environment, the system applied NB classifier firstly supported by discretization and feature selection method to show the difference between the traditional NB classifier and the new model HNB classifier. Two methods are used to select the best feature (Info Gain and Gain ratio) and by used two dataset (KDD cup 99 and NSL KDD datasets) that are used to evaluate the performance of the system. The experiential result show that the proposed system based on HNB classifier enhance the accuracy of detect DoS attack where reach to 100% in three test dataset that are different in size and content by use KDD cup 99 dataset and select only twelve features depended on gain ratio as feature selection, while when used NB classifier the accuracy of detect DoS is equal (94, 97, 98) in three different test dataset. In NSL KDD dataset the accuracy of detect DoS reach to 90% for three test dataset based on HNB classifier and by select 10 features based on GR method, while when used NB classifier is equal to (88, 87, 86) for three test dataset.

Keywords: Intrusion detection system, data mining, multiclass classification.

* University of Technology, Computer Science Department, Baghdad, Iraq.

1. Introduction

Cloud computing provides a service in form of data storage and computing power, through the Internet with little efforts for management, release and resource allocation[1]. Cloud Computing has some similarities with distributed systems, through usage network environment features. It is a modern design of computer services. Therefore the security is considered as one of important matter in this environment. Since a large number of end users connected to the network within times [2]. One of the major security challenges in a cloud environment is the detection of any attacks and intrusions. Cloud network exposes to high risk of attacks and one of these attacks that lead to violation in security is DDoS or DoS attack[3].

Intrusion detection (ID) is the process of monitoring the instances that happen in a network or computer system, then analyzing them for detecting the attacks, its attempts to compromise the confidentiality, availability, and integrity or to bypass the security mechanism of network or computer[4]. ID methods are classified into Signature based detection also called misuse based detection where employs a priori knowledge of attack signatures. The signatures are manually constructed by security experts analyzing previous attack patterns and used to match with incoming traffic to detect intrusions. Another method is Anomaly based detection where uses a different method to detect the attack. It identifies any kind of network connection violating the normal connection as a threat [5].

Data mining (DM) is a task of data analysis where computer driven algorithms are used to find the essential pattern from the data. Data mining involves four classes of the tasks; Classification, Association rule learning, Regression and Clustering. Classification is used to determine the category of new events on the training dataset that containing tuples or instances whose category membership is known [6].

One of the most popular classification methods is NB classifier which applied to many domains, including intrusion detection; NB classifier has strong independence assumption since it depends on applying Bayes theorem. The independence assumption means that the probability of one feature doesn't effect on the probability of the other features. NB classifier can be defined as Eq. (1), where P(C) is the probability of class while

P(a1,a2,..,An|c) is the probability of values in features within a class, the assumption of independence of the features is defined in Eq. (2) [7].

$$\mathrm{c(E)} = \arg\max_{c \in C} P(c) P(a_1, a_2, \ldots, a_n | c) \quad (1)$$

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

There are several studies discussed the importance of eliminating the independence assumption of NB classifier in last years, one of these studies introduced HNB classifier, this new model depends on building another layer which represents a hidden parent for every attribute. The hidden parent (A_{hpi}) is used together the weighted influences from all other attributes (A_i), where i j= 1, 2,… n and i is not equal to j, and P(C) is the probability of class. Joint distribution is defined as Eq. (3), while the hidden parent defined as Eq. (4), and HNB classifier is defined as Eq. (5) [8].

$$P(A_{1,\ldots,} A_2 | \mathrm{C}) = \mathrm{P(C)} \prod_{i=1}^{n} P(A_i | A_{hpi,} C) \quad (3)$$

$$\mathrm{P}\left(A_i \middle| A_{hpi}, C\right) = \sum_{j=1, j \neq i}^{n} W_{ij} * P(A_i | A_j, C) \quad (4)$$

$$c(E) = \arg\max_{c \in C} P(c) \prod_{i=1}^{n} P(a_i | a_{hpi}, c) \quad (5)$$

The weights W_{ij} is calculated by using conditional mutual information (CMI) between every two attributes A_i and A_j as viewed in Eq. (6), The CMI is defined as Eq. (7) [8].

$$W_{ij} = \frac{I_p(A_i; A_j | C)}{\sum_{j=1, j \neq i}^{n} I_p(A_i; A_j | C)} \quad (6)$$

$$I_p\left(A_i; A_j \middle| C\right) = \sum_{a_i, a_j, c} P\left(a_i, a_j, c\right) log \frac{P(a_i, a_j | c)}{P(a_i | c) P(a_j | c)} \quad (7)$$

Feature selection is an important data processing step applying before a training phase, the process of reduction attributes space leads to simplify the use of different visualization technique and a better understandable model. There are two common approaches are used for feature reduction, the first method is information gain (IG) that evaluating the attribute worth by use entropy with respect to the class. Entropy is often used in information theory measure; it is defined as a measure of systems unpredictability. The higher entropy appears in an attribute that

has more information content. The Entropy of each feature is defined as Eq. (8) where a is a value of feature, and a =1, 2, ..,n. The Information needed to classify D after using A for divide D into n partitions is viewed in Eq. (9). Information gained by branching an attribute A as in Eq. (10) [9].

$$\text{Info D} = \text{H(A)} = -\sum_{a=1}^{n} \text{P(a)} \log_2 \text{P(a)} \quad (8)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * I(D_j) \quad (9)$$

$$\text{Gain (A)} = \text{Info D} - \text{Info}_A\text{(D)} \quad (10)$$

The second feature selection that used in the proposed system is Gain Ratio (GR) that modifies the information gain to allow for uniformity and breadth of the values of attributes, GR method choose the attribute by taking into account the number and size of values. It enhances IG by taking into consideration the substantial information, where the substantial information is the entropy of distribution of events into branches as Eq.(11). This value generated by splitting the training data set as in Eq. (12) [10].

$$Split\ Info_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \quad (11)$$

$$\text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{Split Info (A)}} \quad (12)$$

The KDD Cup 99 dataset consist of 10% of the original dataset that consisting of 494,020 records every record contains 41 features and class feature labelled either attack or normal. It has 19.69% normal and 80.31% attack. The NSL KDD has the same features as KDD Cup 99 dataset. It's selected records of the complete KDD cup 99 and solve the inveterate problems of the KDD CUP 99 dataset, the class feature contains 21 kinds of attacks within four types: DOS, Probe, R2L attacks and U2R attacks [11].

Cloud Network compromises the integrity, confidentiality, and availability of computer systems and resources. One of the essential requirements in a security of cloud network is availability that appears as a DoS attack; is considered as one of e the dangers attack among numerous attacks in the cloud network, its make services unavailable for indefinite period of time by overload the system with useless traffic [12].

2. Related Works

1. **Dr. Mukherjeea S. Et al., 2012**, they propose a method for Feature Vitality Based Reduction Method model for feature selection and comparison with three methods (IG, GR, and Correlation based Feature Selection), they discussed that; to build effective and efficient IDS, it importance reduces features. They applied NB classifier by used NSL KDD dataset. Experimental results indicate that selected some of Features lead to enhance the performance of IDS [7].
2. **Koc L. et al., 2012**, they discussed that the HNB binary classifier model can be used to solve intrusion detection problem. They used NB classifier and structurally extended NB methods augmented with discretization and feature selection and compared the performance with the HNB classifier as an IDS, they use KDD cup 99 dataset. The experimental results proved that HNB classifier enhances the performance of the system in term of error rate, misclassification, and accuracy of detecting DOS attacks, where the accuracy of detect DoS reach to 0.99 [13].
3. **Koc L. et al., 2015,** they explained the need to implement DM methods in IDS, they reviewed NB classifier and then introduced HNB classifier to solve the problem of independence of attributes assumption. They used KDD Cup 99 dataset to evaluate the system and CONS feature selection method, the results indicates that the proposed system enhance the performance in terms of accuracy and error rate, where the accuracy of detect normal and attack events reach to 0.93 [14].

3. Proposal Network Intrusion Detection System

The proposed system is offline NIDS in traditional/cloud environment based on NB/HNB Classifiers. The first step will demonstrate how to normalize the KDD Cup 99 and NSL KDD Dataset, Also it will illustrates how to discrete the continuous feature in both datasets into specific ranges, and then use two feature selection algorithms (IG and GR) to remove unrelated features, and finally apply NB and HNB multiclass classifiers to detect attacks in traditional/cloud environment in both datasets. Figure (1) depicts the general structure of the proposed NIDS.

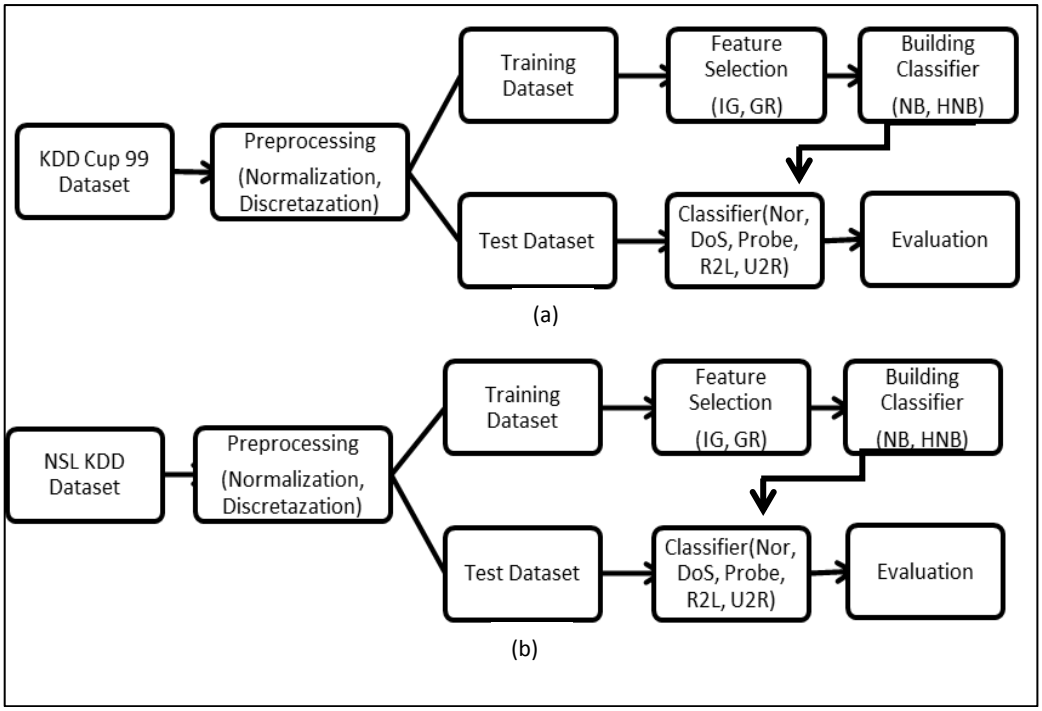

Figure (1) : General structure of proposal NIDS
(a) Block Diagram for KDD Cup 99 dataset
(b) Block Diagram for NSL KDD dataset

3.1. Normalization

After obtained KDD Cup 99 and NSL KDD datasets from internet in text format, the first step converts it to access database, then apply normalization process to the continuous features by use Min-Max Normalization algorithm, see Eq.(13), the normalization process improve effectiveness and performance of the system by making the values of feature within range from 0 to 1 [10].

$$Value\ X = \frac{ValueX - Min}{Max - Min} \quad (13)$$

3.2. Discrete continuous Feature

The KDD Cup 99 and NSL KDD Datasets consist of discrete and continuous feature, hence discretization is used to convert the continuous feature to discrete to increase speed and ensure the effectiveness of the

system and also to solve the problem of NB classifier when new value appeared in test dataset that didn't appear in learning phase.

3.3. Feature Selection Methods

Feature selection is consider one of the most essential and important preprocessing steps in data mining methods, it's used to remove the redundant and unrelated features in a huge database like NSL KDD and KDD Cup 99, and to improve the effectiveness of the system by reducing the consuming time and selecting the best features. The proposed system used information gain and gain ratio as feature selection algorithms.

3.4. Training and testing

In learning phase, the proposed system will apply NB multiclass classifier and HNB multiclass Classifier on both KDD Cup 99 and NSL KDD Datasets to detect attacks in traditional and cloud networks. Proposal used 4000 records for training phase with 2169 DOS, 388 probes, 173 R2L, 35 U2R and 1235 normal in both datasets.

Testing phase used three samples in KDD Cup 99 Dataset to evaluate and validate the performance of the system; first with 1200 records, second with 600 records and third with 900 records. While in NSL KDD Dataset the testing samples that have been used are 1028 records and the other two samples used to validate the performance are (795 and 567 records).

3.4.1. Naive Bayes Classifier

After pre-processing the dataset and select the subset of best features, the next step is building the classifier, in this proposed NIDS will begin with NB classifier to discover the best subset of features through selecting the subset which gives higher detection rate and accuracy with NB**.** Two problems appeared in NB classifier:

- The first problem in NB Classifier solved by discretization process to avoid appeared new values in test dataset that didn’t appear in the training dataset.
- The second problem that occurred in NB classifier is zero probability that happens when a frequency of value in one class or more is zero that leads to making the result of test zero since the posterior probabilities used multiplication operand. The zero probability is solved by using Laplace filter (also called add one) where adding one to the

frequency of values and adding a number of values in feature to the frequency of class in the training dataset.

NB classifier will be tested by the steps mention below to find the best subset of selected features and best algorithm used for feature selection:

1. Test all of Features 41 features (in both KDD Cup 99 and NSL KDD dataset).
2. Test best 20 features selected by IG and GR (in both KDD Cup 99 and NSL KDD Dataset).
3. Test best 12 features selected by IG and GR (in both KDD Cup 99 and NSL KDD Dataset).
4. Test best 10 features selected by IG and GR (in NSL KDD Dataset).

3.4.2 Hidden Naive Bayes Classifier

After applying NB as a classifier and find the best subset of features in both datasets (12 features selected by GR in KDD cup 99 dataset and 10 features selected by GR in NSL KDD dataset), then HNB will be applied on these features to compare the performance of the NB Classifier and HNB Classifier and how will be enhancing the performance of the system to detect DoS attacks in cloud network. HNB classifier solves the problem of conditional independence between features by building hidden parent for each value. One problem appeared in training phase that in hidden parent equation when multiplying the weight with probability$(A_i|A_j, C)$, they are some of these probabilities equal to zero that leads to the same problem in NB Classifier that solved by Laplace filter also. Algorithm (1) describes the proposed system based on NB/HNB classifier in details.

Algorithm (1): Proposed system based on NB/ HNB classifiers
Input: Training and Testing dataset (KDD Cup 99 10% or NSL KDD Dataset) **Output:** Classification the test dataset in traditional environment and in cloud network
Begin **Step1**:**Preprocesing the dataset** 1) Normalize the continuous feature in both training and testing datasets 2) Discrete the continuous features in training and testing datasets **Step2: Feature selection**

1) Apply info gain in training dataset and select the best 20, 12, 10 features
2) Apply gain ratio in training dataset and select the best 20, 12, 10 features

Step3: Naïve bays classifier

In training phase do the following

1) Find the probability of all class in training dataset
2) Find the probability of each value within classes for all features

In test phase do the following:

1) For each record in testing dataset find the probability of each value with class in training dataset
2) Multiply the probability of each value in the record as Eq.(2)
3) Use the multiplication result of point 2 to multiply by the probability of class
4) Select the maximum value result from point 3 to classify the recod as Eq.(1)

Evaluation

In traditional environment consider all kinds of attacks

Find accuracy, DR, ER, confusion matrix

In cloud environment detect DoS attacks

Find accuracy of DoS attack

Step4:apply NB classifier with all feature

Apply NB classifier with best 20 features based on GR in KDD cup 99 and NSL KDD

Apply NB classifier with best 12 features based on GR in KDD cup 99 and NSL KDD

Apply NB classifier with best 10 features based on GR in NSL KDD

Apply NB classifier with best 20 features based on IG in KDD cup 99 and NSL KDD

Apply NB classifier with best 12 features based on IG in KDD cup 99 and NSL KDD

Apply NB classifier with best 10 features based on IG in NSL KDD

Step5: find the best subset of features that given high accuracy based on Naïve bayes

Step6: apply HNB classifier on these subset as the following steps:

In training phase

1) Find the probability of each class in training dataset
2) Find Conditional mutual information between every two features as Eq.(7)
3) Find the weight between every two features as Eq.(6)
4) Great hidden parent by use Eq.(4)

In testing phase

1) For each record in testing dataset find the hidden value of each value with class in training dataset
2) Multiply the hidden value of each value in record as Eq.(3)
3) Use the multiplication result of point 2 to multiply by the probability of class
4) Select the maximum value result from point 3 to classify the record as Eq.(5)

Evaluation

In a traditional environment consider all kinds of attacks
Find accuracy, DR, ER, confusion matrix
In cloud environment detect DoS attacks
Find accuracy of DoS attack
End

3.5. Evaluation of the proposed system

The evaluation of the proposed system performance is given in terms of; Accuracy of detecting attacks vs. normal events(accuracy binary), Accuracy of detecting four kinds of attacks, Accuracy to detect DoS attack in cloud network, Detection rate (DR), error rate(ER) and confusion matrix are taken in consideration. Where Tp (true positive), Tn (true negative), Fp (false positive) and Fn (false negative)

$$\mathrm{DR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{14}$$

$$\mathrm{ER} = \frac{\mathrm{FP+FN}}{\mathrm{TP+TN+FP+FN}} \tag{15}$$

$$\text{Accuracy binary} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{16}$$

$$\text{Accuracy of class} = \frac{\text{number of samples of a class correctly classified}}{\text{number of samples of the class}} \tag{17}$$

$$\text{Accuracy of multiclass} = \frac{\sum \text{number of samples of a class correctly classified}}{\text{total number of test samples}} \tag{18}$$

3.5.1. Evaluation of Proposed System using KDD Cup 99 dataset

The training dataset uses 4000 records, while the testing dataset is used three datasets each one contains different number of records

selected randomly from KDD cup 99 dataset. The results of the proposed NIDS will be viewed as follow:

1- NIDS based on Naïve Bayes Classifier by use KDD Cup 99 Dataset.

2- NIDS based on Hidden Naïve Bayes Classifier by use KDD Cup 99 Dataset.

3.5.1.1. NIDS based on Naïve Bayes Classifier by use KDD Cup 99 Dataset

The KDD cup 99 dataset selected their data from traditional networks, in this section will present the result of the proposed system by taking into consideration all kinds of attack. Table (1) shows the evaluation of classification in three KDD cup 99 testing datasets with five different subsets of features (NB with all features, NB with Gain ration by select 20 features, NB with GR by select 12 features, NB with IG by select 20 features, and NB with IG by select 12 features).

Table (1): Performance measure of KDD cup 99 Dataset based on NB classifier

DS	No. Feature & FS	Acc multiclass	Acc binary	DR	ER	Acc DOS	Acc Probe	Acc R2L	Acc U2R	Acc Normal
Test1	41	0.92	0.98	100	0.01	0.90	100	0.95	0.25	0.93
Test1	20 IG	0.92	0.99	0.98	0.01	0.90	0.98	0.73	0.25	0.99
Test1	12 IG	0.91	0.98	0.98	0.02	0.89	100	0.73	0	0.96
Test1	20 GR	0.91	0.97	0.98	0.025	0.90	0.98	0.78	0.25	0.93
Test1	12GR	0.91	0.97	0.97	0.03	0.94	0.89	0.47	0.50	0.94
Test2	41	0.91	0.98	100	0.01	0.92	0.90	0.77	0	0.93
Test2	20 IG	0.92	0.96	0.96	0.03	0.93	0.91	0.80	0	0.96
Test2	12 IG	0.93	0.97	0.99	0.02	0.95	0.91	0.72	0	0.92
Test2	20 GR	0.91	0.97	0.99	0.02	0.92	0.91	0.75	0	0.93
Test2	12GR	0.92	0.96	0.97	0.03	0.97	0.83	0.38	0	0.93

Test3	41	0.96	0.98	100	0.01	0.97	0.96	0.98	0	0.95
Test3	20 IG	0.96	0.98	0.97	0.01	0.95	100	0.98	0	0.99
Test3	12 IG	0.95	0.98	0.99	0.01	0.96	100	0.92	0	0.95
Test3	20 GR	0.96	0.98	100	0.01	0.97	100	0.96	0	0.95
Test3	12GR	0.93	0.97	0.97	0.02	0.98	0.87	0.49	0.62	0.95

As shown in table (1), the results of accuracy binary and accuracy multiclass are close to each other, but the importance of the proposed system is the accuracy of detects DoS attack that widely appeared in a cloud network where is best when selected 12 features by use gain ratio algorithm.

3.5.1.2. NIDS based on Hidden Naïve Bayes Classifier by use KDD Cup Dataset

After applied NB classifier with different subsets of features according to the two feature selection algorithms and determine the best subset of features. Now will apply HNB classifier on these subsets of features (12 features selected by gain ratio algorithm). Table (2) shows the evaluation of the proposed system in traditional network that demonstrates the accuracy of detecting all kinds of attacks, detection rate, error rate and accuracy of each class.

Table (2): Performance measure of KDD cup 99 Dataset based on HNB classifier

DS	Accuracy multiclass	Accuracy binary	DR	ER	DOS	Probe	R2L	U2R	Normal
Test1	0.94	0.97	0.96	0.02	100	0.89	0	0	100
Test2	0.92	0.97	0.97	0.02	100	0.83	0	0	0.96
Test3	0.93	0.96	0.95	0.03	100	0.87	0	0	0.99

Table (2) views how HNB classifier enhances the accuracy of detecting DoS attack that reaches to 100% in three samples of testing datasets. Also show the accuracy of detecting U2R and R2L attacks is zero, but it detected as another kind of attack see confusion matrix in

tables (3),(4), and (5), for example; Test1 12 of R2L attack is detected it as a DoS attack.

Table (3): Confusion Matrix for Test1

	Normal	DOS	probe	R2L	U2R
Normal	157	0	0	0	0
DOS	0	342	0	0	0
probe	0	8	66	0	0
R2L	11	12	0	0	0
U2R	3	1	0	0	0

Table (4): Confusion Matrix for Test2

	Normal	DOS	probe	R2L	U2R
Normal	226	7	0	0	0
DOS	0	515	0	0	0
probe	0	18	93	0	0
R2L	17	13	6	0	0
U2R	0	5	0	0	0

Table (5): Confusion Matrix for Test3

	Normal	DOS	probe	R2L	U2R
Normal	324	2	0	0	0
DOS	0	680	0	0	0
probe	0	16	117	0	0
R2L	28	25	0	0	0
U2R	7	1	0	0	0

3.5.2. Evaluation of Proposed System using NSL KDD dataset

The proposed system using NSL KDD dataset is also used 4000 samples in the training phase and three samples of testing dataset contain different records and size selected randomly from NSL KDD dataset. The results of the proposed NIDS will be viewed as follow:

1- NIDS based on Naïve Bayes Classifier by use NSL KDD Dataset.

2- NIDS based on Hidden Naïve Bayes Classifier by use NSL KDD Dataset.

3.5.2.1. NIDS based on Naive Bayes Classifier by use NSL KDD Dataset

In this section will present the result of the proposed system in a traditional and cloud networks by finding the accuracy of all kinds of attacks. Table (6) shows the evaluation of classification in three samples of NSL KDD testing datasets with seven subsets of feature (NB with all features, NB with Gain ration by select 20 features, NB with gain ratio by select 12 features, NB with gain ratio by select 10 features, NB with IG by select 20 features, NB with IG by select 12 features, and NB with IG by select 10 features).

Table (6): Performance measure of NSL KDD Dataset

DS	No. Feature & FS	Acc multiclass	Acc binary	DR	ER	Acc DOS	Acc Probe	Acc R2L	Acc U2R	Acc Normal
Test1	41	0.68	0.88	0.98	0.1	0.67	0.98	0.60	0.16	0.63
Test1	20 IG	0.68	0.82	0.90	0.17	0.67	100	0	0	0.63
Test1	12 IG	0.80	0.83	0.91	0.16	0.86	100	0.20	0	0.64
Test1	10 IG	0.80	0.84	0.88	0.15	0.86	100	0.10	0	0.66
Test1	20 GR	0.79	0.84	0.92	0.14	0.85	0.98	0.80	0	0.66
Test1	12GR	0.76	0.83	0.91	0.16	0.88	0.63	0.70	0.50	0.63
Test1	10GR	0.74	0.89	0.99	0.1	0.88	0.35	0.70	0.33	0.61
Test2	41	0.69	0.90	0.98	0.07	0.66	100	0.70	0.18	0.69
Test2	20 IG	0.68	0.84	0.90	0.1	0.64	100	0.05	0	0.69
Test2	12 IG	0.80	0.85	0.91	0.11	0.85	0.99	0.41	0	0.69
Test2	10 IG	0.80	0.85	0.91	0.11	0.86	100	0.23	0	0.70
Test2	20 GR	0.81	0.85	0.92	0.1	0.86	100	0.82	0	0.71
Test2	12GR	0.78	0.85	0.91	0.11	0.89	0.64	0.70	0.54	0.70
Test2	10GR	0.73	0.89	0.99	0.07	0.87	0.35	0.29	0.36	0.66
Test3	41	0.76	0.91	0.97	0.08	0.64	100	0.62	0.15	0.78
Test3	20 IG	0.71	0.86	0.89	0.13	0.64	100	0.04	0	0.78

Test3	12 IG	0.82	0.86	0.90	0.13	0.86	0.99	0.33	0	0.78
Test3	10 IG	0.82	0.86	0.90	0.13	0.85	100	0.20	0	0.77
Test3	20 GR	0.83	0.87	0.92	0.12	0.86	0.98	0.87	0	0.78
Test3	12GR	0.79	0.87	0.91	0.12	0.89	0.56	0.62	0.61	0.78
Test3	10GR	0.76	0.92	0.99	0.07	0.86	0.35	0.62	0.38	0.76

In table (6) The evaluation indicates the best result of detecting attacks and normal is occurred by use only 10 features by applied gain ratio, While in cloud network the accuracy for each class determine that the accuracy of detecting DoS attack in Test1 is high when using best 12 or 10 features by applied IG but the other kinds of attacks cannot detect it, that leads to indicate the best result when select 10 or 12 features by use GR, while in Test2 and Test3 the results of select 10 and 12 features are close to each other.

3.5.2.2. NIDS based on Hidden Naïve Bayes Classifier by use NSL KDD Dataset

HNB classifier applied to the best subset of features (10 features selected by use gain ratio algorithm). The evaluation of performance by using three testing datasets is shown in Table (7) (accuracy of multiclass, accuracy binary, error rate (DR), detection rate (ER), and the accuracy of every kinds of class, the result indicates the accuracy of detecting DoS attack is best when use 10 features by applies GR algorithm.

Table (7): performance measure of NSL KDD Dataset

DS	Accuracy multiclass	Accuracy binary	DR	ER	DOS	Probe	R2L	U2R	Normal
Test1	0.83	0.92	0.90	0.07	0.90	0.29	0	0	100
Test2	0.82	0.92	0.90	0.06	0.90	0.29	0	0	100
Test3	0.83	0.93	0.90	0.06	0.90	0.28	0	0	100

Table (8), (9) and (10) views the confusion matrix for test (1, 2 and 3) of NSL KDD dataset by use only 10 features that selected by GR algorithm which achieves best accuracy in detecting DoS attack. Table (7) view that; the accuracy of detect DoS is 90 % and the accuracy of detect normal instance is 100%, while the accuracy of detect the other attacks

(probe, R2L, and U2R) is low, but it's detect it as a DoS attack, look at tables (8), (9) and (10).

Table (8): Confusion Matrix for Test1

	Normal	DOS	Probe	R2L	U2R
Normal	157	0	0	0	0
DOS	30	296	0	0	0
probe	0	48	20	0	0
R2L	0	10	0	0	0
U2R	3	3	0	0	0

Table (9): Confusion Matrix for Test2

	Normal	DOS	Probe	R2L	U2R
Normal	233	0	0	0	0
DOS	40	394	0	0	0
probe	0	71	29	0	0
R2L	1	16	0	0	0
U2R	6	5	0	0	0

Table (10): Confusion Matrix for Test3

	Normal	DOS	Probe	R2L	U2R
Normal	330	0	0	0	0
DOS	50	489	0	0	0
probe	0	87	35	0	0
R2L	3	21	0	0	0
U2R	7	6	0	0	0

4. Conclusions

The Proposed system indicates the important to use NIDS in cloud network to detect DoS attack that consider the most harmful attack in a network that effect on the availability of the resource, Normalization and discretization processes makes the proposed system more efficient, To enhance the accuracy of proposed system and reduce the consuming time suggests use IG and GR algorithms as a feature selection. Using these algorithms raise the accuracy result of classification, as shown in tables (1) and (6), The accuracy of NB classifier supported by gain ratio is better than using all features or use NB Classifier with IG, The proposed system achieves high accuracy when select twelve feature using GR in KDD Cup 99 dataset, while in NSL KDD dataset is best to select ten features using GR, The Proposed System based on NB classifier gives high accuracy by use KDD Cup 99 Dataset as shown in table (1), but when using HNB as a classifier the result of detect DoS become excellent as shown in table (2), HNB Classifier is more complicated than NB Classifier but is more efficient, Use KDD Cup Dataset gives high accuracy than when using NSL KDD Dataset.

References

[1] Gupta S., Kumar P. and Abraham A., "A Profile Based Network Intrusion Detection and Prevention System for Securing Cloud Environment", Hindawi Publishing Corporation International Journal of Distributed Sensor Networks, 2013.

[2] Alsharafat W. S. and Abdullah H., "Classifier System in Cloud Environment to Detect Denial of Service Attack ", International Journal of Computer Applications Vol., 85, No. 14, 2014.

[3] VidhyaV.," A Review of DOS Attacks in Cloud Computing", Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 5, Ver. II, 2014.

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

[5] Jamadar V.,Jabiulla B., Rakesh S. and Sadanand P., " Denial of Service(DoS) attack incidents and defense mechanisms " Journal of Emerging Technologies and Innovative Research (JETIR) Volume 2, Issue 5 2015.

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

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

[8] Koc L. and Carswell A. D., "Network Intrusion Detection Using a HNB Binary Classifier", UKSIM-AMSS International Conference on Modelling and Simulation, 2015.

[9] 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.

[10] Ibrahim H. E., Badr S. M. and Shaheen M. A.," Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", International Journal of Computer Applications, Vol. 56, No.7, 2012.

[11] Sahu S. K., Sarangi S. and Jena S. K., "A Detail Analysis on Intrusion Detection Datasets", National Institute of Technology, Rourkela, 2014.

[12] Alsharafat W. S. and Abdullah H., "Classifier System in Cloud Environment to Detect Denial of Service Attack ", International Journal of Computer Applications Volume 85 – No 14, January 2014.

[13] Koc L., Thomas A. Mazzuchi and Sarkani S., "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Elsevier Ltd., 2012.

اقتراح استخدام مصنف NB/HNB لبناء نظام كشف التطفل الشبكي

د. سكينه حسن هاشم الباحث: حفصه عادل

المستخلص: في هذا البحث تم الاشارة الى أن الهجمات المحتملة في الشبكة التقليدية والسحابية تكون من قبل DoS الذي يؤثر على متاحيه المصادر, لحل هذه المشكلة تم اقتراح استخدام مصنف HNB لتحسين نسبة كشف DoS في الشبكة السحابية مع الاخذ بنظر الاعتبار البيئة التقليدية , حيث تم تطبيق المصنف NB أولا مدعوما بعملية تجزئة البيانات و اختيار الصفات لتوضيح الفرق بين NB التقليدي و HNB. حيث تم استخدام طريقتين لعملية اختيار الصفات وهما (info Gain, Gain ratio) وباستخدام قاعدتي البيانات (KDD Cup 99 , NSL KDD) التي استخدمت لتقييم أداء النظام. حيث أظهرت النتائج أن النظام المقترح بالاعتماد على HNB حسن نسبة كشف DoS حيث وصلت النسبة الى 100% باستخدام ثلاث قواعد بيانات لفحص النظام والتي كانت مختلفة في المحتويات والحجم باستخدام قاعدة البيانات KDD Cup 99 وباختيار أثنى عشر صفات بالاعتماد على تقنية GR بينما وصلت نسبة اكتشاف DoS الى (94,97,98) في ثلاث قواعد بيانات لفحص النظام عندما تم استخدام NB كمصنف. أما في قاعدة البيانات NSL KDD وصلت نسبة اكتشاف DoS الى 90% في ثلاث قواعد بيانات لفحص النظام بالاعتماد على HNB كمصنف وباختيار عشر صفات بالاعتماد على طريقة GR, بينما عندما تم استخدام NB كمصنف وصلت النتائج الى (88,87,86) في ثلاث قواعد بيانات لفحص النظام.

الكلمات المفتاحية: نظام كشف التطفل, تعدين البيانات, التصنيف المتعدد.

الجامعة التكنولوجية، قسم علوم الحاسوب، بغداد، العراق