ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

An Integration of Binary Bat Algorithm and Naïve Bayes Classifier for Intrusion Detection in Distributed Environment

Varuna S[1], Ramya R[2]

Assistant Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India[1, 2]

Abstract: A computer network is a collection of computer systems and other hardware devices that are linked together through communication channels to facilitate communication and resource-sharing among a wide range of users. Even though static defence mechanisms such as firewalls and software updates can provide a reasonable level of security, more dynamic mechanisms such as Intrusion Detection Systems (IDSs) should also be utilized. Different detection techniques can be employed to search for attack patterns in the data monitored. Misuse detection and anomaly detection are the most widely used techniques. But they have their own drawbacks. To overcome those issues, hybrid methods are used. Hybrid classifiers are able to provide improved detection accuracy, but usually have a complex structure and high computational costs. Hence a new intrusion detection model is introduced in which feature selection is carried out and then with the selected features classification is performed. The binary bat algorithm is used for feature selection in Hadoop environment and then the samples with selected features are trained and tested using naïve bayes classifier. Since it is carried out in a distributed environment, the execution time is greatly reduced and the detection rate is improved.

Keywords: Intrusion Detection System, Hybrid classifier, Naive bayes, Binary bat optimization.

I. INTRODUCTION

The purpose of network security is to maintain the network and its component parts secure from unauthorized access and misuse. Networks are vulnerable because of their inherent characteristic of facilitating remote access. Therefore, it is vital for any network administrator, regardless of the size and type of network, to implement stringent security policies to prevent potential losses.
Intrusion detection[1] complements the protective mechanisms to improve the system security. Moreover, even if the security mechanisms can protect information systems successfully, it is still desirable to know what intrusions have happened, so that we can understand the security threats and risks and thus be better prepared for future attacks.
Intrusions in an information system are the actions that violate the security policy of the system. Intrusion detection is the process used to identify intrusions.

Intrusion detection techniques are traditionally categorized into two methodologies:

A. Anomaly detection - An intruder's behaviour is noticeably different from that of a normal user, and statistical models are used to aggregate the user's behaviour and distinguish an attacker from a normal user.

B. Misuse detection - Misuse detection or Signature based detection is very effective against known attacks, and it depends on the receipt of regular updates of patterns and hence will be unable to detect unknown previous threats or new releases.

These two conventional methods have their own drawbacks, and hence to resolve the disadvantages, hybrid intrusion detection methods have been proposed[2]. The detection performance of the hybrid intrusion detection system depends on the combination of these two different detection methods.

II. LITERATURE REVIEW

There are three different methods to combine the anomaly detection model and misuse detection model: anomaly detection followed by misuse detection, parallel use of misuse detection and anomaly detection, and misuse detection followed by anomaly detection.

Sang Hyun Oh et al.,[3] proposed an anomaly intrusion detection method by clustering normal user behaviour. A detection method which utilizes a clustering algorithm for modelling the conventional behaviour of a user's activities has been used. Clustering eliminates the inaccuracy caused by statistical analysis. Hence the frequent activities of the user are modelled more accurately than the statistical analysis.
Sheng Yi Jang et al.,[4] proposed a Clustering Based method for Unsupervised Intrusion Detection (CBUID) where the clusters are formed from unlabelled training

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

datasets and labeled as normal or anomalous by their outlier factors. Inho Kang et al.,[5] proposed a new one-class classification method with differentiated anomalies to enhance the performance of intrusion detection for harmful attacks.

Yinhui Li et al.,[6] proposed a method that uses support vector machine and gradual feature removal method for intrusion detection. Dr.Saurabh Mukarjee et al., [7] proposed Feature Vitality Based Reduction Method (FVBRM) where features are reduced one at a time, the resultant dataset is then used for the training and testing of the classifier, this process continues until it performs better than the original dataset in terms of performance, known as Feature- Vitality Based Reduction Method.

Kok Chin Khor et al.,[8] proposed an approach for improving detection rates on rare attack categories. Chun Guo et al.,[9] proposed a hybrid method, Distance Sum based Support Vector Machine (DSSVM). Distance sum is the correlation between each data sample and cluster centers. Jin Qian et al.,[10] proposed a parallel attribute reduction algorithm using MapReduce. Three different parallelism strategies of attribute reduction and how the reduction computations can be transformed into map and reduce operations is compared is presented.

Douglas Rodrigues et al.,[11] proposed a wrapper approach for feature selection based on Bat Algorithm and Optimum-Path Forest (OPF).Ping Lv et al.,[12] proposed hierarchical attribute reduction algorithms for big data using MapReduce. Rong-Fang Xu et al.,[13] proposed a method for dimensionality reduction by feature clustering. The instances are grouped and clusters are formed. Then one new feature is extracted from each cluster through a weighted combination of the instances.

III.PROPOSED METHODOLOGY

A Hadoop cluster is specifically designed for storing and analysing huge amounts of unstructured data. A Hadoop cluster is essentially a computational cluster that divides the data analysis workload across multiple cluster nodes to process the data in parallel.

A. Description of Dataset

The dataset used to evaluate the proposed method for intrusion detection is NSL-KDD which contains selective records from the original KDD dataset. The NSL-KDD dataset contains 125973 records. Each record represents a network connection described by 41 features: seven nominal features and 34 continuous features and a label specifying the status of this record as either normal or one of 39 specific attack types. The record types in NSL-KDD can be categorized as either normal class or one of four attack classes, remote-to-local (R2L), denial-of-service (DoS), user-to-root (U2R), and Probe (Prb).

B. Architectural Framework

Feature selection is done on the entire dataset using the binary bat algorithm. As a result the number of features in the dataset can be reduced. This forms the new training data and the new test set is formed with the selected features. Then the training set is given as input for Naïve Bayes classifier and then tested using the test set.

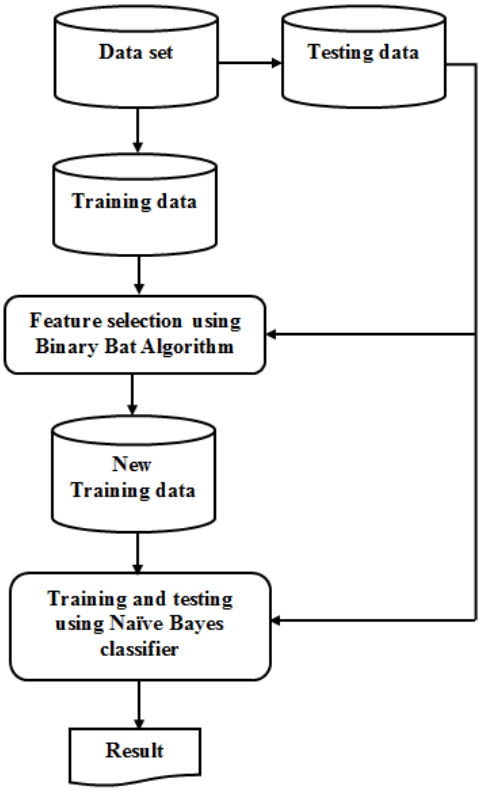

Fig. 1. Architectural Framework

The implementation of the proposed work consists of the following modules:

1. Feature selection using Binary Bat Algorithm
2. Training and testing using Naïve Bayes classifier

C. Feature Selection Using Binary Bat Algorithm

Hadoop is an open source framework for developing distributed applications that can process very large amount of data. It is a platform that provides both computational capabilities and distributed storage. Hadoop is a distributed master-slave architecture that consists of the Hadoop Distributed File System (HDFS) for storage purpose and Map-Reduce for computational capabilities. The special capabilities of Hadoop are data partitioning and parallel computation of large datasets. Its storage and computational capabilities improve with the addition of hosts to the cluster, and can reach volume sizes in the petabytes on clusters with several hundreds of hosts.

Since time plays an important role in intrusion detection systems, great concern should be provided for reducing the time. Hence the number of features should be reduced.

Feature selection, also known as attribute selection, is the process of selecting a subset of relevant features. The assumption when using a feature selection technique is that the dataset contains many redundant or irrelevant features. The features that contribute greatly to the result

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

should be selected. Here feature selection is done using Binary Bat Algorithm in Hadoop environment.

I. BAT Algorithm:

Bat algorithm (BA) is a heuristic algorithm that imitates the echolocation behaviour of bats to perform global optimization. This algorithm carries the search process using artificial bats as search agents mimicking the natural pulse loudness and emission rate of real bats[14]. A binary version of BA, named BBA, is used.

The BBA will have artificial bats navigating and hunting in binary search spaces by changing their positions from ''0'' to ''1'' and vice versa. In BA, an artificial bat has a position vector X, velocity vector V, and frequency vector F, which are updated during the course of iterations

$$V_i(t+1) = V_i(t) + (X_i(t) - Gbest)F_i \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (2)$$

where Gbest is the best solution attained so far and Fi indicates the frequency of i-th bat which is updated in each course of iteration as follows:

$$F_i = F_{min} + (F_{max} - F_{min})\beta \quad (3)$$

where β is a random number of a uniform distribution in [0,1]. It is clear that different frequencies encourage artificial bats to have variety of propensity to the best solution. These equations could guarantee the exploitability of the BA.

A v-shaped transfer function and position updating rule are proposed in order to do this

$$V\left(v_i^k(t)\right) = \left|\frac{2}{\pi}\arctan\left(\frac{\pi}{2}v_i^k(t)\right)\right| \quad (4)$$

$$x_i^k(t+1) = \begin{cases} \left(x_i^t(t)\right) & \text{if rand} < V\left(v_i^k(t+1)\right) \\ x_i^k(t) & \text{rand} \geq \left(v_i^k(t+1)\right) \end{cases} \quad (5)$$

where $x_i^t(t)$ and $v_i^k(t)$ indicate the position and velocity of i^{th} particle at iteration t in k^{th} dimension. The balancing between the techniques is controlled by the loudness (A) and pulse emission rate (r). These two elements are updated as follows:

$$A_i(t+1) = \propto A_i(t) \quad (6)$$

$$r_i(t+1) = r_i(0)[1 - \exp(-\gamma t)] \quad (7)$$

where α and γ are constants. Eventually, Ai will equal zero, while the final value of ri is r(0). Note that both loudness and rate are updated when the new solutions are improved to ensure that the bats are moving toward the best solutions.

The pseudocode of BBA is as follows[15]:

Input: r=0.9, A=0.5,Number of bats,max
Output: Best solution
Begin
Initialize the bat population
fitness1=fitness of initial bats
min_fit=fitness value of bat which is minimum
gbest=bat with minimum fitness values
while(t<max)
Adjust frequency and velocity
Calculate transfer function
If(T>rand)then
Generate new bats
End
If(rand>r)then
Update newbats with gbest
End
fitness2=fitness of newbats calculated
If(fitness1<fitness2&&rand>A)then
Update initial bat & reduce loudness,increase pulse rate
End
If(fitness2<min_fit)then
Update gbest
End
End
End

Each dataset is given to feature selection algorithm and various features are selected. Different features are selected for each separated dataset.

12 features are selected from the entire dataset. The selected features are given to next module.

D. Classification using Naïve Bayes Algorithm

In this stage, new training datasets are used that contains only the selected features from the previous module. These datasets are given into Naïve Bayes classifier for training and then they are tested using the test set.

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification[16]. It is a probabilistic model and it allows to capture uncertainty about the model in a principled way by determining probabilities of the outcomes.

A classification is a method to group data based upon the similarities in their features. The goal of classification is to accurately predict the target class for each instance in the data. A classification task begins with a data set in which the class labels are known. In the model training process, a classification algorithm finds relationships between the values of the predictions and the values of the target. Different classification algorithms use different techniques for finding relationships. These relationships are summarized in a model, which can then be applied to a different data set in which the class labels are unknown. Classification models are tested by comparing the predicted values to known target values in a set of test data.

IV. RESULTS AND DISCUSSION

A. Parameters for Evaluation

True Positive (TP) - True positive rate, also called as sensitivity or the recall rate in some fields measures the proportion of actual positives which are correctly identified.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

True Negative (TN) - An event when no attack has taken place and no detection is made.
False Positive (FP) - An event signalling an IDS to produce an alarm when no attack has taken place.
False Negative (FN) - The system allows an actual intrusive action to pass as nonintrusive behaviour.

The following are the parameters considered for evaluating the performance of the proposed idea.

I. Detection Rate - The detection rate is defined as the number of intrusion instances detected by the system (True Positive) divided by the total number of intrusion instances present in the test set.
II. False Positive Rate - Reducing the false positive rate improves the performance of the system.
III. Accuracy - Accuracy of the existing system can be calculated using the formula below.

B. Result Analysis
Size and distribution of attacks in the training and test data is listed In Table 1.

TABLE 1 SIZE AND DISTRIBUTION OF ATTACKS

Class	Size of training data	Size of test data
Normal	67343	9711
Prb	11656	2421
R2L	995	2754
DoS	45927	7456
U2R	52	200
Total	125973	22542

The binary bat algorithm is implemented in the cluster consisting of three machines and the execution time is compared with that of a single node cluster in Fig 2.
The experiments were carried out in both single node and multi node hadoop cluster.

TABLE 2 FEATURES SELECTED BY BBA

	Single node	Multi node
Features selected	14	12

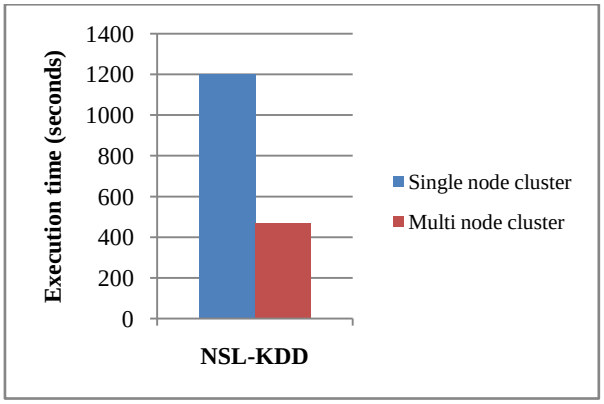

Fig. 2. Comparison of execution time in single and multi node cluster

After the features are selected, the reduced dataset is trained and tested using naïve bayes classifier and their detection rates are calculated. Fig 3 compares the detection rate of each type of attack with the existing method.

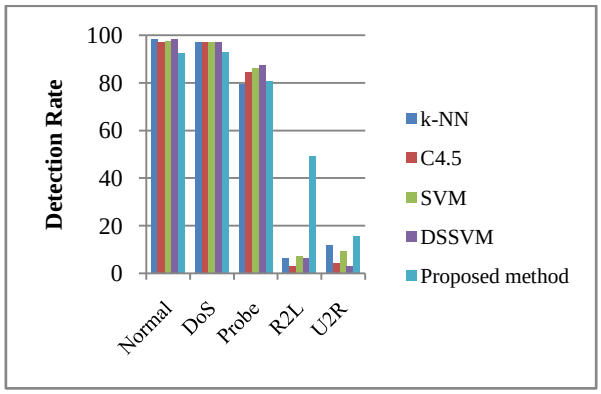

Fig. 3. Comparison of detection rates of existing and proposed method

TABLE 3 COMPARISON OF DETECTION RATES OF EXISTING AND PROPOSED METHOD

Class	k-NN	C4.5	SVM	DSSVM	Proposed method
Normal	98.3	97.0	97.7	98.4	92.24
DoS	97.0	96.8	97.2	97.2	92.89
Probe	79.4	84.3	86.1	87.5	80.83
R2L	6.5	3.0	7.2	6.3	**49.23**
U2R	11.8	4.4	9.2	3.1	**15.5**

Table 3 lists the percentage of detection rate for each class when classification is performed using different intrusion detection techniques like k-NN, C4.5, SVM and DSSVM. The performance of the proposed method is compared with the existing methods in terms of detection rate.
The other intrusion detection systems have higher detection rate for normal traffic and DoS and Probe attack. But the proposed method has higher detection rate for R2L and U2R attacks.
Since Naïve Bayes classifier is a probabilistic model and it considers each feature independently where the presence of a feature is not related to the presence of any other feature, it provides considerably higher detection rates for attack categories like R2L and U2R.

V. CONCLUSION

Hadoop is the distributed computing environment which is used for processing huge volume of data. If the data are distributed equally among the nodes then the execution time for the mapreduce job is decreased. Since in intrusion detection time plays an important role, the binary bat algorithm executed in Hadoop improves the performance of the system. Naïve bayes classifier is then used for training and testing the selected features. In naïve bayes classifier the presence of one feature is unrelated to the presence of any other feature. Hence the detection rates for

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 2, February 2017

R2L and U2R attacks have been greatly improved to 49.23% and 15.5% respectively.

REFERENCES

[1] Ashoor, Asmaa Shaker, and Sharad Gore (2011), 'Importance of Intrusion Detection system (IDS).'International Journal of Scientific and Engineering Research, Vol. 2 no. 1, pp.1-4.
[2] Kim, Gisung, Seungmin Lee, and Sehun Kim (2014), 'A novel hybrid intrusion detection method integrating anomaly detection with misuse detection.'Expert Systems with Applications, Vol. 41 no. 4, pp. 1690-1700.
[3] Oh, Sang Hyun, and Won Suk Lee (2013), 'An anomaly intrusion detection method by clustering normal user behavior.'Computers & Security, Vol. 22 no. 7, pp. 596-612.
[4] Jiang, ShengYi, Xiaoyu Song, Hui Wang, Jian-Jun Han, and Qing-Hua Li (2010), 'A clustering-based method for unsupervised intrusion detections.'Pattern Recognition Letters, Vol. 27 no. 7, pp. 802-810.
[5] Kang, Inho, Myong K. Jeong, and Dongjoon Kong (2012), 'A differentiated one-class classification method with applications to intrusion detection.'Expert Systems with Applications, Vol. 39 no. 4, pp. 3899-3905.
[6] Li, Yinhui, Jingbo Xia, Silan Zhang, Jiakai Yan, Xiaochuan Ai, and Kuobin Dai(2012), 'An efficient intrusion detection system based on support vector machines and gradually feature removal method.' Expert Systems with Applications, Vol. 39 no. 1, pp. 424-430.
[7] Mukherjee, Saurabh, and Neelam Sharma (2012), 'Intrusion detection using naive Bayes classifier with feature reduction.' Procedia Technology, Vol. 4, pp. 119-128.
[8] Khor, Kok-Chin, Choo-Yee Ting, and SomnukPhon-Amnuaisuk (2012), 'A cascaded classifier approach for improving detection rates on rare attack categories in network intrusion detection.' Applied Intelligence, Vol. 36 no. 2, pp. 320-329.
[9] Guo, Chun, Yajian Zhou, Yuan Ping, Zhongkun Zhang, Guole Liu, and Yixian Yang (2014), 'A distance sum-based hybrid method for intrusion detection. 'Applied intelligence, Vol. 40 no. 1, pp. 178-188.
[10] Qian, Jin, Duoqian Miao, Zehua Zhang, and XiaodongYue (2014),'Parallel attribute reduction algorithms using MapReduce. 'Information Sciences, Vol. 279, pp. 671-690.
[11] Nakamura, Rodrigo YM, Luis AM Pereira, K. A. Costa, Douglas Rodrigues, João P. Papa, and X-S. Yang (2012), 'BBA: A binary bat algorithm for feature selection.' In Graphics, Patterns and Images (SIBGRAPI), 25th SIBGRAPI Conference, pp. 291-297.
[12] Qian, Jin, Ping Lv, XiaodongYue, Caihui Liu, and Zhengjun Jing (2015), 'Hierarchical attribute reduction algorithms for big data using MapReduce.' Knowledge-Based Systems, Vol. 73, pp. 18-31.
[13] Xu, Rong-Fang, and Shie-Jue Lee (2015), 'Dimensionality reduction by feature clustering for regression problems. 'Information Sciences, Vol. 299, pp. 42-57.
[14] Yang, Xin-She, and Xingshi He (2013), 'Bat algorithm: literature review and applications. 'International Journal of Bio-Inspired Computation, Vol.5 no. 3, pp. 141-149.
[15] Mirjalili, Seyedali, Seyed Mohammad Mirjalili, and Xin-She Yang (2014), 'Binary bat algorithm. 'Neural Computing and Applications, Vol. 25 no. 3-4, pp. 663-681.
[16] http://www.statsoft.com/textbook/naive-bayes-classifier.