International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 04 Issue 10
September 2017

# Intrusion Detection System using Supervised Learning for Cloud Environment

[1]Sowmya Koneru
[1]Assistant professor, Andhra Loyola Institute of Engineering and Technology, Vijayawada, Andhra Pradesh, India
konerusowmya@gmail.com

**Abstract: This** paper presents a study the effect of classification algorithmson Cloud environment.To analyze the proposed model, four supervised learning algorithms are considered. These classifiers are Naïve Bayes classifier, Neural Networks, Logistic Regression and SMO. A benchmark dataset called CIDD1 is used to test the characteristics of the classifier and finally the conclusions are drawn.

## I. INTRODUCTION

cloud computing is known by more and more people due to its advantages such as high scalability, high flexibility, pay-per use facility and low operational cost[3]. The security plays a vital role in the cloud environment. Many of the researchers are working in the field of intrusion detection on cloud environment. The basic requirements of an intrusion detection system are its early detection and efficiency. Machine learning algorithms will be helpful, especially pattern recognition kind of problems. Bayesian belief networks, K-Nearest Neighbor classifiers, Neural Networks are some of the classifiers that can be applied on those datasets like KDD cup dataset, CIDD datasets etc. Especially when the class labels are already available classification algorithms yields better accuracies than clustering.

The fully distributed and open structure of cloud computing and services becomes an attractive target for potential intruders.Maintaining cloud security is a major challenge for the cloud services. Even many of the classification algorithms like distance based classifiers that yields better accuracies but consumes much computational time. Since an early detection is one of the main characteristics of intrusion detection system, it is necessary to build a model that detects attackers early and accurately.

Many of the researchers are working on these area based on available benchmark datasets like intrusion datasets. In proposed approach four supervised learning techniqueslike Artificial Neural Networks, logistic regression will be applied on CIDDS( Cloud Intrusion Detection Data set)and analyzed the results.

The objective of this paper is of twofold:

- To implement foursupervised learning algorithms on CIDDS data set.
- To analyze and compare these classifiers in terms of two parameters i.e., Building Time of the classifiers and accuracies of the classifiers.

**Classification:**Classification is one of the machine learning methods that are used for pattern recognition. Tree based classifiers, distance based classifiers, Support vector machines and neural network based classifiers are some examples of classification algorithms. Classifiers are well known as supervised learning algorithms i.e., class label dependent learner.

**Naïve Bayes classifier:**Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

**Multi layered Perceptron:**Neural networks, or more precisely artificial neural networks, are a branch of artificial intelligence[4]. Multilayer perceptrons form one type of neural network. Training a multilayer perceptron is the procedure by which the values for the individual weights are determined such that the relationship the network is modelling is accurately resolved.

**Sequential Minimal Optimization or SMO**: Training a support vector machine needsa solution of a largeQP optimization problem. SMO breaks this problem into a sequence of smallest QP problems[5]. These problems are solved analytically, which avoids using a time-consuming numerical QP optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. Because matrix computation is avoided, SMO scales somewhere between linear and quadratic in the training set size for various test problems, while the standard chunking SVM algorithm scales somewhere between linear and cubic in the training set size.

**Logistic Regression**:Logistic regression isa machine learning algorithmknown for the function used at the core of the method, the logistic function.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

**Performance Measure:**To measure the performance of various classifiers there are available with many measures like accuracy, recall, sensitivity, specificity etc. [7].

The confusion matrix is the outcome of the classifier, that consists of True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN) values of the classifier predictions. The following are formulas for some of the classification measures.

**Sensitivity** is the ratio of Number of Correctly classified normal sample as a normal and Total number of Normal Samples.

$$\text{Sensitivity} = TP/(TP + FN)$$

**Specificity** is the ratio of the number of attacks correctly classified by any classifier as attacks and total number of attacks.

$$\text{Specificity} = TN/(TN + FP)$$

**Accuracy** is the total number of attacks and normal samples that are correctly classified by a classifier and total number of samples[6].

$$\text{Accuracy} = (TN + TP)/(TN+TP+FN+FP)$$

## II. RELATED WORK

Massimo Meneganti et al.[1] proposed a fuzzy logic based classification to detect anomalies and further applied fuzzy neural networks to find anomalies in the cloud system.

Pei-Te Chen et al. [2]suggested thesecurity auditors concept, to identify the system weaknesses and modify the tested packets using fingerprints that can be detected and recognized by IDS.

Mehmood, Yasir, et al.[3] has analyzed existing cloud intrusion detection systems based on their performance, type, positioning, data source and attacks they detect. Authors also listed limitations of each detection systems. Finally suggested some of the detection systems based on the security requirements of the cloud computing.

Liu, Bingwei, et al. [8] presented a complete system for sentiment mining on large datasets using a Na¨ıve Bayes classifier with the Hadoop framework. The paper was implemented the NBC on top of Hadoop framework, with additional modules to automate the experiment. We also provide the implementation details for converting a classifier to Hadoop program, of which any machine learning algorithm could be replaced for the NBC.

## III. METHODOLOGY

The proposed model comprises of four phases as follows.
Data preprocessing phase for preparing cloud data ready for the classification. This phase under gone into cleaning, transformation steps.
In the second phase feature selection is applied on dataset.

In the third phase various classifiersare implemented on cloud dataset.
Finally results will be analyzedIn the first phase data preprocessing is applied on the cloud data set. Figure 1 visualizes the methodology.

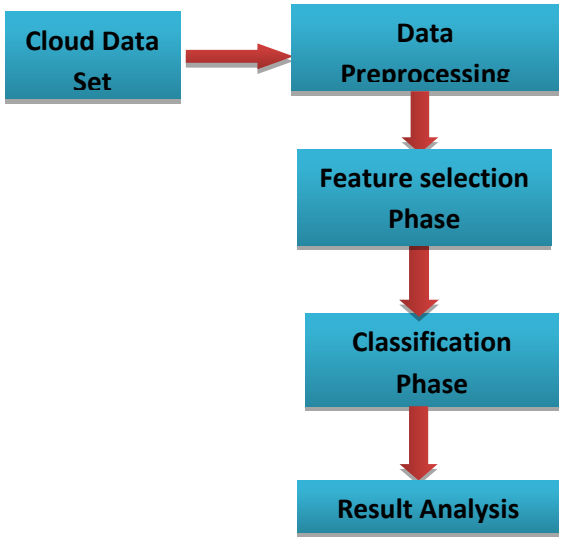


**Figure 1: Proposed Methodology**

Data Preprocessing: In data preprocessing phase data transformation to convert the entire dataset into numeric, as some of the features of the data set are of categorical. Z-score normalization algorithms are applied on CIDD1 data set.

In the next phase in the pre-processed dataset is undergone into feature relevance phase, in which irrelevant features like source and destination IP addresses are removed. After feature selection phase the data is undergone into 4 different classifiers i.e., Naïve Bayes classifier, Multi layered Perceptron, SMO and Logistic Regression. Theseclassifiers are observed by their accuracy, sensitivity, specificity and total building time of the classifier model. Finally the results were drawn.

## IV. RESULTS AND DISCUSSION

The experiment was done on Weka 3.8.2 toll and considered one day dataset of the CIDD-1 dataset. The sample size of the dataset is about 1,56,950.The dataset is a two class labeled i.e., Normal and attacker and consists of 12 features. Four classifier algorithms have undergone for 10 fold cross validation.Table 1 shows accuracy and total build time of the four classifiers.

**Table 1: Classifier Build time and accuracies of Algorithms**

| Classifier | Total build time (Seconds) | Accuracy |
|---|---|---|
| **Naïve Bayes** | 0.12 | 75.11% |
| **Multi layered Perceptron** | 83.96 | 83.97% |
| **SMO** | 121.16 | 88.46 % |
| **Logistic Regression** | 4.49 | 92.06 % |

Figure 2 and figure 3 presents the graphs for total build times and accuracies of four classifiers.

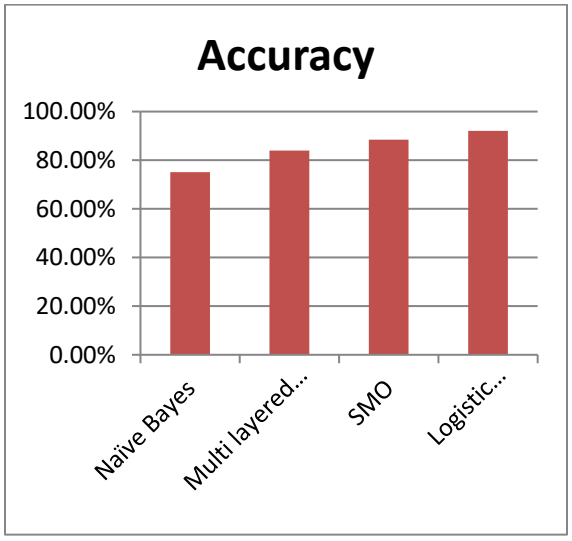


**Figure 2:Accuracies of the algorithms**

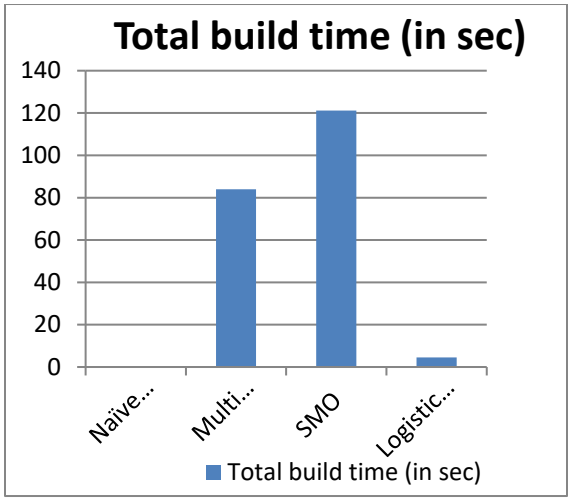


**Figure 3: Total Build time of the algorithms**

From table 1, figure2 and figure 3 it is observed that the logistic regression has shown better accuracy about 92% with a tolerable building time of 4.49 seconds whereas Naïve Bayes classifier exhibits a poor accuracy of 75.11%.

## V. CONCLUSION

This paper analyzed the effects of four classifiers Naïve Bayes classifier, Neural Networks, Logistic Regression and SMO on a benchmark cloud dataset i.e., CIDD1. Among these classifiers logistics regression showed good results in terms of accuracy as well as total build time. This paper CIDD1 was considered to test the characteristics of the classier on cloud environment. A realtime frame work can be created and can be tested and is proposed as a future work.

## REFERENCES

[1] Massimo Meneganti, Francesco S. Saviello, and Roberto Tagliaferri, "Fuzzy Neural Networks for Classification and Detection of Anomalies", IEEE transactions on neural networks, vol. 9, no. 5, pp. 848-861 september 1998.

[2] Pei-Te Chen, Chi-Sung Laih, "IDSIC: an intrusion detection system with identification capability", Springer-Verlag, pp.185-197, June 2007.

[3] Mehmood, Yasir, et al. "Intrusion detection system in cloud computing: challenges and opportunities." *Information Assurance (NCIA), 2013 2nd National Conference on*.IEEE, 2013.

[4] Gardner, Matt W., and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences." *Atmospheric environment* 32.14-15 (1998): 2627-2636.

[5] Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).

[6] Kuncheva, Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy." *Machine learning* 51.2 (2003): 181-207.

[7] Adams, Niall M., and David J. Hand. "Improving the practice of classifier performance assessment." *Neural computation* 12.2 (2000): 305-311.

[8] Liu, Bingwei, et al. "Scalable sentiment classification for big data analysis using naive bayes classifier." *Big Data, 2013 IEEE International Conference on*. IEEE, 2013.

[9] McCarty, John A., and ManojHastak. "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression." *Journal of business research* 60.6 (2007): 656-662.