International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 5 Issue 4, April-2016, Impact Factor: 1.544

Intrusion Detection System: A Statistical Analysis to Find the Best Detector

Ms. Krupa Bhavsar[1], Dr. Jigar Patel[2]

[1]Assistant Professor, DCS, Ganpat University
[2]Associate Professor, KIRC - MCA Dept., Kalol

ABSTRACT

The communication has shifted to the wireless or we can say that more prone to data leak systems so to stop the intrusion we require some intrusion detection system (IDS). These system can be network based or they can be host based. In any of the IDS we cannot say that every result it provides is authentic one so to get the best results we need techniques through which we can minimize the probability of the false Intrusion detection. In this paper we will review the Naïve Bayes, KNN and Firefly technique and will compare them for the best.

Keywords: IDS, Firefly, Naïve Bayes, KNN.

1. INTRODUCTION

Intrusion detection system (IDS) can be device (Hardware) or software application that keep a keen eye on the systems for any of the malicious activity. There are multiple types of IDS available for example if we have to secure the network we can use network based IDS for host machines host based IDS and many more. The main advantage of IDS is that it can detect the external as well as internal intrusion unlike firewall which is only for external intrusion.

The limitations of IDS are as follows:

a. The noise, bad packets failed DNS message and other local packets can trigger the high false alarm rate.
b. The false alarm rate is higher than the real attack sometimes it is so low that the real attacks are missed or ignored by the IDS.
c. If the signatures and versions are constantly changing then the update information should be present else old signatures can lead to false alarm.
d. For signature based IDS, till the signature of the application is found malicious the IDS will take it as non-malicious.
e. In the IDS if the packet received is encrypted then it is termed as safe and can lead to intrusion.
f. Masking of the addresses is an important problem as the mentioned address could be fake.

Due to the above mentioned limitations we need a better IDS so that the real intrusion can be minimized. The IDS is generally divided into two major approaches which are as follows:

i. Anomaly Detection:

In this method the anomaly is detected in the user behavior that is the user has a certain profile with limited access and use if the IDS has detected some different behavior like extended use or breaking of the limitations then we can say that the intrusion is done. The system based on this approach are IDES and NIDES.

ii. Misuse Detection:

It is also called as the signature detection. Here the signature of the application packets etc are matched. The signature characteristic and values are checked if the matches the constraints then the application or the packet is accepted. STAT and IDIOT are the system based on this approach.

2. NAÏVE BAYES

The graphical model which is most widely used is the Bayesian networks because it can handle uncertain information. The Bayesian network is divided into two components.

i. Graphical Component:

A graphical component is composed of a directed acyclic graph. Here the vertices of the graph represent events and the edges are the relations between them.

ii. Numerical Component:

It consists the conditional probability distribution of different nodes with respect to its parent in the DAG.

The naïve Bayesian is a very simple form of the Bayesian network which is composed of the DAG graph with only one parent. The parent node represents the unobserved node and there can be multiple child nodes which are termed as the observed nodes with the strong assumption of independence among child nodes in the context of their parent. To deal with the classification problem we require Naïve Bayesian networks. Here the classification is ensured by considering the parent node to be a hidden variable and the child node specify the different attributes of the node object. Therefore if the system is in training mode so in the presence of the set we should only compute the conditional probabilities since each structure of the set is unique.

The computation can be summarized as follows:

a. For the discrete attributes we calculate the conditional probabilities, which are computed from the frequencies by counting how many times each attribute value pairs occurs with each value of the parent node.
b. In the graph the continuous attributes are usually handled by assuming that they have a normal/Gaussian probability distribution. Since it is normal the values would be same throughout.

In the Naïve Bayesian IDS. In the algorithm first we have to find the prior probability for the given intrusion detection dataset that is the KDD99 [2] cup data set, then we have to find the class conditional probability for the dataset. Once we have the prior and the conditional probabilities we have to find the highest classifier probability and on this base we will find detection rate and the false positive for the intrusion data set.

The prior probability is calculated by its number of occurrences in each class dataset. The class conditional probability can be estimated by counting how often each attribute value occurs in the class in dataset D.

Procedure: Decision Tree
Input: Dataset D
Output: DA, FP For Attack Data
Do
Take the Class CL From D.
For each attribute value
Remove the noise from the dataset.
Calculate the prior probability P(Cj) for each class Cj in dataset.

$$D: P(C_j) = \frac{\sum t_i \rightarrow c_j}{\sum_{i=1}^{n} t_i}$$

End For
For each attribute value
Calculate the class conditional probabilities P (Aij|Cj) for each attribute values in dataset D:

$$P(A_{ij} \mid C_j) = \frac{\sum_{i=1}^{n} A_i \rightarrow c_j}{\sum t_i \rightarrow c_j}$$

End For
End Do
Do
Multiply the prior probability and class conditional probability.
End Do
Do
Consider the class with the highest classifier probability.
End Do
Repeat steps 2to4 until all attribute at their highest probability.

3. KNN (K –NEAREST NEIGHBOR)

The KNN or the K- Nearest neighbor is another false alarm determination technique which is based on the learning by the analogy. It means the system is trained using by comparing the given test tuple with training tuples which are similar to each other.

The training tuples have n attributes therefore each tuple represents a point in the n- dimensional space. In this way all the training set tuples are stored in the dimensional space. When a random tuple is passed the KNN classifier searches the dimensional pattern space for the tuples similar to the input tuple. These k number of tuples found similar are known as the k-nearest neighbor of the unknown tuple inputted.

Closeness is the Euclidean distance between the tuples. The Euclidean distance between two points or tuples X1=(x11, x12,…, x1n) and X2=(x21, x22,…, x2n) obtained from equation below.

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}$$

The basic steps of the k-NN algorithm are; • To compute the distances between the new sample and all previous samples, have already been classified into clusters; • To sort the distances in increasing order and select the k samples with the smallest distance values; • To apply the voting principle. A new sample will be added (classified) to the largest cluster out of k selected samples.

The KNN algorithm is very attractive because of its simplicity with the considerable better performance. Therefore we can reduce the computational cost involved. It has an intelligent system to eliminate the samples or we can say the tuples. The removal of the tuples does not lead to the less number of tests and the smaller decision area.

4. FIREFLY

The firefly is a hybrid algorithm which is robust in finding the optimization problem. This algorithm is n nature inspired that is the flashing lights of the fireflies. The algorithm is based on three rules which are based on the real life characteristics of the real fireflies. The rules are as follows:

a. All the fire flies are same and they will always follow the most attractive and the brighter one.
b. The degree of the attractiveness of a firefly is proportional to its brightness which decreases as the distance other firefly increases.
c. If there is no brighter or more attractive firefly then a particular one, then it will move randomly.

If we have to relate it to the optimization technique consider the flashing light is associated with the fitness function then we can find the efficient optimal solutions. For the solution as in the fireflies we have to use the attractiveness, movement and distance they are define as follows.

i. Attractiveness:

In this algorithm the form in which attractiveness is considered is the function. The function is given by the following monotonically decreasing function

$$\beta(r) = \beta_0 * exp(-\gamma r_{ij}^m) \text{ with } m \geq 1$$

Where, r is the gap between two fireflies.

β is the attractiveness in the starting when distance r=0
γ is an absorption coefficient which controls the decrease of light intensity.

ii. Distance

The distance between two fireflies i & j, at positions x_i and x_j. It can be defined as a Cartesian.

$$r_{ij} = \| x_i - x_{ji} \| = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2}$$

Where $x_{i,k}$, is the Kth component of the spatial coordinate x_i of the ith firefly and d is the number of dimensions we have, for d=2 , we have

$$r_{ij} = \sqrt{(x_i - x_j)^2 - (y_i - y_j)^2}$$

However, the calculation of a distance r can also be defined using other distance metrics, based on the nature of problem, such as manhattan distance.

iii. Movement

The movement of the firefly i which is attracted by a more attractive. Firefly j is given by is given by:

$$x_i = x_i + \beta_0 * \exp(-\gamma r_{ij}^2) * (x_j - x_i) + \alpha * (rand - \tfrac{1}{2})$$

Where the first term is the current position of a firefly, the second term is used for considering a firefly's attractiveness to light intensity seen by adjacent fireflies and third term is used for the random movement of fireflies in case there are no brighter ones. The coefficient α is a randomization parameter determined by the problem of interest. Rand is a random number generator uniformly in the distributed space [0,1].

CONCLUSION

In this paper we are reviewing the three optimization algorithm which will help us in creating the IDS with less probability to create the false alarm. We will implement and compare each of the techniques and will find the best technique to create a better IDS.

REFERENCES

[1]. "Naive Bayesian Networks in Intrusion Detection Systems" by Nahla Ben Amor, Salem Benferhat and Zied Elouedi.
[2]. "Enhanced Naïve Bayes Algorithm for Intrusion Detection in Data Mining" by Shyara Taruna R. and Mrs. Saroj Hiranwal in IJCSIT
[3]. "Comparison of Classification Methods Based on the Type of Attributes and Sample Size" by Reza Entezari-Maleki, Arash Rezaei, and Behrouz Minaei-Bidgoli
[4]. "Using K – nearest neighbor classifier for Intrusion detection" by Yihua Liao.
[5]. "Supression of Low Frequency Oscillations Using Hybrid Optimization Technique" by Mohit Bhagat, Baljit Singh.
[6]. "Intrusion Detection Model Based on Data Mining Technique" by Sadia Patka.