A Hybrid Supervised and Unsupervised Learning Approach for Node Classification

[1]Alekhyasuma,[2]P.Rajasekhar
[1]Final MTech student, [2]Assistant professor
[1,2]Department of CSE, Avanthi institute of engineering and technology, Vizag, AP

Abstract:

Malicious node identification is always an interesting research issue in the field of network security. Various approaches like statistical, distance and density based techniques (trust measures, SVM Base approaches, classification mechanisms) introduced by the various researchers. In this paper we are proposing a novel cluster based approach for detecting the anomaly or outlier node while communicating to the destination node, by initially computing the clusters and followed by the positive and negative probabilities.

I. INTRODUCTION

A distributed sensor networks (Dsn) can be defined as set of spatially scattered intelligent sensors designed to obtain the measurements from the environment, abstract relevant information from the data gathered, and to derive appropriate inferences from the information gained. Distributed sensor networks depend on multiple processors and process information from multiple processes. The major task of dsn is to process data[1], possible noise corrupted, acquired by various sensors and to integrate it. DSNs may be deployed in hostile areas where communication is monitored and nodes are subject to capture and surreptitious use by an adversary. Hence, DSN requires cryptographic protection of communication, sensor capture detection and sensor disabling, reduces uncertainty in it, and produce abstract interpretations of it.

Currently, there has been increasing interest on the development of Dsns for the process of information gathering. Availability of new technologies, these networks are economically feasible. The increased complexity of today's information gathering tasks has created a demand for such networks. These tasks are time-critical and rely on reliable delivery of accurate information. Thus, the search for efficient and fault-tolerant architecture of dsn has become an important research area in computer science[2].

For instance, it should be possible to combine the information given by infrared sensors with microwave radars. No single sensor or sensor cluster has the information to solve the entire problem. The common idea of setting up a global processor that receives all the information from the sensors, solves the entire problem, and sends the relevant parts of the solution to the sensors is really not practicable. Both data collection and control have to be logically and geographically distributing necessitating the sharing of information and the use of cooperative problem solving approaches[3].

DSn is basically system of connected, cooperating and generally diverse sensors that are spatially dispersed. Three important facts are emerged from such a framework:

1. Each sensor can see some but not all of low level activities performed by the sensor networks.
2. Data is perishable, in the sense that information value depends upon the time required to acquire and process it.
3. There should be limited communication among the sensor processors, so that communication computation trade-off can be made.

II. RELATED WORK

In the typical DSN, each node needs to fuse the local information with the data collected by the other nodes, so that an updated assessment is obtained. Maintaining consistency and eliminating redundancy is the two important considerations. The problem of determining what should be communicated is more important than how communication is to be effected. It is easy to see the different classes of information warrant

different degrees of reliability and urgency. Network nodes are equipped with wireless transmitters and receivers using antennas which may be Omni directional (isotropic radiation), highly directional (point to point), possibly steerable, or some combination thereof. At a given point in time, depending on the node's positions and their transmitter and receiver coverage patterns, transmission power levels and co-channel interference levels, a wireless connectivity in the form of a random, multi-hop graph or "ad hoc" network exists between the nodes. This ad hoc topology may change with time as the nodes move or adjust their transmission and reception parameters[4].

A large number of important applications depend on sensor networks interfacing with the real world. These applications include medical, military, manufacturing, transportation, safety and environmental planning systems. Many have been difficult to realize because of problems involved with inputting data from sensors directly in to automated systems. Sensor fusion in the context of distributed sensor networks has emerged as the method of choice for resolving these problems.

Sensor networks vs Ad hoc networks:

- Number of sensors is expected to be orders of magnitude bigger.
- Sensors may not have global identification.
- Sensors are power/CPU/memory constrained.
- Sensors are densely deployed.
- Sensors are prone to failure.
- Possibly very frequent topology changes.
- Sensor uses broadcast, ad- hoc uses point to point.

We proposed a novel and productive trust calculation system with naive Bayesian classifier by examining the new operators data with existing specialists data, by characterizing the feature sets or attributes of the specialists. This methodology demonstrates ideal results than the customary trust calculation approaches[5][6].

III. PROPOSED WORK

We are proposing a productive internet traffic grouping over log information or preparing dataset which comprises of source ip-address or name, Destination ip-address and port number, kind of convention and number of parcels transmitted from source to destination. At the point when a hub associates if recovers the meta information i.e. testing dataset and advances to the preparation dataset .both preparing and testing datasets CAN Be sent to Bayesian classifier for examining the conduct of the associated hub.

We proposed a novel and productive trust calculation system with naive Bayesian classifier by examining the new operators data with existing specialists data, by characterizing the feature sets or attributes of the specialists. This methodology demonstrates ideal results than the customary trust calculation approaches.

In our methodology we proposes a productive arrangement based methodology for breaking down the anonymous clients over network traffic and figures the trust measures in view of the preparation data with the anonymous testing data. Our engineering contributes with the accompanying modules such as Analysis agent, Neighborhood node, Classifier and data collection and preprocess as takes after

Clustering:

Log data can be clustered based on the maximum similarity between the data records. Initially k number of centroids can be selected and computes maximum similar records with respect to all centroids and places the data record in cluster which has maximum similarity and continues the same process until a maximum number of iterations.

K means clustering:

1: Select K points as initial centroids for initial iteration

2: until Termination condition is met (user specified maximum no of iterations)

3: Measure the similarity between the data point and centroid

4: Assign each point to its closest centroid to form K clusters

5: Recompute the centroid within individual clusters

6 .Continue steps from 2 to 5

Classification:

For optimal performance classifies input node with suitable cluster data instead on entire dataset. Initially computes the maximum similarity with the centroids of the clusters and places the input record with respect to cluster holder and then computes the probability of anomaly status (i.e positive and negative probability).

Naïve Bayesian Classification:

Algorithm to classify malicious agent

Sample space: set of agent

H= Hypothesis that X is an agent

P(H/X) is our confidence that X is an agent

P(H) is Prior Probability of H, ie, the probability that any given data sample is an agent regardless of its behavior

P(H/X) is based on more information, P(H) is independent of X

Estimating probabilities

P(X), P(H), and P(X/H) may be estimated from given data

Bayes Theorem

P(H|X)=P(X|H)P(H)/P(X)

Steps Involved:

1. Each data sample is of the type

$X=(x_i)$ i =1(1)n, where xi is the values of X for attribute Ai

2. Suppose there are m classes C_i, i=1(1)m.

X Î Ciiff

$P(C_i|X) > P(C_j|X)$ for 1£ j £ m, j¹i

i.e BC assigns X to class C_i having highest posterior probability conditioned on X

The class for which $P(C_i|X)$ is maximized is called the maximum posterior hypothesis.

From Bayes Theorem

3. P(X) is constant. Only need be maximized.

 If class prior probabilities not known, then assume all classes to be equally likely

 Otherwise maximize

$P(C_i) = Si/S$

Problem: computing $P(X|C_i)$ is unfeasible!

4. Naïve assumption: attribute independence

$P(X|C_i) = P(x_1,\ldots,x_n|C) = PP(x_k|C)$

5. In order to classify an unknown sample X, evaluate for each class C_i. Sample X is assigned to the class CiiffP(X|C_i)P(C_i) > P(X|C_j) P(C_j)

IV. CONCLUSION

We are concluding our research work with efficient hybrid approach of clustering and classification mechanisms, entire training dataset can be initially clustered based on the similarity and then computes the similarity between the centroids and testing samples and then applies naïve Bayesian classification for analyze the input node behavior.

We can improve our concluded work by enhancing the classification approach, In classification based approach, analysis fails when testing sample of data not available in training dataset or new data sample and classification fails when data is inconsistent or not available for specific attributes . By improving these two features we can enhance the performance of current intrusion detection system

REFERENCES

[1] L. Eschenauer and V. Gligor, “A Key-Management Scheme forDistributed Sensor Networks,” Proc. Ninth ACM Conf. Computerand Comm. Security (CCS ’02), pp. 41-47, 2002.
[2] A. Perrig, R. Szewczyk, J. Tygar, V. Wen, and D. Culler, “SPINS:Security Protocols for Sensor Networks,” Wireless Networks, vol. 8,no. 5, pp. 521-534, 2002.
[3] D. Hong, J. Sung, S. Hong, J. Lim, S. Lee, B. Koo, C. Lee, D.Chang, J. Lee, K. Jeong, H. Kim, J. Kim, and S. Chee, “HIGHT: ANew Block Cipher Suitable for Low-Resource Device,” Proc.Eighth Int’l Workshop Cryptographic Hardware and EmbeddedSystems (CHES ’06), pp. 46-59, 2006.
[4] P. Kamat, Y. Zhang, W. Trappe, and C. Ozturk, “EnhancingSource-Location Privacy in Sensor Network Routing,” Proc. IEEE25th Int’l Conf. Distributed Computing Systems (ICDCS ’05), pp. 599-608, 2005.
[5] C. Ozturk, Y. Zhang, and W. Trappe, “Source-Location Privacy inEnergy-Constrained Sensor Network Routing,” Proc.Second ACMWorkshop Security of Ad Hoc and Sensor Networks (SASN ’04), pp. 88-93, 2004.
[6] L. Eschenauer and V. Gligor, “A Key-Management Scheme forDistributed Sensor Networks,” Proc. Ninth ACM Conf. Computerand Comm. Security (CCS ’02), pp. 41-47, 2002.

BIOGRAPHIES

AlekhyaBarigada studyingm.tech(information technology) from avanthi institute of engineering and technology,vizagaffliated to jntukakinada from 2013-2015. She completed B.tech(computer science & engineering) from gokul institute of technology&science,diridi,bobbili, affliated to jntukakinada from 2009-2013.

P.Rajasekhar received his **M.SC** degree in COMPUTER SCIENCE from ANDHRA UNIVERSITY, VIZAG. Presently he is working as HOD in Computer science and engineering department in Avanthi institute of engineering and technology, vizag. His research interests include data mining, software engineering, E-commerce.