ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 389 - 391 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

A Hybrid Intrusion Detection System for Identification of Anonymous Behavior

[1]K.Sasidhar,[2]N.Sagar Pavan Kumar
[1]Final M.Tech Student, [2]Assistant Professor
[1,2]Dept of Computer Science and Engineering
[1,2]Pydah College of Engineering, Visakhapatnam, AP, India.

Abstract: Servers maintain firewall log data of in and out traffic for Intrusion Detection and prevention mechanisms .Even though various approaches available to detect and prevent unauthorized user or access, they are not optimal. In this paper we are proposing a hybrid approach of Classification and Signature mechanism. Classification analyzes behavior of node and signature mechanism maintains authentication of node. Our experimental result shows optimal results than traditional approaches

Index Terms: Intrusion Detection, Classification, Wireless sensor networks

INTRODUCTION

Various approaches available for identify the unauthorized behavior of the incoming nodes like with their trust measures like direct trust, indirect trust and reputation metric, these metrics always maintained globally so network cannot directly depend on it. Main drawback with the Signature based IDS mechanisms are pattern must be continuously updated and difficult to identifying the new pattern. Direct classification techniques make more time complexity while classifying the network traffic of in and out data flows.

Anomaly detection mechanism traditionally works with either signature based approaches or trust based approaches or statistical based approaches or probability based approaches. Traditional approaches not optimal while comparing with static attributes and retrieval of trust computational values from third parties or data rating calculated by intermediate nodes. Even though classification based techniques works they are suffering from mismatched feature set selection and major issue is no semantic comparison.

For example, consider the application of online anomaly detection to syndrome surveillance, where the goal is to detect disease outbreak at the earliest possible instant. Imagine that we monitor two variables: max daily temp and numfever.max daily temp tells us the maximum outside temperature one given day, and num fever tells us how many people were admitted to a hospital emergency room complaining of a high fever. Clearly, max daily temp should never be taken as direct evidence of an anomaly. Whether it was hot or cold on a given day should never directly indicate whether or not we think the main drawbacks with the traditional approaches are redundant patterns leads to inaccurate results. Traditional trust metrics, data ratings may not give the optimal solutions. Identify unauthorized behavior from huge traffic log data takes more time complexity than clustered data.

RELATED WORK

Let **R** be a p x p sample correlation matrix computed from n observations on each of p random variables, $X_1,X_2, \ldots,X_p$. If $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$ are the p eigen value-eigenvector pairs of $\mathbf{R},\lambda_1 \geq \lambda_{2,} \ldots \geq \lambda_p \geq 0$,then the i^{th} sample principal component of an observation vector $x= (x_1 ,x_2 ,x_3 ,\ldots\ldots x_p)'$is $y_i=e_i' z= e_{i1} z_1+e_{i2} z_2+\ldots..+e_{ip}z_p$ i=1,2,3….,p(4) (8)
Where $e_i=(e_{i1},e_{i2},e_{i2}\ldots\ldots\ldots\ldots e_{ip})'$ is the i^{th} eigen vector and $z_i=z_1,z_2,z_3\ldots\ldots z_p$ is the vector of standardized observations defined as $z_k=x_k-x_k/\sqrt{s_{kk}}$, k=1,2,3,4…….p
where x and s_{kk} are the sample mean and the sample variance of the variable X_k. The i_{th} principal component has sample variance $\lambda_i$and the sample covariance of any pair of principal components is 0. In addition, the total sample variance in all the principal components is the total sample variance in all standardized variables Z_1, Z_2, …, Z_p, i.e.,$\lambda_1,\lambda_2,\lambda_3\ldots\ldots\ldots\ldots.\lambda_p=p$
This means that all of the variation in the original data is accounted for by the principal components[5][6].

PCA has been applied to the intrusion detection problem as a data reduction technique, not an outlier detection tool. It is our interest to use PCA to identify attacks or outliers in the anomaly detection problem. Though graphical methods are effective in identifying multivariate outliers, particularly when working on principal components, they may not be practical for real-time detection applications. Applying an existing formal test also presents a difficulty since the data need to follow some assumptions in order for the tests to be valid, e.g., the data have a multivariate normal distribution. Thus, we develop a novel anomaly detection scheme based on the principal components that can be applied in real time and does not impose too many restrictions on the data[7].

Following the anomaly detection approach, we assume that the anomalies are qualitatively different from the normal instances. That is, a large deviation from the established normal patterns can be flagged as attacks. No attempt is made to distinguish different types of attacks. To establish a detection algorithm, we perform PCA on the correlation matrix of the normal group. The correlation matrix is used because each feature is measured in different scales. It is important that the training data are free of outliers before they are used to determine the detection criterion because outliers can bring large increases in variances, co-variances and correlations. The relative magnitude of these measures of variation and co-variation has a significant impact on the principal component solution, particularly for the first few components. Therefore, it is of value to begin a PCA withal robust estimator of the correlation matrix. One simple method to obtain a robust estimator is multivariate trimming[8][9]. First, we use the Mahalanob is metric to identify the 100γ% extreme observations that are to be trimmed.Beginning with the conventional estimators **x** and **S**, the distanced$_i^2$={I,2,3….,n} for each observation $\mathbf{x}_i$ =(i=1,2,…,n) is computed. For a given γ (0.005 in ourexperiments), the observations corresponding to the γ*n largest values of d$_i^2$=1,2,…..,n are removed.

PROPOSED WORK

We are proposing an improved cluster based anomaly detection mechanism for identify unauthorized or malicious behavior, Log of firewall data or synthetic training dataset and it consists of source ip address or node name, destination ip address or node name, destination port number, protocol type and number of packets transmitted from source node to destination node. When an incoming node connects destination node, it retrieves the Meta data i.e. testing sample and forwards to the training dataset, after clustering of training dataset. In our approach initially training samples can be clustered based on the similarity between the data records and centroids which are randomly selected from the records of log data up to a maximum number of user specified iterations and then input sample can be forwarded to final cluster centroids to compute positive and negative probability.

We are considering a synthetic dataset which contains in and out nodes log data of Source ip-address or name, destination ip-address or name, type of protocol, port number, number of packets transmitted along with their anomaly status, this data can be clustered based on similarity between the data records and then input sample can be forwarded to centroids of clusters.

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: Recomputed the centroid within individual clusters
6 .Continue steps from 2 to 5

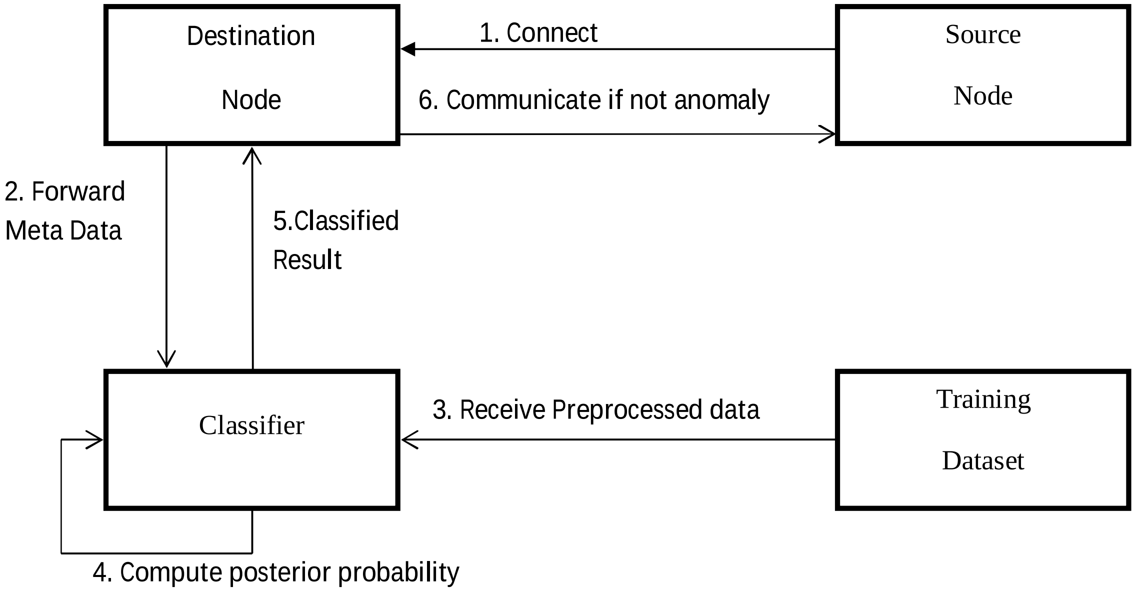

Fig 1 : proposed Architecture

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 a node

P (H/X_i) is our confidence that X_i is an incoming node

P(H) is Prior Probability of H and it is probability that any given training sample is an agent regardless of its anomaly or not anomaly behavior

P(H/X) is a conditional probability and P(H) is independent of X

Estimating probabilities

P(H), $P(X_i)$ and $P(X_i/H)$ may be estimated from given training and testing data samples

$P(H|X_i)=P(X_i|H)*P(H)/P(X_i)$

Steps Involved:

1. Each training data sample is of attribute type

X= (x_j) j =1(1….n), where x_j is the values of X for attribute A_j

2. Suppose there are m decision classes C_j, j=1(1…m).

$P(C_i|X) > P(C_j|X)$ for 1<= j <= m, j>i

i.e. classifier assigns X to decision class C_j having highest posterior probability conditioned on testing sample X

The decision class for which $P(C_j|X)$ is maximum is known as maximum posterior hypothesis of the sample.

From Bayes Theorem

3. $P(X_i)$ is constant and Only need be maximized.

□ if class initial probabilities not known prior then we can assume all decision classes to be more equally likely decision classes

□ Otherwise maximize the samples$P(C_i)$ = Si/S

4. Naïve assumption for attribute independence

$P(X|C_j) = P(x_1,\ldots..,x_m|C) = PP(x_k|C)$

5. To classify an unknown testing sample X_i, compute each decision class C_i and Sample X is assigned to the class

iff (Prob$(X_i|C_i)P(C_i)> P(X_i|C_j) P(C_j)$).

CONCLUSION AND FUTURE WORK

We have been concluding our current 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 for Distributed Sensor Networks," Proc. Ninth ACM Conf. Computer and 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: A New Block Cipher Suitable for Low-Resource Device," Proc. Eighth Int'l Workshop Cryptographic Hardware and Embedded Systems (CHES '06), pp. 46-59, 2006.
[4] P. Kamat, Y. Zhang, W. Trappe, and C. Ozturk, "Enhancing Source-Location Privacy in Sensor Network Routing," Proc. IEEE 25th Int'l Conf. Distributed Computing Systems (ICDCS '05), pp. 599- 608, 2005.
[5] C. Ozturk, Y. Zhang, and W. Trappe, "Source-Location Privacy in Energy-Constrained Sensor Network Routing," Proc. Second ACM Workshop Security of Ad Hoc and Sensor Networks (SASN '04), pp. 88- 93, 2004.
[6] L. Eschenauer and V. Gligor, "A Key-Management Scheme for Distributed Sensor Networks," Proc. Ninth ACM Conf. Computer and Comm. Security (CCS '02), pp. 41-47, 2002.
[7] 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.
[8] 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: A New Block Cipher Suitable for Low-Resource Device," Proc. Eighth Int'l Workshop Cryptographic Hardware and Embedded Systems (CHES '06), pp. 46-59, 2006.
[9] A. Bogdanov, L. Knudsen, G. Leander, C. Paar, A. Poschmann, M. Robshaw, Y. Seurin, and C. Vikkelsoe, "PRESENT: An Ultra-Lightweight Block Cipher," Proc. Ninth Int'l Workshop Cryptographic Hardware and Embedded Systems (CHES '07), pp. 450-466, 2007. [10] K. Mehta, D. Liu, and M. Wright, "Location Privacy in Sensor Networks against a Global Eavesdropper," Proc. IEEE 15th Int'l Conf. Network Protocols (ICNP '07), pp. 314-323, 2007.