IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Distributed Intrusion Detection System using Bayesian learning and Apache Mahout

Ronak Agrawal[1], Ganesh Talekar[2], Akshaysingh Chandel[3], Shriganesh Munde[4], Deeplakshmi Zingade[5]

BE, Computer, AISSMS IOIT, Pune, India [1,2,3,4]

ME, Computer, AISSMS IOIT, Pune, India [5]

Abstract: If the distributed intrusion detection system do not exist then detecting the intrusions in the networks would have become tedious task and the risk would have been increased. In market there are plenty of different intrusion detection systems available but no one of the distributed type. Many of them make use of pattern matching and different complex algorithms but this makes the system slow. So for proper Distributed intrusion detection system using Bayesian learning in Apache mahout using concepts like K-means clustering, Neuro-Fuzzy, Artificial neural network, Bayesian Learning, Apache Mahout to overcome the flaws like lagging in the detection of the intrusions, fail to detect the intrusion, harm to the system or system failure.

Keywords: K-means, Artificial Neural Network, Neuro-Fuzzy, Bayesian Network, Apache Mahout.

I. INTRODUCTION

For the detection of the intrusions in the distributed network we make use of the key-concepts like k-means clustering, Artificial neural network, Neuro-fuzzy logic, Bayesian learning and Apache Mahout. As till now there no such kind of the distributed system available in the market we as an initiative process trying to keep this idea forward so that it will be helpful for getting more secured from the different sever attacks.

K-means is a clustering algorithm which can be used to form a cluster and handle a large dataset in the cluster. The main reason to make use of k-means is one of the simplest algorithm which uses unsupervised learning method to solve known clustering issues.

Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected neurons working in parallel to solve a specific problem. Neural networks are trained by example. An artificial neuron works almost in the same as that biological one does. Neural networks do not perform miracles. But if used sensibly they can produce some amazing results.

The reason why we make use of fuzzy logic is it is easy to understand. It can be built on top of the experience of experts. It is flexible. It is based on natural language. It is tolerant of imprecise data. In direct contrast to neural networks, which take training data and generate opaque, impenetrable models, fuzzy logic lets you rely on the experience of people who already understand your system. It can be built on top of the experience of experts.

We are using Bayesian learning to make the decision on the basis of the probability which is calculated using the Bayesian algorithm for the probability of the occurrences of the intrusions in the network and make the systems more secured with restricting its occurrence again.

A Mahout-based collaborative filtering engine takes users' preferences for items ("tastes") and returns estimated preferences for other items. Mahout provides a rich set of components from which you can construct a customized recommender system from a selection of algorithms. Mahout is designed to be enterprise-ready; it's designed for performance, scalability and flexibility. Mahout is designed for performance. Mahout provides the building blocks from which you can construct the best Recommender for your application.

In this paper section II is dedicated for Literature survey and section III for the Conclusion and section III for the References.

II. LITERATURE SURVEY

[1] Puts forward an idea of stochastic search approaches, including background algorithm. This problem is described by two algorithms. Evolutionary algorithm and Markov Chain Monte Carlo algorithm. By using this approaches to find local maxima. They combine both approaches and produced evolutionary Markov Chain Monte Carlo. Our approach evolves structure and the missing data. They compare our stochastic algorithms and show they all produce accurate results.

[2] Describes a parallel algorithm for the score-based optimal structure search of Bayesian networks. In this approaches they can illustrate optimal solution. This algorithm is used for Dynamic Programming (DP) algorithm which is the fastest algorithm for the optimal structure search method. Recent limitation is a space-time trade-off, our proposed algorithm used direct parallelization of the original DP algorithm. They can find total time and space complexity is O (n σ+12 n). They can also demonstrate optimal structure search for a 32-node network without any constraints, which is the largest network.

[3] Discusses learning Bayesian network structures from data based on score functions that are decomposable. In this properties they can reduce the time and memory costs.

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

This algorithm is based on branch and bound technique. These method illustrate Minimum Description Length, Akaike Information Criterion and Bayesian Dirichlet Criterion. This method produced global optimality solution.

[4] Describes parallel algorithm. They can use directed acyclic graph. In this graph mention n random variables and a set of m observations of each of the n variables, the Bayesian network (BN) structure learning problem. In this approach they used three basic operations. Firstly Mapping to an n-dimensional Hypercube, then Partitioning into k-dimensional Hypercubes and Pipelining Hypercubes. In this paper, they find work-optimal and communication-efficient. They presented a parallel algorithm for exact structure learning that is work-optimal and scalable.

[5] Explains of Bayesian networks. This method find probability distribution along a directed acyclic graph. In this method described graphical d-separation and independence. The two main approaches, search and score versus constraint based. In this article presents some considerations for constraint based algorithms based on recent data analysis, indicating a variety of situations where the faithfulness assumption does not hold.

[6] Introduces LP Relaxation. They can use linear programming relaxation. They can be used for finding the highest scoring Bayesian network structure from data. In this method they can use basic approaches Dual Bound Optimization and Branch and bound. The combinatorial global constraint that the graph structure has to be acyclic. They cast the valid acyclic structures. In this relaxing method, they main bound approximation to the polytope and iteratively tighten it by searching over a new class of valid constraints. Results that method is competitive or faster than alternative exact methods based on dynamic programming.

[7] Expresses Bayesian networks using Constraints. This is used two constraints method like Parameter Constraints and Structural constraints. This paper exact learning of Bayesian network structure from data and expert's knowledge based on score functions that are decomposable. In this properties that strongly reduce the time and memory costs. A branch and bound algorithm is presented that integrates parameter and structural constraints with data in a way to guarantee global optimality.

[8] Narrates Bayesian Network Structure Learning and Deterministic Relations. One of the main hypotheses on which rely most Bayesian network (BN) structure learning algorithms. They can use two phases of deterministic relations. First phase is learning the BN's skeleton and Second phase is orientation and refinement. This paper introduces a new hybrid algorithm, combining a constraint-based approach with a greedy search that includes specific rules dedicated to deterministic nodes that significantly reduce the incorrect learning.

[9] Introduces Bayesian Network Structure based on Dirichlet Prior versus Data. In the Bayesian approach to structure learning of graphical models, the equivalent sample size (ESS) in the Dirichlet prior over the model parameters was recently shown to have an important effect on the maximum-a-posteriori estimate of the Bayesian network structure. They find that the presence of an edge in a Bayesian network is favored over its absence even if both the Dirichlet prior and the data imply independence. They focus on realistic ESS-values, and provide an analytical approximation to the 'optimal' ESS-value in a predictive sense.

[10] Describes Inductive Transfer for Bayesian Network Structure Learning. They present an algorithm for learning Bayes Net structures that takes advantage of the similarity between tasks by biasing learning toward similar structures for each task. Heuristic search is used to find a high scoring set of structures, where the score for a set of structures is computed in a principled way.

[11] Mentions Bayesian Learning approaches, including distributed Host based Intrusion Detection Network (HIDS) collaboration system. This paper is described by two intrusion detection system. Host based intrusion detection system (HIDS) and Network based intrusion detection system (NIDS). This approaches are used to evaluate both the false positive (FP) rate and false negative (FN) rate. By using HIDS system most benefit is collaboration knowledge and accuracy of detection system. Acquaintance selection as the process to find the list of collaborators to minimize false decision cost and maintenance cost.

[12] Describes Real-Time Network Detection System Using Machine Learning. A real time anomaly detection framework based on Apache Storm that uses machine Learning algorithms. In this paper three basic operations like data collection, data pre-processing and invocation of a detection algorithm. Apache's Hadoop distributed File system (HDFS) is reliable, scalable, distributed computing based system. This machine learning technique is supervised learning techniques while simultaneously achieving High-accuracy performance

III. PROPOSED SYSTEM

In this section, we describe our framework of intrusion detection using Bayesian network and apache mahout for the network data with the below mentioned steps as shown in figure 1.

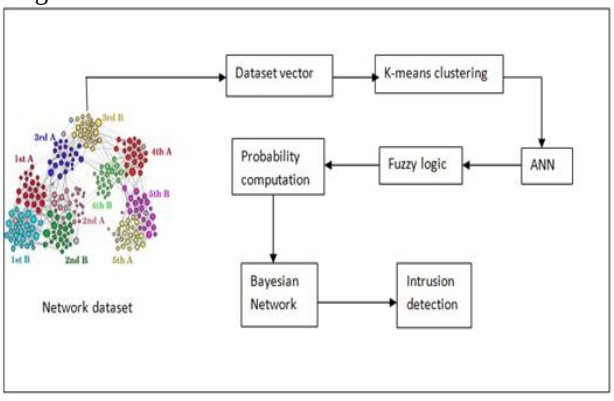

Fig 1: Overview of the proposed work

DOI 10.17148/IJARCCE.2016.5190

IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 1, January 2016

Step 1: Here in this step system uses the real data of the network that was collected over the live router and stored in MS excel format with the entire network data packet attributes. System uses almost 50000 such packet data for the further evaluation.

Step 2: In this step traditional K- means algorithm is used to identify the random cluster centroids. Then based on the distance calculation according to the centroids by using Euclidian distances data is been converged to get the fresh clusters at the end. And these clusters are in the form of abstract format which help further to apply the artificial neural network.

Step 3:-ANN: Artificial intelligence systems are working on the basis of biological neurons in the brain. So biological neurons are come to a conclusion by inter communicating with the many neurons. In the same way in our proposed theory mean and standard deviations are calculated for the attributes of the clusters which are formed using K-Means in the previous step.

$$f(\mu) = \frac{\sum_{i=1}^{\infty}(A_i)}{n} \quad \text{.....1}$$

$$f(\delta) = \sqrt{\frac{\sum_{i=1}^{\infty}(\mu - A_i)}{n}} \quad \text{.....2}$$

Then by using this mean and standard deviation our system identifies the ranges based on the equation 3 and 4.

$$\text{Min range} = f(\mu) - f(\delta) \quad \text{.....3}$$

$$\text{Max range} = f(\mu) + f(\delta) \quad \text{.....4}$$

Then new and hidden neuron clusters will be formed based on the ranges of equation 3 and 4. These new clusters are more crisp and having perfect sharp values in the given ranges. The working pattern of ANN can be depicted in the figure no 2.

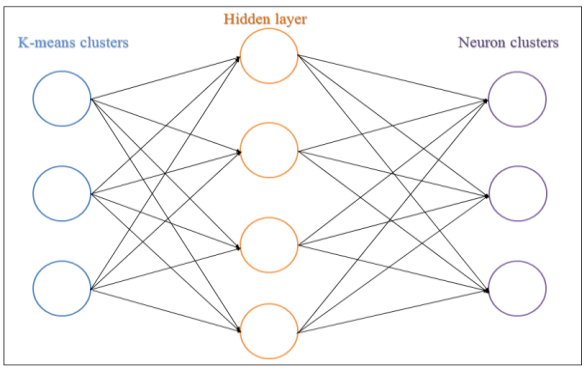

Fig 2. Working Model of ANN

Step 4: Fuzzy Logic is a form of abstract classification theory which works on the basis of crisp values like Very low, low, medium, high and very high. Fine clusters which are formed by the ANN now feed to the fuzzy logic for the classification of the data for the DOS attack. Based on the crisp values and the cluster weights data for different attacks been classified in abstract level.

Step 5: Probability Computation: Here in this step the probability of the attribute for the respective network attack is computed based on the interest ratio of the attribute with respect to the each cluster formed by the ANN and fuzzy as denoted in equation 5.

$$I(r,j) = \frac{P^a(X_{ir}=1 | T_i \in C_j)}{P^a(X_{ir}=1)} \quad \text{.....5}$$

Where
r = Attribute
j= j^{th} Cluster
C_j= Cluster
P^a = Probability factor
T_i = Attack factor

Step 6: Here in this step Bayesian network probability model is used to identify the network attack which is enhanced by the Apache mahout tool. Apache mahout is an effective tool for the collaborative filtering through which system can learn the model for its future performance.

The Bayesian probability can be calculate using the equation 6.

$$\begin{aligned} P(X_1, X_2, \dots, X_n) &= P(X1) \times P(X2|X1) \times \dots \times P(X_n|X_1, \dots, X_{n-1}) \\ &= \prod_i P(X_i|X_1, \dots, X_{i-1}) \quad \text{.....6} \end{aligned}$$

IV. CONCLUSION

This paper puts light on many different methodologies and aspects of Distributed intrusion detection system using Bayesian learning and Apache Mahout. So as a generalisation view of these cumulative study shows no any system or methodology is perfect in providing the solution for Distributed intrusion detection system.

So as an initiative process this paper puts forward an idea of proper Distributed intrusion detection system using Bayesian learning in Apache mahout using key-concepts like K-means clustering, Neuro-Fuzzy, Artificial neural network, Bayesian Learning, Apache Mahout to overcome the flaws like lagging in the detection of the intrusions, fail to detect the intrusion, harm to the system or system failure.

ACKNOWLEDGMENT

It gives us great pleasure in presenting the paper on 'Distributed Intrusion Detection System Using Bayesian Learning and Apache Mahout'. I would like to take this opportunity to thank **Prof. D.S.Zingade** for giving us all the help and guidance we needed. In the end our special thanks to **Prof. S.P.Pimpalkar**, **Prof. Amol Kalugade, Prof. S.N.Zaware** for providing various resources such as laboratory with all needed software platforms, continuous Internet connection, for Our Project. I am really grateful to them for their kind support and their valuable suggestions were very helpful.

REFERENCES

[1] Algorithms James, W. Myers Kathryn, Blackmond Laskey and Tod S. Levitt, "Learning Bayesian Networks from Incomplete Data with Stochastic Search Algorithms Learning Bayesian Networks from Incomplete Data with Stochastic Search", Society for industrial& applied mathematics(SIAM), 2006.

[2] Yoshinori Tamada, Seiya Imoto and Satoru MiyanoParallel "Algorithm for Learning Optimal Bayesian Network Structure", ACM, 2011.
[3] Cassio P. de,Campos,Qiang Ji Olga Nikolova, Jaroslaw Zola, and Srinivas Aluru "Efficient Structure Learning of Bayesian Networks using Constraints", ACM, 2011.
[4] Cassio P. de,Campos,Qiang Ji Olga ,Nikolova, Jaroslaw Zola, and Srinivas Aluru "A Parallel Algorithm for Exact Structure Learning of Bayesian Networks", IOWA State university, 2012.
[5] Timo, J. T. Koski (Stockholm) and John M. Noble (Warsaw), "A Review of Bayesian Networks and Structure Learning", Mathematica Applicanda, 2012.
[6] Tommi Jaakkola, David Sontag, Amir Globerson and Marina Meila "Learning Bayesian Network Structure using LP Relaxations", NYU Computer science, 2010.
[7] Cassio P. de Campos,Zhi Zeng and Qiang Ji,"Structure Learning of Bayesian Networks using Constraints", ACM, 2009.
[8] Ahmed Mabroukd, Christophe Gonzales, Karine Jabet-Chevalier and Eric Chojnaki, "An Efficient Bayesian Network Structure Learning Algorithm in the Presence of Deterministic Relations", DESIR(Paris), 2014.
[9] Harald Steck, "Learning the Bayesian Network Structure: Dirichlet Prior versus Data", Cornell University Library, 2012.
[10] Alexandru Niculescu-Mizil and Rich Caruana, "Inductive Transfer for Bayesian Network Structure Learning", Springer, 2012.
[11] Carol J Fung, Jie Zhang and Raouf Boutaba, "Effective Acquaintance Management based on Bayesian Learning for Distributed Intrusion Detection Networks", IEEE Transactions on network and service management, VOL. 9, NO. 3, September 2012
[12] Shuai Zhao, Mayanka Chandrashekar, Yugyung Lee, Deep Medhi, "Real-Time Network Anomaly Detection System Using Machine Learning", International Conference on the Design of Reliable Communication Networks (DRCN), 2015.
[13] Apache Mahout https://mahout.apache.org/

BIOGRAPHIES

Ronak Agrawal, pursuing B.E Computer Engineering in Savitribai Phule Pune University. He is also interested in Computer Networking, Apache Hadoop and Operating System.

Ganesh Talekar, pursuing B.E Computer Engineering in Savitribai Phule Pune University. He is also interested in Database Management System and Android Based Application.

Akshaysingh Chandel, pursuing B.E Computer Engineering in Savitribai Phule Pune University. He is also interested in Embedded Operating System.

Shriganesh Munde, pursuing B.E Computer Engineering in Savitribai Phule Pune University. He is also interested in Parallel & Distributed Computing and Data Mining.

Deeplakshmi Zingade, completed her M.E from Savitribai Phule Pune University in 2013. She received her B.E from Rashtrasant Tukadoji Maharaj Nagpur University in 2003.She is currently working as Assistant Professor in AISSMS IOIT Pune

DOI 10.17148/IJARCCE.2016.5190