



AN EFFICIENT ALGORITHM FOR CLUSTERING INTRUSION ALERT

ADELINA JOSEPHINE D¹, ANUSHIADEVI R², LAKSHMINARAYANAN T R³

¹ PG Student, Department of Computer Science and Engineering, Sastra University, Thanjavur.

² Asstt. Professor, Department of Information & Communication Technology, Sastra University, Thanjavur.

³ Professor, School of Computing, Sastra University, Thanjavur.

Email: leenajose1986@gmail.com, anushiadevi@it.sastra.edu, trln@cse.sastra.edu

ABSTRACT

Intrusion Detection System is an emerging technology for detecting the unauthorized users and malicious behavior in a system. Alert supervision is tedious in intrusion system, so Meta alerts are created. Meta alerts are generated for appropriate clusters and they form a generalization of alerts. The objective is to identify origin of these alerts. In this paper, we propose a hybrid clustering algorithm which is applied to the data set to cluster the alert. Online alert aggregation is applied to this data which identifies the intruder. Redundant data are filtered during the process of clustering and aggregation, which substantially reduces the false positive rate. From the observed false positive, the origin of the alert are reduced.

Keywords: *Intrusion Detection, Clustering, Meta alert, Root-cause*

1. INTRODUCTION

Protecting our data in the internet is a great threat. Intruders and hackers make an attempt to pilfering our data. Intrusion can be viewed in many traditions. Most of them have their own conventions to find attacks. Intrusion process begins when an unauthorized person endeavors to access data. Rapid growth of threat to data causes evolution of Intrusion Detection System (IDS). Many kinds of Intrusion Detection systems are available. James Anderson was first to recommend that the audit trail should be monitored for threats.

A generic intrusion detection model was first proposed by Dorothy Denning [21]. It is a standard method, which uses a rule-based pattern matching to find the behavior of the jarring. It generates a statistical model and can split the audit logs into system calls and report the information. Generally, there are Physical IDS and Digital IDS. Physical IDS are used for particular area say a building or room whereas Digital IDS are used to produce

alarm for computer network or computers. Intrusion detection usually provides the same function as an alarm system. Misuse and Anomaly are the two common approaches used by IDS.

Misuse Detection follows a pattern to known attacks and therefore a major concern is to create library for the known attacks. Depending upon the known intrusion in library, pattern matching-based intrusion detection works. Pattern matching intrusion detection is capable of finding the common attacks and security problems as patterns are known. Events are monitored and it is more efficient than statistical analysis. Performance depends on the size of the database used. Machine learning is not utilized in pattern matching systems. Pattern databases need to be updated for finding the new type of attacks and complexities to enlarge for new signature types. Exterior events of data are used in misuse-based conditional probabilistic analysis.

Predefined rule set are used to identify the attacks in misuse expert systems. The rule database can be distorted for different operating systems and it



needs to be formulated and tested frequently. Attacks are represented as a sequence of transitions in misuse state transition analysis and it has to satisfy some requirements in order to perform transition. Complex attack types are not identified in state transition analysis. Misuse Keystroke monitoring capture keystroke and analyzes for known attacks. It is application program independent.

Model-based misuse detection system uses “anticipator”, “planner”, “interpreter” to calculate intrusion probabilities and alert is raised if threshold value exceeds. Thus Misuse detection system cannot categorize new type of attacks unless a signature is written. Challenging of resources is another issue in misuse detection system. Anomaly detection follows deviated pattern from expected performance. Probabilistic algorithms are used to analyze the collected data.

Anomaly based intrusion detection causes false positive and false negative, therefore some threshold value is needed to remove outliers. Maintaining profiles is the major concern in statistical anomaly detection. Updating user profile is important for adaptive learning. Feature selection uses subset measure to identify attack which relay on intrusion types. Brute-force and genetic adaptive learning are the methods used to find optimum features.

Predictive Pattern Generation uses rule-based pattern to predict future events. Training and updating the data are straightforward. Besides that it increases number of false positive and false negative. Fine data training is needed for recognizing anomalous patterns in neural networks. They deal with noisy data and large amount of time is required to train data. Bayesian statistical system categorizes unaffiliated data into large number of data classes. It is statistical technique that relay on probabilistic values.

Anomaly based intrusion detection system is expensive. Misuse and anomaly can be pooled together to fabricate a more robust intrusion detection systems. Firewall also provide security but difference is that all traffic will not pass through it and it will not signal to attack inside network. Moreover it has very weak application level protections. IDS can also be classified depending upon the data they inspect. Application based IDS inspect the behavior of an application program say

log files. Host based IDS are designed to protect host where it reside. It checks whether attack is success or failure.

Host based IDS are suited for encrypted and switched environments. They monitor unambiguous activities of the system and do not entail any additional hardware for installation. Network based IDS are designed to monitor network traffic. It is cheaper and easy to fix. Network-IDS respond quickly to real time and unsuccessful attacks.

The inputs to the Multi-network IDS usually come from admin domain [2]. Using Intrusion Detection Method one can easily identify whether the system has been attacked or not. The intention of IDS is to find problem with security policy and to document the thread. IDS need high level of security acquaintance for maintenance.

Upgrading better IDS is always challenging. There is no customary method for assessment of better IDS. Many IDS are expensive and they are reliant on the environment where they reside. The rest of this paper is arranged as follows. In the next section, literature work is discussed with different approaches in intrusion detection. Section 3 describes the methodologies. The rest of this paper is arranged as follows. In the next section, literature work is discussed with different approaches in intrusion detection. Section 3 describes the methodologies that are adopted followed by the experimental results. The last section concludes the paper with the future work.

2. LITERATURE WORK

Alert correlation process detects elevated level of attack without violating security rules in a condensed way. Mainly comprehensive approach to correlation process in introduced in [3] [4]. Alarm correlation process distinguishes the similarities between the alert. Instead of using clustering algorithm, the one of the correlation process uses a categorization of alerts. Aggregations of alerts into predefined clusters are so called alert clustering. No clustering algorithm is used to group the alerts [5] [6]. In [7], the construction of an Alert classifier based on analyst feedback machine learning. It evaluates system performance and avoids redundancy of alerts.

Only few percent of false positive error is reduced and numeric values should be adjusted depend on

input data. Attribute wise operator are used to fuse the alert and also require a large number of parameters [8] [9]. Three different approaches are also used to fuse the alerts. First approach grouping is based on IP address and supplementary approaches are based on supervised learning technique [10]. In TACC the alarm are pipelined in order. Alerts are grouped into meta-alert in PAC. Alarm clustering is a NP complete problem. Root cause is responsible for fabrication of alerts.

Attribute –oriented induction algorithms (AOI) are used to find large number of clusters [11]. CURE algorithm paved the way for Offline algorithm, their domino effects are limited to numeric values and number of clusters involved should be set manually [12]. Alert clustering is focused to reduce false positive alarm. Data mining approaches for Intrusion Detection has been growing in recent years. Self Organizing Structure (SOM) sculpts the normal behavior. Rule based Decision Support System (DSS) was designed to interpret results for both anomaly and misuse detection [15]. Two data mining approaches are recommended in [16] which include Artificial Neural Network (ANN) and Support Vector Machine (SVM). SVM with tfidf had better performance rather than ANN with simple frequency based scheme.

Binary classification problem was formulated by Zhang and Shen [17] using SVM and various text processing techniques are also employed for intrusion detection. SVM for intrusion detection has proved to have good training speed and scalability. A different approach for clustering is represented in [18] debits about autoassociator neural network (AA-NN) which discriminate the diverse type of alerts. Alerts with same reconstruction error are assigned to same cluster. AA-NN needs an offline training segment and training data set. To cluster the alert, reconstruction error has to be manually adjusted.

Bayesian Network can also be used for intrusion detection system. Bayer's rule provides a mean of combining anomaly detection and pattern recognition. They are attack specific and build a decision network on individual attack. The size of network increases as the attack type increases. Bayesian method uses probabilities to find fraud investigators. Decision tree induction is classification algorithm used for development of IDS. It has high detection and high operational speed [19] [20]. K-means and fuzzy c-means are

clustering methods used for intrusion detection. The main difficulty is that symbolic values are not clustered which produces inaccuracy results.

3. METHODOLOGY

3.1 OVERALL ARCHITECTURE

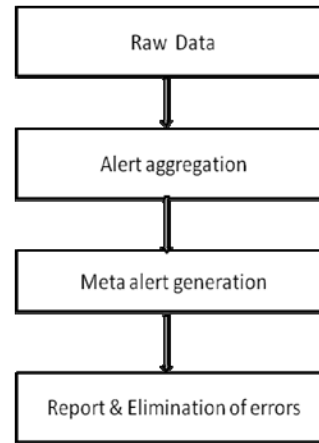


Fig1. Overall architecture of IDS

Fig 1 shows the overall architecture of the proposed system. The raw data are confined to valuable information and the alert aggregation combines the alert belonging to the specific attack instances and thus the Meta alert is generated. False positive and true positive errors are identified and origin of alert are identified and removed.

3.2 META ALERT

Related alerts are grouped together as alert correlation process to form Meta alert. Meta alert is a sub set of related alert that represent the single alert. Thus Meta alert will produce an abstraction of the alert reports, which includes the details of all alerts. It is similar to an alert that refer to all possible alerts which are merge to produce a Meta alert. The intention of creating Meta alert is that it summarizes all alert into a single alert instance. In short, Meta alert represent an abstract view of alert that are produced.

Intrusion Detection Message Exchange Format is standard format to encode the intrusion alerts. Raw alerts are transformed into IDMEF format in order to have standardized attribute value to alerts. IDMEF contains information like node, user, network service, source and destination IP address.

Meta alert can also be merged with relevant other alerts to form a hierarchical formation. Root node represent Meta alert and leaf node represent reference of merged alert. Whenever a new alert is raised, it compare alert with current Meta alert and it is merged with an appropriate group.

Fig 2 .shows IDMEF alert .Irrelevant alert are considered as another Meta alert until similar alerts get fused together. The process continues till all alerts get fused or merged to form a meta-alert. Inappropriate alerts are maintained separately and are ignored.

```
<IDMEF –Message Version="0.1">
<Alert ident="15000"impact="unknown">
<Create Time stamp tps=" 0x0, 0x1C121F15 ">
<Source><Node>
<address> 172.16.19.130 </address>
</Node></Source>
<Target><Node>
<address> 172.16.25.99 </address>
</Node></Target>
</Alert>
</IDMEF>
```

Fig 2. IDMEF Alert

3.3 ELFL ALGORITHM

```
Input : log data of n user {D1,D2,D3..Dn}
Output: K number of clusters
Method:
(1)Initialize leader list with D1
(2)Initialize cluster count k to 1 and i to 2
(3) Repeat
(4) Compare Di with Di-1.....D1.
If pattern matches
append Di to leader list
otherwise
list(k+1) ← { Di}
increment cluster count,k
(5)increment i
(6)Until Dn
(7)Return k
```

Fig3. Leader algorithm

Extensive Leader farthest Linkage algorithm (ELFL) has both advantages of incremental and hierarchical techniques. It is computationally effective. From a data set, single data object is taken as a leader and starts comparing with others. The distance between one cluster and other is equal to greater distance from any member of one cluster to any member of other cluster. Fig 3 shows leader algorithm.

The process repeats until all data set is processed. Processing of data is comparatively quicker and no prior knowledge is needed to know about clusters. Processing time and results will be based on data set used.

3.4 OFFLINE ALERT AGGREGATION

The data stream approach for an offline alert aggregation can be extended to online aggregation. A hand- some of alerts are produced by attack instances. Alert Aggregation is processed by, estimating allotments of alert instances and analyzing the cluster structure. True alerts are wrongly clustered, false alert are not recognized are some of the problematic situation. Offline alert, aggregates only historical data or logs in which intrusion can be detected but preventive measures cannot be taken. Moreover the intrusion detected will not be in real-time. Losses of data, alteration of data are possible. In order to avoid this online alert was proposed.

3.5 ONLINE ALERT

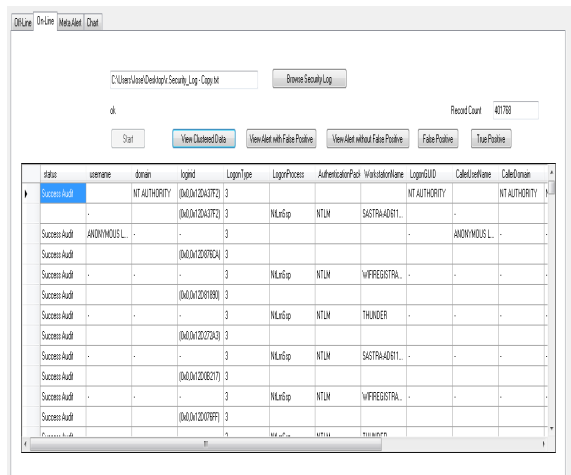
Online aggregation alert is trivial change, over the offline alert. The data collected are sent in a streaming manner. i.e. alerts are clustered systematically and grouped continuously. Every time an irrelevant alert occurs, it intimates the alert. Thereby the corresponding actions can be taken which will obviously find the root cause of the alert and it is eliminated. The process repeat until the entire alerts are grouped and intrusion are found. Since the alerts are repeatedly clustered, the possible occurrence of false positive is reduced substantially. With the occurrence of false positive error the cause of the error can be found out and it can be eliminated.

4. EXPERIMENTAL RESULT

We estimated the performance of alert aggregation technique. It has three phases. First

phase deals with alerts that are created. Repetitions of alerts are avoided by clustering the alerts in second phase. The third phase deals with finding the original cause for the alert and reduce them. It monitors for malicious activities and reports it. When a new alert arrives it starts comparing the alert for similarity measures and starts merging it. When no alert is found matching it is set in queue for future alerts. Merging of alert is fashioned by creating Meta alert. Several alerts against multiple targets from same source are managed by merging alerts.

The ELFL algorithm is tested with log data which contains the login information of a university. The log files contain visiting records of the user within a day. After preprocessing we check for their status information and it is taken as a one data set to cluster rather than multiple partitioning. ELFL algorithm produces the cluster by applying leader algorithm and final cluster is produced by farthest linkage algorithm.



#	status	username	domain	login	login type	login process	authentication protocol	workstation name	login GUID	cabinet name	cabinet name
Success Audit	-	NT AUTHORITY\	NT AUTHORITY	(0A01N120A3F2)	3				NT AUTHORITY		NT AUTHORITY
Success Audit	-	ANONIMOUS.L		(0A01N120A3F2)	3	MLatip	NTLM	BASTRA-0851			ANONIMOUS.L
Success Audit	-			(0A01N120A3F2)	3						
Success Audit	-			(0A01N120A3F2)	3	MLatip	NTLM	WFFREGSFR			
Success Audit	-			(0A01N120A3F2)	3	MLatip	NTLM	THUNDER			
Success Audit	-			(0A01N120A3F2)	3						
Success Audit	-			(0A01N120A3F2)	3	MLatip	NTLM	BASTRA-0851			
Success Audit	-			(0A01N120A3F2)	3						
Success Audit	-			(0A01N120A3F2)	3	MLatip	NTLM	WFFREGSFR			
Success Audit	-			(0A01N120A3F2)	3						

Fig 4. Online alert aggregation

Online alert aggregation is shown in fig 4. Depending upon their status, the corresponding information about the user is clustered. Clustering value will change according to the data set being used. The main advantage is that it starts clustering as the individual data set until the data set ends rather than clustering whole. Fig5 shows the valuation result of root cause analysis. The data set are normalized and they are clustered and true positive, false positive errors are identified. From the experimental result it is shown that the root cause are reduced. Data stream alert aggregation of

offline alert deals only with the outdated log files in which security to the data is not guaranteed.

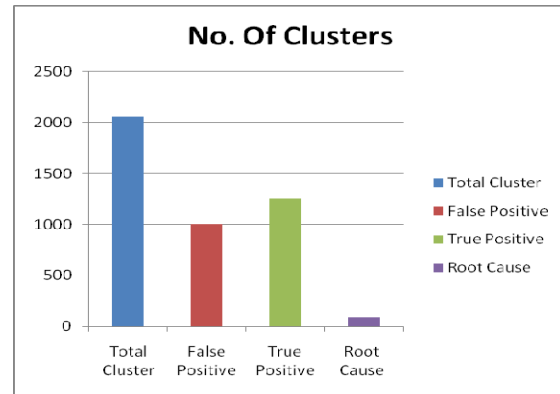


Fig 5 Evaluation of root cause

Offline alert shows only the result of archaic information. Upgraded version of offline alert is online alert. Each stream of data is check now and then to regulate the flow of information without intruder. If any abnormal flow of information persists they are termed as intruders. The user information, login time are compared .The origin of the intruders is identified and they are discarded by which false positive error is also reduced.

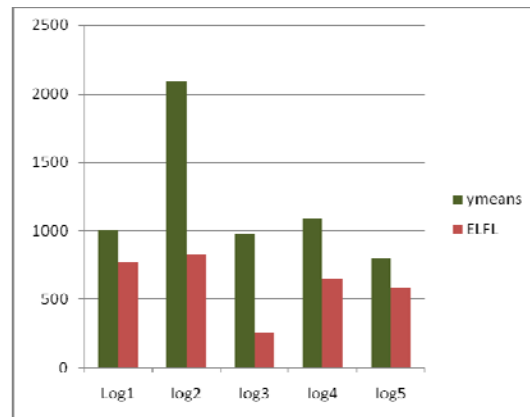


Fig 6 .Performace analysis graph

As many clustering algorithm are widely used in various application and considered as more versatile than many other algorithm so we decided to choose y-means algorithm to compare with the implemented work. The initial clusters are compared as whole in y-means whereas the implemented algorithm compares with the data object chosen. Fig 6 shows the performance evaluation of the algorithm.



Sample log data are taken and made to run in both algorithms. An experimental result shows that the proposed ELFL algorithm shows better results compared to y-means.

5. CONCLUSION AND FUTURE WORK

Clustering of alerts is done using ELFL algorithm which combines the features of incremental and hierarchical clustering .It speed up the cluster thereby improving the performance. The data stream aggregated approach reduces the occurrences of false positive rate by continuously monitoring the incoming alerts. From the observed false positive error, the origin of the alert are detected and eliminated.

In future work is planned to develop a technique which interacts with IDS tools to make the correlation process effortless. In addition, IDS is designed to inspect the intruder based on their attack types and restrict them to perform further actions.

REFERENCES

[1] Alexander Hofmann, Bernhard Sick "Online Intrusion alert aggregation with generative Data Modeling" IEEE transactions on dependable and secure computing, VOL.8, No.2 March- April 2011.

[2] Julia Allen, Alan Christie, William Fithen , John McHugh, Jed Pickel, Ed Stoner "State of Practice of Intrusion Detection Technologies" Technical report CMU/SEI-99-TR-028 ESC- 99- 028

[3] Norazah Aziz (GCIA), "Intrusion Alert correlation" Cyberspace Security lab, MIMOS Berhad.

[4] F.Valeur, G.Vigna, C.Krugel, and R.A Kemmerer, "A Comprehensive Approach to Intrusion Detection Alert Correlation"IEEE Trans Dependable and Secure Computing, Vol.1, No.3, pp.146-169, July-Sept 2004.

[5] Deli, Z.Li, and J.Ma,"Processing Intrusion Detection Alerts in Large-Scale Network",

Proc.Intl symp.Electronic Commerce and Security.pp.545-548, 2008.

[6] F.Cuppens,"Managing Alerts in a Multi-Intrusion Detection Environment," Proc.17th Ann.Computer Security Applications conf. (ACSAC '01),pp.22-31,2001

[7] Tadeusz Pietraszek," Using Adaptive Alert Classification to Reduce False Positives in Intrusion Detection," IBM Zurich Research Laboratory, S: aumerstrasse 4, CH-8803 R: uschlikon, Switzerland.

[8] T.Pietraszek," Alert Classification to Reduce False Positives in Intrusion Detection,"PhD Dissertation, Universitat Freiburg, 2006.

[9] F.Autrel and F.Cuppens,"Using an Intrusion Detection Alert Similarity Operator to Aggregate and Fuse Alerts,"Proc.Fourth Conf.Security and Network Architecture, pp.312-322, 2005.

[10] O.Dain and R.Cunningham,"Fusing a Heterogeneous Alert stream into Scenarios," Proc.2001 ACM Workshop Data Mining for Security Application, pp.1-13, 2001.

[11] K.Julisch," Using Root Cause Analysis to Handle Intrusion Detection Alarms," PhD Dissertation, Universitat Dortmund, 2003.

[12] M.S.Shin, H.Moon, K.H.Ryu, K.Kim and J.Kim, "Applying Data Mining Techniques to Analyze Alert Data". Web Technologies and Applications, X.Zhou, Y.Zhang, and M.E Orłowska eds.pp.193-200, Springer, 2003.

[13] Melanie Rose Rieback, dr.M.E.M.Spruit, ir. R.Prins, "The Meta-Alert Correlation Engine", Msc Thesis Report, TUDelft, July 2003.

[14] G.Giacinto, R.Perdisci, and F.Roli, "Alarm Clustering for Intrusion Detection Systems in Computer Networks," Machine Learning and Data Mining in Pattern



- Recognition, P.Perner and A.Imiva, eds.pp.184-193, Springer, 2005.
- [15] O.Depren, M.Topallar, E.Anarim and M.K.Ciliz, "An Intelligent Intrusion Detection System (IDS) for anomaly and misuse Detection in computer networks", Expert Systems with Applications, vol.29, issue 4, pp.713-722, 2005.
- [16] W.H.Chen, S.H.Hsu, and H.P. Shen,"Application of SVM and ANN for Intrusion Detection", Computers & Operations Research, vol. 32, issue 10, pp. 2617-2634, 2005.
- [17] Z.Zhang, and H.Shen,"Application of Online-training SVMs for real-time intrusion detection with different considerations", Computer Communications, vol.28, issue12, pp.1428-1442, 2005.
- [18] R.Smith, N.Japkowicz, M.Dondo, and P.Mason,"Using Unsupervised Learning for Network Alert Correlation," Advances in ArtificialIntelligence, R.Goebel, J.Siekmann, and W.Wahlster, eds. pp. 308-319, Springer, 2008.
- [19] L.Brieman, J. Friedman, R.Olshen, and C.Stone,"Classification of regression trees", Wadsworth Inc, 1984.
- [20] Sapna S.Kaushik and P.R.Deshmukh," Comparison of approaches to implement Intrusion detection system", International Journal of Computer Science and Communication vol.2, January-June 2011, pp. 45-48
- [21] Dorothy Denning," Intrusion Detection Model", IEEE Transactions on software Engineering, Vol. SE-13, NO. 2, February 1987, 222-232.