IMPROVED K-MEANS CLUSTERING ALGORITHM TO DETECT NETWORK INTRUSIONS

K.S. Anil Kumar

ABSTRACT

Intrusion Detection Systems are integral part of system's security and are one of the fastest technologies within the security space. Various approaches to Intrusion Detection are currently being used, but they are relatively ineffective. The major problem in intrusion detection system research is the speed of detection - good measure of performance since they measure what percentage of intrusions the system is able to detect. The basic K-Means clustering algorithm is inefficient on high data sets due to its unbounded convergence of cluster centroids. So for removing this problem we have adopted an improved optimum cluster initialization algorithm to obtain effective and efficient crisp clusters in Intrusion Detection System. The technique is tested using multitude of background knowledge sets in DARPA network traffic datasets.

Keywords : *Intrusion Detection System (IDS), K-Means Clustering, DARPA dataset*

I. INTRODUCTION

Computer networks are one of those unique gifts of modern science and the rapid proliferation of computer networks has changed the prospect of network security and as the network advanced, intrusions and misuses followed. Now-a-days intrusion detection systems have become a standard component in security infrastructures.

Intrusion as generally described is an act of trespassing or infringing the integrity, confidentiality or preventing the availability of a resource [1]. Intrusion Detection Systems detects unauthorized or malicious attacks over a computer system that occurs primarily through network. These attacks can compromise the security and trust of a system. Intrusions refer to the network attacks against vulnerable services, data-driven attacks on applications, host-based attacks like privilege escalation, unauthorized logins and access to sensitive files.

Researchers have developed Intrusion Detection Systems (IDS) capable of detecting attacks in several available environments. Categorized broadly based on their patterns of detection, IDSs can be classified as misuse detectors or anomaly detectors. Misuse detectors rely on comprehending the patterns of known attacks [2, 3], while anomaly detection exploits user profiles as the basis of detection, and brands the characteristics of the deviant from the normal ones as intrusion [2, 3, 4, 5].

We have used K-Means algorithm to cluster the data into normal and intrusion packets. The basic K-Means clustering algorithm is inefficient on high data sets due to its unbounded convergence of cluster centroids. So choosing the proper initial centroids is the key step of the basic K-Means clustering algorithm. To make-out the issues, we have developed an improved K-Means

Associate Professor, Department of Computer Science
Sree Ayyappa College, Eramallikkara,
Chengannur, Kerala (India)
E-mail : ksanilksitm@gmail.com

clustering algorithm to obtain effective and efficient crisp clusters in intrusion detection inference system [6]. The study has been based on the selected fields of DARPA dataset.

II. Rational of The Study

This section gives a brief introduction of the techniques used in the proposed work.

A. K-Means Clustering Algorithm

Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. A clustering algorithm attempts to find natural groups of components (or data) based on similarities. The clustering algorithm also finds the centroid of a group of data sets. The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters[7].

K-Means clustering is a technique that classifies objects in to K number of groups based on their attributes. Obviously K is a positive number. The cluster centroid is calculated first. Then the grouping is done by minimizing the sum of squares of distances between the data and the corresponding cluster centroid. The object of K means clustering is to classify the data by analyzing the traits and then organizing them in accordance to their attributes. The reason why K-Means clustering has been chosen for the algorithm is that the datasets we analyze need to be categorized into just two clusters, normal and intrusion. And hence the value of K can be simply defined as '2' [8,9].

K.M. Faraoun and A. Boukelif [10] have proposed that K-means is one of the simplest unsupervised learning algorithms. The unsupervised intrusion detection is more appropriate for anomaly detection than classification based intrusion detection methods. This is due to the dynamic nature of network traffic intrusions. Mrutyunjaya et.al [11] have also pointed out that K-means clustering is a well known Data Mining algorithm that has been used in an attempt to detect anomalous user behavior, as well as unusual behavior in network traffic.

B. Improved Optimum Cluster Initialization Algorithm

Witcha Chimphlee,et.al [12] have reported two problems that are inherent to K-means clustering algorithms. The first is determining the initial partition and the second is determining the optimal number of clusters. Zhang Chen et.al [13] has proposed a new concept for selecting the number of clusters and their concept is not suited for large volume of data and inefficient in cluster initialization also.

M. B. Al-Daoud [14] proposed an algorithm for cluster initialization; this algorithm is based on finding a set of medians extracted from an attribute with maximum variance. But in this algorithm the partitioning of dataset into K clusters is not sound. The optimum cluster initialization only calculates the maximum of variance and divides the cluster based on the maximum of variance. This algorithm does not suite for dividing the data objects of maximum variance into desired subsets. So we have modified the model for choosing the appropriate initial centroids. This improved algorithm is based on finding the K centroids from a set of means extracted from the variance. The pseudo code for the improved optimum cluster initialization algorithm is given in Figure : 1.

Improved optimum cluster initialization algorithm

Input: DARPA Dataset with n data object with d dimensions

Output: Set of 2 initial centroids

Procedure

Step 1: compute the variance of each attribute
Step 2: Sort it out the variance in any order
Step 3: Find the mean from the set of variance
Step 4: Find the difference of distance from the mean to its immediate neighbor variance
Step 5: Divide the data objects of variance into 2 subsets based on the mean
Step 6: The mean is included in any one of the 2 subsets which have minimum distance to immediate neighbor variance.
Step 7: Find the mean of each subsets.
Step 8: Use the corresponding data objects of each mean as initial cluster centroids

Figure 1 : The pseudo code for the improved optimum cluster initialization algorithm

Figure : 2 shows the functioning of improved K-Means clustering Algorithm. This algorithm has been used to train the datasets, which may contain normal and anomalous traffic without labeling them as such in advance. At the end of the K-means training, the K cluster centroids are generated and the algorithm is ready for classifying traffic. The fields in the DARPA datasets are scrutinized and categorized as intrusion and normal datasets by applying K-means clustering technique.

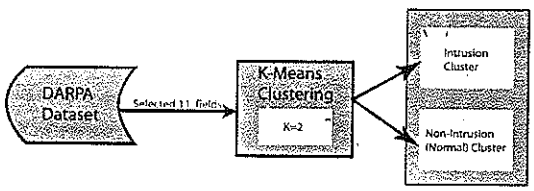

Figure 2: Bifurcation of data using K-Means clustering

This algorithm is efficient on large data sets due to the bounded convergence of the cluster centroids. The improved optimum cluster initialization technique used in K-Means algorithm results crisp clusters in Intrusion Detection System.

III. Methodology

K-means clustering is a technique that classifies objects in to K number of groups based on their attributes or features. This algorithm takes 2.703 seconds for completing the clustering operation. The cluster centroid is calculated first. Then the grouping is done by minimizing the sum of squares of distances between the data and the corresponding cluster centroid.

The number of connection records used in the training dataset is 5000 records from the DARPA using simple random sampling method. Each record is described by 11 attributes and a labeled attribute which specifies the status of connection records as either normal or intrusions. The

labeled attribute is considered as 12th attribute in the training data set. Since the labeled attribute specifies the status of 5000 records as either normal or intrusions, it is separated into two clusters- cluster 1 and cluster2. The training dataset is clustered into Intrusion and Non-Intrusion clusters by applying K-Means technique and it is verified with actual data.

Table 1: Clustering result statistics

Labeled data (DARPA) /K-Means Clustering	Cluster-1 (intrusion data)	Cluster-2 (non intrusion data)
	No. of records	No. of records
Actual data (DARPA)	1668	3332
K-Means Clustering	1633	3367

Table 1 shows the difference between actual data and K-Means clustered data. The actual data clearly identified 1668 Intrusion data in cluster-1 and 3332 Non-Intrusion data in cluster-2. But the K-Means clustering algorithm clustered the dataset into two; 1633 in cluster-1 and 3367 in cluster-2, which is very close to the actual data classification.

IV. Experimental Results

A. K Means Performance

The efficiency of the K-Means clustering algorithm is analyzed by using the performance indicators such as True Positive Rate, True Negative Rate, False Positive Rate, False Negative Rate, Detection Rate and Overall accuracy.

Once these partitions are performed, every attribute in the normal and deviant cluster is probed to understand its characteristics. Analyzing the intrinsic nature of these intrusion data can give a clear picture of the factors that signify the abnormality. This knowledge can further help to distinguish the regular from the irregular one. A similar process is carried out in parallel to interpret the patterns of normality from the regular packets.

Table 2 revealed that the observed result of K-Means Clustering follows the expected result. This algorithm has achieved a True Positive Rate of 97.90, a True Negative Rate of 98.92, a False Positive Rate of 1.08, a False Negative Rate of 2.10, a detection Rate of 99.00 and an Overall accuracy of 98.58.

Table 2 : K-Means Result

Performance Indicators	Result
True Positive Rate	97.90
True Negative Rate	98.92
False Positive Rate	1.08
False Negative Rate	2.10
Detection Rate	99.00
Overall Accuracy	98.58

B. Measure of K-Means Cluster Asymmetry

The characteristics of intrusion cluster and non intrusion cluster has been statistically analyzed with the help of minimum value, maximum value, mean, mode, standard deviation skewness and kurtosis. It is noted that the figures exhibits a deviation in all attributes of both the intrusion cluster and non-intrusion cluster. Table 3 and Table 4 revealed the asymmetry spectra of Intrusion cluster and Non-Intrusion Cluster in K-Means clustering. The patterns thus obtained from the analysis makes explicit that the nature of intrusion and non-intrusion data are not same.

Table 3 : Characteristics data of Intrusion Cluster

Cluster-1 (Intrusion)											
Fields	1	2	3	4	5	6	7	8	9	10	11
mean	2.003	0.499	50.020	48.683	47.575	0.519	4.476	505.965	495.919	517.231	501.927
mod	2.000	0.000	31.000	11.000	54.000	1.000	3.000	674.000	339.000	775.000	595.000
standard deviation	0.812	0.500	28.888	28.927	29.925	0.500	2.852	294.951	287.490	285.493	286.725
Skewness	-0.006	0.004	-0.013	0.032	-0.019	-0.075	0.035	0.000	0.042	-0.038	0.039
Kurtosis	-0.999	-1.013	-0.975	4.438	0.211	-1.047	-1.307	-1.307	-1.290	-1.213	-0.592

Table 4 : Characteristics data of Non Intrusion Cluster

Cluster-2 (Normal)											
Fields	1	2	3	4	5	6	7	8	9	10	11
mean	2.015	0.509	49.229	49.701	47.808	0.494	4.564	513.731	511.590	509.543	505.008
mod	2.000	1.000	23.000	78.000	38.000	0.000	9.000	934.000	881.000	47.000	865.000
standard deviation	0.811	0.500	28.577	28.684	29.870	0.500	2.893	286.802	286.650	288.718	293.584
Skewness	-0.027	-0.036	0.031	-0.005	-0.011	0.025	-0.015	-0.026	-0.021	-0.004	0.012
Kurtosis	-1.48	-2	-1.171	-1.175	-1.289	-2	-1.23	-1.21459	-1.19986	-1.21984	-1.23252

V. Conclusion

The solution crafted with an object to create a powerful intrusion detection approach proved worthwhile. Convergence of Improved optimum cluster initialization algorithm technique with K-Means Clustering Algorithm has helped to achieve a robust architecture. Comprehensive analysis of the characteristics of the abnormal and even the normal packets helped recognition of their patterns and discrimination efficiently.

References

[1] Heady R., Luger G., Maccabe A., and Servilla M. 1990. The architecture of a Network level intrusion detection system, Technical Report, CS90-20, University of New Mexico, Albuquerque, NM 87131.

[2] Denning D. (1987) *"An Intrusion-Detection Model,"* IEEE Transactions on Software Engineering, Vol. SE-13, No. 2, pp.222-232

.[3] Kumar S., Spafford E. H. (1994) *"An Application of Pattern Matching in Intrusion Detection,"* Technical Report CSD-TR-94-013. Purdue University.

[4] Ryan J., Lin M-J., Miikkulainen R. (1998) *"Intrusion Detection with Neural Networks,"* Advances in Neural Information Processing Systems, Vol. 10, Cambridge, MA: MIT Press.

[5] Terran lane, Carla E. Brodley, Temporal Sequence Learning and Data Reduction for anomaly Detection, Vol. 2, No. 3, August 1999, pp. 295- 331.

[6] Masayuki Murakami, Nakaji Honda. A study on the modeling ability of the IDS method: A soft computing technique using pattern-based information processing, International Journal of Approximate Reasoning, Volume 45 , Issue 3 August 2007, Pages 470-487.

[7] Marimuthu,A.Shanmugam.A, Intelligent progression for anomaly intrusion detection, 6th International Symposium, PP: 261-265, Jan 2008.

[8] W.Lee, S.J. Stolofo, *"Data mining approaches for intrusion detection"* in proc. Of 7th USENIX Security Symposium, San Antonio, TX

[9] K.S. Anil Kumar and Dr. V. NandaMohan, *"Novel Anomaly Intrusion Detection Using Neuro-Fuzzy Inference System"*, International Journal of Computer Science and Network Security, vol.8, no.8, pp.6-11 , August 2008.

[10] K. M. Faraoun and A.Boukelif, Neural Networks Learning Improvement using the K-Means Clustering Algorithm to Detect Network Intrusions, International Journal of Computational Intelligence, 2007,pp161-168

[11] Mrutyunjaya Panda & Manas Ranjan Patra, SOME CLUSTERINGALGORITHMSTO ENHANCETHE PERFORMANCE OF THE NETWORK INTRUSION DETECTION SYSTEM, Journal of Theoretical and Applied Information Technology, 2005 - 2008 ,pp710-716

[12] Witcha Chimphlee,et.al. *"Un-supervised clustering methods for identifying Rare Events in Anomaly detection"*, in Proc. Of World Academy of Science, Engg. and Tech (PWASET), Vol.8, Oct2005, pp.253-258.

[13] Zhang Chen, Xia Shixiong, *" K-means Clustering Algorithm with improved Initial Center"*, Second International Workshop on Knowledge Discovery and Data Mining, 2009 IEEE,pp790-793

[14] Moth'd Belal Al-Daoud: A New Algorithm for Cluster Initialization. World Academy of Science, Engineering and Technology 4, pp 74-76, 2005

Author Biography

Dr. K S Anil Kumar, received the B.E. degree in Electrical and Electronics Engineering from the Mangalore University, Karnataka, India, in 1996. He received the M Tech Degree from University of Kerala and the Ph.D. degree in Technology Management (Information Security Management) from the University of Kerala, Thiruvananthapuram, India, in 2003 and 2012, respectively.

In 1996, he joined the Dewpartment of Computer Science, Sree Ayyappa College, affiliated to University of Kerala, as a Lecturer, and in 2011 became an Associate Professor. He had also served as Head, e-Governance, Kerala State IT Mission, Government of Kerala during the period 2007 to 2009. His current research interests include Information Security, e-Governance, Security Management and Technology Management. Dr. Kumar is a member of the

Institution of Electronics and Telecommunication Engineers (IETE). He is also a Member of the Academic Council, University of Kerala.

He was the General Convener of the International Conference on Free Software hosted by the IT Department, Government of Kerala (ICFS2008) held in Thiruvananthapuram. He was also the Chairman of the Academic Committee of the Centre for Development of Imaging Technology(CDIT), Kerala. He is one of the Visiting Faculty member of Indian Institute of Information Technology and Management -Kerala (IIITM-K).