A Novel Multi-Threaded K-Means Clustering Approach for Intrusion Detection

Vidit Pathak,
Information Technology Dept.
NITK Surathkal, Karnataka, India
viditpathak@gmail.com

Dr. AnanthanarayanaV. S.
Information TechnologyDept.
NITK Surathkal, Karnataka, India
anvs@nitk.ac.in

Abstract— **Due to the proliferation of high-speed internet access, more and more organizations are becoming vulnerable to potential cyber-attacks. An intrusion is defined as any set of actions that compromise the integrity, confidentiality or availability of a resource. Intrusion Detection System (IDS), as the main security defending technique, is widely used against malicious attacks. IDS system should be good enough to detect existing attacks as well as novel attacks at high speed. Thus to fulfil these requirements a new novel Multi-Threaded K-Means clustering approach has been used which has resulted in high detection rate and low false alarm rate. A subset of KDD99 Data set has been used as an input dataset for experiments.**

Keywords- Intrusion Detection System (IDS), Data Mining, K-Means algorithm, KDD99 Data Set

I. Introduction

Intrusion detection as defined by the SysAdmin, Audit, Networking, and Security (SANS) Institute; is the art of detecting inappropriate, inaccurate, or anomalous activity [1]. Today, intrusion detection is one of the high priority and challenging tasks due to the high and rapid growth in network. Intrusion Detection System (IDS) is a component of the information security framework. Its main goal is to differentiate between normal activities of the system and suspicious or intrusive behaviour. Traditional instance-based or rule based IDS can only be used to detect known intrusions, since these methods classify instances based on what they have learnt from labelled data. Thus we need a technique for detecting known intrusions as well as new and un-known types of intrusions. A method that offers promise in these tasks is anomaly detection. Anomaly detection detects anomalies in the data (i.e. data instances in the data that deviate from normal or regular ones). It also allows us to detect new types of intrusions, because these new types will be deviations from the normal network usage. However, an accurate system that cannot handle large amount of network traffic and is slow in decision making will not fulfill the purpose of an IDS. Thus a faster algorithm is needed to handle a high data coming into network at constant rate.

Most current approaches for detecting intrusions utilize some mathematical and intelligent methods and tools, including decision tree system [2], artificial neural network [3], SVM [4], genetic algorithm [5] and so on. Recently, there has been an increased interest in data mining based approaches to build intrusion detection models [6-9]. Even approaches based on K-Means are also discussed in [10-12]

In this paper Clustering technique is used for anomaly detection. Clustering is the process of dividing the object in the groups such that objects within a group be similar to one another and different from the objects in the other group [13]. In this paper a novel Multi- Threaded K-Means approach has been proposed. The result obtained by new approach has been compared with K-Means algorithm. Proposed approach is able to detect novel attacks also. Sometimes two attacks are very closely related that it is almost not possible to differentiate that packet is of which type. At this time proposed approach flags it as both possible attack types where other approaches classifies it into any one type resulting in wrong decision sometimes.

II. KDD99 Dataset

The KDD Cup 1999 Intrusion detection contest data [14] was prepared by DARPA Intrusion detection evaluation program by MIT Lincoln Laboratory. Lincoln Labs set up an environment to acquire nine weeks of raw TCP dump data for a local-area network (LAN) simulating a typical U.S. Air Force LAN. They operated the LAN as if it were a true Air Force environment, but peppered it with multiple attacks.The raw data was processed into connection records. Most of the researchers use this KDD99 data set as input to their approaches. There are main 4 attacks in KDD99 dataset.

1) Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine [15]. E.g. Ping of Death, Smurf etc.

2) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine [15]. E.g. Multihop, Phf etc.

3) User to Root Attack (U2R): is an attack in which attacker starts out with access to a normal user account on the system and is able to exploit some vulnerability to gain root access in system [15]. E.g. Perl, Rootkit etc.

4) Probe Attack: is an attempt to gain access to a computer and its files through a known or probable weak point in the computer system [16]. E.g. Portsweep, Nmap etc.

978-1-4673-2008-5/12/$31.00 ©2012 IEEE

III. Methodology

When clustering is done on a data set having each data point of N-dimensional; to cluster the data into some K clusters same attributes of each data point are considered. This scenario is shown in Fig. 1. It shows the general scenario where all N attributes of a data point are considered to cluster it into Type-1, Type-2, Type-3, ..., Type-K. That is all the attributes are used to find the distance between data point and a cluster.

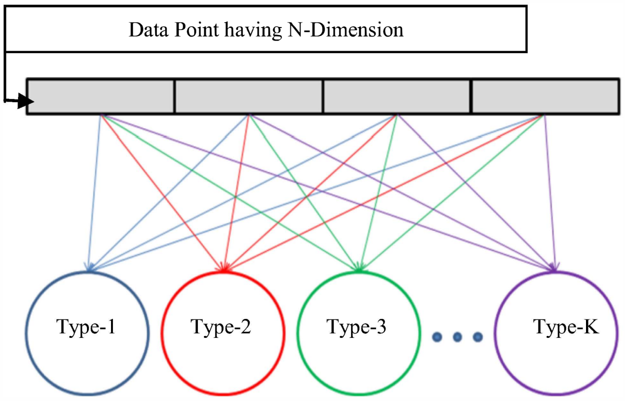

Figure 1. General Clustering Scenario

But in some particular applications you need only p attributes to cluster data into one cluster and some q attribute to cluster into another. This scenario is shown below in Fig. 2. The Fig. 2 shows that to check that data point belongs to Type-1 or not we need to consider only those attributes which are labeled as X. Same way to check whether it is of Type-2 or not we need to consider attributes with Y labeled and labeled with Z attributes are considered to check whether it is of Type-3 or not.

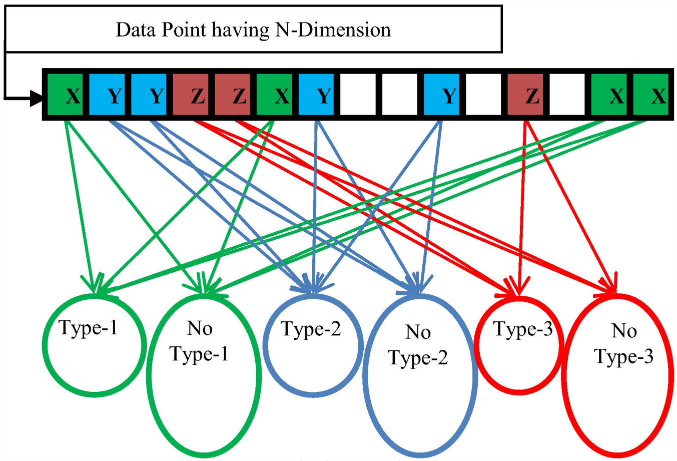

Figure 2. Typical Clustering Scenario

The same concept has been used here to detect different types of attacks. There is no need to consider all 41 attributes present in KDD99 data set to cluster them into different types of attack. As mentioned in [17] to cluster the data only some attributes out of total 41 attributes are needed. Table I shows that which attributes should be considered to cluster data point into particular cluster.

TABLE I. Reduced Attribute Set

Different Class	Attributes To Be Considered
Normal	{1,3,5-10,14,15,17,20-23,25-29,33,35,36,38-40}
DoS	{1,3,5,6,23-28,32,33,35,36,38-41}
Probe	{3,5,6,23,24,32,33}
U2R	{5,6,8,15,16,18,32,33}
R2L	{3,5,6,21,22,24,32,33}

A. Proposed Multi-Threaded K-Means Approach

In our approach as KDD99 data set has been used as input, it is known that it contains 4 types of main attack. So we have taken total 6 threads to run in parallel. Out of these, 5 threads (Clustering threads T_i, where i=1 to i=5) are used to cluster the data and 6th thread (Coordinator thread T_c) is used to take decision and classify the results. To cluster data point into any particular attack as mentioned in Table I, there is no need to consider all 41 attributes of KDD99 data set; we can take only some attribute and cluster that data point as a particular attack or not that attack. Thread T_i (where i=1 to i=5) will take only needed attributes to cluster data point as mentioned in Table I.

Thread T_1 takes only attributes needed to cluster it into Normal or No Normal clusters.

Thread T_2 takes only those attributes which are needed to check whether the data point is DoS attack or No DoS.

T_3 thread is responsible to check whether data point is a Probe attack or not so it takes needed attributes accordingly.

Thread T_4 takes those attributes needed to check whether data point is R2L attack or No R2L.

Thread T_5 takes attributes needed to cluster data point into either U2R attack or No U2R.

All these 5 threads run in parallel. Whenever thread T_i clusters the data point it puts the result into a shared result matrix R [N][T-1] (where N is number of data points in a data set and T is total number of threads). R stores the result of each data point that each thread T_i has clustered it into which cluster. The Coordinator thread T_c's task is to conclude from the result matrix R that each data point is of which type of attack and put the classified result into classification matrix C [N][1]. Fig. 3 shows the flow chart of Clustering threads T_i and Fig. 4 shows the flow chart of Coordinator thread T_c.

Fig.3 shows that each T_i is responsible to cluster the data point into any one of two allotted clusters. Thus each T_i has two initial centroids C_1, C_2 which have exactly same number of attributes as are needed to cluster the data point into allocated clusters. Here all six threads run in parallel. Synchronization points that are where T_c has to wait for the results put by T_i and where T_i has to wait for T_c are shown in the flow charts. Sometimes it may possible that same data point can be clustered into more than one attack type.

When a data point is clustered into No Normal cluster by T_1 and remaining thread T_2, T_3, T_4, T_5 cluster it into No DoS, No Probe, No R2L and No U2R respectively at that time Coordinator thread T_c considers it as an attack in classification

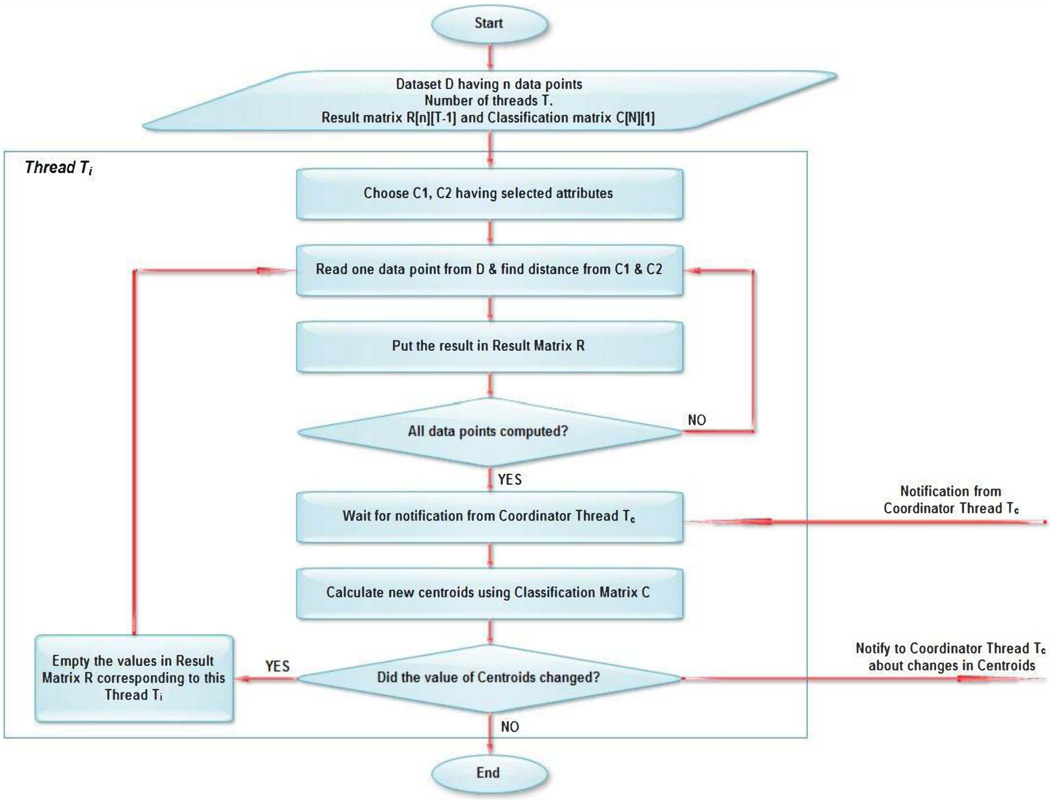
Start
Dataset D having n data points
Number of threads T.
Result matrix R[n][T-1] and Classification matrix C[N][1]
Thread T_i
Choose C1, C2 having selected attributes
Read one data point from D & find distance from C1 & C2
Put the result in Result Matrix R
All data points computed?
NO
YES
Wait for notification from Coordinator Thread T_c
Notification from Coordinator Thread T_c
Calculate new centroids using Classification Matrix C
Empty the values in Result Matrix R corresponding to this Thread T_i
YES
Did the value of Centroids changed?
Notify to Coordinator Thread T_c about changes in Centroids
NO
End

Figure 3. Clustering Thread T_i's Flow Chart

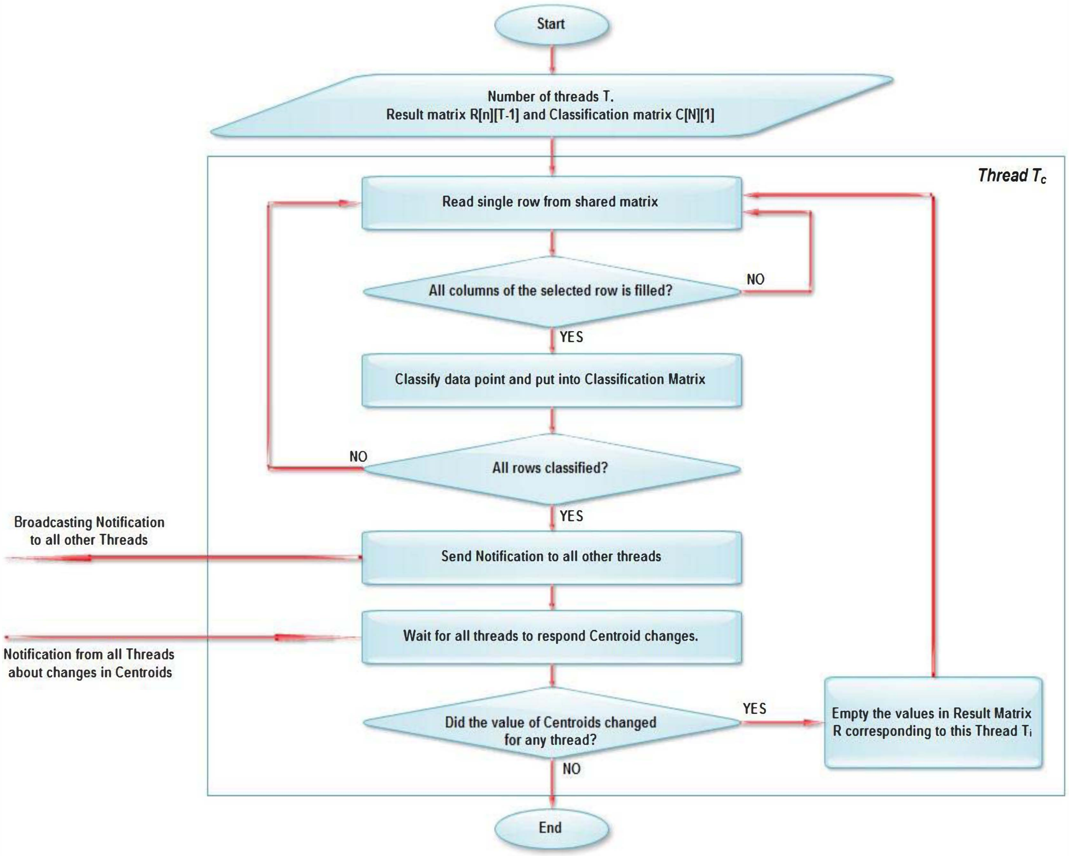
Start
Number of threads T.
Result matrix R[n][T-1] and Classification matrix C[N][1]
Thread T_c
Read single row from shared matrix
All columns of the selected row is filled?
NO
YES
Classify data point and put into Classification Matrix
NO
All rows classified?
YES
Broadcasting Notification to all other Threads
Send Notification to all other threads
Wait for all threads to respond Centroid changes.
Notification from all Threads about changes in Centroids
Did the value of Centroids changed for any thread?
YES
Empty the values in Result Matrix R corresponding to this Thread T_i
NO
End

Figure 4. Coordinator Thread T_c's Flow Chart

matric C. Further that particular data point is considered as an anomaly and that is manually checked by a system administrator for new type of attack.

IV. Experiments And Results

A subset of KDD99 10% training data set has been used for experiments. The subset (Data Set) has total 20,000 randomly chosen records. It has 11,000 Normal, 8,175 DoS, 675 Probe, 134 R2L and 16 U2R attack records. Symbolic attributes of data set like protocol_type, service and flag are converted into integer value from [0, N-1] (where N = total available symbols in particular attribute).The labelling information is used after our experiments are performed, to evaluate the results. All the experiments are done using C#.net Version 4.5 and SQL Server 2008 is used to store data sets. Our experiments are done on a system having 4GB RAM and two Intel® Core™ i5 CPU of clock frequency 3.2GHz and 3.33GHz.Each processor has two cores.

The comparison between K-Means and proposed approach is shown in Table II. It shows the Detection Rate (DR), False Positive Rate (FPR) and False Negative Rate (FNR). Detection rate shows how perfectly the given attack is detected. False positive occurs when the system classifies an action as an intrusion while it is a legitimate action. False negative occurs when an intrusion action has occurred but the system considers it as a nonintrusive.

TABLE II. Overall Comparison Between K-Means and Proposed Approach

	DR	**FPR**	**FNR**
K-Means	95.12%	8.3%	0.18%
Multi-Threaded K-Means	99.18%	1.34%	0.16%

As shown in Table II proposed Multi-Threaded K-Means has better detection rate and low false positive-negative rate. Detail analysis of K-Means and proposed Multi-Threaded K-Means has been shown below in Table III. Table III shows the detail for each particular attack.

TABLE III. Detailed Analysis of Result

		Normal	**DoS**	**Probe**	**R2L**	**U2R**
K-Means	**DR**	91.70%	100%	99.85%	61.19%	37.5%
	FPR	-------	0.36%	0.03%	7.96%	0.37%
	FNR	-------	0.00%	0.19%	5.95%	50.00%
Multi-Threaded K-Means	**DR**	100%	100%	99.85%	50.00%	43.75%
	FPR	-------	0.00%	0.00%	0.13%	1.34%
	FNR	-------	0.00%	0.14%	5.22%	50.00%

V. Conclusion

Proposed approach has resulted in overall high detection rate and low false alarm rate. It is flexible enough to detect new attacks in the network. Sometimes it is difficult to make out exactly that particular packet is of which type and at this time proposed approach flags that packet as two or more attacks based on result matrix R so that it can be manually taken care where other approaches cluster them into any one type and may fail.

Proposed approach can be applied to different similar applications where there is no need to consider all the attributes to cluster. Even our proposed approach can be applied when it is difficult to distinguish that particular data point belongs to which cluster or when a data point can belongs to two or more cluster.

References

[1] SANS Institute—Intrusion Detection FAQ, http://www.sans.org/ resources/idfaq/, 2010.

[2] X. B. Li, "A scalable decision tree system and its application in pattern recognition and intrusion detection," Decision Suppport Systems, 41,pp.112-130, 2005.

[3] Y. Yu, J.H. Wang, J. Zhang, "Researches on intrusion detection system based on RBF and Elman hybrid neural network," Microelectronics & Computer (in Chinese), 8,pp.154-157,2009.

[4] D. Zhang, F. Ren, K. Zhao, "A SVM-based system for on-line unsupervised intrusion detection,"Journal of Jilin University (Science Edition) (in Chinese), 2,pp. 323-328,2009.

[5] H.L. Guo, Y. Tan, D.L. Zhang, "Application of genetic algorithm in rule extraction of intrusion detection," Journal of Harbin Institute of Technology(in Chinese), 1,pp.248-250,2009.

[6] P. Tadeusz, T. Axel, "Data mining and machine learning-towards reducing false positives in intrusion detection," Information Security Technical Report, 10,pp.169-183,2005.

[7] Mrudula Gudadhe, Prakash Prasad and Kapil Wankhade," A New Data Mining Based Network Intrusion Detection Model" Int'l Conf. on Computer & Communication Technology (ICCT'10).

[8] Q. Zhou, F. Y. Zhao, C. J. Wang, "Study of data mining in intrusion detection," PR & AI, 4, pp.520-526,2008.

[9] J. H. Gu, L. J. Sun, "Application research of data mining technology to intrusion detection," Computer Technology and Development(in Chinese), 9,pp.243-245,2006.

[10] Pang-Ning Tan, Vipin Kumar, Michael Steinbach,"Introduction to Data Mining", Copyright © 2006 by Pearson Education, Inc. ISBN-978-81-317-1472-0.

[11] Meng Jianliang, Shang Haikun and Bian Ling , "The Application on Intrusion Detection Based on K-means Cluster Algorithm", 2009 International Forum on Information Technology and Application DOI 10.1109/IFITA.2009.34.

[12] Z. Muda, W. Yassin, M.N. Sulaiman, N.I.Udzir, "Intrusion detection based on K-Means clustering and OneR Classification", 2011 7th International Conference on Information Assurance and Security (IAS) 978-1-4577-2155-7/11.

[13] Gerhard Münz, Sa Li, and Georg Carle, "Traffic anomaly detection using k-means clustering", In Proc. of Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und Verteilten Systemen, 4. GI/ITG-Workshop MMBnet 2007,Hamburg, Germany, September 2007.

[14] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[15] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani," A Detailed Analysis of the KDD CUP 99 Data Set",Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Application (CISDA 2009).

[16] Raj Basu, Robert K. Cunningham, Seth E. Webster, Richard P. Lippmann," Detecting Low Profile Probes and Novel Denial-of-Service Attacks", Proceedings of the 2001 IEEE Workshop on Information Assurance, New York, USA-IEEE,2001

[17] Ye Qing, Wu Xiaoping and Huang Gaofeng , "An Intrusion Detection Approach based on Data Mining" : 2010 2nd International Conference on Future Computer and Communication.