Integration Bat Algorithm with k-means for Intrusion Detection System

Somayeh Sedghi, Mirkamal Mirnia
Faculty of Computer & Electrical Engineering, Tabriz Islamic Azad University

Abstract
Intrusion detection is a new network security mechanism for detecting, prevent-ting, and repelling unauthorized access to a communication or computer network. Intrusion detection systems (IDS) play a crucial role in maintaining a safe and secure network. One technical challenge in intrusion detection systems is the curse of high dimensionality. To overcome this problem, we propose a feature selection phase, which can be generally implemented in any intrusion detection system. In this work bat algorithm with k-means integrated to improving Intrusion Detection System accuracy. Experiments on KDD Cup 99 data set address that our proposed method results in detecting intrusions with higher accuracy, especially for remote to login (R2L) and user to remote (U2R) attacks.

Keywords
component; Intrusion detection, Bat Algorithm, clustering, k-measn

1. Introduction

As information systems have become more comprehensive and a higher value asset of organizations, intrusion detection systems have been incorporated as elements of operating systems, although not typically applications. Intrusion detection involves determining that some entity, an intruder, has attempted to gain, or worse, has gained unauthorized access to the system[1].

Intruders are classified into two groups. External intruders do not have any authorized access to the system they attack. Internal intruders have at least some authorized access to the system. Internal intruders are further subdivided into the following three categories. Masqueraders are external intruders who have succeeded in gaining access to the system and are acting as an authorized entity. Legitimate intruders have access to both the system and the data but misuse this access (misfeasors)[2]. Clandestine intruders have or have obtained supervisory (root) control of the system and as such can either operate below the level of auditing or can use the privileges to avoid being audited by stopping, modifying, or erasing the audit records[3].

In the last three years, the networking revolution has finally come of age. More than ever before, we see that the Internet is changing computing as we know it. The possibilities and opportunities are limitless; unfortunately, so too are the risks and chances of malicious intrusions[4].

It is very important that the security mechanisms of a system are designed so as to prevent unauthorized access to system resources and data. However, completely preventing breaches of security appear, at present, unrealistic. We can, however, try to detect these intrusion attempts so that action may be taken to repair the damage later. This field of research is called Intrusion Detection[5].

Classification of IDS essentially falls under two models: the misuse or signature-based model and the anomaly model[6].

The misuse or signature-based is the most-used IDS model. Signatures are patterns that identify attacks by checking various options in the packet, like source address, destination address, source and destination ports, flags, payload and other options. The collection of these signatures composes a knowledge base that is used by the IDS to compare all packet options that pass by and check if they match a known pattern.

The anomaly model tries to identify new attacks by analyzing strange behaviors in the network. To make this possible, it first has to ``learn" how the traffic in the network works and later try to identify different patterns to then send some kind of alert to the sensor or console. IDS made using this model have higher tendency for raising false alarm, as they often suspicious about all network behavior irrespective of malicious or legitimate[7].

With the rapid progress in the network-based technology and applications, the threat of spammers, attackers and criminal enterprise has also grown accordingly. The 2005 annual computer crime and security survey showed that the total financial losses caused by all kinds of network viruses/intrusions for respondent companies were about US $130 million. Furthermore, according to other studies, an average of twenty to forty new vulnerabilities that existed in networking and computer products was detected every month[8].

The rest of the paper is organized as follows. In Section 2, the previous works is reviewed. In section 3, we describe the proposed method. The experimental results and the comparison with a set of algorithms from the literature are presented in Section 4. Finally, in Section 5, we draw a conclusion and discuss the perspectives of development of this work.

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

2. Related works

A computer system should provide confidentiality, integrity and assurance against denial of service. However, due to increased connectivity (especially on the Internet), and the vast spectrum of financial possibilities that are opening up, more and more systems are subject to attack by intruders. These subversion attempts try to exploit flaws in the operating system as well as in application programs and have resulted in spectacular incidents like the Internet Worm incident of 1988[9].

We thus see that we are stuck with systems that have vulnerabilities for a while to come. If there are attacks on a system, we would like to detect them as soon as possible (preferably in real-time) and take appropriate action. This is essentially what an Intrusion Detection System (IDS) does. An IDS does not usually take preventive measures when an attack is detected; it is a reactive rather than pro-active agent. It plays the role of an informant rather than a police officer[7].

The most popular way to detect intrusions has been by using the audit data generated by the operating system. An audit trail is a record of activities on a system that are logged to a file in chronologically sorted order. Since almost all activities are logged on a system, it is possible that a manual inspection of these logs would allow intrusions to be detected. However, the incredibly large sizes of audit data generated (on the order of 100 Megabytes a day) make manual analysis impossible. IDSs automate the drudgery of wading through the audit data jungle. Audit trails are particularly useful because they can be used to establish guilt of attackers, and they are often the only way to detect unauthorized but subversive user activity.

Intrusions can be divided into 6 main types[10]

1. Attempted break-ins, which are detected by atypical behavior profiles or violations of security constraints.
2. Masquerade attacks, which are detected by atypical behavior profiles or violations of security constraints.
3. Penetration of the security control system, which are detected by monitoring for specific patterns of activity.
4. Leakage, which is detected by atypical use of system resources.
5. Denial of service, which is detected by atypical use of system resources.
6. Malicious use, which is detected by atypical behavior profiles, violations of security constraints, or use of special privileges.

However, we can divide the techniques of intrusion detection into two main types.

Currently there are two basic approaches to intrusion detection. The first approach, anomaly detection, attempts to define and characterize correct static form of data and/or acceptable dynamic behavior. In effect, it searches for an anomaly in either stored data or in the system activity. IDS utilizing anomaly detection include Tripwire [Kim93], Self-Nonself, and NIDES[11].

The second approach, called misuse detection, involves characterizing known ways to penetrate a system in the form of a pattern. Rules are defined to monitor system activity essentially looking for the pattern. The pattern may be a static bit string or describe a suspect set or sequence of events. The rules may be engineered to recognize an unfolding or partial pattern. IDS utilizing misuse detection include NIDES [Anderson95], MIDAS [Sebring88], and STAT[12].

In the past decades, a great number of intrusion detection has been proposed to detect anomalies . Next generation intrusion detection export system (NIDES) was one of the few intrusion detection systems, which could operate in real-time for continuous monitoring of user activity or could run in a batch-mode for the periodic analysis of the audit data. the audit data (Anderson et al., 1994), (Anderson et al., 1995). It generates the profile by using statistical measurement. Audit Data Analysis and Mining (ADAM) is one of the known data mining projects in an intrusion detection, which uses a module to classify the abnormal event into false alarm or real attack . It is an online network-based IDS, which used two data mining techniques, association rule and classification. An ADAM was one out of the seven systems tested in the 1999 DARPA evaluation.

Boukerche et al. (2004) used natural immune human systems to detection anomaly in the computer network. Authors applied the proposed scheme to extract significant features of the immune human system and then map these features within a software package designed to provide security of a computer system and to identify irregular activities according to the usage log files[12].

3. Propsed method

In this work the k-means algorithm and bat algorithm is integrated for Intrusion Detection System.

There are several clustering techniques, including hierarchical clustering, minimum spanning tree, and k-means clustering. The k-means clustering algorithm is used in the case study described in this paper. This is one of the simplest and most popular clustering algorithms. The algorithm finds k distinct classes of jobs, or clusters, when

given a specific workload data set. The algorithm finds the midpoint, or centroid, of each cluster and assigns each job to its nearest centroid. The algorithm is initialized by providing: 1) the number of desired clusters, k, and 2) initial starting point estimates of the k centroids. There is no commonly accepted or standard "best" way to determine either the number of clusters or the initial starting point values. The resulting set of clusters, both their number and their centroids, depends on the specified choice of initial starting point values. This case study shows that the number of clusters and the final cluster centroid values differ based on the initial starting values used when performing k-means clustering[13].

Given a data set of workload samples, a desired number of clusters, k, and a set of k initial starting points, the k-means clustering algorithm finds the desired number of distinct clusters and their centroids. A centroid is defined as the point whose coordinates are obtained by computing the average of each of the coordinates (i.e., feature values) of the points of the jobs assigned to the cluster [2]. Formally, the k-means clustering algorithm follows the following steps[14].

1. Choose a number of desired clusters, *k*.
2. Choose *k* starting points to be used as initial estimates of the cluster centroids. These are the initial starting values.
3. Examine each point (i.e., job) in the workload data set and assign it to the cluster whose centroid is nearest to it.
4. When each point is assigned to a cluster, recalculate the new *k* centroids.
5. Repeat steps 3 and 4 until no point changes its cluster assignment, or until a maximum number of passes through the data set is performed.

In this paper for optimization of cluster center we use Bat algorithm. The bat set of rules (BA) became initially brought in perivous work which has been related to benchmark functions, consequently BA performs particle swarm optimization and genetic algorithms. BA has been correctly implemented to hard optimization problem consisting of motor wheel optimization hassle, clustering problem, in conjunction with[8]famous engineering optimization duties. BA shows within the said literature own attracted the authors to pick this set of rules for attributes reduction assignment. Bats are animals that have wings and consist of the capability of echolocation (also referred to as biosonar).

Echolocating animals produce refers to to the surroundings and additionally concentrate to the echoes of these calls. Those echoes will be used to discover and become aware of the gadgets. Among all the the bat species, microbats use echolocation broadly. In microbats, echolocation is a kind of sonar acquainted with perceive target, keep away from close to limitations within the dark, and locate roosting crevices. At some stage in echolocation those sorts of microbats emit a string of brief, high-frequency sound effects after which pay interest for the echo that bounces backside from the surroundings gadgets as illustrated in parent 1. With this echo a bat can decide an object's dimensions, shape, direction, duration, and movement. Even as the bats fly close to to their prey, the pace of pulse emission can accelerate as a great deal as 2 hundred pulses per 2d. A persevering with frequency in every pulse can be observed. The wavelengths of a pulse have been within the comparable order of their prey sizes. The loudness for trying to find prey is more than whilst homing toward the prey50. To vicinity it in another way, the loudness decreases even at the same time as acquiring closer to the victims.

4. Experimental results

In order to evaluate the high performance of proposed method, performance test is illustrated for the proposed algorithm, and the proposed results are compared with that of other state-of-the-art approaches.

Whereas the KDD Cup 99 training set contains more than four million data points and such a large data set cannot be fed to an SVM in the training phase, we randomly selected 6480 records from five classes as the training data and 6703 records as the evaluation data.

The data set used in these experiments is ''KDD Cup 1999data'' (kddcup data set. kdd.ics.uci.edu//databases/kddcup99/kddcup99.htm), a well-known set of intrusion evaluation data.The raw training data was processed into about five million connection records. A connection is a sequence of TCP packets starting and ending at some well defined times. Each record is unique in the data set with 41 continuous and nominal features plus one class label. In this paper, the nominal feature such as protocol (tcp/udp/icmp), service type (http/ftp/telnet/y) and TCP status flag (sf/rej/y) have been converted into a numeric feature. The method is simply by replacing the values of the categorical attributes with numeric values. For example, the protocol-type attribute in KDD Cup, 1999.

ROC curves are shown in Figs. 1 and 2 in terms of the number cluster center of the purposed method.

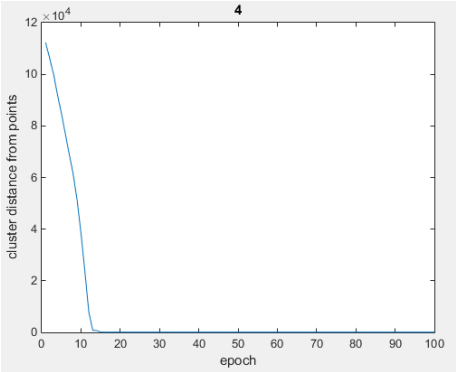

Fig. 1 Result of clustering

Moreoer, result of clusteng and cluster center for proposed method is shown in table 1

Table 1: Cluster center

x	y	Z
3.8053	0.9707	0.6424
29.6835	8.6757	0.9201
0.4277	0.5171	44.6689
0.3170	34.8172	0.8157

Moreover optimizing chart for 8 cluster is shown in Fig 2.

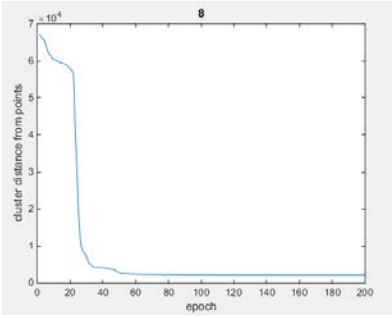

Fig. 2 result to 8 cluster

In table 2 proposed methods has been compared with previous work.

Table 2: proposed method compared with pervious works

Proposed method	K-means and BAT algorithm	K-means	Algorithm
2200.4	2300.4	2409.9	Error

5. Conclusion

Intrusion detection is a new network security mechanism for detecting, prevent-ting, and repelling unauthorized access to a communication or computer network. Intrusion detection systems (IDS) play a crucial role in maintaining a safe and secure network. The term anomaly-based intrusion detection describes a class of techniques that attempt to classify network traffic as either normal or anomalous. It mainly involves binary classification of selected audit data and other aspects of a system. The success of an intrusion detection system depends on how well it succeeds in maximizing its detection accuracy while minimizing its false alarm rate. Because of their success in practice at spotting new and unfamiliar attacks, anomaly-based intrusion detection systems have remained a heavily researched topic in the IDS community.

In this work bat algorithm with k-means integrated to improving Intrusion Detection System accuracy. Experiments on KDD Cup 99 data set address that our proposed method results in detecting intrusions with higher accuracy, especially for remote to login (R2L) and user to remote (U2R) attacks

References

[1] Aburomman, A.A. and M.B. Ibne Reaz, A novel SVM-kNN-PSO ensemble method for intrusion detection system. Applied Soft Computing, 2016. 38: p. 360-372.
[2] Amin Hassanzadeh , Ala Altaweel , and R. Stoleru, Traffic-and-resource-aware intrusion detection in wireless mesh networks. Ad Hoc Networks, 2014. 21: p. 18-41.
[3] Chan, G.-Y., C.-S. Lee, and S.-H. Heng, Defending against XML-related attacks in e-commerce applications with predictive fuzzy associative rules. Applied Soft Computing, 2014. 24: p. 142-157.
[4] Eesa, A.S., Z. Orman, and A.M.A. Brifcani, A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Systems with Applications, 2015. 42(5): p. 2670-2679.
[5] Maleh, Y., et al., A Global Hybrid Intrusion Detection System for Wireless Sensor Networks. Procedia Computer Science, 2015. 52: p. 1047-1052.
[6] Devi, R., et al., Implementation of Intrusion Detection System using Adaptive Neuro-Fuzzy Inference System for 5G wireless communication network. AEU - International Journal of Electronics and Communications, 2017. 74: p. 94-106.
[7] Singh, R., H. Kumar, and R.K. Singla, An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Systems with Applications, 2015. 42(22): p. 8609-8624.
[8] Qin, T., et al., Robust application identification methods for P2P and VoIP traffic classification in backbone networks. Knowledge-Based Systems, 2015. 82: p. 152-162.
[9] Yi, G., et al. Multi-agent Intrusion Detection System Using Feature Selection Approach. in Intelligent Information

Hiding and Multimedia Signal Processing (IIH-MSP), 2014 Tenth International Conference on. 2014.
[10] Mohammed, M.N. and N. Sulaiman, Intrusion Detection System Based on SVM for WLAN. Procedia Technology, 2012. 1: p. 313-317.
[11] Pal, D. and A. Parashar. Improved Genetic Algorithm for Intrusion Detection System. in Computational Intelligence and Communication Networks (CICN), 2014 International Conference on. 2014.
[12] Mabu, S., et al., An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011. 41(1): p. 130-139.
[13] Bandyopadhyay, S., et al., Integration of dense subgraph finding with feature clustering for unsupervised feature selection. Pattern Recognition Letters, 2014. 40: p. 104-112.
[14] Jain, A.K., Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010. 31(8): p. 651-666..