International Journal of Embedded and Wireless Communication Systems
Volume1, Issue 3 2025

A CLUSTERING ALGORITHM FOR DETECTING DDoS ATTACKS IN NETWORKS

Sannanuddin[1], Syed Zabiulla Hussaini[2], Nouman Hussain[4]

Assistant Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology.

Research Article

Date of Submission: 20-04-2025 **Date of Acceptance: 17-06-2025** **Date of Publication: 23-07-2025**

Abstract—Intrusion detection systems are crucial to maintaining network security as the number of connected machines increases. Data mining techniques have recently become more significant in solving network security problems, such as network intrusion detection, which is a difficult job in network security. The goal of intrusion detection systems is to find assaults with a low false alarm rate and a high detection rate. Signature-based detection is used in the most popular and commercially accessible intrusion detection techniques. However, their techniques are very expensive and time-consuming to get the signatures, and they are unable to identify unexpected incursions that do not match the signatures. Numerous academics have put forward a variety of methods based on unsupervised learning approaches to address the issues. In order to identify intrusions and increase the detection rate while keeping a low false positive rate, we describe in this work a unique clustering-based intrusion detection technique called unsupervised anomaly detection, which trains on unlabeled data. The 2000 DARPA Intrusion Detection Scenario Specific Data Set was used to assess our approach.

Index terms: Anomaly detection, heuristic clustering, true positive rate, false positive rate.

1. INTRODUCTION

Local area networks and wide area networks are two new application areas for computer networks that have evolved due to the rapid growth of computer-based technologies.This became a major weakness for the community and an appealing target for exploitation. This infrastructure's security has emerged as the exclusive focus of study. Network intrusion detection systems are becoming a common part of security systems. "The process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions, defined as attempts to compromise the confidentiality, integrity, availability, or to bypass the security mechanisms of a computer or network" is known as intrusion detection.

In network intrusion detection, there are often two kinds of assaults. Each occurrence in a data collection is classified as "normal" or "intrusion" in misuse detection, and a learning algorithm is taught using the labeled data. By creating models of typical behavior, an anomaly detection method automatically identifies any deviations and marks them as suspicious. assaults may be divided into four primary kinds [8]: Denial of Service (DoS) assaults, in which the attacker overloads a computer or memory resource to the point that it cannot process valid requests, A remote to user (R2L) attack is a type of attack in which an attacker sends packets to a machine via a network, then takes advantage of the machine's vulnerability to unlawfully obtain local access for a user. A user-to-root exploit is a type of attack in which an attacker begins with access to a regular user account on the system and is able to take advantage of a vulnerability to obtain root access. A kind of attack called as probing involves an attacker searching a network for information or known weaknesses.

The massive amount of current and freshly emerging network data that has to be processed is the primary driver behind the use of data mining techniques for intrusion detection systems. A network

gathers enormous amounts of data every day. A number of data mining methods, including association rules, clustering, and classification, are showing promise in obtaining various insights for intrusion detection.

The main benefit of unsupervised anomaly detection (UAD) algorithms is their capacity to evaluate unlabeled data and identify intrusions that would otherwise go undetected. Finding a "natural" grouping in a collection of patterns, points, or objects without being aware of any class labels is the aim of data clustering, also known as unsupervised learning.

We need a method for identifying new and unidentified forms of intrusions as well as for identifying intrusions when our training data is unlabeled. Using a dataset of network connections, we assessed our clustering technique. We looked at network data from DARPA 2000 [9], a well-known and extensively used intrusion assault data collection. According to our experimental findings, the suggested method's detection rate regularly beats other algorithms that have been published in the literature, particularly when it comes to the false positive rate.

RELATED WORKS

Because network-based computer systems are becoming more and more essential to contemporary civilization, network system security is more crucial than ever. IDS (Intrusion Detection Systems) is an active security system that looks for intrusions in secure architecture. The U.S. Department of Defense established security objectives for audit mechanisms in the 1970s, one of which was permitting the identification of efforts to circumvent security measures.

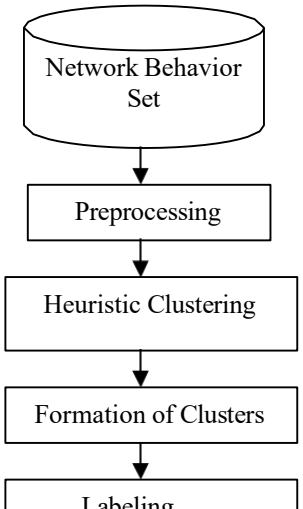
There have been many distinct kinds of intrusions in recent years, and multiple detectors are required to identify both typical and unusual network activity. Since they may benefit existing intrusion detection systems in a number of ways, several clustering algorithms have lately drawn interest in related research.

The DARPA 1998 data set of network connections [3] and actual network data are used to assess a number of current supervised and unsupervised anomaly detection strategies and their variants using established assessment methods.

Anomaly detection has recently been used to detect harmful activity in computer systems, assaults in computer networks [1], and abuse in online sites [6], [7]. Because of their high detection accuracy at low false positives, a more recent class of anomaly detection systems—such as artificial neural networks [8], Kohonen's self-organizing maps [9], fuzzy classifiers, and others [7]—have gained popularity. Numerous search techniques, including statistical analysis, deviation analysis, rule induction, neural abduction, establishing linkages, correlations, and grouping, are used in data mining for security knowledge [2].

The use of clustering algorithms for efficient intrusion detection is another data mining method [6].However, the current approaches that rely on K-Means or K-Medoids have two drawbacks when it comes to clustering huge network datasets: the inability to handle character properties in network transactions and the quantity of cluster dependencies. The amount of cluster dependence among them indicates that the value of K is crucial to the clustering outcome and need to be determined prior to clustering.Trained a Hidden Markov Model [7] mplemented on the trained datasets that he used to train instance-based learner. Fixed- width and k-nearest neighbor clustering techniques [4] to connection logs looking for outliers, which represent anomalies in the network traffic. A more recent class of IDS developed using machine learning techniques like artificial neural networks, Kohenen’s self organizing maps, fuzzy classifiers, symbolic dynamics [10], and multivariate analysis.

This paper presents the idea of applying data mining techniques to intrusion detection systems to maximize the effectiveness in identifying attacks, thereby helping the users to construct more secure information systems.



Contributions of the paper

The rest of the paper is organized as follows. In Sect. 3, we present Heuristic clustering algorithm. In Sect. 4, we describe the details of our experiment and present the results and their analysis. Finally, we present concluding remarks and suggestions for future study.

2. INTRUSION DETECTION SYSTEM

Clustering is one of unsupervised learning techniques to group data instances into meaningful subclasses. In this section, we describe the Heuristic Clustering Algorithm, and illustrate how to apply this algorithm to generate detection models from audit data.

The Overall clustering process of the clustering algorithm and labeling as follows.
is composed of two phases: Clustering Process and Labeling (shown in Fig 3.1).In the first phase , partitions the training instances into clusters using heuristic clustering. In the second phase using labeling detects the normal and anomaly instances.

Figure 3.1 Framework of IDS

2.1 Data preprocessing

Depending on the monitoring data and the data mining algorithm, it may be necessary or beneficial to perform cleaning and filtering of the data in order to avoid the generation of misleading or inappropriate results. Extract nine attributes values from the dataset and store into the database.Attributes are Source IPaddress, destination IPaddress, Protocol, Source Port, destination port, Sequence number, Acknowledgment number, length, window size.

2.2 Clustering Algorithm Notations of Some Terms

Some notations needed in heuristic clustering algorithm are
Notation1: Let H= {H_1, H_2 ,...., Hm} be a set of
attribute values, the m is number of attribute values. Some attribute values are duration, src- bytes, dest-bytes, flag, etc..,
Notation 2: Let H = H_N U H_S and $H_N \cap H_S = \varphi$, where H_N is the subset of numerical attribute (e.g., no of bytes), and H_S is the subset of character attribute. (e.g., service, Protocol).
Notation 3: Let, $e_i = (h_{i1}, h_{i2,,...,}h_{im})$, e_i is a record, the m is number of attribute values and h_{ij} is the value of H_m.
Notation 4: E = {e_1, e_2 ...e_n}, E is the set of records; n is the number of packets.

Heuristic Clustering Algorithm (HCA)

In the HCA algorithms we don't need fix the value of K in the beginning of clustering. We use the heuristic clustering technology to determine the number of cluster automatically. We compute the similarity between e_i and every center of cluster: Sim(e_i,Cj) and the similarity between every center of cluster Sim(C), if the minimal of Sim(e_i,C_j) is more than the minimal of Sim(C) , we create a new cluster, and the center of cluster is e_i , otherwise we insert the e_i into C_j.

Step 1. Confirm two initial cluster centers by algorithm search ( ).
Step 2. Import a new record e_i. Repeat 3 to 5 until no more records.
Step 3. Compute the similarity by algorithm Similar (), and find Min (Sim (e_i,c_j)), Min (Sim (C)).

(3) For i=1 to *l* do m=Count_m(m_i)
// m=center {m_1,m_2, m_3..... m_l}
(4) m1=m , m2=max (Sim (m, mi))

The method of computing similarity

The audit data consists of numerical attribute (e.g. the receive Bytes) and character attribute. (E.g. Protocol).But whether k-Means or K- medoids all lack the ability of dealing with character attribute. The HCA algorithm resolves this problem by processing the character attribute using the method of attribute matching. The similarity of character attributes:

let e_i and e_j be two records in the E. all containing m attributes (including P character attributes), the n_{hik} and n_{hjk} is the number of h_{ik} and h_{jk} respectively.

$$Sim^{P}(e_i,e_j) = \sum_{k=1}^{p} \frac{(n_{hik} + n_{hjk})}{(n_{hik} * n_{hjk})} * A$$

if (h_{ik}=h_{jk}) then A= 0 else A=1.

The similarity of numerical attribute (to the numerical attribute, still use the classical Euclidean distance to computer similarity)then C = {e }, C= {C , C ... C},

$$(e_i, e_j) = \sum_{k+1}^{i} | hik - hjk | 2$$

create a new cluster, and the e_i is the center of the new cluster.

Else C_j=C_j U {ei}, insert e_i into C_j.

The initial center of clustering

In the beginning of clustering, we should confirm two initial center of clustering by the algorithm Search ( ).

Algorithm: Search_m (E,l).

Import data set E,the number of sampling l Output: initial center m_1,m_2.

(1) Sampling E, get S_1,S_2,..,S_l

(2) For i=1 to *l* do mi=Count_m(S_i)

(3) The similarity of two records (including similarity of numerical attribute and similarity of character attribute)

Sim (ei,ej)) = Sim^{N}(ei,ej)+ Sim^{P}(ei,ej)

The center of cluster

A cluster are represented by its cluster center .In the HCA algorithm, we use the algorithm Count ( ) to compute the cluster center. The center of a cluster is composed of the center of numerical attributes and character attribute.Let P= (P_N + P_S), and P= (P_1, P_2,…., P_m)

where P_N is the center of numerical attribute, the Ps is the center of character attribute,

EXPERIMENTS AND RESULTS

In this section, we present the results of the

i= 1,2,…., p (p<= m)
Heuristic clustering method compare it with
K-Means over the dataset 2000 DARPA . In heuristic clustering the relations between

The h_{ji} is the numerical attribute. The P_S is a frequent character attribute set which consists of q (q=m-p) most frequent character attribute.

Labeling

We propose a method to detect anomaly that does not either depend on the population ratio of the clusters. In our labeling method, we assume that center of a normal cluster is highly close to the initial cluster center v*h* which are created from the clustering.

In other words, if a cluster is normal, the distance between the center of the cluster and v*h* will be small, otherwise it will be large.Thus, we first, for each cluster center C*j* ($1 \leq j \leq k$), calculate the maximum distance to v*h*. We then calculate the average distance of the maximum distances. If the maximum distance from a cluster to v*h* is less than the maximum average distance, we label the cluster as normal. Otherwise, label as attack.

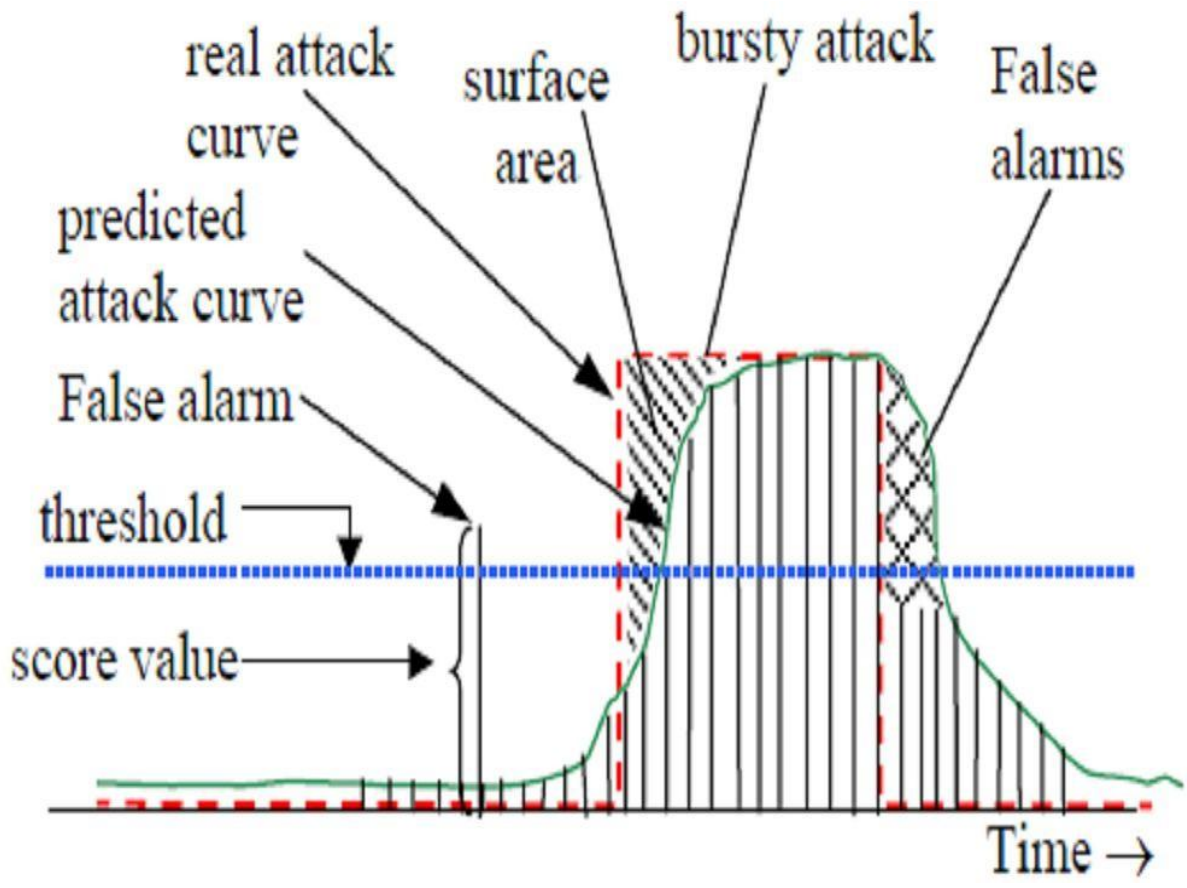


Fig 4.1Assigning Scores to IDS

categorical and continuous features are handled naturally, without any forced conversions (k- means) between these two types of features. Fig
3.2 shows how can we detect the anomaly instances.

We use the measures for evaluating the performance
1. TPR or recall is the percentage of anomaly instances correctly detected,
2. FPR is the percentage of normal instances incorrectly classified as anomaly,
3. "Precision" is the percentage of correctly detected anomaly instances over all the detected anomaly instances,
4. "Total accuracy" or "accuracy" is the percentage of all normal and anomaly instances that are correctly classified.

The identification of normal and abnormal classes has been represented in the form of a matrix called Confusion matrix as given in Table 4.1.

Table 4.1 Confusion Matrix

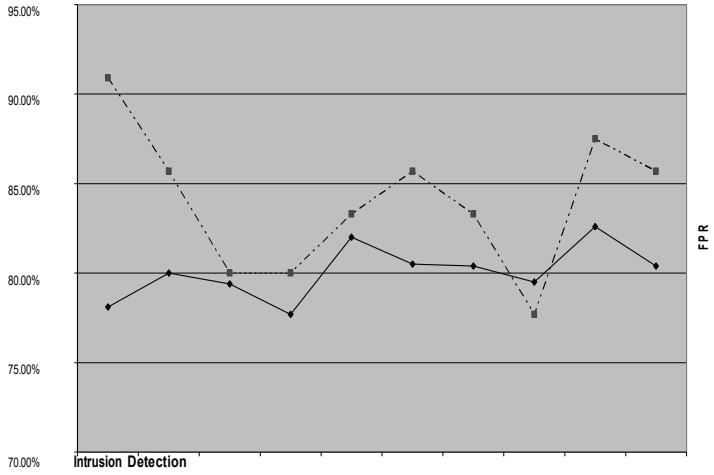
Actual \ Predicted Class	**Normal**	**Abnormal**
Normal	True positive (TP)	False negative (FN)
DDoSl	False positive (FP)	True negative (TN)

The detection rate is computed using the equations

Detection Rate = TN / (FN + TN) * 100 False Positive Rate = FP / (TP + FP) * 100 Error Rate =

(FP + FN) / (TP + FP +
FN + TN) *100

The results obtained from heuristic clustering based intrusion detection accuracy is 93.05 percent at a false-positive-rate of 3.08 percent on 10% subset of the training dataset 2000 DARPA.Table 4.2 depicts the results obtained from the DARPA by using the heuristic Clustering approach. For each iteration the training instances taken for testing and the corresponding Detection



accuracy, false positive rate is shown in fig 4.2.

Table 4.2 Results for anomaly detection using heuristic clustering.

Iteration	Training instances	Detection Rate	False Positive Rate	Error Rate
0	415	90.9%	2.8%	3.4%
1	820	94.7%	2.5%	2.9%
2	1200	89.0%	4.4%	5.4%
3	3000	87.0%	6.4%	7.5%
4	5200	89.3%	2.4%	3.1%
5	6130	89.7%	1.9%	2.7%
6	7117	88.3%	3.1%	3.9%
7	8045	90.7%	6.1%	7.3%
8	9300	87.5%	2.8%	3.6%
9	11245	91.7%	5.3%	5.4%

Fig 4.2 Results of Detection rate with heuristic clustering

3. CONCLUSIONS AND FUTURE WORKS

In order to increase the efficacy of spotting assaults and assist users in building more secure information systems, this study integrates the concept of using data mining methods to intrusion detection systems via network intrusion detection with heuristic clustering. Our algorithm's primary benefit is that it handles the

relationships between continuous and categorical characteristics in the 2000 DARPA data set naturally, without requiring any forced conversions (k-means) between these two feature types.

Future research must create a new benchmark dataset for intrusion detection and confirm the suggested clustering algorithm's performance on actual data.

Extracting more factors and creating a sophisticated detection system to accommodate these many variables are potential future initiatives in this field of study. Multiple benchmark data sets might be used to evaluate the new system.