Intrusion Detection Based On Clustering Algorithm

Nadya El MOUSSAID 1, Ahmed TOUMANARI 2, Maryam ELAZHARI 3

1, 2, 3 ESSI, National School of Applied Sciences, Ibn Zohr University, Agadir, Morocco

nadya.el.moussaid@gmail.com, atoumanari@yahoo.fr, maryam.ensa@gmail.com

Abstract - The traditional Intrusion detection systems have been used long time ago, namely Anomaly-Based detection and Signature-based detection but have many drawbacks that limit their performance. Consequently the main goal of this paper is to use data mining techniques including classification using clustering methods to overpass these defects. This classification will be done by using k-means algorithm. Therefore we have improved k-means to overcome its limits specially the cluster’s number initialization. The experiment results of the work done on KDD’99 dataset shows the performance of the improved k-means in detecting new attacks with more than 90% for Dos and R2L also more than 60% for probe and U2R.

Keywords – Intrusion detection, data mining, classification, clustering, k-means, KDD Cup’99.

I. INTRODUCTION AND REVIEW

Currently, operating systems are the weak links. In fact, they are currently unable to provide effective means of protection for themselves but also for the applications they host. The predominant model protection is discretionary model where end users define their resource rights. With this model, it is impossible to guarantee security objectives. Mandatories approaches are intended to enable the defining of security objectives so that the operating system can ensure it. These proposed approaches only partially meet the problems. On the other hand, they do not treat a subset of the required security objectives. Therefore the security properties serve as the first line of defense. And Intrusion Detection serves as the second. The concept of intrusion detection system was introduced in 1980 by James Anderson [1]. Later Denning in 1987 published a model of intrusion detection [2].

The intrusion detection systems (IDS) using the traditional methods are limited in detecting unknown intrusion behaviour, or update the profile in real time and the maintenance will be heavy, also the rate of false positives and false negatives is high compared to IDS using data mining approach, further limits of traditional IDS listed in [3]. Thus several research focus to resolve these issues by using the data mining technology into the IDS, which makes them automatically produce accurate detection model from a large number of audit data.

Data mining is a set of techniques and methods in the field of statistics, mathematics and computer science to extract, identify valid, novel, potentially useful, and ultimately understandable patterns in massive data. Various data mining’s methods and algorithms have been used such as classification tree and support vector machines for intrusion detection [4], Genetic Algorithms, Neural Networks [3], and Clustering [5-7], all these methods helps to provide a good level of security to the systems from external and internal attacks, also from new attacks. Detecting attacks is an essential need in networks. The data set used for network anomaly detection well-known as KDD Cup 1999 [8].

K-means is one of the simplest unsupervised learning algorithms for finding patterns in unlabeled data with many dimensions (number of attributes). It’s a simple and easy approach to categorize and analyze a given data set through a certain number of clusters suppose k clusters fixed a priori. Among the limits of the algorithm, we quote: - presence of outliers, since “mean” is not a robust statics, - dealing with huge number of dimension and huge number of data items can be problematic because of time complexity, also we have to fix the number of clusters a priori and the results depend on the value of k [9]. To overcome the shortcomings of the K-means, clustering approach for intrusion detection combine K-means algorithm with other related clustering algorithms, such as in [7], this approach expected to automatically partition a data set into a reasonable number of clusters so as to classify the instances into “normal” clusters and “abnormal” clusters. In [6], the authors have proposed a network anomaly detection system by cascading K–Means and C4.5 decision tree algorithm. The proposed method is divided into two phases, first one is Training Phase and the second is Testing Phase. During the training phase, the k-Means-
based anomaly detection method is first applied to partition the training space into k disjoint clusters. Then, C4.5 decision tree is trained with the instances in each k-Means cluster. The testing phase is subdivided into two steps. In the first step: the selection phase, compute the Euclidean distance for every testing instance and find the closest cluster. Compute the decision tree for the closest cluster. In the second step: classification phase, apply the test instance \( Z_i \) over the C4.5 decision tree of the computed closest cluster and classify the test instance \( Z_i \) as normal or anomaly.

An improved fuzzy C-means algorithm (IFCM) [11] is used to intrusion detection. This approach reduces the infection of isolated point by means of weighting the degree of membership for objects to be clustered, and avoids the subjectivity in choosing the number of clustering by introducing the function of validity. Other way to improve the anomaly detection using clustering approach is presented in graph-based intrusion detection algorithm by using outlier detection method that based on local deviation coefficient (LDCGB) [5]. Compared to other intrusion detection algorithm of clustering, this algorithm is unnecessary to initial cluster number. LDCGB uses graph-based cluster algorithm (GB) to get an initial partition of data set which is depended on parameter of cluster precision rather than initial cluster number. On the other hand, because of this intrusion detection model is based on mixed training dataset, so it must have high label accuracy to guarantee its performance. Therefore, in labeling phrase, the algorithm imposes outlier detection algorithm of local deviation coefficient to label the result of GB algorithm again. This measure is able to improve the labeling accuracy. The detection rate and false positive rate are obtained after the algorithm is tested by the KDDCup99 data set.

The structure of the paper is as follow: the next section will describe KDD’99 dataset. The methodology is mentioned in Section III. The clustering for intrusion detection system is on section IV. The experiments analyses and results are shown in Section V. Conclusion in Section VI.

II. KDD’99 DATASET

The KDD’99 was simulated in a military network environment and used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to learn a predictive model or a classifier capable of distinguishing between legitimate and illegitimate connections in a computer network. This data set contains one type of normal data and 24 different types of attacks that are categorized into four types such:

- **Denial of Service Attack (DoS):** a type of attacks on a network that flood it with useless traffics by the consumption of resources and memories.
- **Users to Root Attack (U2R):** the attacker login a normal user account on the system with intent to get to root access to the system.
- **Remote to Local Attack (R2L):** is when the attacker attempt to get a local access as a user of a machine on a network which he doesn’t have any right to access to system.
- **Probe Attack:** This attack is about collecting information from a network of computers for a later use.

<table>
<thead>
<tr>
<th>Attack types</th>
<th>List of attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoS</td>
<td>Back, land, neptune, pod, smurf, teardrop</td>
</tr>
<tr>
<td>U2R</td>
<td>Buffer_overflow, load module, Perl, root kit</td>
</tr>
<tr>
<td>R2L</td>
<td>Ftp_write, guess_pwd, imap, multihop, phf, spy, warezclient, warezmaster</td>
</tr>
<tr>
<td>Probes</td>
<td>Satan, ipsweep, nmap, portsweep</td>
</tr>
</tbody>
</table>

The TCP/IP records information were summarized into connections and each connection instance contains 41 features. Well-known the number of normal connection is widely great than abnormal connection and it value is different.

III. METHODOLOGY

The current and traditional IDS have many drawbacks, and to surmount it, many researches use data-mining, among the drawback I quote:
Intrusion Detection Based On Clustering Algorithm

- **False positives**: a false positive occurs when normal attack is mistakenly classified as malicious and treated accordingly. The solution is to investigate and review the IDS configuration to prevent the false positive from occurring again.

- **False negatives**: A false negative occurs when an attack or an event is either not detected by the IDS or is considered benign by the analyst. Ordinarily the term false negative would only apply to the IDS not reporting an event.

- **Updates lag**: the main issue occurs to Signature-Based Intrusion Detection System is the update lag. In other words, will be always a lag between the appearance of new thread and the IDS’s updates.

- **Data size**: The amounts of data the analyst can efficiently analysis. Among the data-mining techniques, there are the clustering algorithms such as k-means algorithm. K-means is the simplest unsupervised learning algorithm, also is the least costly in time complexity used in intrusion detection field. But it has several lacks previously mentioned. Thus, to overcome some of k-means algorithm shortcomings, namely the problem of the cluster’s number initialization which we will initial by calculating the similarity.

IV. CLUSTERING FOR INTRUSION DETECTION SYSTEM

A. **Cluster’s number initialization**

The following steps explain the calculated similarity:

- **Step1.** Calculate the maximum and the minimum for each feature fi.

- **Step2.** Calculate the similarity,

  \[
  \text{The upper limit: } U = \frac{\text{Max}(f_i) + \text{Min}(f_i)}{2}, \\
  \text{The lower limit: } L = \frac{\text{Max}(f_i) - \text{Min}(f_i)}{2}
  \]

- **Step3.**

  FOR i=0 to n
  IF \( f_i \in [L, U] \) THEN
  IT = True
  ELSE
  IT = False
  END IF.
  END FOR.

- **Step4.** Compare and compute the similarity of each connection features; each connection record with the same similarity is a cluster.

  Step 4 is repeated until there is no connection record.

- **Step5.** Cluster with number of connection records less than 10% of total numbers of connection records are grouped in one cluster; often contain abnormal connection records but also some of the normal one.

B. **Applying clustering in Intrusion detection**

After the cluster’s initialization, we have applied the k-means algorithm to eliminate false positive and false negative detection by calculating the Euclidian distance and the density of each connection records.

The primary idea of k-means is to define k centroids, one for each cluster. These centroids should be placed in an intelligent and smart manner, because a different location causes different results. So, the best way to have better result is to place them as much as possible far away from each other. The coming step is to associate each point from a given cluster with the nearest centroid. We use k-means clustering to find natural groupings of similar alarm records. Any records that are apart from normal cluster indicate unusual activity that may be belonging to a new attack [9] [10].

The following figure, sum up the steps of k-means algorithm previously cited:
Algorithm 1: K-means algorithm

Input: set of connection records $X = \{x_1, x_2, ..., x_n\}$
Number of clusters calculated using similarity method: $k$
Limit of iterations: $MaxIteration$

Output: set of centroids $C = \{c_1, c_2, ..., c_k\}$
Set of labels of $C$: $L = \{l_1, l_2, ..., l_k\}$

For $c_i \in C$ such as $i \in \{1 ... k\}$ do
  $c_i \leftarrow x_i \in X$
End For

For $x_i \in X$ such as $i \in \{1 ... n\}$ do
  $d_i \leftarrow Distance(x_i, c_j) \mid j \in \{1, ..., k\}$
End For

Changed $\leftarrow$ false;
Iter $\leftarrow 0$;

Repeat
  For $c_i \in C$ such as $i \in \{1 ... k\}$ do
    UpdateCluster($c_i$);
  End For
  For $x_i \in X$ such as $i \in \{1 ... n\}$ do
    $\text{minDist} \leftarrow Distance(x_i, c_j) \mid j \in \{1, ..., k\}$
    If $\text{minDist} \neq d_i$ then
      $d_i \leftarrow \text{minDist}$;
      Changed $\leftarrow$ true;
    End If
  End For
  Iter $\leftarrow$ Iter + 1;
Until changed = true and iter $\leq$ $MaxIteration$

For $c_i \in C$ such as $i \in \{1 ... k\}$ do
  $l_i \leftarrow Labeling(c_i)$;
End for

Figure 1: k-means algorithm

After clustering phase is done, and all records are in the exact cluster, we give to each cluster a label to choose which one is normal or abnormal. For that we calculate the percentage of abnormal connection $\theta$.

IF nembre of cluster elements $\leq N \times \theta$ THEN
  Label = anomaly
ELSE
  Label = normal
END IF.

V. EXPERIMENT

After the labeling phase is done, we test the detection model built during training dataset on test dataset; this section presents the results achieved.
A. Experiment data set
KDD’99 dataset contains 41 features which are divided into 34 nominal and 7 numeric features. The 41 features are not all useful, according to [12]; only 13 features have been selected to be significant features to reduce the noise, data set dimension thus time complexity.

B. Data transformation
In this phase we convert nominal features to numeric values to make it appropriate to be applied in k-means algorithm, since k-means use only numeric values. The following table shows the transformation has been used.

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCP protocol</td>
<td>1</td>
</tr>
<tr>
<td>UDP protocol</td>
<td>2</td>
</tr>
<tr>
<td>ICMP protocol</td>
<td>3</td>
</tr>
<tr>
<td>Flags</td>
<td>4 to 14</td>
</tr>
<tr>
<td>Services</td>
<td>15 to 82</td>
</tr>
</tbody>
</table>

C. Data Preprocessing:
The initial measurements of numeric features are incoherent since the different values of it do not fit together properly, for example the values of feature “dst_bytes” ranges from 0 to more than 2290000, while the feature “same_srv_rate” only ranges from 0 to 1. Therefore, the main aim of the preprocessing phase is to normalize features values to appropriate range. Otherwise the features with great values dominate the feature with the small values, and this will affect the clustering thus the anomaly detection.

The following equations are used for the normalization phase:

- Calculate the mean absolute deviation $S_f$
  
  First we calculate the mean of each feature:
  
  $$m_f = \frac{\sum_{i=1}^{n} x_{if}}{n}$$
  
  Than the absolute deviation:
  
  $$S_f = \frac{1}{n} \left( |x_{1f} - m_f| + |x_{2f} - m_f| + \cdots + |x_{nf} - m_f| \right)$$

  Where $x_{1f}, x_{2f}, x_{3f}, \cdots, x_{nf}$ are $n$ measuring values of each feature.

- Calculate the standardized measurement:
  
  $$Z_{if} = \frac{x_{if} - m_f}{S_f}$$

D. Results
In the experiments, we compare the performance of our improved k-means with FCM and Hierarchical. The next table shows the results of initial clustering using similarity methods.

<table>
<thead>
<tr>
<th>False positive</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>30%</td>
<td>15%</td>
</tr>
</tbody>
</table>

The similarity method is able to detect a certain number of attacks but the false positive rate is still high. For this reason we apply k-means algorithm to raise the detection rate of new attacks.

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Dos (%)</th>
<th>Probe (%)</th>
<th>R2L (%)</th>
<th>U2R (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved k-means</td>
<td>97.59</td>
<td>65.61</td>
<td>90.70</td>
<td>60.55</td>
</tr>
<tr>
<td>FCM</td>
<td>80.63</td>
<td>56.0</td>
<td>80.82</td>
<td>60.07</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>75.0</td>
<td>45.47</td>
<td>60.49</td>
<td>63.75</td>
</tr>
</tbody>
</table>
We conclude from the table of comparison result the detection new attacks rate between the improved k-means, FCM and Hierarchical, that the performance of the improved k-means is better than the performance of FCM and Hierarchical in detecting Dos and R2L attacks, but for Probe and U2R attacks, the rate of detection is low compared to other category of attacks because this attacks categories act as normal behavior which make the detection difficult.

VI. CONCLUSION

One of the aims of using data-mining in intrusion detection systems is to facilitate the analyzing of great amount of data. Therefore, in this paper, we introduce k-means which is one of many data-mining techniques, but we also present the improvements done by initializing the cluster’s numbers which is one of the limits of this algorithm to make it more efficient in overcoming some lacks of the traditional intrusion detection system, and also make it more intelligent and unsupervised.

Experimental results on the KDD’99 dataset shows the efficiency of the proposed methods using k-means algorithm. Moreover the time complexity of this algorithm is low compared to other clustering algorithms, and also does not consume a lot of the machine’s resources such as memory and CPU requirement for large number of records. Our further work will be extended to improve the algorithm to augment the detection rate for probe and U2R by improving the performance of the clustering, i.e. make a deep study of the attacks and their compare it with a normal connection.

REFERENCE