Detecting Cyber Crime by Analyzing Users Data

Sheveta Vashisht, Manveer kaur, Richa Sapra, Mandeep Singh
Lovely Professional University, Phagwara
shevetavashisht@ymail.com, mkr987@yahoo.co.in, sapra.richa20@gmail.com, mandeep.13742@lpu.co.in

Abstract: There has been an increase in the number of attacks on computer system and networks. Along with the rapid popularity of the internet, cyber crime rate also got rapidly high. Indeed banks, corporations, insurance companies, casinos etc. are increasingly mining data about their employees in view of detecting cyber crime. The critical infrastructures upon which our communities, states and nation rely are increasing dependent on computer systems and networks and are thus also increasingly vulnerable to cyber attacks on them. The main objective of this paper is to construct an efficient cyber crime detection system which is adaptive to the behavior changes of the user and the data. This paper presents a security application which is a two stage cyber crime detection system and not only detect the cyber crime but also reduce the false alarm rate.

Keywords: Decision tree, Cyber Crime, Bootstrapping, Clustering, Classification.

I. Introduction
Globalization increases the electronic commerce over the internet and cyber security is emerging as one of the major concern for any organization. Cyber Crime means that the illegal activities are committed through the use of computers and the internet. Cyber Crime can basically divide into two major categories. One in those take the network as criminal object such as trespassing, destructing the network system etc. The others are those using the network to commit crime such as fraud, eroticisms, illegal trade etc.

The internet has created fertile ground for cyber crimes. Information of violence pornography, fraud can be seen everywhere on internet. According to a statistics report conducted by researchers, the most common proportion of illegal network usage cases in sequence are: Internet pornography, Internet fraud, trafficking in illicit goods, intimidation and extortion, illegal intrusion, insult and slander.

The proliferation, ubiquity and increasing power of computer technology has increased data collection storage and manipulation. As data sets have grown in size and complexity, direct hands on data analysis has increasingly been augmented with indirect, automatic data processing. This is being aided by data mining. Data mining is the process of applying various methods such as clustering, decision tree etc. to data with the intention of uncovering hidden patterns. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. Data mining proposes several classification techniques which can be effectively applied to detect fraudulent transactions [1],[2],[3],[4].

The objective of this paper is to develop a security application for an efficient cyber crime detection system by combining different classifying and clustering techniques. The proposed system is a two stage cyber crime detection system which analysis the current and the historical data to identify the anomaly using bootstrapping at first stage. In second stage to reduce the false alarm rate suspected data are checked and made sure that the detected data come under cyber crime or not.

The remainder of this paper is organized as follows. Section 2 gives the overview of the techniques used in the proposed architecture. The proposed system architecture is described in section 3. Section 4 describe the implementation process and results are discussed in section 5. Conclusion and future scope is given in section 6.

II. Overview of Techniques Used
The propose method makes use of k-means clustering algorithm and bootstrapped ID3 classification algorithm. Proposed adaptive cyber crime detection system has two stages. First stage calculates the matching score with the genuine data set and second stage detects the illegal data by comparing it with the history database.

A. Definition and notations
Let D be the database which contains N transactions’={d1,d2,d3, …..dik, ….dN}, dik are any two records in D. GD: Genuine Data set, CD: Crime Data set. GD D = D, K: Number of clusters.

1. Square error E:
It is the measure of cluster similarity, which is based on the distance between the object and the cluster mean.

2. Entropy:
It is a measure of the impurity in a collection of training samples.
3. Information Gain:
It is the impurity degree of the parent table and weighted summation of impurity degrees of the subset tables.
4. Profile Score (PS) and Deviation Score (DS):
PS is the similarity measure with the genuine data set. DS is the similarity measures with the Crime data set.

B. K-means Clustering Algorithm
This approach uses K-means clustering algorithm [5] to determine the transaction amount clusters. The number of clusters k is fixed a priori. The grouping is performed by minimizing the sum of squares of distances between each data point and the centroid of the cluster to which it belongs. First it randomly selects k of the objects, each of which initially represents a cluster mean. For each of the remaining objects, an object is assigned to the cluster to which it is most similar based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. The square error (E) is calculated as
\[ E = \sum_{i=1}^{k} \sum_{p \in c_i} (p - m_i)^2 \]
Where \( p \) is the point representing a given object; and \( m_i \) is the mean of the cluster \( C_i \).

C. Decision Tree
Decision trees are powerful and popular tools for classification and prediction. There are a variety of algorithms for building decision trees. Using any of the traditional algorithms such as C4.5, ID3 etc. construct a decision tree \( T \) from a set of training dataset. In this work ID3 is used, ID3 is based on the concept learning system algorithm. ID3 extracts the best attribute from the training set which separates the given samples. ID3 stops if the attribute perfectly classifies the training set; otherwise it recursively operates on the \( m \) partitioned subsets to get their “best” attribute. The algorithm uses the concept of entropy for the selection of the attribute. Entropy is used to determine which node to split next. If a target attribute takes on \( c \) different values, then the entropy of set \( D_1 \) is defined as
\[ \text{Entropy}(D_1) = \sum_{i=1}^{c} -p_i \log_2 p_i \]
where \( p_i \) is the probability of \( D_1 \) belonging to class \( i \). Information gain is a measure of the effectiveness of an attribute \( A \) in classifying the training data and is defined as
\[ \text{Gain}(D_1, A) = \text{Entropy}(D_1) - \sum_{v \in \text{values}(A)} \frac{|D_v|}{|D_1|} \text{Entropy}(D_v) \]
values \( (A) \) is the set of all possible values for attribute \( A \), \( D_v \) is the subset of \( D_1 \) for which attribute \( A \) has value \( v \). For each non-terminal descendant node, the process of selecting a new attribute and partitioning the training samples is now repeated. This process continues until either every attribute has already been included, or the training samples associated with this leaf node all have the same target attribute value.

D. BOAT Algorithm
BOAT is a scalable algorithm that can incrementally update a decision tree when the training dataset changes dynamically [6]. Instead of rebuilding, BOAT allows us to “update” the current tree to incorporate new training data. All other decision tree algorithm requires separate database scan for each level of the tree. But BOAT algorithm constructs several levels of the tree in a single scan over the database. Thus BOAT eliminates the speed and scalability limitations of other decision tree algorithms.

BOAT algorithm takes a sample \( D_1 \) from the training database \( D \) which fit in the memory. Now using bootstrapping technique take small samples \( S_1, S_2, S_3, \ldots, S_N \) with replacement of \( D_1 \) and construct decision trees \( ST_1, ST_2, ST_3, \ldots, ST_m \) using any of the traditional tree constructing algorithms for each of the samples. When using bootstrapping we randomly extract a new sample of \( N \) transactions out of \( m \) sampled data, where each transaction can be selected at most \( t \) times. By doing this several times, we create alternative versions of data. Increasing the number of samples can reduce the effects of sampling errors. For each node \( n \), check whether the bootstrap splitting attributes are identical; if not delete \( n \) and its sub tree in all bootstrap trees. Then combine the sample trees and find each node’s exact split point and make sure that the split point in the nodes is not outside the confidence interval. Final tree \( T \) is constructed from complete training database \( D \) in traditional way. Any difference between \( T \) and sample tree \( T_1 \) is detected, refine it to arrive at the final tree. The level of confidence can be controlled by increasing the number of bootstrap repetitions.

E. Attributes
A1- Erotic Material
A2- Number of times the proxy server is used
A3- Malicious Code Presence
A4- Password Violations
III. Proposed System Architecture
The basic motivation to the proposed approach is to improve the cyber crime detection from the dataset, since the presence of crime data is less and sparse. For this a two stage fraud detection system which combines BOAT decision tree classification and K-means clustering techniques is used. In the first stage, the current data is compared with the historical data stored in the database GD and compute the profile score PS. If any deviation from the legal data is observed, it passes to the second stage. Second stage confirms the deviation that, it is due to criminal activity or due to short term change in behavior by comparing it with the database CD and calculating the deviation score DS.

The algorithm for the proposed system is as follows:

START:
1. Select the Sample data
2. Apply the k clustering algorithm and the desired clusters are formed.
3. Now apply the attributes on the obtained clusters.
4. The decision tree now calculates the PS
5. If the PS is greater than threshold, go to 6 step else go to 7 step
6. The user is genuine user and no cyber crime is committed
7. Calculate the DS.
8. If the DS is greater than threshold then go to step 9 else to step 10.
9. User has committed cyber crime.
10. Generate the Alarm.

END

IV. Implementation Details
Implementation of the proposed cyber crime detection system has three phases, first recovery of the data, second training and testing model and third validation and analysis. Experiments are conducted on real world data of few users as well as on synthetically generated data of various users with different kind of usage behavior. The collected data is distributed into three categories, two third of genuine data GD and fraud history data FD are used for training and remaining datasets are used for prediction and cyber crime detection. The various implementation steps are

Step1: Collect the User data
Companies are not ready to share information of their employees. Therefore the performance of the proposed system has been tested on five real users’ data and on synthetically generated data by a simulator. For describing the implementation steps we selected three legal and two crime data sets which are given in table I.

<table>
<thead>
<tr>
<th>Users</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>H</td>
</tr>
<tr>
<td>U2</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>U3</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>U4</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>M</td>
</tr>
<tr>
<td>U5</td>
<td>M</td>
<td>H</td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>L</td>
</tr>
</tbody>
</table>

Step 2: Select the Training Sample
The profile size can be chosen appropriately. In the case of existing crime detection large profile size increases the accuracy but same time it increases the training time. But in the present case profile size does not affect the speed of operation of the bootstrapping ID3 classifier.

Step 3: Data Cleaning
Select only the required attributes and discard others. Sample transaction data used for training after cleaning is given in table I.

Step 4: Data Transformation
Classify the data into two cluster-cluster 0(genuine user) and cluster1 (illegal data detected) using k-means clustering algorithm.

This work selects k as two. After executing k-means algorithm on the sample transactions in table I, we get the cluster which is given in table II.

<table>
<thead>
<tr>
<th>Users</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>H</td>
<td>L</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>H</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>U2</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>U3</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>Cluster 0</td>
</tr>
<tr>
<td>U4</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>U5</td>
<td>M</td>
<td>H</td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>Cluster 0</td>
</tr>
</tbody>
</table>

From the above observation it is clear that U3 and U5 are in cluster0 and U1,U2,U4 are in cluster1. Now the users in cluster1 are under further investigation to reduce the false alarm rate. The figure 1 is the screen shot of the k-means algorithm’s output. The figure 2 shows the screen shot of the cluster distribution. The blue color represents the data that belong to cluster 0 and the
orange color represents the data belonging to cluster 1.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Cluster#</th>
<th>Full Data</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>0.207</td>
<td>0.352</td>
<td>0.0487</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>0.4413</td>
<td>0.0404</td>
<td>-0.0019</td>
<td></td>
</tr>
<tr>
<td>f</td>
<td>0.4444</td>
<td>0.0722</td>
<td>-0.0403</td>
<td></td>
</tr>
<tr>
<td>g</td>
<td>0.6211</td>
<td>0.0217</td>
<td>-0.1122</td>
<td></td>
</tr>
<tr>
<td>h</td>
<td>0.1159</td>
<td>0.1295</td>
<td>0.0719</td>
<td></td>
</tr>
</tbody>
</table>

Time taken to build model (full training data) = 0.02 second.

--- Model and evaluation on testing set ---

Clustered Instances
0 268 (76%) 1 93 (24%)

**Step 5: Training of Model using Classification**

Now when the complete data is clustered into two groups and in order to reduce the false alarm rate and for the detection of the cyber crime we pay attention on the datasets belonging to cluster 1. We train the model using the classification ID3 algorithm and the bootstrapping technique. We use the bootstrapping technique for the formulation of the decision tree and thus result is feed into the ID3 classifier. The figure 3 shows result of the classification. The accuracy obtained is 94.75% and 5.283% comes under false alarm class.

**Figure 1: Cluster Formulation.**

**Figure 2: Cluster Distribution Plot**

**V. Results**

In every model, the accuracy and the cost analysis plays an important role in the acceptance of that model for the application. It’s applicable for the proposed system as well. Table III shows main result of the implementation for the user data described in table I. Figure 5 shows the graph of the cost benefit analysis of the hard and soft classes. The graph obtained is straight line, which signifies the accuracy of the classes. The accuracy of the classification is 91.66%.

**Table III. Final Output**

<table>
<thead>
<tr>
<th>User</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>Cluster</th>
<th>Class</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>U1</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>H</td>
<td>C1</td>
<td>Soft</td>
<td>Alert</td>
<td></td>
</tr>
<tr>
<td>U2</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>C1</td>
<td>None</td>
<td>Genuin</td>
</tr>
<tr>
<td>U3</td>
<td>L</td>
<td>M</td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>C0</td>
<td>None</td>
<td>Genuin</td>
</tr>
<tr>
<td>U4</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>C1</td>
<td>Hard</td>
<td>Crime</td>
<td></td>
</tr>
<tr>
<td>U5</td>
<td>M</td>
<td>H</td>
<td>M</td>
<td>M</td>
<td>H</td>
<td>L</td>
<td>C0</td>
<td>None</td>
<td>Genuin</td>
</tr>
</tbody>
</table>

**Figure 4: Class matrix**

**Figure 5: Cost Benefit Curve**

**VI. Conclusion and Future Scope**

Cyber detection is important in today’s internet environment. The combination of facts such as the extensive growth of internet, the vast financial possibilities opening up in electronic trade, and the lack of truly secure systems make it an important field of research. An effective online cyber crime detection system should be able to discover both known and new attacks as early as possible. The detection process should be self-adjustable to allow the system to deal with the constantly changing nature of online attacks. The hybrid of the anomaly and misuse detection models can improve cyber...
crime detection and securely permit genuine transaction. This research uses a scalable algorithm for constructing decision tree incrementally for detecting the cyber crime where the training data set changes dynamically. Bootstrapping constructs several levels of the tree in only one scan over the training database, resulting in high performance gain than the existing decision tree algorithms. The accuracy of the proposed work is 94.67 % and it efficiently detects the false rate anomalies. This research focused on user level anomaly and misuse detection. In the future to achieve highly secure transaction we will extend this system for distributed level cyber crime detection also by profiling the system behavior.

References