ISSN:0975-9646

Adaptive Algorithm for Cyber Crime Detection

Manveer Kaur [1], Sheveta Vashisht [2], Kumar Saurabh[1]

[1] *Ramgarhia Institute of Technology, Phagwara*
[2]*Lovely Professional University, Phagwara*

Abstract: The technology changes so fast, but is security catching up? Will security ever drew level? The speed at which technology is moving has left massive gaps in safety. Along with the rapid popularity of the internet, cyber crime rate also got rapidly high. With the rapid development of the internet, the network structure has becomes larger and more complicated and attacking methods are more sophisticated as well. Cyber security is the major concern of every organization. Indeed banks, corporations, insurance companies, casinos etc are increasingly mining data about their employees in view of detecting cyber crime. The main objective of this paper is to construct an efficient cyber crime detection system which is adaptive to the behavior changes by combining the data mining techniques. The proposed system is a two stage cyber crime detection system which is based on the analysis of the user data in first stage and in second stage detects the false alarm.

Keywords: C4.5, Decision tree, bootstrapping, clustering, classification, bootstrapping.

I. INTRODUCTION

With the development of the information technology and the computer science, high capacity data appear in every application area .In order to help people analyze and digging out useful information, the generation and application of data mining seem so significance [1].The internet has created fertile ground for cyber crime. Information of violence, pornography fraud can be seen everywhere on the 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.

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 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[2] 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.

The main objective of this is to develop a security application for detection of cyber crime. The proposed system is two stage cyber crime detection systems which collects and analyze the data and in second stage reduce the false alarm rate.

The remained of this paper is organized as follows. Section II describes the basic idea related to the proposed work. In the section III proposed system is introduced and section IV gives the implementation details followed by result in section V. In the last section V concludes the paper.

II. BASIC IDEA

The proposed method make use of k-means clustering algorithm and C4.5 bootstrapping algorithm.Adaptive cyber crime algorithm has two stages.First age calculates the matching score with the genuine data set and second stage detect the illegal data by comparing it with the history data.

A. Definition and notations

Let D be the database which contains N transactions'={d1,d2,d3……di,dj…..dN}, di,dj 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

The k-means[3] algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers (“means”).
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters [4].

C. C4.5 Decision Tree Algorithm

Decision tree is one of the important analysis methods in classification. It builds its optimal tree model by selecting important association features. While selection of test attribute and partition of sample sets are two crucial parts in building trees. Different decision tree methods will adopt different technologies to settle these problems. Traditional algorithms include ID3, C4.5, CART, CHAID etc.

C4.5 Decision Tree Algorithm: C4.5 is the enhancement of ID3 algorithm that improves the performance. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3 using the concept of information entropy. The training data is a set $S = s_1,s_2,...$ of already classified samples. Each sample $s_i = x_1,x_2,...$ is a vector where $x_1,x_2,...$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1,c_2,...$ where $c_1,c_2,...$ represent the class to which each sample belongs [5].

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recourses on the smaller sub lists. This algorithm has a few base cases.

- All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
- None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
- Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value [6]

D. BOAT Algorithm

BOAT is a scalable algorithm that can incrementally update a decision tree when the training dataset changes dynamically [7].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 D1 from the training database D which fit in the memory. Now using bootstrapping technique take small samples S1, S2, S3, …., SN with replacement of D1 and construct decision trees ST1, ST2, ST3, ..…, STm 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 T1 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
- A5- Excess Privileges
- A6-Data Forwarding

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*

*C*ompanies 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 I:Sample Data

Users	A1	A2	A3	A4	A5	A6
U1	H	L	H	M	H	H
U2	M	H	L	M	L	M
U3	L	M	M	L	L	H
U4	H	H	M	H	L	M
U5	M	H	M	M	H	L

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 C4.5classifier.

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.

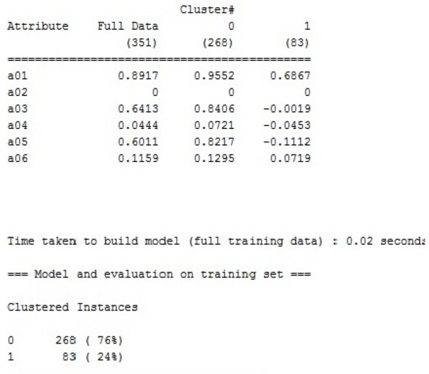
This work selects k as two. After executing k-means algorithm on the sample transactions in table I, we get the cluster which is given in table II.

Table II: Result of Clustering

Users	A1	A2	A3	A4	A5	A6	Cluster
U1	H	L	H	M	H	H	Cluster 1
U2	M	H	L	M	L	M	Cluster 1
U3	L	M	M	L	L	H	Cluster 0
U4	H	H	M	H	L	M	Cluster 1
U5	M	H	M	M	H	L	Cluster 0

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 figure1is 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.



```
                          Cluster#
Attribute     Full Data          0          1
                  (351)      (268)       (83)
=============================================
a01              0.8917     0.9552     0.6867
a02                   0          0          0
a03              0.6413     0.8406    -0.0019
a04              0.0444     0.0721    -0.0453
a05              0.6011     0.8217    -0.1112
a06              0.1159     0.1295     0.0719




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

=== Model and evaluation on training set ===

Clustered Instances

0      268 ( 76%)
1       83 ( 24%)
```

Figure 1: Cluster Formulation.

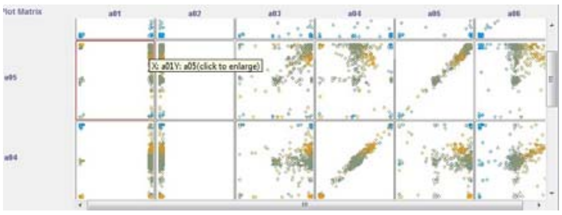
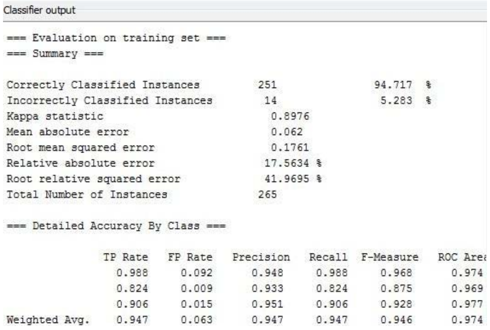


Figure 2: Cluster Distribution Plot

Step 5: *Training of Model using Classification*

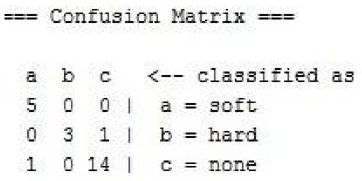
Now when the complete data is clustered into two groups and in order to reduce the false alarm rate and for the detection of the cyber crime we pay attentation on the datasets belonging to cluster 1.We train the model using the classification C4.5 algorithm and the bootstrapping technique. We use the bootstrapping technique for the formulation of the decision tree and thus result is feed into the C4.5 classifier. The figure 3 shows result of the classification. The accuracy obtained is 94.75% and 5.283% comes under false alarm class.



```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         251               94.717  %
Incorrectly Classified Instances        14                5.283  %
Kappa statistic                          0.8976
Mean absolute error                      0.062
Root mean squared error                  0.1761
Relative absolute error                 17.5634 %
Root relative squared error             41.9695 %
Total Number of Instances              265

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
                 0.988     0.092      0.948     0.988     0.968     0.974
                 0.824     0.009      0.933     0.824     0.875     0.969
                 0.906     0.015      0.951     0.906     0.928     0.977
Weighted Avg.    0.947     0.063      0.947     0.947     0.946     0.974
```

Figure 3: Classifier output

The figure 4 shows the formation of the class matrix and the content belonged to which class. Here class soft means that an alert is generated and hard class signify the detection of the cyber crime and the class none means that the user is genuine user.



```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = soft
  0  3  1 |  b = hard
  1  0 14 |  c = none
```

Figure 4: Class matrix

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 96.66%.

Table III. Final Output

User	A1	A2	A3	A4	A5	A6	Clus-ter	Class	Re-mark
U1	H	L	H	M	H	H	C1	Soft	Alert
U2	M	H	L	M	L	M	C1	None	Genuin
U3	L	M	M	L	L	H	C0	None	Genuin
U4	H	H	M	H	L	M	C1	Hard	Crime
U5	M	H	M	M	H	L	C0	None	Genuin

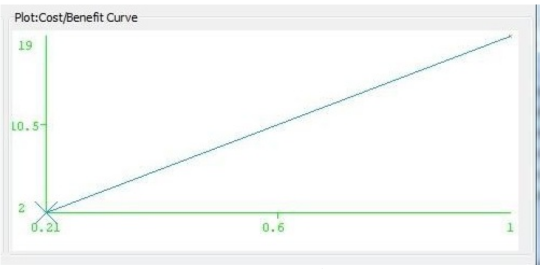


Figure 5:Cost Benefit Curve

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

[1] Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, He Peng, A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", 10th IEEE International Conference on Computer and Information Technology (CIT 2010), pp.2722-2728, 2010.

[2] Alex Berson, Stephen J. Smith, "Data Warehousing, Data Mining, & Olap" by Tata McGraw-Hill.

[3] Teknomo, Kardi ," K-Means Clustering Tutorials".

[4] http://databases.about.com/od/datamining/a/kmeans.htm

[5] http://en.wikipedia.org/wiki/C4.5_algorithm

[6]http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html

[7]JohannesGehrke_,VenkateshGanti, Raghu Ramakrishnany and Wei-Yin Lohz,"BOAT— Optimistic Decision Tree Construction", Proceedings of the 1999 ACM SIGMOD International conference on Management of data, 169-180, 1999.