Proc. of Int. Conf. on Current Trends in Eng., Science and Technology, ICCTEST

Distributed Level Cyber Crime Detection Using an Effective Method in Data Mining

Navya Shridhar[1], Lochan Gowda M[2]
Asst. Prof, Dept. Of CSE, Don Bosco Institute of Technology[1],
Asst. Prof, Dept of CSE, SJB Institute of Technology, Bangalore[2],
Email: navyamay11@gmail.com, lochangowdam@gmail.com

***Abstract*—Due to the rapid popularity of the internet, cyber crime rate also increased.Now a days, people are connected to internet with their device, and are exposed to threats. Data mining is a powerful tool to analyse and detect these crime patterns. Clustering method is extremely being used in many areas like pattern detection, data analysis and so on. The objective of this paper is to construct an efficient cyber crime detection system which fits the constantly changing the attacks.**

***Index Terms*— Clustering, Cyber crime, k-means.**

I. INTRODUCTION

Cyber crimes are the offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm to the victim directly or indirectly, using modern telecommunication networks such as Internet and mobile phones. Such crimes threaten a nation's security and financial health. Meanwhile, many organizations may be vulnerable to cyber crime based on a false sense of security, perhaps even complacency, driven by non agile security tools and processes. These crimes also create the privacy breach when confidential information is lost. Internet frauds, illegal trading, virus spreading, and cyber piracy are some of the familiar cyber crimes. Over the last decade, cyber crime rates have increased exponentially. In this era, cyber crime has transformed into a money spinner business that yields hundreds of millions of dollars and involves lesser risk than traditional crimes.

Data mining and prediction techniques usually provide useful means for crime analysis. Data mining is generally used to analyze large databases and provide meaningful results. It plays a significant role in recognizing and tracking patterns within the data. The advantage of using data mining techniques is that it extracts information from the databases which may not be known to exist.

Clustering techniques have a wide use and importance nowadays. This importance increases as the amount of data grows and the processing power of the computers increases. Clustering applications are used extensively in various fields such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing.

The main purpose of clustering techniques is to group a set of data into different groups, called clusters. These groups may be consistent in terms of similarity of its members. As the name suggests, the representative-based clustering techniques uses some form of representation for each cluster. Thus, every group has a member that represents it. The motivation to use such clustering techniques is the fact that, besides reducing the cost of the algorithm, the use of representatives makes the process easier to understand. There are many decisions that have to be made in order to use the strategy of representative - based clustering. For example, there is an obvious trade - off

Grenze ID: 02.ICCTEST.2017.1.125

between the number of clusters and the internal cohesion of them. If there are few clusters, the internal cohesion tends to be small. Otherwise, a large number of clusters makes them very close, so that there is little difference between adjacent groups. Another decision is whether the clusters should be mutually exclusive or not, that is, if an entity can co-exist in more than one cluster at the same time.

K- means clustering algorithm

K-means clustering algorithm is a method of cluster which groups the data sets into k clusters. It is one of the simplest algorithms that solve the well known clustering problem. The algorithm groups observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based on the observation's nearness to the mean of the cluster. The cluster's mean is then recomputed and the process repeats again.

Here is the working:

1. The algorithm arbitrarily selects k points as the initial cluster centers.
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.

Limitations of k means algorithm:

- It does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
- Sensitive to the selection of initial cluster center.
- There is no evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.

The further algorithms we have discussed in the below section are based on k-means algorithm.

Advanced K- means

Advanced k-means algorithm hires a cluster centre. Choosing a good initial cluster centre will improve solutions and reduce execution time. It randomly choose some subsets within equal number of samples from large data sets. Secondly, partitional algorithm is applied to each subsets to get each center sets of the subsets. Thirdly, gather these center sets and apply the partitional algorithm again to obtain the most proper center set. To get a more initial points, we repeat the partitional algorithm 2 times by fewer sample sets. Finally run the partitional algorithm with the most feasible center set as the initial seeds and original large data sets.

Limitations of advanced k-means

- It doesnot deal with large data sets and it will cost more time.
- It deals with the problem of assigning number of desired cluster as input.

K-means++

The k-means++ algorithm includes few types of clustering algorithms, this method overcome some of the problems associated with the definition of initial cluster-centres for k-means clustering. Instead of updating cluster centroid after all points have been assigned to a cluster, the centroid can be updated incrementally, after each assignment of a point to a cluster. It requires zero or two updates to cluster centroids at each step.

Limitation of k-means++

- Often updating centroid incrementally introduces an order dependency. In other words, the clusters produced may depend on the order in which the points are processed. Although this can be addressed by randomizing the order in which the points are processed.
- Incremental updates are slightly more expensive.

K–medoid

k-medoids algorithm is a clustering algorithm relates to k-means algorithm and the medoid shift algorithm. k-medoid algorithm breaks up the dataset into groups and attempts to reduce the distance between points in a cluster and a point designated as the center of that cluster. Unlike k-means, k-medoids chooses datapoints as centers (medoids) and works with an arbitrary metrics of distances between datapoints. Medoids are more resistant to outliers and noise.

Limitations of k-medoid

- The k-means method is based on the centroid techniques to represent the cluster and it is sensitive to outliers.
- Data object with an extremely large value may disrupt the distribution of data.

II. PROPOSED SYSTEM

As we got to know that each flavours of k-means has its own limitations. There is a need of an efficient clustering algorithm to detect the cyber crime that deploys in a distributed environment.
It would be a great idea, if we combine all the above clustering algorithm advantages to make an efficient clustering algorithm. Analysis of mass data sets in distributed system is another approach of this system. The most popular distributed processing tool "hadoop" employs to provide scalability to the system. Deploying distributed system is not only the motto. Also to solve the security issues in distributed environment.
Hadoop is a popular tool to distributed system for data storage which is highly scalable. It is due to its file system called Hadoop Distributed File System (HDFS). Hadoop can run application on system with thousands nodes. It distributes files among nodes and allow system to work even though the node failure occur. This approach reduces the risk of system failure.
MapReduce is a software for distributed processing of large data sets on the computing nodes. MapReduce process large scale data records in clusters. This programming model is based on two function which are map() and reduce() function Map function perform task as master node which takes input and divide into smaller sub modules and distributed among slave nodes. Slave node further divide sub modules that again lead to hierarchical tree structure. Slave node process base problem and passes result to master node. The Map Reduce arranges all intermediate nodes together and then send to reduce function for producing final output. Reduce function collect all intermediate node result and combines and put together to form output.

III. CONCLUSION

Varieties of k-means algorithm is popular and advantageous but some of its limitations makes it somewhat difficult. There is a need of efficient algorithm to overcome the drawbacks of these varieties of k means algorithm. To handle the scalability issue, we can extend the system for distributed level cyber crime detection using hadoop and map reduce over large data set. We intend to provide security to our distributed detection system.

REFERENCES

[1] Adaptive Algorithm for Cyber Crime Detection, Manveer Kaur et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,4381 – 4384
[2] Distributed Document Clustering Using K-Means, Volume 4, Issue 11, November 2014 ISSN: 2277 128X.
[3] Advanced Methods to Improve Performance of K-Means Algorithm, Volume 12 Issue 9 Version 1.0 April 2012.
[4] An Intelligent Document Clustering Approach to Detect Crime Patterns, The 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013).
[5] Evolving limitations in K-means algorithm in data mining and their removal, IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011.
[6] Cyber Crime and Security- Challenges and Security Mechanisms, International Journal of Engineering Trends and Technology (IJETT) – Volume 36 Number 7- June 2016.
[7] Adaboost and SVM based cybercrime detection and prevention model, Artificial Intelligence Research, December 2012.