International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection

Jithin Mathew[1], S. Ajikumar[2]

[1]PG Scholar, Department of M.Sc(Software Engineering), PSN College of Engineering & Technology, Tirunelveli, Tamilnadu, India

[2] Research Supervisor, Department of M.Sc(Software Engineering), PSN College of Engineering & Technology, Tirunelveli, Tamilnadu, India

ABSTRACT

An intrusion detection system is software that monitors a single or a network of computers for malicious activities that are aimed at stealing or censoring information or corrupting network protocols. Most technique used in today's intrusion detection system are not able to deal with the dynamic and complex nature of cyber-attacks on computer networks. Even though efficient adaptive methods like various techniques of machine learning can result in higher detection rates, lower false alarm rates and reasonable computation and communication cost. With the use of data mining can result in frequent pattern mining, classification, clustering and mini data stream. This survey paper describes a focused literature survey of machine learning and data mining methods for cyber analytics in support of intrusion detection. Based on the number of citations or the relevance of an emerging method, papers representing each method were identified, read, and summarized. Because data are so important in machine learning and data mining approaches, some well-known cyber data sets used in machine learning and data mining are described for cyber security is presented, and some recommendations on when to use a given method are provided.

Keywords : Local Area Network, Wide Area Network, Metropolitan Area Networks, Close Circuit Television, Security through Obscurity GPS, Global Positioning System, Point Of Access, Network Intrusion Detection System

I. INTRODUCTION

Recommendation The Machine learning, Data Mining methods are described, as well as several applications of each method to cyber intrusion detection problems. The complexity of different machine learning and data mining algorithms is discussed, and the paper provides a set of comparison criteria for machine learning and data mining methods and a set of recommendations on the best methods to use depending on the characteristics of the cyber

Problem to solve Cyber security is the set of technologies and processes designed to protect computers, networks, programs, and data from attack, unauthorized access, change, or destruction. Cyber security systems are composed of network security systems and computer security systems. Each of these has, at a minimum, a firewall, antivirus software, and an intrusion detection system .Intrusion detection system s help discover, determine, and identify unauthorized use, duplication, alteration, and destruction of information systems. The security breaches include external intrusions attacks from outside the organization and internal intrusions.

There are three main types of cyber analytics in support of intrusion detection systems: misuse-based, anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel attacks. Anomaly-based techniques model the normal network and system behaviour, and identify anomalies as deviations from normal behaviour. They are appealing because of their ability to detect zero-day attacks. Another advantage is that the profiles of normal

activity are customized for every system, application, or network, thereby making it difficult for attackers to know which activities they can carry out undetected. Additionally, the data on which anomaly-based techniques alert can be used to define the signatures for misuse detectors. The main disadvantage of anomaly-based techniques is the potential for high false alarm rates because previously unseen system behaviours may be categorized as anomalies.

This paper focuses primarily on cyber intrusion detection as it applies to wired networks. With a wired network, an adversary must pass through several layers of defence at firewalls and operating systems, or gain physical access to the network. However, a wireless network can be targeted at any node, so it is naturally more vulnerable to malicious attacks than a wired network. The Machine learning and data mining methods covered in this paper are fully applicable to the intrusion and misuse detection problems in both wired and wireless networks. The reader who desires a perspective focused only on wireless network protection is referred to papers such as Zhang et al. , which focuses more on dynamic changing network topology, routing algorithms, decentralized management, etc.

II. METHODS AND MATERIAL

Related Work

The authors SongnianLi, Suzana Dragicevic, et al. in [6] made review on various geospatial theory and methods used to handle geospatial big data. Given some special attributes, authors considered that customary data taking controlling methodologies and techniques are lacking and the following domains were recognized as in requirement for further advancement and examination in the control. This incorporates the advancements in calculations to manage real-time analytics and to support ongoing flooding data, as well as improving new spatial indexing techniques. The improvement of theoretical and methodological ways to deal with transfer of big data from illustrative and parallel research and applications to ones that investigates easygoing and illustrative connections.

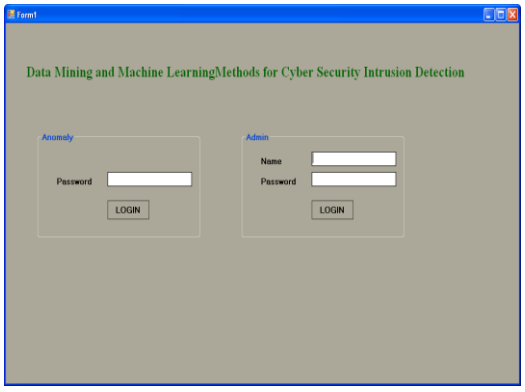
In [13] Yuehu Liu, Bin Chen et al. have proposed another technique for overseeing gigantic remote sensing image data by utilizing HBase and MapReduce framework. At first they have divided the actual image into various tiny pieces, and store the blocks in HBase, which is dispersed in a gathering of hubs. They have used MapReduce programming model on handling the stored blocks, which can be simultaneously executed in a group of hubs. The hubs in Hadoop cluster have no prerequisites for high performance and accuracy so that they can be exceptionally inexpensive. Besides, as a result of the high versatility of Hadoop, it is anything but difficult to add new hubs to the cluster, which was normally exceptionally troublesome in general ways. Finally they notice that the speeds of data commerce and processing increase because the cluster of HBase grows. The outcomes demonstrate that HBase is extremely reasonable for large image information stockpiling and handling.

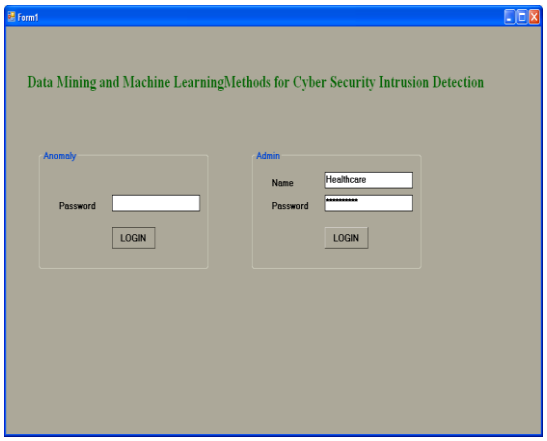
The authors Chaowei Yang, Michael Goodchild et al. in [14] have projected a replacement paralleling storage and access methodology for big scale NetCDF scientific information that is enforced dependent on Hadoop. The recovery technique is actualized dependent onMapReduce. The Argo data is utilized to exhibit the proposed technique. The execution is looked at under a disseminated domain taking into account PCs by utilizing distinctive data scale and diverse task numbers. The examinations result demonstrates that the parallel strategy can be utilized to store and retrieve the vast scale NetCDF productively.

Big data has turned into a noteworthy center of worldwide interest that is progressively pulling in the acknowledgment of the educated community, industry, government and other association. The incremental development in volume and changing

III. RESULTS AND DISCUSSION

The implementation results can be shown as figure below





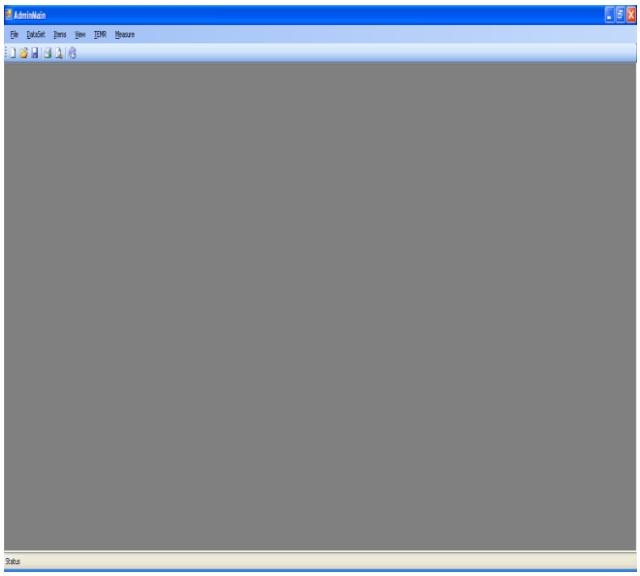
Envisioning and checking the nature of data. There are wide assortments of methods accessible and adjusted to imagine, dissect, control and composite big data to make this sort of data volume reasonable. Some of these procedures are data fusion, cluster analysis, network analysis, crowd sourcing, Association rule learning, machine learning and etc. In this section we have covered some of these techniques and their challenges briefly.

A. Data Fusion

Conventional data processing sometimes consider data from one domain. In this big data era, everyone has to make wide selection of datasets from totally different sources in several domains. Each of these datasets comprise of various strategies such as alternate representation, measurements, scale, dissemination, and consistency. Removing the force of information from numerous diverse (however conceivably associated) data sets is an extraordinary arrangement in big data research, which incorporates basically isolating big data from customary data mining undertakings. Which itself prompts propelled procedures that can comb data fusion and conventional data fusion contemplated in the database group [10].

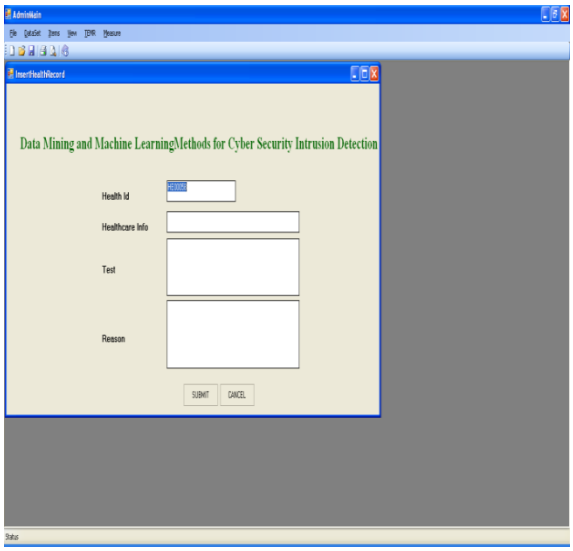
B. Crowdsourcig

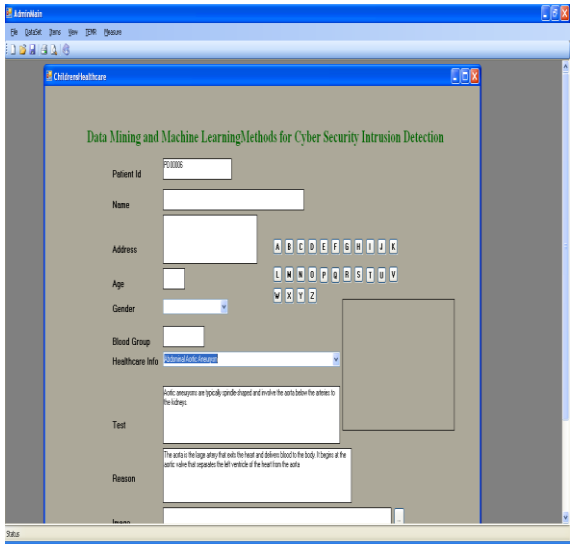
The term crowd sourcing means to data acquirement by vast and various gatherings of individuals, who much of the time are not prepared measurer and who don't have exceptional PC learning, utilizing web innovation. In this way, these information are exchanged to and put away in a typical computer architecture e.g. a focal or a combined database, or in a distributed computing environment. The ensuing undertaking of programmed data incorporation and handling are vital to produce additional data.



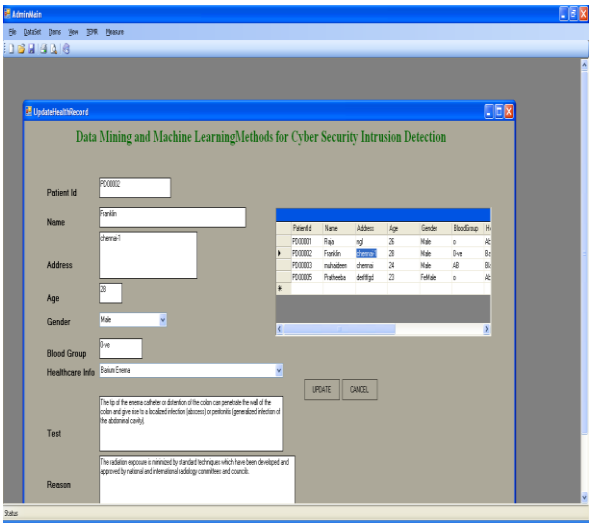
An assortment of data mining techniques can be applied to find associations and regularities in data, extract knowledge in the forms of rules and predict the value of the dependent variables. Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), Bagging algorithm, K- nearest neighborhood (KNN), Support vector machine (SVM) etc. Data mining is an important step of knowledge discovery in databases (KDD) which is an iterative process of data cleaning, integration of data, data selection, pattern recognition and data mining knowledge recognition. KDD and data mining are also used interchangeably. Data mining encompasses association, classification, clustering, statistical analysis and prediction. Data mining has been widely used in areas of communication, credit assessment, stock market prediction, marketing, banking, education, health and medicine, hazard forecasting, knowledge acquisition, scientific discovery, fraud detection, etc but data mining holds significant presence in every field of medical for the diagnosis of several diseases such as diabetes, skin cancer, lung cancer, breast cancer, heart disease, kidney failure, kidney stone, liver disorder, hepatitis etc. Data mining applications include analysis of data for better policy making in health, prevention of various errors in hospitals, detection of fraudulent insurance claims early detection and prevention of

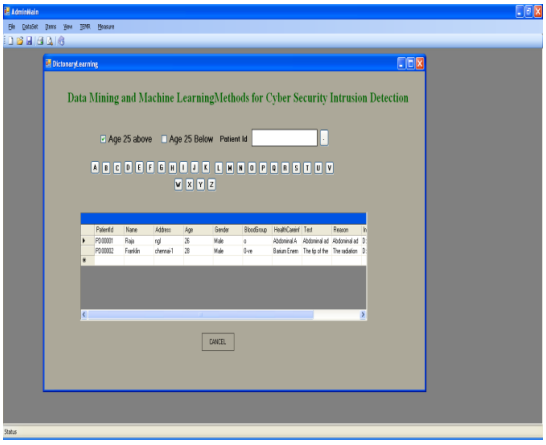
various diseases, value for more money, saving costs and saving more lives by reducing death rates.

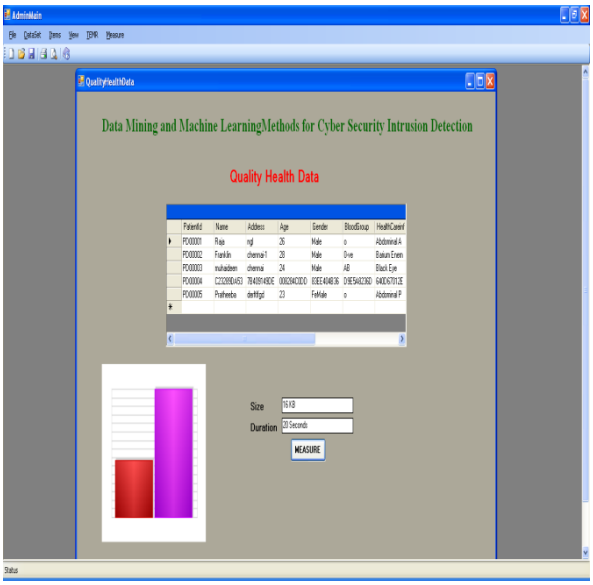




With a wired network, an adversary must pass through several layers of defence at firewalls and operating systems, or gain physical access to the network. However, a wireless network can be targeted at any node, so it is naturally more vulnerable to malicious attacks than a wired network. The Machine learning and data mining methods covered in this paper are fully applicable to the intrusion and misuse detection problems in both wired and wireless networks. The reader who desires a perspective focused only on wireless network protection is referred to papers such as Zhang et al. , which focuses more on dynamic changing network topology, routing algorithms, decentralized management, etc.







The atmosphere science assumes critical part in investigating and extemporizing individuals' living surroundings and shielding from catastrophic event too. NetCDF has been broadly utilized as a part of physical, marine and air sciences [14].It is applicable to many more fields in future because of its unified data format. As there is a fast increment in

information scale, parallel access of NetCDF information got to be one of the prompt interests. Map Reduce based method for parallel access and storage of massive NetCDF data are more efficient. At the point when contrasted with other parallel programming models like MPI, MapRedce standard manages parallel access of information consequently by performing two critical operations, for example, Map and Reduce.

IV.CONCLUSION

In proposed work the prediction and prevention of various medical diseases is done using PCA, Canny edge operator along with some pre- processing and post- processing steps. Firstly edge detection is done then feature extraction is done to get the optimized no. of feature to classify between infected and non-infected diseases. Following steps will be followed to get the proposed disease prediction model. The proposed system has been fully implemented (in matlab 2010) and tested with real CT scan images. The objective is to support efficient image data processing and feature extraction. Obviously, to deal with real image data, the image processing tool must possess important characteristics such as being noise- tolerant, efficient, practical, and convenient to use. The aim of this research was to detect features for accurate images. An assortment of data mining techniques can be applied to find associations and regularities in data, extract knowledge in the forms of rules and predict the value of the dependent variables. Common data mining techniques which are used in almost all the sectors are listed as: Naive Bayes, Decision Tree, Artificial neural network (ANN), Bagging algorithm, K- nearest neighborhood (KNN), Support vector machine (SVM) etc. Data mining is an important step of knowledge discovery in databases (KDD) which is an iterative process of data cleaning, integration of data, data selection, pattern recognition and data mining knowledge recognition. KDD and data mining are also used interchangeably. Data mining encompasses association, classification, clustering, statistical analysis and prediction. A steeper Subthreshold Slope (SS) is obtained compared to conventional CMOS, because of the better electrostatic control and absence of doping. Besides the reduction of the leakage current, the multigate topology of the FinFET also increases the drain-source saturation current of the device with a factor two at the same bias condition [3]. In very thin (or narrow) multigate devices, such as a FinFET, volume inversion takes places. In volume inversion charge carriers are not confined near the (Si-SiO_2) interface, but throughout the entire body of the device. Therefore the charge carriers experience less interface scattering. As a result an increase of the mobility and transconductance is expected in multigate devices. The multiple gate structure of the FinFET reduces the short channel effects. To further improve the control over the channel.

V. REFERENCES

[1]. Zhenlong Li, Chaowei Yang, Baoxuan Jin, Manzhu Yu, Kai Liu, Min Sun, Matthew Zhan, "Enabling Big Geoscience Data Analytics with a Cloud-BasedMapReduce-Enabled and Service-Oriented Workflow Framework", Research Article, Plos One, DOI:10.1371/journal.pone.0116781 March 5, 2015

[2]. Duffy DQ, Schnase JL, Thompson JH, Freeman SM, CluneTL,"Preliminary Evaluation of MapReduce for High- Performance Climate Data Analysis", NASA new technology report white paper, 2012.

[3]. Santiago A.Nunes, Luciana A.S. Romani, Ana M.H. Avila, "Analysis of Large Scale Climate Data: How Well Climate Change Models and Data from Real Sensor Networks Agree?", 22nd international conference on world wide web,New York, USA, pp.517-526,ACM,ISBN:978-1-4503-2038-2,2013.

[4]. Yang C, Goodchild M, Huang Q, Nebert D, Raskin R, "Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?", International Journal of Digital Earth, pp. 305-329,Vol. 4, No. 4, July 2011.

[5]. Vatika Sharma, Meenu Dave, "SQL and NoSQLDatabases", International Journal of Advanced Research in Computer Science and Software Engineering,pp. 20-27, volume 2, Issue 8, august2012, ISSN:2277 128X.

[6]. Songnian Li, SuzanaDragicevic ,FrancescAntón Castro , Monika Sester, Stephan Winter,ArzuColtekin, Christopher Pettit, "Geospatial big data handling theory and methods: A review and research challenges",

ISPRS Journal of Photogrammetry and Remote Sensing, pp. 119–133, Volume 115, May 2016.

[7]. Tong Zhang, Jing Li , Qing Liu , Qunying Huang, "Cloud- Enabled Remote Visualization Tool for Time Variant Climate Analytics", journal of Environmental Modelling&Software,Science Direct, pp. 513–518, Volume 75, January 201

[8]. GemaBello-Orgaza,JasonJ.Jungb, DavidCamacho, "Social big data: Recent achievements and new challenges", Journal of Information Fusion, ScienceDirect, pp. 45–59,Volume 28, March 2016.

[9]. Stefano Nativi , Paolo Mazzetti , Mattia Santoro , FabrizioPapeschi , Max Craglia ,Osamu Ochiai, "Big Data challenges in building the Global Earth Observation System of Systems", Journal of Environmental Modelling & Software,ScienceDirect,pp. 1–26, Volume 68, June 2015.

[10]. Yu Zheng, "Methodologies for Cross-Domain Data Fusion: An Overview", IEEE Transactions on Big Data, pp. 16 – 34, Volume: 1, Issue: 1, TBD-2015-05-0037, March 2015.

[11]. Yu Zheng, "Crowdsourcing geospatial data", ISPRS Journal of Photogrammetry and Remote Sensing, ScienceDirect,pp.550– 557, Volume 65, Issue 6, November 2010.