Int. Jour. of App. Sc. and Eng. (IJASE) **2**(2) : Dec 2014, 117-125

Advances in intrusion detection systems with applications to data mining

Sandeep Kumar

Haryana College of Technology and Management, Kaithal

Corresponding author: deepumehla56@gmail.com

Abstract

With the introduction of smart phones with advanced computing and storage capabilities users experience novel kind of security threats. Conventional preventive mechanisms like encryption, authentication alone don't seem to be enough to produce adequate security for a system. So, we tend to need sensible Intrusion detection systems which will improve security and substantially reduce the cellular phones computing resources. In this work we tend to plan an intrusion detection procedure that with efficiency detects intrusions in mobile phones with an application to Data Mining. To remove overhead of processing from the mobile phones we used network based approach. We build a neural network classifier that may be trained for every user using its call logs. Application will run on phone of the user and collects concerned data of the user and sends them over to the remote server. Results indicate the effectiveness of our methodology to observe intrusions and outperformed existing Intrusion detection strategies with about ninety five percent detection rate.

Keywords: Overhead, strategies, intrusion, sensible.

The advanced cell phone communication gadgets like touch/smart cellular phones are ever changing the manner in which we tend to communicate with others and use it as data repository. The evolution started from simple land line phones to cell phones and then to hi-tech compact hand held computers. Today they are not just being used as voice communication devices. In addition to being used for browsing internet these devices can receive and send email, send and receive multimedia messages, send and receive data by connecting to other electronic devices. They are also furnished with operating system, text editors, spreadsheet editors, database processors and many other useful applications. The capabilities of mobile devices is ever evolving, their usage for process and store important information tend to substantially increase. Presently more than 350 million users use smart cellular phones and this figure likely to increase to more than a billion in coming one or two years that's about one sixth of the population of world and at par with the population of Indian sub-continent.

Wicked usage, attacks and damages have been on the rise as more and more cell phones particularly smart phones are put into use. The attacks on the Internet have become both more creative and easier to execute because of the ubiquity of the Internet and the popularity of easy-to-use operating systems and development environments. There are several penetration points through which intrusions may be introduced in a network. For instance, malicious created IP packets can break down a client; even susceptible operating system software can be misused to create an invalid root shell. These security pressures have exploited all kinds of networks ranging from traditional computers to point-to-point and distributed networks. These security threats have also demoralized the vulnerable protocols and operating systems extending attacks to operating system on various kinds of applications, such as database and web servers. Almost all the popular operating systems regularly publish updates, but due to the poorly administered machines, uninformed users, a majority of targets and ever-present software bugs has allowed exploits to remain ahead of patches.

As these devices will permit third party applications to run on them they are prone to various threats/attacks like viruses, malware, worms and Trojan horses. In addition to that a mobile device can start communication to other devices through any of its interfaces and also communicate to wide variety of wireless networks. The intrusion prevention methods such as encryption, authentication alone cannot provide complete security of the system. Also existing desktop based Intrusion Detection applications may not be good for mobile systems by virtue of the memory and battery consumption. Therefore, we need to explore with existing intrusion detection systems by not only enhancing the security of these devices but also shifting the processing cost from host to network.

Literature Survey

Although the works in intrusion detection systems (IDS) started in the early 80's but major work was started in mid and late 1990s along with the explosion of the Internet. In the beginning research was in the field of intrusion detection often focused on host-based solutions, but the radical growth of networking changed the later efforts to be concentrated on network-based systems. The tools and techniques focused here was reflected a core of active research that had happened in the last two decades. Several studies have indeed been published in the past [1, 2, 3, 4, 5, 6], but the growth of IDSs has been such that a lot of IDSs have appeared in the meantime.

James P Anderson is recognized as the first person to document the need for automated audit trail review to support security goals for the US Department of Defense in 1978. He authored the Reference Monitor concept *Computer Security Technology Planning Study 2* in a planning study for US Air Force and this report is considered to be the seminal work on IDS. Anderson also authored a paper *Computer Security Threat Monitoring and Surveillance* [7] in 1980 and this is generally considered to be the first real work in the area of intrusion detection. The

research papers recommend taxonomy of classifying internal and external threats to computer systems. He points out that when a violation occurs, in which the attacker attains the highest level of privilege, such as root or super user in UNIX, there is no reliable remedy. He also focused on the problems associated with masqueraders for which he proposes that some sort of statistical analysis of user behavior, capable of determining unusual patterns of system use, might represent a way of detecting masqueraders. This was tested in the next milestone in IDS, the IDES project. The US Navy's Space and Naval Warfare Systems Commands (SPAWARS) in 1984 financed a project to research and develop a model for real-time IDS and *Dorothy Denning* and *Peter Neumann* came up in 1988 with the Intrusion Detection Expert System (IDES) model. The uncommon or unusual traces of traffic were referred to as *anomalous* and the assumptions made in this project served as the basis for many intrusion detection research and system prototypes of the late 1980s. The IDES model is based on the use statistical metrics and models to describe the behavior of caring users. The IDES prototype used hybrid architecture, comprising an anomaly detector and an expert system. The anomaly detector used statistical techniques to characterize abnormal behavior. The expert system used a rule-based approach to detect known security violations. The expert system was included to mitigate the risk that a patient intruder might gradually change his behavior over a period of time to defeat the anomaly detector. This situation was doable as a result of the anomaly detector tailored to gradual changes in behavior to reduce false alarms. Denning's analysis paper on AN Intrusion Detection Model [8] in 1986 illustrates the model of a period intrusion-detection skilled system capable of detection break-ins, penetrations, and different styles of laptop abuse. The model relies on the hypothesis that security violations may be detected by observance a system's audit records for abnormal patterns of system usage. The model includes profiles for representing the behavior of subjects with regard to objects in terms of metrics and applied math models, and rules for effort data concerning this behavior from audit records and for detection abnormal behavior. The model is freelance of any specific system, application surroundings, system vulnerability, or variety of intrusion, thereby providing a framework for a general purpose intrusion-detection skilled system. My paper is considered to be the stepping-stone for all the further works in this field. In the following years, an ever-increasing number of research prototypes are explored. Several of these efforts will be looked at in brief and more details are available in [9].

In the year 1984, the US Navy's SPAWARS funded a research project *Audit Analysis* at Sytek and the proto-typed system utilized data collected at shell level of a UNIX machine running in a research environment. In 1985 an internal research and development project named *Discovery* started at TRW and this monitored the TRW's online credit database application and not the operating system for intrusions and misuse. Almost the same time, *Multics Intrusion Detection and Alerting System (MIDAS)* was developed by the National Computer Security Center to monitor NCSC's Dockmaster system, which is a highly secure operating system. The MIDAS was designed to take data from Dockmaster's answering system audit log and used a hybrid

analysis strategy, combining statistical anomaly detection with expert system rule-based approaches. In 1989, *Wisdom and Sense* from Los Alamos National Laboratory and *Information Security Officer's Assistant (ISOA)* from Planning Research Corporation were developed. In the year 1990 Kerr and Susan reported all the experimental as well as actually implemented IDSs in the Datamation report titled *Using AI to improve security.* In the same year, an audit trail analysis tool *Computer Watch* was developed by AT&T and was designed to consume operating system audit trails generated by UNIX system. An expert system was used to recapitulate system security relevant events and a statistical analyzer and query mechanism allowed statistical characterization of system-wide events. *Network Audit Director and Intrusion Reporter (NADIR)* was developed in 1991 by the Computer division of Los Alamos National Laboratory to monitor user activities on the Integrated Computing Network (ICN) at Los Alamos. NADIR performs a combination of expert rule-based analysis and statistical profiling. The US Air Force in 1992 funded the research for the *Distributed Intrusion Detection System (DIDS)* [10]; a major initiative to integrate host and network based monitoring approaches. Until the year 1990, intrusion detection systems were mostly host-based and then in1990 NSM extended intrusion detection to the network environment. USTAT, real-time IDS for UNIX [11] was introduced in 1993 by Ilgun and Koral. USTAT is a state-transition analysis tool for UNIX. This is a UNIX specific implementation of a generic design STAT, state-transition analysis tool. Helman and Paul in 1993 came up with a paper on *Statistical foundations of audit trail analysis for the detection of computer misuse* where the modeling of computer transactions is done, as generated by two stationary stochastic processes, the normal process and the misuse process. In the year 1994, Crosbie and others advocated the use of *autonomous agents* in order to improve the scalability, maintainability, efficiency and fault tolerance of an Intrusion Detection System [12]. Christoph and Gray in 1995 has expanded NADIR to include processing of audit and activity records for the Cray UNICOS operating system and called *UNICORN: misuse detection for UNICOS* [13]. Kosoresow and Hofmeyr in 1997 published a paper on *Intrusion Detection via System Call Traces* [14]. In the year 1998, Anderson and others offered an innovative approach to intrusion detection, by incorporating information retrieval techniques into intrusion detection tools. Huang and others in the year 1999 introduces a large scale distributed ID architecture based on IDS agents and collaborative attack strategy analysis which creates an opportunity for IDS agents to pro-actively look ahead for data most pertinent to current case development. In the year 2000, Ning et al. presented the research paper on *Modelling requests among cooperating IDSs* [15]. Luo and otehrs in 2000 published their paper on *Mining fuzzy association rules and fuzzy frequency episodes for Intrusion Detection* [16]. Luo and others published another paper on *Fuzzy frequent episodes for real-time intrusion detection* in 2001. Data mining methods including association rule mining and frequent episode mining have been applied to the intrusion detection problem. In the year 2002, Mukkamala and Srinivas suggested the use of neural networks and support vector machines in intrusion detection. Their research paper on *Intrusion detection using neural networks and support vector machines* describes these approaches

to intrusion detection and also compares the two methods. Also in year 2002 Krugel and others presented *a* paper *Service specific anomaly detection for network intrusion detection*. Kemmerer and Richard in 2003 presented a paper on *Internet security and intrusion*

detection which highlights the principal attack techniques that are used in the Internet today and possible countermeasures. In particular, intrusion detection techniques are analyzed in detail. In the year 2003, Ling and Jun in the paper *Novel immune system model and its application to network intrusion detection* analyzes the techniques and architecture of existing network Intrusion Detection Systems, and probes into the fundamentals of Immune System (IS), a novel immune model is presented and applied to network IDS, which is helpful to design an effective IDS. In the year 2003, Xiang and others in their paper *Generating IDS attack pattern automatically based on attack tree* illustrate the generation of attack pattern automatically based on attack tree. Ye and Nong in the year 2004 had a paper on *Robustness of the Markov-chain model for cyber-attack detection*. This paper presents a cyber-attack detection technique through anomaly-detection, and discusses the robustness of the modeling technique employed. Xu and Ming in the paper *Two-layer Markov chain anomaly detection model* in 2005 propose, on the basis of the current single layer Markov chain anomaly detection model, a new two-layer model. Zhao et al. [17] in 2005 have proposed a misuse detection system and anomaly detection system that encode an experts knowledge of known patterns of attack and system vulnerabilities as if-then rules. In the year 2009, [18] published a paper discussing about enhancement of IDS using sensor fusion techniques.

In this research endeavor I am trying to focus on designing a good IDS using data mining techniques by using network based approach which substantially eases the overhead of processing from the cell phone system and find out occurrence of intrusion if any.

Architecture of Proposed System

Our proposed a novel Data Mining method for observing intrusion in cellular phones is shown in figure 1.1. The worms which will create calls from the cellular devices with no knowledge to the user of that device, they may send text messages/make calls involving cash transfer requests. Our planned systems are helpful in finding those attacks and can send an alert instantly to the user rather than waiting for a week or two. We have a tendency to really collect the decision logs of the user to train that user's classifier. The classifiers are trained using call logs of the user. At a specified time of every day this app will upload the logs to the server from his cellular device. In this way call logs are used to already trained classifier. The previous classifier is also aware of the normal call logs. In this way finally trained classifier scrutinizes the uploaded logs and detects and abnormality. In case of any abnormality client is altered through emails or text messages. A mechanism is introduced by which only valid abnormalities are considered and are used for raising any alarms.

This Network based mostly approach can take away the overhead of process from the cellular

smart phones that saves the limited battery power and therefore the computing power for other functions. In this work we have evaluated Naïve-Bayes neural network classifier and SVM Classifier to classify the decision logs to decide which classifier offers optimum performance. Weka, an open source application a knowledge Mining tool is employed for the aforementioned purpose. We collected logs of 4 users of various sizes to assess the scale of the dataset on the performance. We have used Ten-Fold cross validation to calculate the performance of the classifier.

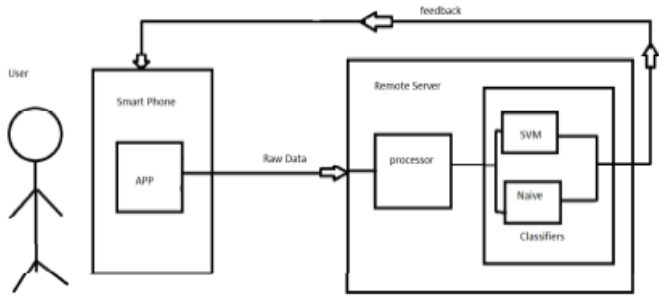

Figure1.1 Architecture of Proposed System

The Experiment

Out of many attributes from the call logs, we've thought of the subsequent four attributes that we have undertaken and will be sufficient for the classification.

1. Day of the week (DOW) i.e., Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday
2. Length of Call (LC)-in minutes
3. Nature of the call (NC)-Interstate, Intrastate, International
4. Call Timing (CT)-Morning, Afternoon, Evening, Night, Mid-Night

The attribute DOW points to the actual day of the week when the call log was created, LC indicates you the length of the call in minutes, NC tells you whether a call is within the state, out of state, or international, CT is used to find at what time of the day a call log was made.

After cleaning and preprocessing the call logs dataset, we then tend to add some known fraud information into this dataset and so build models for every user using Naïve-Bayes and SVM. Each Naïve-Bayes classifier and SVM is used for classification to see which classifier yields higher performance. We tend to introduce some more known fraud data into the existing logs while training the classifier. By accomplishing several experiments on both SVM and Naïve-Bayes we tend to get the following results.

In this paper ten-fold cross validation is employed within the field of machine learning to

work-out how accurately a learning algorithm program are ready to predict information that it absolutely was not trained on. The training data-set is arbitrarily divided into ten sub-parts, first nine sub-parts are employed for training the classifier in question and the last tenth sub-part is employed as testing dataset. This method is repeated till all sub-parts are covered and the performance is calculated for the total of all the ten sub-parts.

In these experiments the outcomes are given label either as positive (YES) or negative (NO) category. There are four likely out-comes in our experiment. If the out-come from this experiment is YES and the original value is also YES, then it is known as a true positive (TP); however if the actual result ids NO then it is known as false positive (FP). Whereas, a true negative (TN) has happened when the calculated out-come and the actual out-come are NO, and false negative (FN) when the observed outcome is NO while the actual value is YES.

For ROC curve, we need solely true positive rate and false positive rate. True positive rate indicates a classifier is successful in classifying positive cases correctly in all the positive samples existed during the experiment. False positive rate indicates how many in-correct positive out-comes happen in all the negative results existed while performing the experiment. False positive rate and True positive rate are computed by using equations given below.

TPR (True Positive Rate) = TP / (TP + FN)

FPR (False Positive Rate) = FP / (FP + TN)

The best prediction method in the ROC curve is indicated by a point in the upper left-corner or by the coordinate (0, 1). This indicates 100% sensitivity and specificity that means no false negatives and false positives respectively. This curve is used to examine the performance of the classifier for each user. Figure 1.2 represents the ROC curve for all the users using Naïve-Bayes.

The results of Naïve-Bayes classifier is shown in Table 1 and SVM in Table2.

Table 1. The performance of Naïve-Bayes classifier

User	Percentage of Correctly Classified Instances	Percentage of Incorrectly Classified Instances
User-I (290 Instances)	89.5833	10.4167
User-II (578 Instances)	91.39	8.68
User-III (866 Instances)	91.667	8.33
User-IV(1156 Instances)	91.840	8.1597

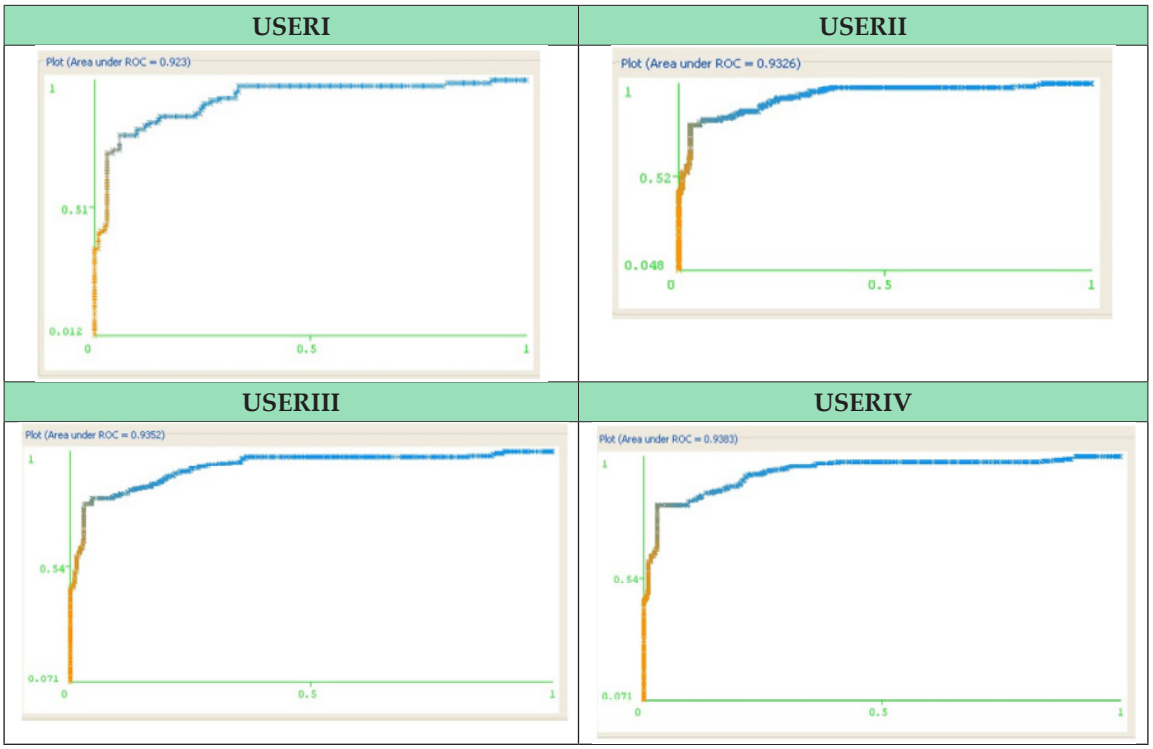

Figure 2. ROC Curve for user I to IV

Table 2. The performance of SVM classifier

User	Percentage of Correctly Classified Instances	Percentage of Incorrectly Classified Instances
User-I (290 Instances)	92.3611	7.638
User-II (578 Instances)	95.13	4.8611
User-III (866 Instances)	95.833	4.167
User-IV(1156 Instances)	96.18	3.8194

Conclusion

In this paper we proposed an intrusion detection method for smart/cellular phones with application to data mining followed by its performance using real time call log data of different users. In this experiment we computed the out-come of two classifiers Naïve-Bayes and SVM. Our results show that the classifiers in question are good in finding intrusions with accuracy which is about 95 percent. But to be more specific SVM out-come is better than the Naïve-Bayes. We also investigated the impact of dataset size on the outcome and to our surprise the outcome was good irrespective of the size of dataset. We have shown in this experiment that

our network based method substantially reduces the computing overhead from the cellular phones.

Also our experiment clearly shows that our is better technique from the previous techniques in detecting intrusions. This research can be further extended by including other features like user location, logs of emails, SMS logs, operating system level events etc. for scrutinizing purposes. A Graphical User Interface can be developed on this through which user can see his/her previous records and intrusions happened.

References

[1] H. Debar, M. Dacier, and A. Wespi, A revised taxonomy of Intrusion Detection Systems, Research Report, IBM, 1999.

[2] M. Esmaili, R. S. Naini, B. Balachandran, J. Pieprzyk, Case-based reasoning for intrusion detection, 12th annual computer security applications conference, pp. 214-223, 1996.

[3] S. Northcutt, J. Novak, Network Intrusion Detection, New Riders/Pearson, Indianapolis, IN, third edition, 2003.

[4] D. E. Denning, An intrusion detection model, IEEE Trans. S. E., SE-13(2), pp. 222-232, 1987.

[5] R. Weber, Information systems control and audit, Upper Saddle River, NJ: Prentice Hall, 1999.

[6] T. F. Lunt, A survey of intrusion detection techniques, Comput. Security, vol. 12, no. 4, pp. 405-418, June 1993.

[7] James P. Anderson, Computer Security Threat Monitoring and Surveillance, Technical report, James P. Anderson Co., Fort Washington, PA., April 1980. on Software Engineering, vol. SE-13, pp. 222-232, February 1987.

[8] D.E. Denning, An Intrusion-Detection Model, IEEE Transactions on Software Engineering, vol. SE-13, pp. 222-232, 1987.

[9] Bace R., Intrusion Detection, Macmillan Technical Publishing, 2002.

[10] May Grance, The DIDS (Distributed Intrusion Detection System) prototype, Proceedings of the Summer USENIX Conference, 227-233, San Antonio, Texas, 8-12 June 1992.

[11] Ilgun, Koral, USTAT:a real time IDS for Unix, Proceedings of the 1993 IEEE Computer Society Symposium on research insecurity and privacy, 1993.

[12] Mark Crosbie, Gene Spafford, Defending a Computer System using Autonomous Agents, Technical report No. 95-022, COAST Laboratory, Department of Computer Sciences, Purdue University, March 1994.

[13] Christoph, Gray G, UNICORN: misuse detection for UNICOS, Proceedings of the 1995 ACM/IEEE Supercomputing Conference, Dec. 1995.

[14] Andrew P. Kosoresow and Steven A. Hofmeyr, Intrusion Detection via System Call Traces, IEEE Software, 14(5), pp. 24-42, September /oct 1997.

[15] Ning,Wang X.S, Jajodia S, Modelling requests among cooperating IDSs, Computer Communications, v 23, n 17, Nov, 2000.

[16] Luo, JianXiong, Mining Fuzzy association rules and fuzzy frequent episodes for intrusion detection, International Journal of Intelligent systems, v15, n 8, Aug, 2000.

[17] J. L. Zhao, J. F. Zhao, and J. J. Li, Intrusion Detection Based on Clustering Genetic Algorithm, International Conference on Machine Learning and Cybernetics IEEE, Guangzhou, pp. 3911-3914, 2005.

[18] Ciza Thomas and N. Balakrishnan, Mathematical Analysis of Sensor Fusion for Intrusion Detection Systems, Proceedings of the International Conference on Communications and Networking, 97, 2009.