Cluster Based Anomaly Detection in Wireless LAN

P.Kavitha[#1], M.Usha[*2]

[#]Associate Professor , Department of Information Technology, Adhiyamaan College of Engineering Hosur, Tamil nadu, India

[*2]Professor, Department of Computer Science and Engineering, Sona College of Technology Salem, Tamil nadu, India

Abstract— Data mining methods have gained importance in addressing computer network security. Existing Rule based classification models for anomaly detection are ineffective in dealing with dynamic changes in intrusion patterns and characteristic. Unsupervised learning methods have been given a closer look for network anomaly detection. We investigate hierarchical clustering algorithm for anomaly detection in wireless LAN traffic. Since there is no standard datasets available to do research in wireless network, we simulated a wireless LAN using NS-2 and the traces are used to observe the traffic patterns. Our study demonstrates the usefulness and promise of the proposed approach which uses hierarchical cluster based framework for anomaly detection in wireless computer networks to produce low false positive alarm and high detection rate also compared with the real time wireless traffic. This system can help Wireless network management system to quickly identify the attacks, which extends the system administrators security management capabilities and improve the integrity of the information security infrastructures.

Keywords— Anomaly detection, Wireless Network, Data mining, Clustering , Wireless LAN Traffic data.

I. INTRODUCTION

Wireless networking is revolutionizing the way people work and play. By removing physical constraints commonly associated with high-speed networking, individuals are able to use networks in ways never possible in the past. Students can be connected to the Internet from anywhere on campus. Family members can check email from anywhere in a house. Neighbours can pool resources and share one high-speed internet connection. Over the past several years, the price of the wireless networking equipment has dropped significantly. Wireless NICs are nearing the price of their wired counterparts. At the same time, the performance has increased dramatically. Wireless networking is a double-edged sword. Wireless users have many more opportunities in front of them, but those opportunities open up the user to great risk. The risk model of the network security has been firmly entrenched in the concept that the physical layer is at least somewhat secure. With wireless networking[6,15], there is no physical security. The radio waves that make wireless networking possible are also what make wireless networking so dangerous[19]. An attacker can be anywhere nearby listening to all the traffic from the wireless network,-in the parking lot across the street, or on the hill outside of the town.

II. RELATED RESEARCH WORK

Clustering is a well known technique which includes statistics, machine learning, databases and visualization. Dokas ,Ertoz kumar ,Srivasta[16] developed algorithms using outlier detection schemes. They conducted experiments on KDDCUP 99 dataset and concluded that LOF approach was the most promising technique for detecting novel intrusions. Zhang and Nath[1,2,18] first presented a distributed intrusion detection and response architecture for wireless ad hoc networks, which provides an excellent guide for the later works. A data mining approaches to network intrusion detection provides an opportunity to learn the behaviours of network users by mining the data trails of their activities. While recent research e.g., Clustering, MADAM ID[16] , ADAM , MINDS , have investigated data mining for intrusion detection, considerable challenges remain unexplored. This involves intrusion detection models for wireless networks not requiring hard-to-get training data in wired network environment[9,12], as well as intrusion detection that has no prior knowledge of relationships between attack types and attributes of the network audit data. One of the research in wired IDS by Zhong et al[3],Multiple centroid based unsupervised Online K-Means clustering algorithm for intrusion detection, is with an effective self-labelling heuristic for detecting attack and normal clusters of network traffic audit data. Some of the drawbacks of this Zhong et al. work are: they used only metrics available in the recorded wireless logs rather than all that are theoretically required to model common wireless attacks. While these methods can detect anomalies that cause unpredicted changes in the network traffic, they may be deceived by attacks that increase their traffic slowly. Our work can detect anomalies regardless of the speed of the network traffic.An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. TYPES OF WIRELESS ATTACKS

Wireless intrusions belong to four broad categories, namely: (1)Passive attacks (2) Active attacks (3) Man-in-the-middle (MITM) attack (4) Jamming attacks.

A Passive attack (e.g., war driving) occurs when someone listens to (or eavesdrops) on network traffic[11,12]. Armed with a wireless network adaptor that supports promiscuous mode, the eavesdropper can capture network traffic for analysis using easily available tools. Active attacks launched by hackers who access the network to launch these active attacks include unauthorized access, Denial of Service (DoS) and Flooding attacks like (SYNchronized) SYN Flood attacks, and (User Datagram Protocol) UDP Flood attacks. There are generally two types of approaches taken toward network intrusion detection: Anomaly detection and Misuse detection. In misuse detection[11], each network traffic record is identified as either normal or one of many predefined intrusion types. A classifier is typically then trained to discriminate one category from another, based on network traffic data. On the other hand, anomaly detection amounts to training models for learning normal traffic behavior and then classifying, as intrusions, any network behavior that significantly deviates from the known normal network traffic patterns. We focus on anomaly detection in this paper.

The general unavailability of benchmark data on wireless attacks (i.e., data with known attack types)calls for unsupervised models for wireless intrusion detection. An unsupervised approach to intrusion detection entails knowledge discovery based solely on the attributes of network traffic records. In anomaly detection normal (good) [19] behavior of users or the protected system is modeled[7], often using machine learning or data mining techniques. During detection new data is matched against the normality model, and deviations are marked as anomalies. Since no knowledge of attacks is needed to train the normality model, anomaly detection may detect previously unknown attacks.

IV. DATAMINING IN INTRUSION DETECTION

Clustering is the method of grouping objects into meaningful subclasses so that the members from the same cluster are quite similar[8], and the members from different clusters are quite different from each other. Therefore, clustering methods can be useful for classifying log data using distance and density function and detecting intrusions.

Classification-based methods[4,14] require training data that contains normal data as well as good representatives of those attacks that should be detected, to be able to separate attacks from normality. Producing a good coverage of the very large attack space (including unknown attacks) is not practical for any network. Also the data needs to be labeled and attacks to be marked.

One advantage of *clustering based methods[2,3]* is that they require no labeled training data set containing attacks, significantly reducing the data requirement. There exist at least two approaches. When doing *unsupervised anomaly detection* a model based on clusters of data is trained using unlabelled data, *normal as well as attacks*. If the underlying assumption holds (i.e. attacks are sparse in data) attacks may be detected based on cluster sizes, where small clusters correspond to attack data. Our problem is primarily an unsupervised learning problem. Clustering is a technique that has been used since a long time to solve problems in several domains. Clustering offers simple solutions that can be hierarchically organized which helps in maintaining the clusters in a well defined manner.

V. PROPOSED WORK

A. Data Collection using NS-2:

Network Simulator (Version 2), widely known as ns-2, is simply a discrete event driven network simulation tool for studying the dynamic nature of communication networks. It is an open source solution implemented in C++ and OTcl programming languages. ns-2 provides a highly modular platform for wired and wireless simulations supporting different network element, protocol (e.g., routing algorithms, TCP, UDP, and FTP), traffic, and routing types. In general, ns-2 provides users with a way of specifying network protocols and simulating their corresponding behaviors. Result of the simulation is provided within a trace file that contains all occurred events.

We designed a wireless LAN with 200 nodes. The Attacks like Denial of Service(DoS),Man-In-The-Middle(MITM),Packet Modification were randomly injected in the network and the traces are observed from the log file.

B. Data Collection using Wireless Network and Feature Extraction:

Real time Wireless traffic of Wi-Fi Lab with 120 nodes and 10 Access points are captured using Wireshark. Three weeks of the traces are observed and the extracted features are given in the table.

Table 1: Features used in the implementation

Feature	Description
SrcMac	The MAC address of the source device.
DstMac	The MAC address of the destination (could be broadcast as well).
NumFrames	The number of frames sent from the source to destination.
AvgFrmSize	The average size of frames in bytes, sent from source to destination.
NumDeauths	The ratio of number of De-authentication frames sent to the number of frames sent.
NumDisassocs	The ratio of number of Disassociation frames sent to the number of frames sent.
NumRetries	The ratio of number of retransmitted frames sent to the number of frames sent.
NumCRCErrs	The ratio of number of error frames sent to the number of frames sent.
AvgSignal	The average signal strength of frames sent

	from source to destination.
AvgNoise	The average noise (in terms of percentage of signal) in frames sent from source to destination.
SeqNo	Sequence Number

C. Hierarchical Clustering Using Clustering Feature Tree:

A data structure called Clustering Feature Tree or CF-Tree similar to B+ tree is used to maintain cluster information[8,21]. The CF-Tree provides an efficient method of organizing clusters and maintaining cluster information via Clustering Feature vectors which store much less information than the actual data but sufficient to calculate all the information that the clustering or anomaly detection algorithm would require. A Clustering Feature vector is a triple that maintains summarizing information about a particular cluster.More specifically, this vector is a (N,Xavg,SS) triple, where N is the number of data elements in the cluster, Xavg is the average vector of the cluster (also called the centroid) and SS is the Squared Sum of the vectors in the cluster .

A CF tree is a height-balanced tree with two parameters: branching factor B and threshold T. Each non-leaf node contains less than B entries of the form [CFi, childi], where i = 1, 2, ...,B, "childi" is a pointer to its i-th child node, and CFi, is the Clustering Feature of the sub-cluster represented by this child. So a non-leaf node represents a cluster made up of all the sub clusters represented by its entries. A leaf node contains at most, B entries, each of the form [CFi], where i = 1, 2, . . B. A leaf node also represents a cluster made up of all the sub-clusters represented by its entries. But all entries in a leaf node must satisfy a threshold requirement, with respect to a threshold value T: the radius has to be less than T.

This data structure basically provides a compact way to storing information about the clusters by storing only those parameters which really matter and not every data point. It has been used for representation of clusters in as well.
Given n d-dimensional data vectors Vi in a cluster CFj = {Vi | i = 1...n}
the centroid V_0 and radius R(CFj) are defined as:

$$V_0 = \frac{\sum_{i=1}^{n} v_i}{n}$$

$$R(CF_j) = \sqrt{\frac{\sum_{i=1}^{n} (vi - v0)^2}{n}}$$

R is the average distance from member points in the cluster to the centroid and is a measure of the tightness of the cluster around the centroid. For every CFj we store the corresponding radius R(CFj) , the average vector X(CFj) and the centroid V0(CFj).

The distance between a data point and a cluster denoted by CFi is the d-dimensional Euclidean distance between the data point and the centroid of the cluster denoted by CFi. The distance between two clusters is calculated by calculating the Euclidean distance between their corresponding centroids.

The parameter T is used for checking two threshold conditions ,While entering a data point into the CF-Tree,

1. R(CFj) ≤ T
2. Distance(vi,CFi) ≤ T

VI. EXPERIMENTAL RESULTS

A.Data Preprocessing: A basic data pre-processing techniques like sampling and filtering are applied for the sake of easy and smooth operation of the experiments. Samples with known and unknown attacks are merged together so as to render two types of data viz. Normal and Attack data.

B.Training: During training, CF –tree is created by inserting each data item in the tree. If the size of the tree increases so that the number of nodes is larger than fixed number (M), the tree needs to be rebuilt. The threshold T is increased, all CFs at leaf level are collected and inserted a new into the tree. Now it is not single data points that are inserted but rather CFs. Since T has been increased, old clusters may be merged thereby reducing the size of the tree. If the increase of T is too small, a new rebuild of the tree may be needed to reduce the size below M again.
The CF-Tree thus created is a normality model of the network and will now serve to help detect anomalies and intrusions by classifying the testing data.

C.Testing: When a new data point v arrives,the detection starts with a top down search from the root to find the closest cluster feature CFi. This search is performed in the same way as during training. When the search is done and terminates at a CFi in any of the leaf nodes, the distance D(v,CFi) from the centroid of the cluster i to the new data point v is computed. Informally, if D is small, i.e. lower than the threshold T, v is similar to the data included in the normality model and v should therefore be considered normal. If D is large, v is an anomaly since it is far away from any of the existing normal clusters.

The k-means clustering and CF-tree algorithm are employed on the data sets.The metrics of Anomaly detection like false alarm and detection rate are calculated. The values are tabulated in table 2.

VII. COMPARISION OF RESULTS OF K-MEANS AND CF TREE CLUSTERING:

Table-2

No. of Data Samples	**K-Means Clustering**		**Hierarchical Clustering**	
	FAR%	**DR%**	**FAR%**	***DR%***
1000	28.5	63.73	17.3	93.5
10000	18.2	72.8	13.76	96.3
23000	14.3	76.5	10.3	90.2
36000	10	89.3	5.3	94.3

From the experimental results and the performance evaluation, it is seen that False Alarm Rate (FAR) has been decreased and Detection Rate (DR) has been increased in hierarchical clustering. Hence hierarchical approach is best suitable for clustering than traditional k-means clustering.

VIII. CONCLUSION

The cluster based anomaly detection using K-means and Hierarchical data mining clustering methods are implemented and tested with datasets collected from Wi-Fi-Lab and simulated from NS-2. Hierarchical Clustering using CF-tree stores the essential features of clusters not the actual dataset.The experimental results compared with traditional clustering algorithm .It has been observed that the detection rate is promoted and the false alarm rate is diminished,also it can detect new type of attacks by splitting the clusters.Though many methods are available for intrusion detection, CF tree based clustering methods can yield better results than others.

REFERENCES

[1] Khoshgoftaar, T.M., Nath, S.V., Zhong, S., Seliya, N, "Intrusion detection in wireless networks using clustering techniques with expert analysis", in proc. of the ICMLA 2005:Fourth International Conference on Machine Learning and Applications, pp. 120-125, 2005.

[2] Zhong, S., Khoshgoftaar, T.M., Nath, S.V., "A Clustering approach to wireless network intrusion detection", in proc. Of the International Conference on Tools with Artificial Intelligence, ICTAI 2005, pp. 190-196, 2005.

[3] Zhong, T. M. Khoshgoftaar, and N. Seliya Clustering-based network intrusion detection. International Journal of Reliability, Quality, and Safety Engineering,2007.

[4] P.Kavitha, M.Usha,"Classifier Selection Model for Network Intrusion Detection using Data Mining", CiiT International Journal of Data Mining and Knowledge Engineering, Vol 3,No.12 ,2011.

[5]P.Kavitha, Usha.M,"Detecting Anomalies in WLAN using Discrimination Algorithm" 4th International Conference on Computing, Communication and Networking Technologies - ICCCNT 2013,July 2013.

[6]M. Balazinska and P. Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In The 1st Int. Conf. Mobile Systems, Applications, and Services, 2003

[7]Wireless network intrusion detection system : Implementation and architectural issues – Gianluca Papaleo, 2006.

[8]Jiawei Han, and Micheline Kamber. Data Mining:Concepts and Techniques. Higher Education Press, 2001.

[9]Wired and wireless intrusion detection system : Classification, good characteristics and state-of-the-art . – Tarek S. Sobh, Elsevier, 2005

[10]A real- time network intrusion detection system for large scale attacks based on an incremental mining approach – Ming-Yang Su, Gwo-Jong Yu, Chun-Yuen Lin Elsevier 2008.

[11].Wireless intrusion detection based on different clustering approaches – Athira. M . Nambiar, Asha Vijayan, Aishwarya Nandakumar,A2CWiC 2010.
[12].Wireless intrusion detection : Not as easy as traditional network intrusion detection – Zhiqi Tao, A.B. Ruighaver, 2006.

[13].An intrusion detection Model, Dorothy E. Denning, IEEE 1986.
[14]Effective network intrusion detection using classifiers decision trees and decision rules – G. Meera Gandhi, Kumaravel Appavoo, S.K, Srivatsa , Int. J. Advanced networking and applications – Vol.2, issue-3, 2010.

[15].B. Potter, B. Fleck, 802.11 Security. O'Reilly & Associates Inc, 2003, ch2, pp. 18- 29. G. Held, *Securing Wireless LANs. John Wiley & Sons Ltd, 2003, ch 5*,pp. 113-148.

[16] L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, J. Srivastava, V. Kumar, and P. Dokas. The MINDS-Minnesota Intrusion Detection System in Next Generation Data Mining, chapter 3. MINDs, 2004.

[17] Y. Guan, A. L. Ghorbani, and N. Belacel. Y-means: A clustering method for intrusion detection. In Proceedings of the Canadian Conference on Electrical and Computer Engineering (IEEE CCECE), 2003.

[18] P. Chhabra, A. John, and H. Saran. PISA: Automatic extraction of traffic signatures. In Proceedings of IFIP Networking, 2005.

[19] K. Burbeck and S. Tehrani. ADWICE -Anomaly detection with realtime incremental clustering. In Proceedings of the 7th International Conference on Information Security and Cryptology, Seoul, Korea, 2004.
[20] F. Guo and T. Chiueh. Sequence number-based MAC address spoof detection.In http://www.ecsl.cs.sunysb.edu/tr/TR182.pdf, 2005.

[21]T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: An efficient data clustering method for very large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Montreal,Canada, 1996.

[22]Wireless Communications & Networks Second Edition, William Stallings, PHI 2006.