ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 4, April 2013

Survey of Clustering Algorithm in Wireless Sensor Networks

R.Juliana*, S.Deepajothi#

(*#Assisstant Professor, CSE department, Chettinad College of Engineering and Technology)

***Abstract*— The use of Wireless Sensor Networks in various fields of technology has arose the need for securing the data during routing. Wireless Sensor Networks (WSN) are imparted in various sensational environments and so its security is an important feature to be considered. This paper summarizes the security attacks in Wireless sensor networks and the possible means of avoiding it. The various security routing attacks, such as wormhole attack, sinkhole attack, Sybil attack, Denial of Service Attack, Compromised Node attack, Data Insert Attack are studied. Data Clustering algorithms of Data Mining could be utilized for solving the security attacks. Clustering is a data mining technique used to group similar data in clusters. A study of the Clustering algorithms like k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm, expectation maximization clustering algorithm, Partitioning clustering, Distance-based clustering, K-medians clustering is done. Comparison of the algorithm based on accuracy and dataset is analyzed.**

Keywords— Worm hole, Sink hole, Sybil, Denial of Service, Compromised Node, data clustering, k-means, hierarchical clustering, self-organizing maps, expectation maximization, Partitioning, Distance-based, K-medians

I. INTRODUCTION

A wireless sensor network is a collection of nodes organized in to a cooperative network. Each node consists of processing capability , may contain multiple types of memory , have a RF transceiver , have a power source and accommodate various sensors and actuators. The nodes communicate wirelessly and often self-organize after being deployed in a commercial adhoc fashion. Systems of thousands and even ten thousand and more nodes may be participating. Such systems can revolutionize the way they live and work.

Wireless sensor networks are beginning to be deployed at an accelerated pace. In few years that the world will be covered with wireless sensor networks with access to them by the Web. This new know-how is thrilling with limitless potential for numerous application areas including environmental, medical, military, catastrophe prevention, transportation, surveillance, entertainment, surroundings monitoring, homeland defense and clever spaces. Plenty of these applications need that the sensor network be deployed in an area that is hostile, inaccessible & mission critical. The resource shortage nature of sensor networks & its application domains requires for a secure sensor network.

In this paper, we survey the attacks on sensor networks and the possibilities of avoiding these by the use of Data Mining technique called data clustering

Data mining is the process of discovering meaningful new correlation, patterns and trends by sifting through large amounts of data, using pattern recognition technologies as well as statistical and mathematical techniques. Clustering is a data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Representing data by fewer clusters necessarily loses certain fine details, but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its cluster.

A) Data clustering

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent. Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the numbers of disk accesses are to be minimized. In clustering the objects of similar

properties are placed in one class of objects and a single access to the disk makes the entire class available.

B) **Cluster (data mining and clustering techniques)**

The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources. The main purpose of clustering is to locate information and in the present day context, to locate most relevant electronic resources. The research in clustering eventually led to automatic indexing --- to index as well as to retrieve electronic records. Clustering is a method in which we make cluster of objects that are some how similar in characteristics. The ultimate aim of the clustering is to provide a grouping of similar records. Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre defined classes, whereas in clustering the classes are formed. The term " class" is in fact frequently used as synonym to the term "cluster".

In database management, data clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency of search and the retrieval in database management, the number of disk accesses is to be minimized. In clustering, since the objects of similar properties are placed in one class of objects, a single access to the disk can retrieve the entire class. If the clustering takes place in some abstract algorithmic space, we may group a population into subsets with similar characteristic, and then reduce the problem space by acting on only a representative from each subset. Clustering is ultimately a process of reducing a mountain of data to manageable piles. For cognitive and computational simplification, these piles may consist of "similar" items.

C) Use of Clustering in Data Mining: Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign.

For example, a company those sales a variety of products may need to know about the sale of all of their products in order to check that what product is giving extensive sale and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving fewer sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

II. TYPES OF SECURITY ATTACKS IN WIRELESS SENSOR NETWORKS

Wireless networks are vulnerable to security assaults due to the broadcast nature of the transmission medium. Furthermore, wireless sensor networks have an additional vulnerability because nodes are often placed in a hostile or dangerous surroundings where they are not physically protected.

Denial of service on sensing (DoSS) Attack:
An attacker tampers with information before it is read by sensor nodes, thereby leading to false readings and finally leading to a wrong decision. A DoSS assault usually targets physical layer applications in an surroundings where sensor nodes can be found.

Node capture attack: An attacker physically captures sensor nodes and compromises them such that sensor readings sensed by compromised nodes are inaccurate or manipulated. In addition, the attacker may try to extract essential cryptographic keys (e.g., a group key) from wireless nodes that are used to protect communications in most wireless networks.

Eavesdropping attack: An attacker secretly eavesdrops on ongoing communications between targeted nodes to collect information on connection (e.g., medium access control [MAC] address) & cryptography (e.g., session key materials). Although this assault can be classified in to other categories such as privacy-related assaults, they group it in to this section due to its extreme consequences in the sense that the collected cryptographic information may break the

encryption keys such that the attacker can retrieve significant information.

Denial of sleep attack: An attacker tries to drain a wireless device's limited power supply (sensor devices) so that the node's lifetime is significantly shortened. In general, in the work of a sleep period in which there is no radio transmission, the MAC layer protocol reduces the node's power consumption by regulating the node's radio communications. Thus, the attacker assaults the MAC layer protocol to shorten or disable the sleep period. If the number of power drained nodes is giant , the whole sensor network can be severely disrupted.

Flooding attack: An attacker usually sends a giant number of packets to the access point or a victim to prevent the victim or the whole network from establishing or continuing communications.

Jamming (radio interference) attack: An attacker can effectively cut off wireless connectivity among nodes by transmitting continuous radio signals such that other authorized users are denied from accessing a specific frequency channel. The attacker can also transmit jamming radio signals to intentionally collide with legitimate signals originated by target nodes.

Replay attack: An attacker copies a forwarded packet and later sends out the copies repeatedly and continuously to the victim in order to exhaust the victim's buffers or power supplies, or to base stations and access points in order to degrade network performance. In addition, the replayed packets can crash poorly designed applications or exploit vulnerable holes in poor
method designs.

Selective forwarding attack: A forwarding node selectively drops packets that have been originated or forwarded by sure nodes, and forwards other irrelevant packets in lieu.

Unauthorized routing update attack: An attacker attempts to update routing information maintained by routing hosts, such as base stations, access points, or information aggregation nodes, to exploit the routing protocols, to fabricate the routing update messages, & to falsely update the routing table. This assault can lead to several incidents, including: some nodes are isolated from base stations; a network is partitioned; messages are routed in a loop & dropped after the time to live (TTL) expires; messages are perversely forwarded to unauthorized attackers; a black-hole route in which messages are maliciously discarded is created; & a earlier key is still being used by current members because the rekeying messages destined to members are misrouted or delayed by false routings.

Wormhole attack: An adversary intercepts communications originated by the sender, copies a portion of or a whole packet, & speeds up sending the copied packet through a specialized wormhole tunnel such that the copied packet arrives at the location earlier than the original packet traversed through normal routes. The wormhole tunnel can be created by several means, such as by sending the copied packet through a wired network & at the finish of the tunnel transmitting over a wireless channel, using a boosting long-distance antenna, sending through a low-latency route, or using any out-of-bound channel. The wormhole assault poses lots of threats, to routing protocols & other protocols that heavily depend on geographic location & nearness, & lots of later assaults (e.g., selectively forwarding, sinkhole) can be launched after the wormhole path has attracted a large amount of traversing packets. Readers are referred to for details as well as a mechanism to detect such an assault.

Sinkhole attack: An attacker attracts all nodes to send all packets through or several of its colluding nodes, called sinkhole node(s), so that the attacker (and its colluding group) has access to all traversing packets. To attract the victimized nodes, the sinkhole node is usually introduced as an beautiful forwarding node such as having a higher trust level, being advertised as a node in the shortest distance or short delay path to a base station, or a nearest knowledge aggregating node (in WSNs).

Impersonate attack: An attacker impersonates another node's identity (either MAC or IP address) to establish a connection with or launch other assaults on a victim; the attacker may also use the victim's identity to establish a connection with other nodes or launch other assaults on behalf of the victim. There is several softwares able to reprogramming the devices to forge the MAC & network addresses.

Sybil attack: A single node presents itself to other nodes with multiple spoofed identifications(either MAC or network addresses). The attacker can impersonate other nodes' identities or basically generate multiple arbitrary identities in the MAC and/or network layer. Then the assault poses threats to other protocol layers; for examples, packets traversed on a route consisting of fake identities are selectively dropped or modified; or a threshold-

based signature mechanism that depends on a specified number of nodes is corrupted.

Traffic analysis attack: An attacker attempts to gain knowledge of the network, traffic, & nodes' behaviors. The traffic analysis may include examining the message length, message pattern or coding, & period the message stayed in the router. In addition, the attacker can correlate all incoming & outgoing packets at any router or member. Such an assault violates privacy & can harm members for being linked with messages (e.g., religious-related opinions that are deemed provocative in some communities).The attacker can also perversely link any members with any unrelated connections. If a group of attackers collude to launch any type of assaults, it is often called a collusion assault. For example, the colluding group of attackers orchestrates to collect information to significantly exploit the process, masquerade a legitimate member & send out fault messages on behalf of that member, conjointly mount assaults against other members or network entities, or falsely accuse a legitimate member as an attacker.

III. TYPES OF CLUSTERING METHODS

1. The k-means algorithm

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

The k-means 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.

2. Hierarchical clustering

In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a greedy (A greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage[1] with the hope of finding a global optimum.) manner. The results of hierarchical clustering are usually presented in a dendrogram(A dendrogram (from Greek dendron "tree", -gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering).

In the general case, the complexity of agglomerative clustering is O (n^3),which makes them too slow for large data sets. Divisive clustering with an exhaustive search is O(2^n), which is even worse. However, for some special cases, optimal efficient agglomerative methods (of complexity O (n^2)) are known: SLINK for single-linkage and CLINK for complete-linkage clustering.

The Single Link Method (SLINK)

The single link method is probably the best known of the hierarchical methods and operates by joining,

at each step, the two most similar objects, which are not yet in the same cluster. The name *single link* thus refers to the joining of pairs of clusters by the single shortest link between them.

The Complete Link Method (CLINK)

The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

3. Partitioning methods

The partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In case where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases, e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of the clusters is large, the centroids can be further clustered to produces hierarchy within a dataset.

Single Pass: A very simple partition method, the single pass method creates a partitioned dataset as follows:

1. Make the first object the centroid for the first cluster.
2. For the next object, calculate the similarity, S, with each existing cluster centroid, using some similarity coefficient.
3. If the highest calculated S is greater than some specified threshold value, add the object to the corresponding cluster and re determine the centroid; otherwise, use the object to initiate a new cluster. If any objects remain to be clustered, return to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of order *O(NlogN)* for order *O(logN)* clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

4. The expectation maximization clustering algorithm

Expectation Maximization (EM) is a well-established clustering algorithm in the statistics community. EM is a distance-based algorithm that assumes the data set can be modeled as a linear combination of multivariate normal distributions and the algorithm finds the distribution parameters that maximize a model quality measure, called log likelihood.

EM is chosen to cluster data for the following reasons among others:

- It has a strong statistical basis.
- It is linear in database size.
- It is robust to noisy data.
- It can accept the desired number of clusters as input.
- It can handle high dimensionality.
- It converges fast given a good initialization.

5. Self-organizing map algorithm

The self-organizing map (SOM) is an excellent tool in exploratory phase of data mining. It projects input space on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. When the number of SOM units is large, to facilitate quantitative analysis of the map and the data, similar units need to be grouped, i.e., clustered. In this paper, different approaches to clustering of the SOM are considered. In particular, the use of hierarchical agglomerative clustering and partitive clustering using k-means are investigated. The two-stage procedure--first using SOM to produce the prototypes that are then clustered in the second stage--is found to perform well when compared

with direct clustering of the data and to reduce the computation time.

6. K-medians clustering

In statistics(Statistics is the study of the collection, organization, analysis, and interpretation of data.) and machine learning, k-medians clustering[1][2] is a variation of k-means clustering where instead of calculating the mean for each cluster to determine its centroid, one instead calculates the median. This has the effect of minimizing error over all clusters with respect to the 1-norm(group) distance metric, as opposed to the square of the 2-norm distance metric (which k-means does.)

This relates directly to the k-median problem which is the problem of finding k centers such that the clusters formed by them are the most compact. Formally, given a set of data points x, the k centers c_i are to be chosen so as to minimize the sum of the distances from each x to the nearest c_i.

The criterion function formulated in this way is sometimes a better criterion than that used in the k-means clustering algorithm, in which the sum of the squared distances is used. The sum of distances is widely used in applications such as facility location(location analysis).[3]

Note that this algorithm is often confused with k-medoids, which finds the optimal medoid, not median, for each cluster. (A medoid is an actual point from the dataset; a median is the mathematical median calculated separately for each dimension.)

7. K-medoids

The k-medoids algorithm is a clusteringalgorithm related to the k-means algorithm and the medoidshift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers (medoids or exemplars).

k-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters known a priori. A useful tool for determining k is the silhouette(mp to: navigation, search Silhouette refers to a method of interpretation and validation of clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster). It is more robust to noise and outliers as compared to k-means because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances.

A medoid (Medoids are representative objects of a data set or a cluster with a data set whose average dissimilarity to all the objects in the cluster is minimal.) can be defined as the object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the cluster.

IV. CONCLUSION AND FUTURE WORK

The various Data Clustering Algorithms could be used in solving the various security threats of wireless sensor networks. The future work of implementation would help in finding out the most appropriate Data clustering algorithm.

References:

[1]. (Xiangqian Chen, Kia Makki, Kang Yen, and Niki Pissinou, Sensor Network Security: A Survey, IEEE Communications Surveys & Tutorials, vol. 11, no. 2, year 2009.
[2]. Tahir Naeem, Kok-Keong Loo, Common Security Issues and Challenges in Wireless Sensor Networks and IEEE 802.11 Wireless Mesh Networks, International Journal of Digital Content Technology and its Applications, Volume 3, Number 1, year 2009
[3]. Sen, J. & Ukil, A. (2010). A secure routing protocol for wireless sensor networks. Proceedings of the International Conference on Computational Sciences and its Applications (ICCSA'10), Fukuaka, Japan.
[4]. Zhan, G. ; Shi, W. & Deng, J. (2010). TARF : a trust-aware routing framework for wireless sensor networks. Proceedings of the 7th European Conference on Wireless Sensor Networks (EWSN'10), Coimbra, Portugal.
[5] Osama Abu Abbas ,"Comparisons between Data Clustering Algorithms", The international Arab

Journal of Information Technology, Vol.5,No.3,July 2008.
[6] Odukoya, O.H,Aderounmu, G.A and Adagunodo, E.R. ," An improved Data Clustering Algorithm for Mining Web Documents, Cbafemi Awolowo University, Osun State.
[7] Gu, L., Jia, D., Vicaire, P., Yan, T., Luo, L., Tirumala, A., Cao, Q., He, T., Stankovic, J. A., Abdelzaher, T., and Krogh, B. H. (2005). Lightweight detection and classification for wireless sensor networks in realistic environments, Proc. of ACM SenSys 2005, San Diego, California, USA,
[8] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[9] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, Nov. 1999.
[10] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4.