Energy Efficient Cluster-Based Intrusion Detection System for Wireless Sensor Networks

Manal Abdullah, Ebtesam Alsanee, Nada Alseheymi
Computer Science Department, Faculty of Computing and Information Technology FCIT,
King Abdul-Aziz University KAU,
Jeddah, Saudi Arabia

***Abstract*—Wireless sensor networks (WSNs) are network type where sensors are used to collect physical measurements. It has many application areas such as healthcare, weather monitoring and even military applications. Security in this kind of networks is a big concern especially in the applications that required confidentiality and privacy. Therefore, providing a WSN with an intrusion detection system is essential to protect its security from different types of intrusions, cyber-attacks and random faults. Clustering has proven its efficiency in prolong the node as well as the whole WSN lifetime. In this paper we have designed an Intrusion Detection (ID) system based on Stable Election Protocol (SEP) for clustered heterogeneous WSNs. The benefit of using SEP is that it is a heterogeneous-aware protocol to prolong the time interval before the death of the first node. KDD Cup'99 data set is used as the training data and test data. After normalizing our dataset, we trained the system to detect four types of attacks which are Probe, Dos, U2R and R2L, using 18 features out of the 42 features available in KDD Cup'99 dataset. The research used the K-nearest neighbour (KNN) classifier for anomaly detection. The experiments determine K = 5 for best classification and this reveals recognition rate of attacks as 75%. Results are compared with KNN classifier for anomaly detection without using a clustering algorithm.**

Keywords—wireless sensor networks WSN; intrusion detection ID; clustering protocols; stable election protocol SEP; KDD cup'99; KKN

I. Introduction

Due to their easy and inexpensive deployment features, Wireless Sensor Networks (WSNs) are applied to various fields of science and technology. These applications include to gather information about human activities and behavior, such as healthcare, military surveillance and reconnaissance, highway traffic; to observe physical and environmental phenomena, such as ocean and wildlife, earthquake, pollution, wild fire, water quality; to monitor industrial sites, such as building safety, manufacturing machinery performance, and so on [1]. On the other hand, security in WSNs is an important issue, particularly if they have mission-critical jobs. For example, a confidential patient health record should not be unrestricted to third parties in a healthcare applications. Securing WSNs is critically important in military applications where security crack in the network would cause causalities of the friendly armies in a battlefield [1]. Security attacks against WSNs are categorized into two main branches: Active and Passive. In passive attacks, attackers are normally hidden and either tap the communication link to collect data; or destroy the functioning elements of the network. Passive attacks can be grouped into eavesdropping, node malfunctioning, node tampering/ destruction and traffic analysis types. In active attacks, an adversary actually affects the operations in the attacked network. This effect may be the objective of the attack and can be detected. Active attacks can be grouped into Denial-of-Service (DoS), jamming, hole attacks (black hole, wormhole, sinkhole, etc.), flooding and Sybil types [1].

Solutions to security attacks against wireless sensor networks involve many components such as prevention, detection and mitigation. First, we discuss the intrusion detection components. According to [1], detection means being aware of the attack that is present. So if an attacker manages to pass the measures taken by the 'prevention' step, then it means that there is a failure to defend against the attack. At this time, the security solution would immediately switch into the 'detection 'phase of the attack in progress and specifically identify the nodes that are being compromised. ID systems are used to monitor both user and system activities to analysis any abnormal activity patterns and recognize patterns of typical attacks. In WSN, sensor nodes use batteries as power supply so battery power is a significant resource for sensor devices. The sensor nodes can be installed in an extensive geographical space to observe physical phenomenon with adequate precision and dependability. After installed, the minor sensor nodes are usually unapproachable to the operator. Therefore, conservation of energy and energy efficient routing must be taken into account when choosing a clustering algorithm. Contribution in this paper is to build an intrusion detection system that combines three main features:

- Use an energy efficient cluster-based WSN that guarantee prolong the life time of the single sensor node and the whole network as well. SEP protocol works based on election of the node which have the highest energy within each cluster as a cluster head. This technique has proven to prolong the life time of the network.

- Use of KNN classifier that has the advantage of having simple classifier and reduce the computation of detecting the attacks. Reducing the computation is an important advantage toward saving the network energy in general.

- Use of KDD-NSL[2] dataset that has a specific feature of avoiding the redundant attributes by removing irrelevant and redundant features that are intercorrelated. This technique helps to achieve high detection rate and accurate results.

The rest of the paper is organized as follows. Section 2 discusses the literature review and related works. In section 3, the proposed ID system is introduced. The experimental work is discussed in section 4 and finally in section 5, the paper is concluded.

II. Literature Review

In this section it is required to review the LEACH protocol as basic clustering protocol where it is used to compare the results. The research relies on three main parts which are the SEP cluster-based WSN, the ID system and the classification technique. The three parts are discussed in the following subsections then some related work are introduced.

A. LEACH Clustering protocol: advantages and problems

The core idea of LEACH protocol is to split the whole network into numerous clusters. The cluster head node is arbitrarily selected, the chance of every node to be selected as cluster head is equal, and energy consumption of the entire network is averaged. Thus, LEACH can extend network life-cycle. LEACH algorithm is cyclical; it provides a conception of rounds. Every round contains two states: cluster setup state and steady state. In setup state, it forms cluster in self-adaptive mode and in steady state, it transfers data. The selection of cluster head depends on decision made 0 or 1. If the number is less than a threshold, the node turns into a cluster head for the present round. The threshold is set as shown in formula (1) [3]:

$$T(n) = \begin{cases} \frac{p}{1-p*(r*mod1/p)} & if\ n\in G \\ 0 & else \end{cases} \qquad (1)$$

where P is the preferred percentage of cluster head (e.g. 4 or 5%), r is the present round, and G is the set of nodes that have not been cluster heads in the last 1/p rounds. Using this threshold, every node will be a cluster head at some point within 1/p rounds. Nodes that have been cluster heads cannot become cluster heads for a second rounds 1/(p-1). Each node has 1/p probability of becoming a cluster head in each round. At the end of every round, every normal node that is not a cluster head select the nearest cluster head and joins that cluster to transfer data. The cluster heads combine and compress the information and forward it to the base station, thus it extends the life span of main nodes. In this algorithm, the energy consumption will be assigned uniformly among all nodes and the non-head nodes are turning off as much as possible. LEACH assumes that all nodes are in range of wireless transmission of the base station which is not the case in many sensor deployments. 5% of the entire nodes play as cluster heads in each round. Time Division Multiple Access (TDMA) is deployed for better management and scheduling.

One problem in the traditional LEACH protocol is that the cluster head node is randomly selected [4]. After several rounds, the node with more remaining energy and the node with less remaining energy have same probability to be selected as cluster head. If the node that has less energy is chosen as cluster head, it will run out of energy and die rapidly, so that network's robustness will be affected and network lifetime will be short [5].

B. Stable Election Protocol SEP

The SEP (Stable Election Protocol) preserves a clustering hierarchy. SEP is an improvement over LEACH in the way that it took into account the heterogeneity of networks. In SEP, some of the high energy nodes are referred to as advanced nodes and the probability of advanced nodes to become CHs is more as compared to that of non-advanced nodes[5]. In SEP, the clusters are re-established in every “round”. New cluster heads are selected in every round and as a result the load is well distributed and balanced among the nodes of the network. Furthermore every node transfers to the closest cluster head so as to divide the communication cost to the sink (which is tens of times greater than the processing and operation charge). Just the cluster head has to report to the sink and may consume a large amount of energy, but this happens periodically for every node. In SEP there is an ideal percentage (determined a priori) of nodes that has to become CH in every round, according to [5] we denote this ideal percentage as “Popt”. When the nodes are homogeneous, that means all the nodes in the field have the same primary energy, the SEP protocol assurances that each one of them will become a cluster head exactly once each 1/Popt rounds. According to [5] 1/Popt is denoted as “epoch” of the clustered sensor network. On average, n × Popt nodes need become cluster heads per round per epoch where n is the whole number of nodes. Nodes that are chosen to be CH in the present round can no longer become CH in the same epoch. The probability of non-elected nodes belong to the group G to become a CH growths after every round in the same epoch. This maintains a stable number of CHs per round. The choice is made at the beginning of every round by every node s ∈ G independently where picking an arbitrary number between [0,1]. If the arbitrary number is less than a threshold T(s), then the node turn into a CH in the present round. The threshold is set as in equation (2) [5], where r is the present round number.

$$T(s) = \begin{cases} \frac{Popt}{1-Popt(r\ mod\ 1/Popt)} & if\ s\in G \\ 0 & otherwise \end{cases} \qquad (2)$$

C. KNN Classifier

Nearest neighbor rule is widely used in identifying the category of unknown data point on the basis of its nearest neighbor whose class is already known [6]. In KNN, the nearest neighbor is calculated on the basis of value of k that specifies how many nearest neighbors are to be considered to define class of a sample data point [7]. Success of the KNN classifier depends on the least distant between instance features, which are determined by its distance function such as the ordinal Euclidean distance. The Euclidean distance between points is defined by equation (3) [8]:

$$E(P,Q)= \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}$$

$$= \sum_{i=1}^{n}((p_i-q_i)^2) \qquad (3)$$

Where P = (p_1,p_2,p_3 ,…,p_n) and Q = (q_1,q_2,q_3 ,…,q_n)

D. *Intrusion Detection System*

Proposed ID system detects four types of attacks which are [9]:

- Denial of Service (DOS): Attacker tries to prevent legitimate users from using a service.
- Remote to Local (R2L): Attacker does not have an account on the victim machine, hence tries to gain access.
- User to Root (U2R): Attacker has local access to the victim machine and tries to gain super user privileges.
- Probe: Attacker tries to gain information about the target host.

It is important to note that the test data includes specific attack types not in the training data which make the task more realistic. The datasets have a total number of 24 training attack types, with extra 14 types in the test data only. The name and classifications of the training attack types are listed in table1.

TABLE I. ATTACK TYPES WHICH WILL BE DETECTED BY THE ID SYSTEM

Class	Known attack	Unknow attack
Probe	Ipsweep,nmap, portsweep,satan	Saint, scan
DoS	Back,land,Neptne,pod, smurf,teardrop	Apache2,processtable, udpstorm,mailbomb
U2R	Buffer_overflow,loadmodule, perl,rootkit	Xterm,ps,sqlattack
R2L	ftp_write,guess_passwd, imap,multihop,phf,spy, warezclient,warezmaster	Snmpgetattack,named, xlock,xsnoop,sendmail, httptunnel,worm, snmpguess

E. *Related Work*

Bharti et al (2010)[10] defined clustering as the best technique for intrusion detection, and k-mean clustering is one of the useful ID clustering technique because it gives efficient results in case of datasets. But sometimes k-mean clustering fails to give best result because of class dominance and no-class problems. The ID system is an effective approach to deal with the problems of networks using various neural network classifiers. Sapna et al (2011) [11] stated that network based intrusion detection are the best methods. IDS can be a piece of installed software or a physical appliance. The different types of attacks are normal, Probe attacks, u2R, Dos and R2l attacks. Attacks are generated randomly using a random function. The type of attack generated is classified to be a Probe, R2L, U2R or Dos attack [12].

Jianlinetal (2011) [13] worked on fuzzy clustering analysis. Fuzzy clustering is the most popular research currently. It is one of the most perfect and most widely used theories although the rear some drawbacks for classical algorithms. Aizhonget al (2010) [14] focused on pattern recognition as the best classifier selection to network ID and clustering based selection method. The multiple clusters are selected for a test sample. The purpose of selecting the multiple classifiers is to optimizing the pattern recognition. Ajitetal (2005) [15] explained Expectation-Maximization (EM) technique which used in point guesstimate. Given a set of noticeable variables X and unknown (latent) variables Z we want to estimate parameters q in a model. Sometimes the M-step is a constrained maximization, which means that there are constraints on legal solutions not encoded in the function itself. The method to arrange the set of objects into classes of similar (which are having same behavior) objects, is defined as clustering. Objects are being categorized into two categories, (1) Documents within a cluster should be similar (2) Documents from different clusters should be dissimilar.

III. PROPOSED ID SYSTEM FOR WSNS

The proposed ID system supposes that all nodes are equipped with sensor and radio system. This assumption enables all nodes to be eligible to be chosen as cluster head. Three steps of the methodology as follow: using the training data and its features, we train the system by clustering the four attacks to the cluster which representing the attacks. Another cluster will present the normal state in which there is no attack and all the detected intrusion is legal. Then it comes the role of SEP protocol which calculates the weighted election probabilities of each node to become CH according to the remaining energy in each node. The SEP protocol is shown in figure 1. Then the KNN classifier that is built with function in MATLAB with multiple values of K is used to find out the best detection rate as shown in figure 2. KNN works by choosing *k* cluster centers to coincide with *k* randomly chosen or *k* randomly defined points inside the hyper volume containing the pattern set. Then assign each pattern to the closest cluster center. The last step is to recompute the cluster centers using the current cluster memberships. If a convergence criterion is not met, move to step2 as shown in figure 2. Classic convergence criteria are used as no (or minimal) reassignment of patterns to new cluster midpoints, or minimal reduction in squared error.

IV. EXPERIMENT SETUP

The ID system for WSN is implemented using MATLAB. The network consists of 100 node distributed in area of 50*50 meter with all nodes start with same energy and are equipped with sensor and radio system as mentioned before. Many trails are done to determine some important parameters before running the experiment. First, we need to find out which data set will be used to train the system and detect attacks and also to test the system performance. Second, the features used for the best detection classification rate are determined. Finally, data inside the data set is normalized. Each step will be explained in details in the following subsections.

SEP Protocol Algorithm

1. Force each advanced node to be elected every sub-epoch of length (*1+a x m)/P /(1+a*) rounds
2. Probability of a normal node getting elected as cluster head is *P normal*

$$P\ normal = \frac{\mathrm{P}}{1 + a \times m}$$

$$T(i) = \begin{cases} \frac{P\ normal}{1-P\ normal\times\ (r \bmod \frac{1}{P\ normal})} & if\ i \in G\ normal \\ 0 \quad otherwise \end{cases}$$

3. Probability of an advanced node getting elected as cluster-head is $P\ advanced$

$$P\ advanced = \frac{P}{1 + \mathrm{a} \times \mathrm{m}} \ (1 + a)$$

$$T(i) = \begin{cases} \frac{P\ advanced}{1 - P\ advanecd \times\ (r \bmod \frac{1}{P\ advanced})} & if\ i \in G\ advanced \\ 0 \quad otherwise \end{cases}$$

4. Average number of nodes elected per round = nxP

Fig. 1. SEP Protocol Algorithm.

Classification Algorithm

```
1. Data Feature selection
       Select the appropriate 19 features
2. Data pre-processing and normalization
   a. Select the nominal feature
   b. Calculate the probability using probability
   density function
       $Pr[a \le X \le b] = \int_b^a fX\,(x) dx$
   c. Replace the nominal with numerical value
3. Input : training data set , testing data set , group set , K-value
4. KNN classification
       Class = knnclassify(Sample, Training, Group, k)
5. Compute the detection rate
       Detection rate =
       # normal connections  misclassified as attack /
       total number of normal connections
```

Fig. 2. KNN Classifier Algorithm

A. KDD CUP'99 Intrusion Detection Data Set.

KDD cup '99 is the most widely used data set in network intrusion detection and evaluation [9]. MIT Lincoln Labs prepared and managed the 1998' DARPA Intrusion Detection Evaluation Program to survey and evaluate researches in intrusion detection. A typical set of data which includes a large diversity of intrusions simulated in a military network situation was provided. The 1999 KDD intrusion detection contest uses a version of this dataset. KDD training dataset consists of about 4,900,000 single connection vectors each of which contains 41 features and is labeled as either normal or an attack, with exactly one specific attack type [16]. Attack types fall into four main categories: User to Root; Remote to Local; Denial of Service; and Probe.

B. KDD '99 Features

Features shown in table 2 are grouped into four groups as follows: **Basic Features:** can be derived from packet headers without inspecting the payload. Basic features are the first six features listed in table 2. **Content Features:** Domain knowledge is used to assess the payload of the TCP packets. This contains features such as the number of failed login attempts. **Time-based Traffic Features:** These features are designed to capture properties that mature over a 2 second time-based window. One example of such a feature could be the number of connections to the same host over the 2 second interval; **Host-based Traffic Features:** Utilize a historical window estimated over the number of connections – in this case 100 – as a substitute of time. Host based features are then designed to assess attacks, which distance intervals longer than 2 seconds [17].

TABLE II. LIST OF ATTRIBUTES

Total Attribute NSL_KDD		
Protocol_type	Service	Src_byte
Wrong_fragment	Flag	Num_failed_logins
Logged_in	Root_shell	count
Serror_rate	Srv_serror_rate	Rerror_rate
Same_srv_rate	Diff_srv_rate	Dst_host_srv_count
Dst_host_serror_rate	class	Srv_rerror_rate

C. Data Preprocessing and Normalization

Most classifiers in IDS range, particularly artificial intelligence like KNN, handle only numeric dataset and ignore the symbolic features. Therefore, in this section we present a simple version algorithm that transfers nominal features in KDD dataset into numeric value. Furthermore, after transformation, we normalize the dataset scale for all features into [0,1] to avoid dominance and feature impact.[18].

Step 1: Data Set Transformation:

There are three futures that have character values (protocol type, Service, Flag), which must be converted to numeric values by using Probability Density Function PDF as given by equation (4):

$$\Pr[a \le X \le b] = \int_a^b fx(x)\,dx \qquad (4)$$

Step 2: Data Set Normalization:

Normalization is essential to enhance the performance of intrusion detection system. Normalization phase must be applied on all features on KDD dataset. This paper has used MinMax function given by equation (5). To normalize numeric values to range between MinX and MaxX that are the minimum and maximum values for feature X, first [MinX, MaxX] is converted to new range [New MinX, New MaxX], According to equation (5) each value of V in the original range is converted to a new value.

$$new_v = \frac{v - \min x}{\max x - \min x} \qquad (5)$$

V. RESULTS AND DISCUSSION

The experiment is starting with creating a wireless sensor network using MATLAB, and clustering it using SEP protocol. At first, the energy for each node is calculated and based on calculated energy, we choose the cluster head which of course the nodes with the highest energy according to the SEP protocol. Secondly, the unlabeled patterns of nodes are grouped into clusters based on the distance between the cluster heads and nodes. The nodes join the cluster with closest cluster head. This minimizes the communication energy between the nodes and their cluster head and lead to preserve WSN energy and prolong the lifetime of WSN as a result. As we can see in figures 3 and 4, in each round we cluster the nodes and define a cluster head according to the sensor with the highest remaining energy.

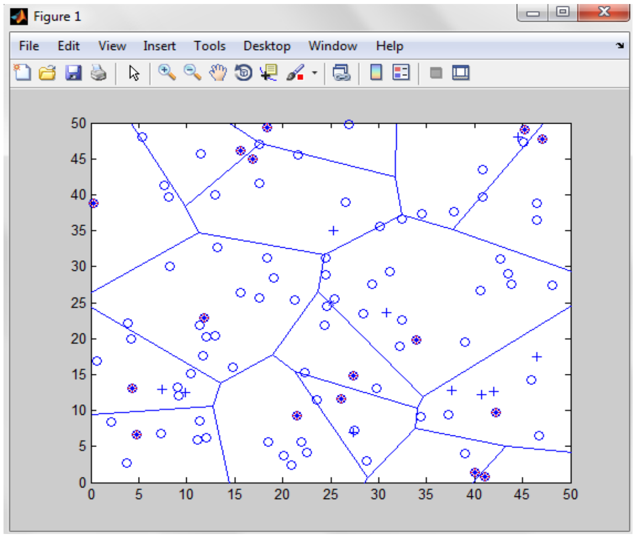

Fig. 3. Clustering 100 nodes using SEP protocol

For IDS, KNN classifier algorithm over KDD99' dataset is used to determine the optimum value of parameter k that reveals the best detection rate as shown in table 3. The experimental results are based on the standard evaluation metric for intrusion detection which is the detection rate.

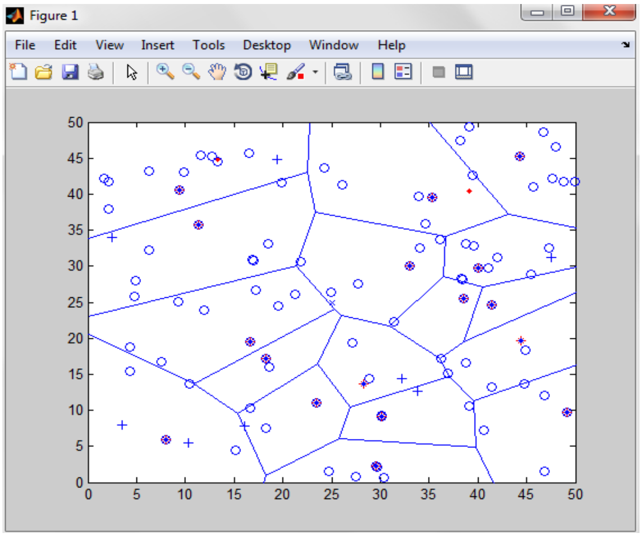

Fig. 4. Assigning new cluster head when the cluster head die

TABLE III. EXPERIMENTAL RESULT

K	DETECTION RATE
1	20.8333%
2	20.8333%
5	75%
10	70.8333%
20	50%
25	33. .8333%

The above table illustrates that, as the value of k increases, the detection rate will be increased until reach the optimal k-value with the highest detecting rate. Then, as the k-value increases, the detection rate will be decreased considerably.

From the table we conclude that the optimum value of k is 5 which results in the highest detection rate of 75%.

Comparing the results of the purposed experiment with other work which is not clustered before classification. The experimental results provide the highest detection rate up to 75%. Figure 5 shows the comparison results. The results also show that the KNN classification without clustering is working better in terms of recognition rate where k-value is less than 5. Although, with k = 5 or greater the KNN classifier with clustering provides the highest recognition rates.

The percentage of recognition rate is decreased with k-value increased for non clustered KNN. This percentage is decreased with increasing k-value for the clustered KNN.

VI. CONCLUSIONS

Intrusion Detection Systems are important tool to detect different types of attacks in WSN which help to monitor the activities and violations in WSN. It's important to consider the energy of the WSN during designing an intrusion detection system. In this paper we have designed an IDS for detecting four types of attacks which are Probe, DoS, U2R and R2L. We have focused on designing energy efficient IDS that preserve the energy of the WSN and prolong the lifetime of the nodes by using the SEP protocol which gives the best results comparing to non clustered network protocols. KDD CUP99' data set has been used for the intrusion detection to give more precise results. The system used KNN classification algorithm to determine the k-value that gives the maximum percentage recognition rate. Then SEP protocol is used for electing cluster head. The system can detect the intrusions with detection percentage rate of 75% at k =5.

As a future work we will consider to use different classification methods to compare with KNN classifier so that we can decide the best classification that works perfectly with the SEP protocol and to gain the maximum detection rate with the longest lifetime for the WSN.

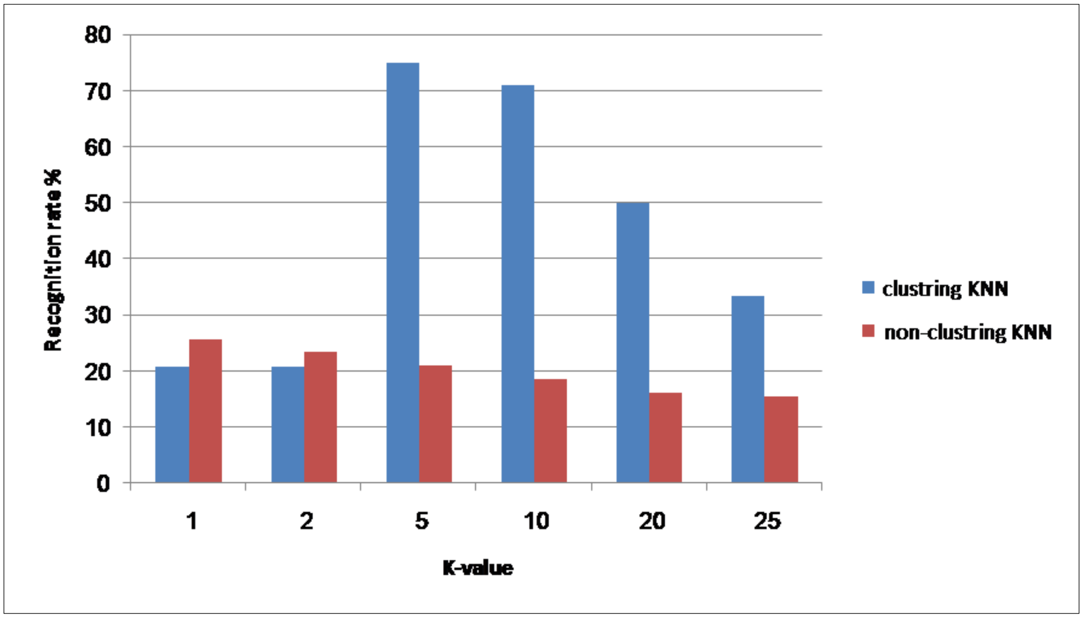

Fig. 5. KNN classification of clustered node compared with KNN classification with non-clustering node

REFERENCES

[1] Ismail Butun, Salvatore D. Morgera, and Ravi Sankar, A Survey of Intrusion Detection Systems inWireless Sensor Networks.

[2] A Detailed Analysis of the KDD CUP 99 Data SetTavallaee, Mahbod; Bagheri, Ebrahim; Lu, Wei; Ghorbani, Ali-A

[3] Qing Bian, Yan Zhang,"Research on Clustering Routing Algorithms in Wireless Sensor Networks," in 2010 International Conference on Intelligent Computation Technology and Automation.

[4] Jianguo SHAN, Lei DONG, Xiaozhong LIAO, Liwei SHAO, Zhigang GAO, Yang GAO Research on Improved LEACH Protocol of Wireless Sensor Networks.

[5] GeorgiosSmaragdakis, Ibrahim Matta, AzerBestavros, "SEP: A Stable Election Protocol for clustered heterogeneous wireless sensor networks".

[6] Uvenir, H. A. &Akkus, A. "KNearest Neighbor Classification on Feature Projections". – ResearchGate

[7] Cover, T. M. &Hart,P. E. "Nearest Neighbor Pattern Classification", IEEE Trans. Inform. Theory, Vol. IT-13, pp 21-27

[8] Deokar, C. (2009). "Weighted K-Nearest Neighbor Algorithms (Solving the Curse of Dimensionality Problem)".

[9] Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. (2005, October). Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion detection datasets.

[10] Kusumbharti, SanyamShukla&Shweta Jain "Intrusion Detection using unsupervised learning".

[11] Sapna S. Kaushik, Dr. Prof.P.R.Deshmukh "Detection of Attacks in an Intrusion Detection System"Amravati India in 2011.

[12] Vladimir Golovko, PavelKachurka, LeanidVaitsekhovich "neural Network Ensembles for IntrusionDetection" Brest State Technical University in 2007.

[13] PengShanguo; Wang Xiwu; ZhongQigen; , "The study of EM algorithm based on forward sampling,"Electronics, Communications and Control (ICECC), 2011 International Conference on , vol., no., pp.4597-4600, 9-11 Sept. 2011 doi: 10.1109/ICECC.2011.6067693

[14] Maria Colmenares& Olaf WolkenHauer, "An Introduction into Fuzzy Clustering",http://www.csc.umist.ac.uk/computing/clustering.htm, July 1998, last update 03 July,2000

[15] http://home.deib.polim.it/matteucc/Clustering/tutorial_html/cmeans.html

[16] Hettich , S., & Bay, S. D. (1999, October 28). KDD Cup 1999 Data.Tavallaee , M., Bagheri, E., Lu, W., &Ghorbani, A. A. (2009). A Detailed Analysis of the KDD CUP 99 Data Set. Proceedings of the IEEE Symposium on computational Intelligence in Security and Defense application (CISDA 2009).

[17] Sathya, S. S., Ramani, R. G., &Sivaselvi, K. (2011). Discriminant analysis based feature selection in kdd intrusion dataset. International Journal of Computer Applications, 31(11).

[18] Salem, M. (2013, April 4). Preprocessing dataset in IDS. Retrieved from http://www.mathworks.com/matlabcentral/fileexchange/41129-preprocessing-dataset-in-ids.

[19] Ibrahim, L. M., Basheer, D. T., &Mahmod, M. S. (2013). A COMPARISON STUDY FOR INTRUSION DATABASE (KDD99, NSL-KDD) BASED ON SELF ORGANIZATION MAP (SOM).