ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 5, May 2016

Distributed Outlier Malicious Nodes Detection Using a Secure Reverse Tracing Technique in a 3-Hop Sets for WSNs

Denita Supriya D, Lakshmikanth S

***Abstract*— Chances of faults occurring during sensing could limit the effectiveness of WSN as pervasive monitoring tool. So Distributed Outlier Detection (DODE) algorithm is proposed to detect data faults/outliers also facilitates fine adjustment of classification accuracy, time complexity and communication complexity, according to application-specific requirements. RSA algorithm that is incorporated ensures that data is securely transmitted to the base station. The proposed algorithm can also effectively detect the malicious nodes that attempt to launch blackhole attacks and any detected malicious node is kept in a blackhole list so that all other nodes that participate to the routing of the message are alerted to stop communicating with any node in that list.**

***Index Terms*—Wireless Sensor Networks, Bayesian Networks, Anomaly, Cooperation network**

I. INTRODUCTION

Sensor arrangement is associated system of detecting nodes scattered spatially to measure and monitor a physical phenomenon. Generally these sensors are arranged by means of remote communications and have constrained preparing, memory, and communication capacities. Taking after with the quick change of equipment and correspondence advancements, numerous present day information securing frameworks are basically programmed conveyed furthermore, ceaseless [1].

A. Reasons for the Occurrence of Outliers

Regardless of their numerous favorable circumstances, for example, adaptability, one of the fundamental downfalls of WSNs is the likelihood of deficiencies happening amid detecting, which could restrain their adequacy as pervasive observing instrument. WSNs are powerless against shortcomings and malign assaults; this in turn causes inaccurate and unreliable sensor readings.

Factors that cause outliers are [14]: 1). Outliers are that data/pattern which differs from the regular set of information. The nature of information set might be affected by missing qualities, commotion and blunder, inconsistent information or copied information. The minimal effort and sensor nodes of low quality have strict asset limitations, for example, energy, computational limit, storage capacity and correspondence data transfer capacity. The usage of inexact sensing devices to report the sensing information from the physical world to the WSNs exhibits low performance as these devices are battery powered their power exhausts. 2).The above interior and outer elements lead to lack of quality of sensor information, which later on badly influences the nature of crude information and collected results. [11]. 3).The chances of errors occurring in sensor devices are more that in traditional network when they are exposed to military and security applications as the number of sensor nodes can reach to extremely high perhaps million nodes depending upon the requirements. So the possibility of error occurrence is more that in traditional network. [13]. 4). Sensor systems develop, the attention on information quality has additionally expanded. Subsequent to the sensor hardware is left exposed to in some cases rude environment; they might come up breakdown amid an arrangement, prompting broken information and awful inferences. [18].

Anomaly is a design that does not coordinate the normal pattern in examined information [1]. The right recognition of anomalies in information gained by a WSN might give helpful data about the condition of the system and about the surrounding environment, e.g., residual WSN lifetime, defected nodes and sudden natural events. Different sorts of information outliers/anomalies that can be detected in WSNs [2]–[4], Tan et al., understands that outlier identification otherwise called peculiarity discovery or deviation recognition, is one of the basic work of information mining alongside probable displaying, cluster examination and association analysis [5]. Han and Kamber contrasted and these other three assignments, exception identification is the nearest to the introductory inspiration driving information mining, which is helpful and intriguing data from a lot of information [6]. As of late, the theme of anomaly recognition in WSNs has pulled in much consideration. By Zhang et al. [7], the five main classes, specifically by which outlier discovery in WSNs can be sorted are: statistical, nearest neighbor, clustering, classification, and spectral decomposition. Statistical methodologies [8], [9] illustrates low computational complexity for the information by assembling mathematical models to determine the likelihood of the readings produced by that model and the readings that falls beneath a given limit are named as anomalies. A statistical methodology requires directed learning and threshold limit.

II. DISTRIBUTED OUTLIER DETECTION ALGORITHM

This section envelops an elaborate description of the proposed algorithm and some techniques adapted to outlier detection and neighborhood selection.

A. Distributed Outlier Detection Algorithm

Identification of anomalies of in-network in WSN facilitates prevention of energy waste due to unnecessary transmissions to the sink is considered as the main objective. Anomaly identification in WSN is carried by method for a distributed arrangement of many BNs [10]. Every BN depends upon collection of cooperative sensor nodes to carry out distributed probabilistic inference; in addition, the subset of shared nodes is selected progressively and in a completely controlled manner, since every node is self-governing. From the perspective of an individual node, DODE comprises of two primary stages, to be specific outlier detection and neighborhood choice. The principal stage, which is followed by every sensing, identifies a probable outlier by teaming up with cooperating nodes and proceeds into the assessment of three measurements: classification exactness, complexities associated with time and communication. The later stage, which is performed occasionally, goal is to recognize the optimum group of neighbors to participate with, and in this way conform to BN structure configuration.

B. Outlier Detection

Utilizing Bayesian systems for anomaly discovery permits to record the probabilistic reliance between arbitrary variables. Nodes in the Bayesian Network which are direct acyclically connected to one another relate to those random variables and causal relation is established through coordinated connections [16]. In DODE, Bayesian networks that are far apart gets connected to each other when their nodes cooperate and every node executes only a partition of the BN that it belongs to in WSN. Every Bayesian Network partition comprises of a set of local observed variables (LOBV) and hidden variables. LOBV are those variables which correspond to the temporal correlation of readings sensed by solitary node and hidden variables correspond to the class of the latest sensory values. Association between two BN portions happens by means of the insertion of shared observable variables, communicating spatial relations among readings assembled by various hubs. This methodology is incorporated in various ways by changing the arrangement of classes to be identified, and local and shared variables.

In this proposition, the arrangement of local observed variables is $L = \{l_1, l_2, l_3\}$ = {inner-ramp, recurrence, variations}, where

l_1 = processed as distinction amidst the two recent readings and r^t is assumed as the current reading,

l_2 = processed as the quantity of back to back recurrences of r^t;

l_3 = processed as the fluctuation of the last K readings;

The hidden variable (c) can corresponds to any of these values: {spike, noise, stuck-at, correct}. The variable that relies upon current readings of node i and node j is called shared observed variable which is computed as

$$S_{i,j} = \{s_{i,j,1}\} = r_i^t - r_j^t.$$

After the Bayesian Networks are constructed the possible class to which the sensory readings of the partaking nodes belong to should be estimated for the outlier detection. For a solitary BN, the greatest estimation of the posteriori probability learnt by proof is obtained by deciding ideal classes $c^* = (c_1^* \dots \dots \dots, c_N^*)$ for a group of sensory readings. And the (MAP) issue is defined below,

$$p(c_1, \dots., c_N | L_1, \dots., L_N, S_{1,2}, \dots., S_{N-1,N}) = \prod_{i_j \in CN\,(i)} p(L_i|c_i) p(S_{i,j}|c_i, c_j) p(c_i) \quad (1)$$

Where CN(i) is expressed as the node i neighbors in the COOP system. Conditional probabilities are given by $p(L_i|c_i)$, $p(S_{i,j}|c_i, c_j)$, and $p(c_i)$ they are processed for every node through logged off regulated learning, by means of a frequentist approach in light of an arrangement of beforehand gathered and arranged readings.

Based on the information obtained by the hierarchical routing protocol and the cooperation (COOP) network structure each sensor node in the maximum- product algorithm plays one of the three key roles

1) Root: Node with no parent.
2) Leaf: Node with no child.
3) Intermediate: node with at least one child and one parent.

Inference algorithm for every leaf is initiated by setting $p(c_i \mid L_i)$ to the initial belief class of last reading of leaf and thus local evidence variable is exploited followed by message passing converge casting procedure. Every node i receives the converge-cast messages from its children nodes at each step of converge casting procedure and sends the below converge-cast message to its parent node j,

$$\mu_{i \to j}(c_j) = max_{Ci}\, \phi(c_i, c_j) \quad (2)$$

Where $\phi\,(c_i, c_j)$ denotes given the local variable of node i, shared variable of node i and node j, including entire shared and local observances of sub-tree which is rooted at node i, and the joint faith about the set of class of the current reading sensed by node i and node j, $\mu_{i \to j}(c_j)$ is the possible set of values of c_j represented by vector dimension.

The matrix $\phi\,(c_i, c_j)$ is obtained as follows,

$$\phi(c_i, c_j) = p(c_i) p(L_i|c_i) p(S_{i,j}|c_i, c_j) \prod_{z \in CN\,(i)/j} \mu_{z \to i}(c_i) \quad (3)$$

Eventually broadcast procedure is started by root node during which leaf and intermediate nodes compute their optimal class provided if they have classes of their respective parent, After obtaining c_j^* that is parent optimal class, optimal label of each child node i is computed as given below

$$c_i^* = argmax_{Ci}\, \phi\,(c_i, c_j^*) \quad (4)$$

Probability of classification error p^i_{err} is computed by each node as given below

$$p^i_{corr} = p\,(c_i^* | L_{1,\dots\dots,} L_N, S_{1,2,\dots\dots,} S_{N-1,N}) = \phi(c_i^*, c_j^*)$$
$$p^i_{err} = 1 - p^i_{corr} \quad (5)$$

Based on all the proofs/evidence obtained from within the cluster p^i_{corr} is the belief $\phi(c^*_i\,,\, c^*_j)$ denotes the probability that the selected class label c^*_i is correct. It is assumed from all the description made above that computation in nodes is

initiated after every reading so to reduce the overall no. of exchange messages a lightly altered version that executes the inference algorithm on data tuples is used rather than opting for the method of buffer storage of consecutive readings (in high rate phenomena).

Figure 1 shows the Flowchart for outlier detection. After the nodes are deployed they are configured with network parameters, every node implements a portion of the Bayesian network in the cooperation network of communication network. If node is a leaf compute the initial belief for the class of last sensory readings by setting distance to zero, if not wait for the converge-cast from the children nodes. For the broadcast phase if it is a root node compute depth setting distance as 0, if not wait for the parent node to and then compute the optimal label and quality metrics.

C. *Neighborhood Selection*

Neighborhood selection is specifically done based on different configurations and requirements. The proposed dynamic and distributed algorithm with the help of values of some quality metrics facilitates every sensor node to select the neighbors to cooperate with and thereby provide optimal structure for the COOP network. The primary objective of the algorithm is to determine a network configuration that provides optimal exchange-off between classification accuracy, complexities associated with time and communication. Locally every sensor node performs neighborhood selection using a technique called Pareto optimization a technique which permits to consider multi-objective functions.

Multiple objective functions are those that are featured by non comparable measurement units and a single solution cannot optimize these functions simultaneously so it becomes impossible to merge them into single objective, this is where Pareto optimization plays a key role in optimizing multi-objectives simultaneously without even compromising on a single functionality. Pareto optimization using the polynomial algorithm finds a set of best results called as Pareto optimal front and the selection of single solution from within the Pareto optimal is based on some application criteria/requirements.

To make decisions like disconnecting from or connecting to any nearby hub each hub in a system can make a decision d to select its neighborhood. Every available decision is evaluated by node for its fitness in the application criteria by means of quality vector Q_d.

If decision d_1 outperforms or Pareto dominates another decision d2 with respect to all the quality metrics considered it means decision d1 relatively provides good trade-off of all the multi-objective function hence superior to decision d2. Pareto dominance of decision d1 over decision d2 is represented by the below equation,

$$d1 \leq d2 \Leftrightarrow \{\forall k = 0, \ldots\ldots, n \Longrightarrow Q_{d1}(k) \leq Q_{d2}(k)\} \quad (6)$$

Where, Q_{d1} = quality vector of decision d_1, Q_{d2} = quality vector of decision d_2. A decision which is best than any other decision is considered as Pareto optimal

$$d^* = \{d_i \in D : \forall d_j \in D, d_j \neq d_i \Longrightarrow d_i \leq d_j\} \quad (7)$$

Now considered the case of imposing some application specific constraints to the quality metrics hence following section is dealing with constrained optimization problems. Let v be the vector of such constraints, decision d is permitted only if d ≤ v, i.e.

$$\forall k = 0, \ldots\ldots, n \Longrightarrow Q_d(k) \leq v(k) \quad (8)$$

Figure 2 shows the flowchart for neighborhood selection. Obtain the classification error, time and distance of child node from parent node, exchange these readings to compute the quality metrics, select best decision out of all the quality vectors and accordingly update the cooperating list, repeat the steps for every T time steps.

III. Experiment and Simulation Results

The adaptive nature with respect to various extents to which the sensed data is corrupted and various compulsions on quality metrics is experimentally evaluated. Consider a COMM network in the form of tree structure with following quantities,

1. Depth = 8
2. Max branching consideration = 3
3. No. of nodes = 86
4. Sampling time (Δt) = 60ms

A. *Standard Case:*

To understand the performance of DODE it is seen in comparison with two stationary benchmark geometric properties and also attempted to check its behavior by imposing various constraints on the metrics by inducting 10%, 25% & 50% quantities of corruption of the overall readings. In the first benchmark G = G', means COMM network matches to the COOP network in CONN network; in this scenario the average taken on communication complexity gives 6 sent messages and average on time complexity is fourteen rounds this amounts to double the no. of COOP neighbors in the CN list of a solitary node. The first benchmark is the upper bound for three metrics considered and denotes 100% active links. The second benchmark D' = ∧ that is no. of nodes is equal to no. of clusters and nodes do not cooperate with each other, this case puts a lower bound on the three metrics denoted by 0% active links gives zero time and communication complexity since temporal correlation is taken among the readings to detect anomalies

In the configuration of 0% active links as there is less communication and time involved and no communication needed to classify readings as nodes perform temporal correlation among the readings outstanding results are obtained in the cases of communication and time complexities. Based on the percentage levels of outlier corruption DODE is appropriately carried out.

ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 5, Issue 5, May 2016

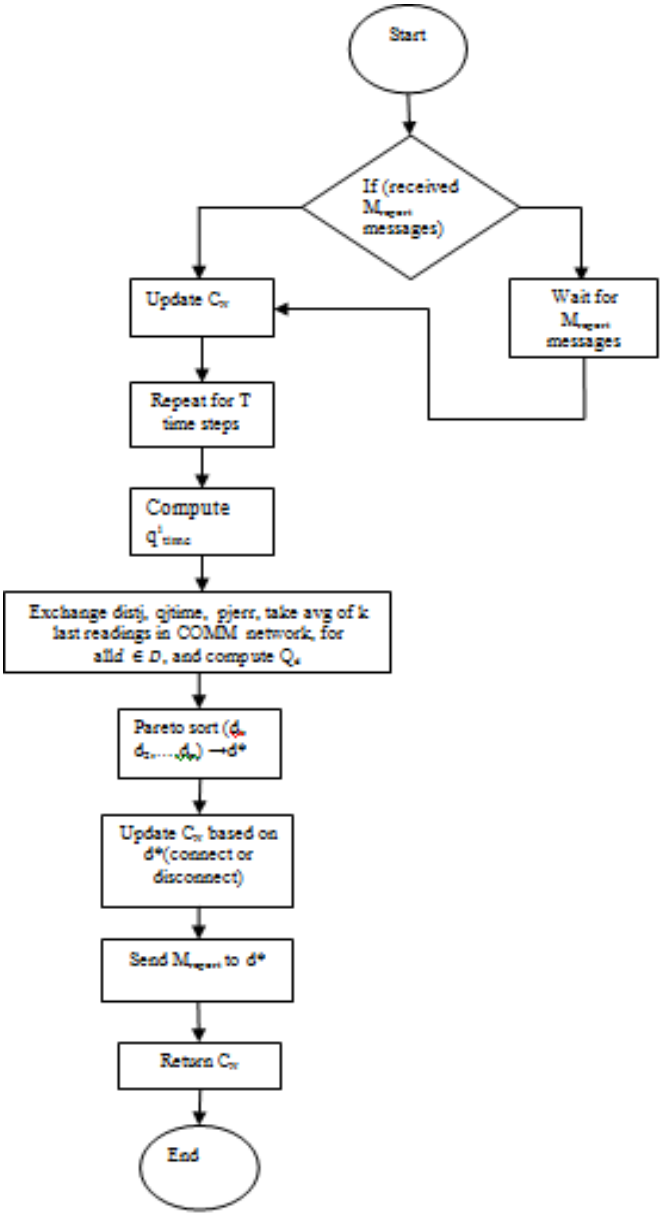

Figure 1: Flowchart for Outlier Detection

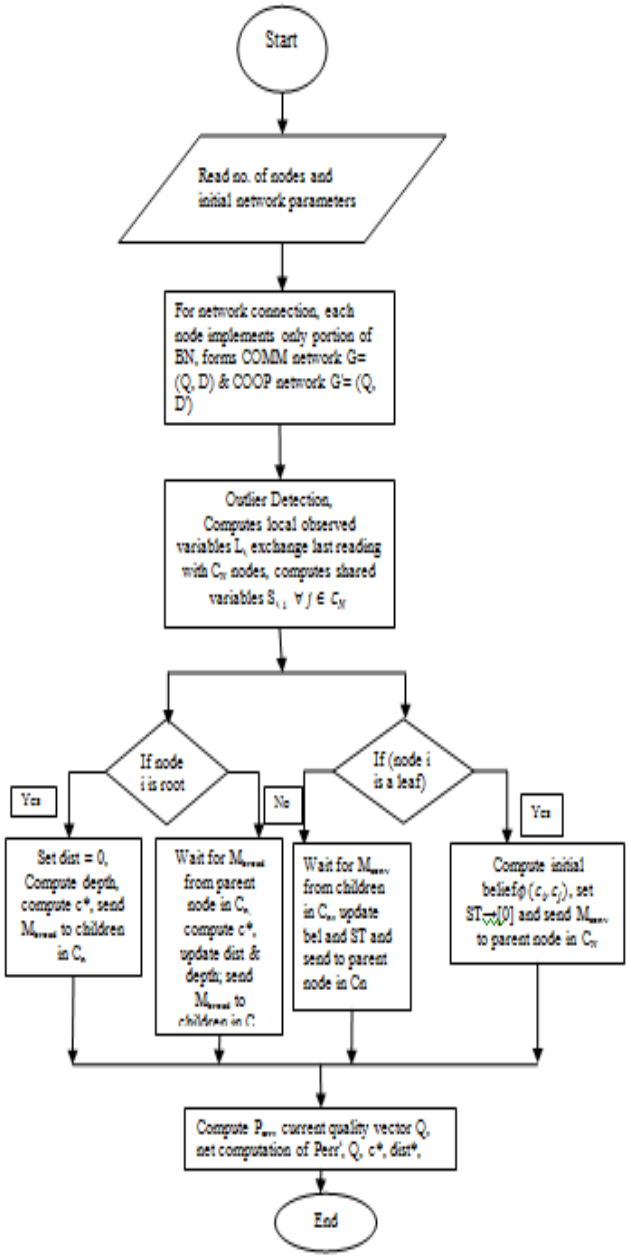

Figure 2: Flowchart for Neighborhood Selection

Consider the two cases where the amount of corruption is 10% and 50%, during the 10% case 96% of the readings are correctly classified, for the 25% amount of corruption 93% of readings are correctly classified, for the 50% accuracy of classification is 85% of the readings are correctly classified. Apparently 0% active links exhibits not much of classification accuracy because temporal correlation of readings are taken so classification is not much needed hence it exhibits best results of communication and time complexities.

Figure 3 shows the classification accuracy in green line, time and communication complexity in blue line for 10% corruption for 100% benchmark in red line and 0% active line in yellow line, as seen in all the above descriptions classification accuracy shows best result at 95% while time and communication complexity is the highest at 95% for the 100% active links because of spatio-temporal correlation where the correlation is taken of all the nodes in BN with respect to single node.

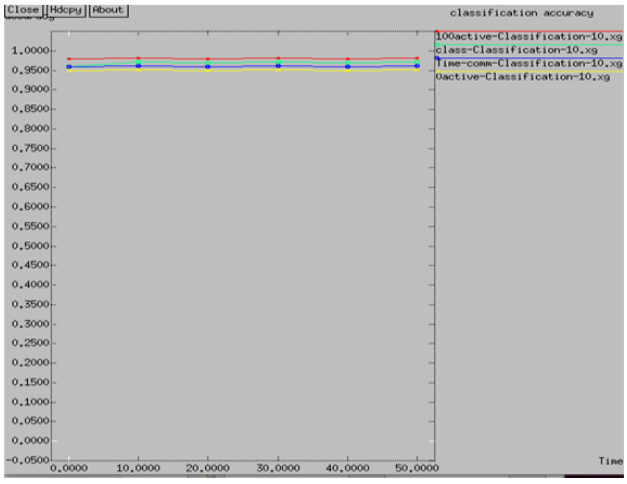

Figure 3: Classification Accuracy for 10% Corruption.

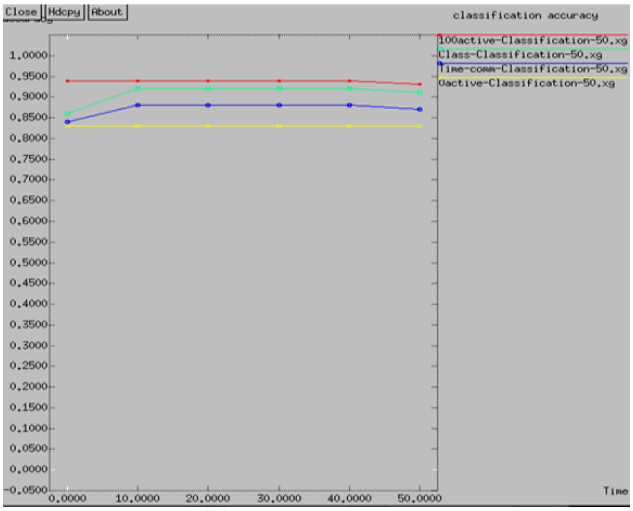

Figure 4: Classification Accuracy for 50% corruption

Figure 4 shows the classification accuracy in green line, time and communication complexity in blue line for 50% corruption for 100% benchmark in red line and 0% active line in yellow line, as seen in the above description classification accuracy shows best result at 86% while time and communication complexity drops to 84% for the 100% active links because of spatio-temporal correlation. Figure 5 shows the performance wrt time and communication complexity when corruption is 10% for 0% active link (which means only temporal correlation that is taking difference of the readings of single node) the time and communication complexity exhibits best result (blue line) and for 0% active link classification accuracy is at the lowest since the difference of the sensory values of single node is taken, so not much effort is needed for classification.

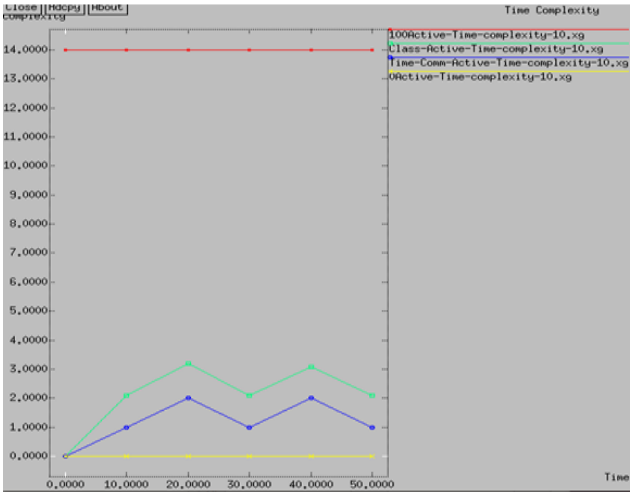

Figure 5: Time & Communication for 10% corruption

Figure 6 shows the outliers detected. Group of nodes with same color form one BN, before the outliers are detected Bayesian network forms a tree structure of every cooperating node, nodes that are near to the lower partition line of window constitutes leaf nodes, nodes closer to the sink node no. 81 are named as root node, in between these two nodes intermediate nodes are formed. Now nodes which have relatively greater energy and that which is capable of communicating with nodes in other Bayesian networks are named as cooperating nodes. All the nodes in the BN send their information to their corresponding root nodes and cooperating nodes to detect outliers, the detected outliers are listed in block list so that no future communications are made by these nodes and also a encrypted message of readings sensed by nodes are send to the sink node no. 81. Figure 7 shows the list of outliers detected in the network these nodes will be block listed and no longer communicated for the future processing.

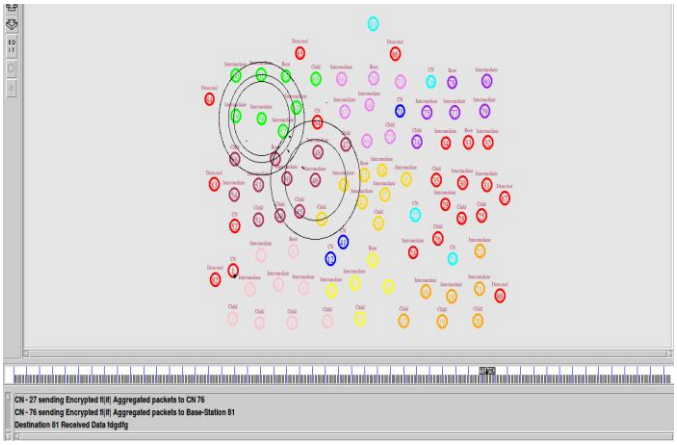

Figure 6: Snapshot of Outlier Detection

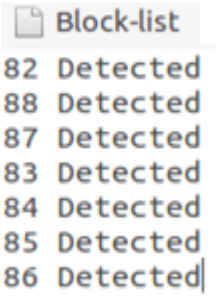

Figure 7: Detected Outliers are Block Listed

IV.CONCLUSION AND FUTURE WORK

A. Conclusion

The proposed work Distributed outlier Detection (DODE) algorithm achieves best results by adapting its behavior to the various amounts of corruption, making optimal decisions using Pareto optimization thereby achieved trade-off between the three conflicting goals with a very subtle compromise in the classification accuracy. Classification accuracy is achieved up to 97% for 100% active link, time complexity and communication complexity up to 2% for 0% active link. The added advantage of incorporating neighborhood selection algorithm in finding high energy cooperating nodes and dynamic construction of distributed Bayesian network aided outlier detection without affecting the performance of system and maintaining the three goals throughout. Therefore gives a remarkable performance in adaptivity, energy consumption and also best optimization of the three conflicting goals compared to non-adaptive approaches.

B. Future Work:

Security can be enhanced for every Bayesian network using error control coding thereby nodes under that Bayesian network becomes completely reliable for future processing.

IV. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 15.1–15.58, 2009.
[2] R. Jurdak, X. Wang, O. Obst, and P. Valencia, "Wireless sensor network anomalies Diagnosis and detection strategies," Intell.-Based Syst. Eng., vol. 10, pp. 309–325, 2011.
[3] A. B. Sharma, L. Golubchik, and R. Govindan, "Sensor faults: Detection methods and prevalence in real-world datasets," ACM Trans. Sensor Netw., vol. 6, no. 3, pp. 23.1- 23.39, 2010.
[4] K. Ni et al., "Sensor network data fault types," ACM Trans. Sensor Networks., vol. 5, no. 3, pp. 25.1–25.29, 2009.
[5] Tan PN. Published "Knowledge discovery from sensor data. Sensors" in 2006.
[6] Han J, Kamber M. "Data mining: concepts and techniques". Morgan Kaufmann; 2001.
[7] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," IEEE Commun. Surv. Tuts., vol. 12, no. 2, pp. 159–170, Apr. 2010.
[8] Y. Zhang et al., "Statistics-based outlier detection for wireless sensor networks," Int. J. Geograph. Inf. Sci., vol. 26, no. 8, pp. 1373–1392,2012.
[9] W. Wu et al., "Localized outlying and boundary data detection in sensor networks," IEEE Trans. Knowl. Data Eng., vol. 19, no. 8, pp. 1145–1157, Aug. 2007.
[10] J. W. Branch, C. Giannella, B. Szymanski, R. Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," Knowl. Inf. Syst., vol. 34, no. 1, pp. 23–54, 2013.
[11] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, "Distributed anomaly detection in wireless sensor networks," in Proc. 10th IEEE Singapore Int. Conf. Commun. Syst., Singapore, 2006, pp. 1–5.
[12] S. Bandyopadhyay et al., "Clustering distributed data streams in peerto-peer environments," Inf. Sci., vol. 176, no. 14, pp. 1952–1985, 2006.
[13] Subramaniam S, Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D. "Online outlier detection in sensor data using nonparametric models". Seoul, Korea: VLDB; 2006, pp. 187–198.
[14] Zhuang Y, Chen L. "In-network outlier cleaning for data collection in sensor networks". In: Proceedings of VLDB; 2006.
[15] Sheng B, Li Q, Mao W, Jin W. "Outlier detection in sensor networks". QC, Canada: MobiHoc; 2007, pp. 219–228.
[16] Zhuang Y, Chen L, Wang X, Lian J. "A weighted moving averagebased approach for cleaning sensor data". IEEE ICDCS; 2007.
[17] Zhang K, Shi S, Gao, Li J. "Unsupervised outlier detection in sensor networks using aggregation tree". In: Proceedings of ADMA; 2007.
[18] Verma V, Kumar S, Harsh K. "Outlier detection of data in wireless sensor networks using kernel density estimation". *Int J Comput Appl* August 2010;5 (7):28–32.

Denita Supriya D *obtained her* B.E. in Electronics and Communication Engineering from ACS College of Engineering, and pursuing MTech in Digital Communication & Networking, Vivekananda Institute of Technology, India. Has published a paper on WSN in International Conference platform and also took part in workshop conference. Her area of interest Computer Networks, Network Security and Wireless Networks

Dr. Lakshmikanth .S Obtained his B.E in Electrical and Electronics and M.Tech(CAID) from Visveshwaraya Technological University, India in the year 2002 and 2007 respectively. He has pursued his Ph.D from Jain University, Karnataka. His field of interest includes signal processing, acoustic noise cancellation and signal designing. He is working as Associate Professor at Vivekananda Institute of Technology, Bangalore, Karnataka. He is member of IE