Sensor Data Fault Detection Techniques in a Wireless Sensor Network

Ujala Arora, Bipan Kaushal
PEC Deemed to be University
(E-mail: aroraujala@gmail.com)

***Abstract*—** Wireless Sensor networks are being extensively employed to monitor changes in the environment in order to take swift and accurate decisions to prevent disasters, coordinate activities in a scenario or report any disturbance or threats. Hence it is extremely necessary to verify the fidelity of the data being sensed by the nodes in the network in order to increase network reliability. In this paper we provide a brief comparison of the various schemes used for outlier detection and their evolution from simple rule based mechanisms to machine learning algorithms.

***Keywords*—** *Wireless Sensor Network; fault detection; spatial correlation; machine learning; outlier detection.*

I. INTRODUCTION

Wireless Sensor Networks(WSN) consist of innumerable sensors spread across the field of interest from where the data is to be gathered .These sensors continuously monitor the environment around them and send the sensed data to a central system for analysis. However as the range of the environment being monitored increases in various applications such as IoT, the number of sensors employed and consequently the amount of data gathered also increases exponentially. In such a scenario sensor data fault detection becomes a herculean task. In order to make the system cost effective, the sensors used may be of low cost and operated in an environment that is neither closely monitored nor controlled which eventually leads to faulty or unreliable sensors.

Faults in sensors can be broadly categorized as follows.

(a) This is the most basic type of fault wherein the sensor cannot deliver the data packet correctly and has been thoroughly investigated and many solutions have also been proposed [1][2][3].

(b) Another major fault observed is in the cases when the sensor successfully delivers the data however the sensed data is incorrect. The identification of these outliers in the collected data can greatly improve the performance of the system in the following ways:

- Improves the energy efficiency of the network thus leading to increased network lifetime by stopping the transmission of garbage values or data.
- Network reliability is considerably improved since malicious data is eliminated. Also the accuracy of the decision making system which could be converted to critical actions is improved and consequently false alarms can also be prevented.
- Online learning and malicious attacks on the network via external agents can be prevented.

II. APPROACHES FOR FAULT DIAGNOSIS

As the data collected by the sensor nodes increases, the probability of encountering faulty data also increases. Owing to the enormous amount of data being generated, improved routing algorithms that reduce data communication, centralize all data being transmitted, and handle all of the data in a base node do not suffice. In order to make the network reliable routing algorithms need to be supplemented with outlier and anomaly detection techniques.

A. Distributed and Centralized Approaches for Fault Diagnosis

Faulty data generated by sensor nodes puts excessive strain upon the limited energy and bandwidth resources of the network. One of the earliest approaches to identify outliers was to exploit the abrupt changes in the readings of nodes [4] [5]. Fault diagnosis can be broadly performed in two distinct ways and one of them is the centralized approach [6] [7]. In this approach the nodes in WSN send their readings to the sink node which will follow a specific algorithm in order to detect the outliers by assimilating all data at a centralized location. One of the centralized approaches for sensor data fault detection is the weighted majority voting scheme [6]. The other approach is the distributed approach which is based on spatial correlation wherein sensors in a defined region are referred to as neighbors and exchange data among themselves to identify the faulty nodes.

B. Distributed Bayesian Algorithm (DBA) For Data Fault Detection

One of the most effective distributed fault detection schemes is the Distributed Bayesian Algorithm (DBA) for data fault detection wherein the concept of border nodes is exploited in order to improve fault detection capabilities. The process of locating data faults is most effective when the sensor before sending out its reading to the sink checks if the readings are correct or not ,thus saving transmission energy and the cost of processing faulty data. Hence in this approach nodes within a certain transmission range are considered to be neighbors and share their fault probability (p) and readings to determine their status [8]. *Fig.1* describes a possible

distribution of good, faulty and border nodes that can be discovered using the DBA approach.

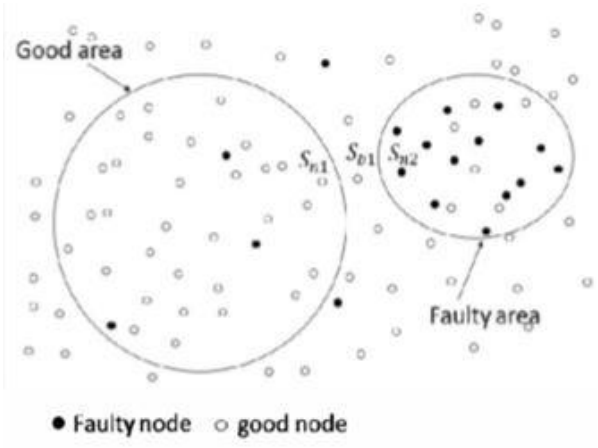

Fig 1.The Distribution of Good, Faulty and Border Nodes

The initial fault probability which is to be used as a prior does not affect subsequent calculations and is assumed to be 0.1. Thus each node has their neighbors' readings and p, which they use to calculate the posterior probability p of the existence of fault and this value is then used to assign a status to the node. This probability is then duly verified in the second step and if it falls below the threshold, the flag for fault detection is raised. The algorithm also conjointly employs the concept of border nodes to eliminate the possibility of faulty nodes which are borderline close to the threshold .Border nodes have low fault probability but their status differs from their neighbors, hence after identifying the border nodes, the border node will send out messages to its neighbor nodes and they will respond with certain confidence values that will determine if the node can be trusted or not.

C. Spatial and Temporal Correlation for Anomaly Detection

Research in the field of fault tolerance of the network reveals that there is a high probability of sensor faults to be stochastically uncorrelated while correct sensor readings are spatially correlated. Spatial correlation is basically based on the notion that in a densely populated network of nodes, neighboring nodes generally detect similar values while observing an identical phenomenon. It can be used to identify anomalies by calculating the variance of the data points in a centralized manner.

Temporal correlation is the counterpart of spatial correlation where instead of geographical distance, time is used as the correlating factor [9]. It is used to express similarity of one signal over time given the assumption that unless extreme conditions the changes observed in incoming data will be gradual and not quite abrupt and this can be used to find anomalous behavior. Temporal correlation is adopted in [9] by applying locally five simple heuristic rules to the data series in order to detect, online, one of the four data fault types (abrupt, noise, stuck at and out of range faults) .Also in case of multiple sensor failures, temporal correlation is more effective compared to spatial correlation.

Many localized threshold based algorithms have been proposed for faulty nodes diagnosis. For instance in Faulty Node Detection (FIND), whenever the event being observed occurs, it ranks the nodes based upon their distance from the event as well as their readings [10]. A node is considered faulty if there is a considerable mismatch between the two ranks. FIND overcomes the limitation associated with spatial correlation since it does not assume similarity between the readings of neighboring nodes. However, the algorithm is information hungry and requires the location of all nodes which may or may not be available.

D. Statistical and Non – Parametric Techniques for Fault Identification

Fault identification techniques in WSN can also be classified as: statistical and non- parametric .The results offered by statistical techniques are better when the data distribution is known a priori [11].These techniques can be used in scenarios where the environment is not subjected to radical changes and is generally stable. The multivariate technique which is based on the chi-square test statistic in which first the parameters are estimated and then the normal state is defined on its basis, consequently any divergence from these estimations is considered to be an anomaly.

However when the environment is subjected to dynamic changes and is not quite stable, non – parametric approach is a proven solution. One example of the non – parametric approach is the rule – based approach that requires the existence of pre – defined rules on the basis of which the data points are classified as normal or anomalies.

III. MACHINE LEARNING FOR SENSOR DATA FAULT DETECTION

With the introduction of machine learning (ML) algorithms which are based upon statistics, the task of extracting properties from the data points and then leveraging them on the data itself to detect faulty points has become quite

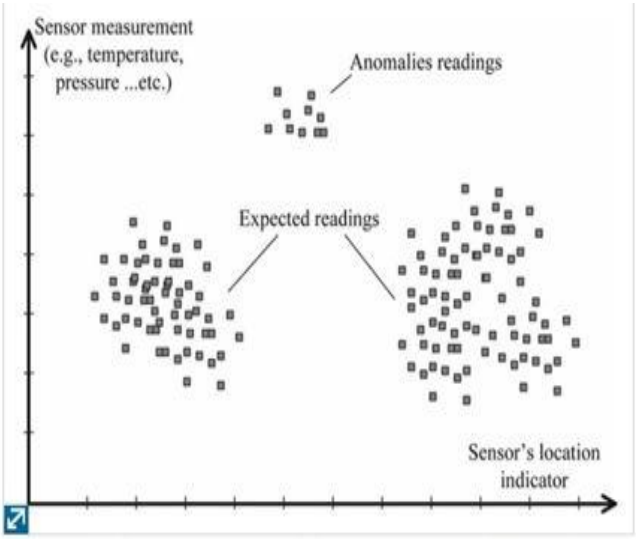

Fig 2.Anomaly Detection Using Machine Learning Clustering and Classification Techniques (Data Set in Euclidean Space)

simple. ML algorithms have been successfully used to find abrupt changes in data patterns, perform qualitative analysis by deriving and manipulating threshold parameters and also tune parameters in order to compensate for the dynamic environment. ML algorithms can be used to solve qualitative as well as quantitative problems; qualitative problems may involve classification and hence are quite similar to faulty data identification as shown in *Fig.2*.

A. Logistic Regression

Logistic regression is a classification algorithm and has been successfully used to detect data faults [12]. Logistic regression (LR) can be used on the basis of spatial correlation and follows the distributed approach thus reducing the cost of communication. LR specifically reduces the complexity of computation in comparison to decision trees since it does not require many parameters and hence is energy efficient [7]. The algorithm is executed in two levels; the learning and execution steps. The nodes send data to the sink node where the learning step is executed, the sink node then transmits the optimized LR algorithm to the nodes in the WSN. *Fig.3* depicts the architecture of LR execution in WSN.

The sink node basically transmits the optimized value of the parameter w to the nodes which then complete the execution step. In the execution step, the algorithm uses binary values 0 or 1 to indicate the status of values generated by the node. This status is decided with respect to a predefined threshold which is based on either institution or experience. The simulation produced high accuracy results for data fault identification with very low computational complexity.

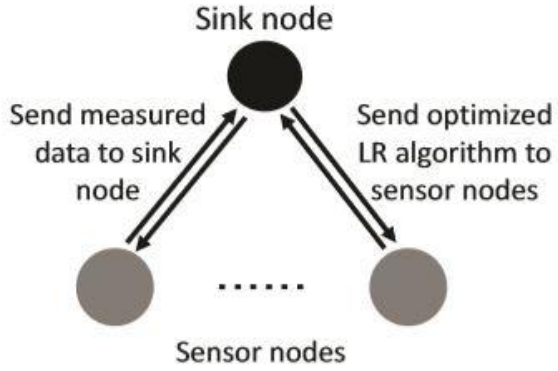

Fig 3.WSN Structure

B. Self Organizing Map (SOM)

However the LR algorithm is a supervised learning algorithm and hence is not dynamic in nature since the threshold is to be pre – determined and hence a certain intuition or knowledge of the data is required. Self Organizing Map (SOM) is an unsupervised classification algorithm that is it does not require a prior knowledge of the data set is not necessary [14]. SOM is basically a type of artificial neural network (ANN) used to identify hidden patterns in the data space. In this approach, a two dimensional map of the problem space is built and also unlike other schemes such as backpropogation which operate on error correcting mechanisms, SOM uses competitive learning. In [3] the authors developed an approach based on SOM and spatial correlation and conjointly proposed a distributed rule based method based on temporal correlation. Alternatively SOM based algorithms have been used to detect attacks in large and complex data sets [15]. The complexity in determining the input weights is a major drawback of this scheme.

C. Support Vector Machines (SVMs)

Support Vector Machine (SVM) based algorithms have been successfully used for anomaly detection because of their efficiency in learning non-linear and complex problems efficiently. One-class quarter-sphere SVM anomaly recognition technique is an alternative to counter the drawbacks of conventional SVM that has high computational requirements [16].This scheme not only reduces computational complexity but also minimizes communication overhead.

D. Alternative Approaches

Bayesian belief networks (BBNs) have been successfully used to develop conditional dependencies among the readings of sensor nodes by exploiting spatial and temporal correlation [17]. K – Nearest Neighbors (k-NN) is an unsupervised classification algorithm that has been used for in-network outlier identification. Both BBNs and k-NN based algorithms can also be used for filling in missing values however these algorithms require large chunks of memory in order to store data for analysis [18]. One class support vector machine classifier has been used to detect black hole attacks and selective forwarding attacks by using routing information, bandwidth and hop count as features to find malicious nodes [19].

IV. CONCLUSION AND FUTURE SCOPE

Statistical and non-parametric approaches that have been extensively used for fault identification require certain pre-assumptions or knowledge base regarding the WSN. In comparison to these techniques ML algorithms are more flexible and adaptive since they can extract features from the environment and then manipulate the thresholds to provide accurate results. ML algorithms are however are computationally expensive since they are quite complex and execution at sensor level is difficult. Sensor nodes in a typical WSN are relatively simple devices and cannot handle such complex computations over extended time intervals. In order to overcome this shortcoming, we could consider distributing these algorithms over the complete WSN. But these algorithms inherently operate in a centralized manner i.e. they need to collect the information centrally for analysis and hence one optimum solution to counter this problem would be to distribute the learning and execution steps. Learning could be executed at the sink node or the central station while the execution could be carried at node level which would not only reduce computational complexity at the node level but also curb the anomalies in the bud. Machine learning can thus be effectively used for anomaly detection, missing value

prediction and to ward off any security threats aimed at the network.

REFERENCES

[1] B. Krishnamachari, and S. Iyengar. "Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks," IEEE Transactions on Computers, vol. 55, pp. 241-250, 2004.

[2] K. Liu, Q. Ma, X. Zhao and Y. Liu . "Self-diagnosis for large scale wireless sensor networks, " Proceedings of IEEE INFOCOM, pp. 15391547, 2011.

[3] Y. Liu, K. Liu, and M. Li. "Passive diagnosis for wireless sensor networks," IEEE Transactions on Networking, vol. 18, pp. 1132-1144, 2010.

[4] T. Palpanas, D. Papadopoulos, V. Kalogeraki V and D. Gunopulos."Distributed deviation detection in sensor networks," SIGMOD Record, vol. 32, pp. 77-82, 2003.

[5] D. Li, K.D. Wong, Y.H. Hu and A.M. Sayeed. "Detection, classification, and tracking of targets," Signal Processing Magazine, vol. 19, pp. 17-29, 2002.

[6] F. Koushanfar, M. Potkonjak, A.S. Vincentelli. "On-line fault detection of sensor measurements," Proceedings of IEEE Sensors, pp. 974-979, 2003.

[7] K. Ni, G. Pottie. "Bayesian selection of non-faulty sensors," Proceedings of IEEE International Symposium onvInformation Theory, pp. 616-620, 2007.

[8] H. Yuan, X. Zhao, and L. Yu, "A distributed bayesian algorithm for data fault detection in wireless sensor networks," in Information Networking (ICOIN), 2015 International Conference on. IEEE, 2015, pp. 63–68.

[9] D. Hamdan, O. Aktouf, I. Parissis, B. Hassan and A. Hiijazi, "Online data fault detection for WSN- Case study," the 3rd International Conference on Wireless Communications, Clerment-ferrand, France,2012.

[10] S. Guo, Z. Zhong, and T. He. "Find: faulty node detection for wireless sensor networks, " Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. pp. 253-266, 2009.

[11] T. Lim, "Detecting anomalies in Wireless Sensor Networks," Qualifying Dissertation, University of York, August 2010.

[12] T. Zhang, Q. Zhao and Y. Nakamoto, "Faulty Sensor Data Detection in Wireless Sensor Networks Using Logistical Regression," 2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW), Atlanta, GA, 2017, pp. 13-18.

[13] M. Abu Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, "Machine learning in wireless sensor networks: Algorithms, strategies, and applications," Communications Surveys & Tutorials, IEEE, vol. 16, no. 4, pp. 1996– 2018, 2014.

[14] S. Siripanadorn, W. Hattagam and N. Teaumroong, "Anomaly detection in wireless sensor networks using self-organizing map and wavelets," International Journal of communications, Issue 3, Volume 4, 2010.

[15] M. A. Alsheikh, S. Lin, D. Niyato and H. P. Tan, "Machine Learning in Wireless Sensor Networks: Algorithms, Strategies, and Applications," in IEEE Communications Surveys & Tutorials, vol. 16, no. 4, pp. 1996-2018, Fourthquarter 2014

[16] S. Rajasegarar, C. Leckie, M. Palaniswami, and J. Bezdek, "Quarter sphere based distributed anomaly detection in wireless sensor networks, "in Proc. IEEE Int. Conf. Commun., 2007, pp. 3864–3869.

[17] D. Janakiram, V. Adi Mallikarjuna Reddy, and A. Phani Kumar, "Outlier detection in wireless sensor networks using Bayesian belief networks," in Proc. 1st Int. Conf. Commun. Syst. Softw. Middleware, 2006, pp. 1–6.

[18] 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, Jan. 2013.

[19] S. Kaplantzis, A. Shilton, N. Mani, and Y. Sekercioglu, "Detecting selective forwarding attacks in wireless sensor networks using support vector machines," in Proc. 3rd Int. Conf. Intell. Sensors, Sensor Netw. Inf., 2007, pp. 335–340.