Online data fault detection in wireless sensor networks

Temporal and Spatial correlations, SOM3D

Mira Sarkis, Dima Hamdan, Bachar El Hassan
LASTRE laboratory, Azm center,
Doctoral School of Sciences and Technology
Lebanese University
Tripoli, Lebanon
gne.mirasarkis@gmail.com
dima.hamdan@ul.edu.lb
bachar_elhassan@ul.edu.lb

Oum El-kheir Aktouf, Ioannis Parississ
LCIS laboratory,
Grenoble INP
Valence, France
oum-el-kheir.aktouf@lcis.grenoble-inp.fr
ioannis.parissis@grenoble-inp.fr

Abstract—**The critical applications of wireless sensor networks, the increased data faults and their impact on decision making reveal the importance of adopting online techniques for data fault detection and diagnosis. Keeping in mind the hardware limitations of sensors, this work focuses on complementary signal processing techniques (temporal, spatial correlation and self organizing map) in order to cover several types of data faults, reduce the misdetection rate and also isolate faults when possible by specifying the defaulting sensors. The methods applied to a real database show that 31.6% of data are faulty by applying SOM3D in conjunction with the spatial correlation. The combination of the above technique in addition to the temporal correlation reduces the misdetection by increasing the detection percentage by 17.6%. SOM3D model also helped identifying the least trustful sensors among the network sensors, this can be helpful when reconciling errors.**

Keywords—wireless sensor networks, data faults detection, diagnostics Self organizing map, spatial correlation, temporal correlation.

I. Introduction

Wireless sensor networks (WSN) consist of a number of nodes which are the association of different types of sensors (temperature, humidity, luminosity, …) equipped with limited resources (battery, memory and processor) and communicating their data via wireless medium in order to serve the application. WSN applications may vary from a simple image recording to critical applications that necessitate an instantaneous control and intervention *i.e.* fire control, house monitoring, intrusion detection, telemedicine [11] … However, data or measures collected by sensor networks are often subject to different sources of errors: environmental fluctuations, free and transient failures, contention and interference of wireless communications [12] or other factors such as wear sensors with time. In consequence, data can be exchanged inaccurately, never exchanged or even lost. The increase deployment of WSN and the increase of data fault rate imply the adoption of a test and diagnostic system to detect, isolate, identify and finally reconciliate faults if possible. In our work, we focus on data faults detection based on complementary signal processing techniques: 1) temporal and spatial correlation 2) Self Organizing Map (SOM). By applying the spatial correlation, we could exploit the existent redundancy among sensors, and the huge number of deployed sensors in a WSN. Spatial correlation is modeled using an artificial neural networks method (ANN), the unsupervised self organized map, which is an intelligent technique that can adapt dynamically to the increasing number of data.

The rest of this paper is organized as follow: an overview of related works, a brief description of the used database, definition of temporal and spatial correlation and a theoretical study of a non supervised data classification method, SOM are presented in section 3. Application of temporal correlation, modelization of spatial correlation and application of SOM in a global architecture are described in section 4. Analysis and interpretation of the methods and results are provided in section 5 to end up with a conclusion and new perspective in section 6.

II. Related Work

Two different approaches are used in the literature: the statistical and the non-parametric techniques. The first one offers better performance when the data distribution is known *a priori* [13]. This can be applied in a stable environment where the sensor distribution is known and always static. The second approach can overcome this last limitation and is applied in the case of unknown and dynamic distributions by updating the parameters recursively [13]. Among the statistical techniques we note the multivariate technique based on the chi-square test statistic [1]: parameters are first estimated and the normal state is defined, any divergence from these estimated parameters is flagged as anomaly. The rule-based approaches are part of the non-parametric techniques. In this case, predefined rules are used to classify data points as anomalies or normal. Rules are applied to monitored data; if they are satisfied then an anomaly is declared. Anomaly detection rule can be based on average receive power and average packet arrival rate [2]. Another algorithm is the cumulative Summation (CUSUM) used to detect abrupt changes in the mean value of a stochastic process

978-1-4673-2489-2/12/$31.00 ©2012 IEEE

[3]. An anomaly is detected by comparing this CUSUM value to a compromised threshold. Spatial correlation is applied by calculating the variance of data in a centralized manner [4] and under a restricted environment of identical and stable nodes in addition to the fact that the supervised parameter changes should be smooth. Temporal correlation is adopted in [5] 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) [5] without the need to exchange data and thus consume battery resources. Parameters are calculated and compared to thresholds. The limitation of this method is in the case of sensors disconnection or destruction. FTEQ [6] performs self-evaluation and cooperative evaluation (distributed and centralized architecture) schemes based on short and long terms spatiotemporal correlation to detect faulty data. A two-tiered architecture is adopted in [7] where a fault analysis takes place at the first level; fault detection based on spatial correlation is at the global level in order to distinguish faults from abrupt changes in the environment. SOM is also adopted in conjunction with wavelets in [8]. Wavelets are used to compress the amount of data and replace them with coefficients at the entry of SOM. This reduces the resources consumption while detecting faults accurately and efficiently.

Our work is motivated by all the above approaches and the objective of our work is to profit from the spatial correlation to modelize data with SOM and apply temporal correlation in both a global and a local architecture in order to maximize the detection rate.

III. Materials and Methods

A. Database description

Our approach is applied to a real database collected from 54 “Mica2Dot” nodes deployed in “Intel Berkeley” laboratory [10] as shown in “Fig. I”. These nodes consist of temperature, humidity and luminosity sensors. Data are acquired each 31s during 34 days; date and acquisition time are also saved.

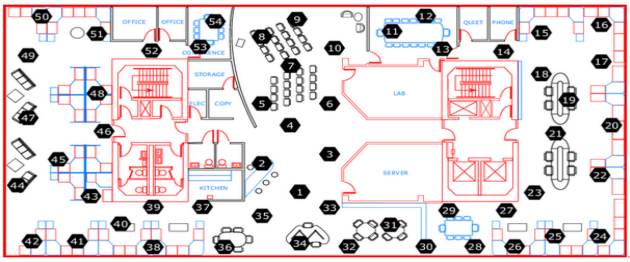

Figure I. Deployment of 54 nodes in Intel Berkeley Laboratory. [10]

B. Temporal Correlation

Temporal Correlation expresses the similarity of one signal over time. In a normal behavior when no exterior events occur, temperature variations should be smooth. Based on this and in a local manner, five simple heuristic rules are applied in [5] as shown in “TABLE I” where Δv, Δt, Δ_{max} (v_{max}), Δ_{min} (v_{min}) and ∑v are respectively data variation, time variation, maximum and minimum thresholds that are obtained based on an offline learning phase and the summation.

C. Spatial Correlation

The huge numbers of deployed sensors in WSNs in addition to the existent redundancy imply that neighbored sensors supervising the same phenomenon should deliver identical values. If this rule isn’t verified at the level of a cluster neighborhood, than spatial correlation is broken and thus we can note the presence of anomalies. Spatial correlation is expressed by the difference or distance between two neighbored sensor values.

TABLE I. Rules applied for fault detection

Fault type	Rule applied
Abrupt	$\Delta v/\Delta t > \Delta_{max}$
Noise	$\Delta v > \Delta_{max}$
Stuck at	$\Delta v > \Delta_{min}$
Out of range	$\sum v > v_{max}$ or $\sum v > v_{min}$

D. ANN method: SOM

SOM is a non supervised classification method that doesn’t require any pre-knowledge of data; instead, neurons compete among themselves and encode the statistical regularities [8] in the weight vector.

SOM is a two full-connected and layered network. Relations between the two layers are expressed by weights (W_{ij}). SOM has the basic objective of projecting the N-dimensional input data to a low-dimensional grid. Based on this grid, each neuron is characterized by neighborhood relations (as “Fig. II” shows) following a neighborhood Gaussian (1)

$$h_c(i,t)= exp(-\frac{||rc(t)-ri(t)||2}{2\sigma 2(t)}) \qquad (1)[8]$$

where rc(t) and ri(t) are the positions of the neuron i and the winner neuron (best matching unit (BMU) or the black dot in “Fig. II”), respectively, and σ(t) is the radius of the neighborhood at time t.

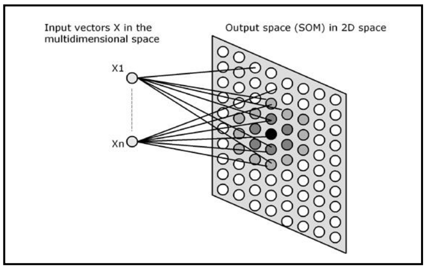
Input vectors X in the multidimensional space
Output space (SOM) in 2D space
X1
Xn

Figure II. An illustration of SOM with hexagonal lattice neighbors belonging to the innermost neuron (black neuron)

At each iteration, a BMU is selected based on the computing of the minimum Euclidian distance between the sample data vector x(t) and all of weight vectors $W_i(t)$ as shown in (2) and then the weight vector is updated following (3) where η_t is the learning rate. The weight update has the higher impact on the BMU and a lower impact as we move away from the BMU.

$$||x(t)-W_{BMU}(t)||=arg\ min||x(t)-W_i(t)|| \qquad (2)\ [8]$$

$$W_i(t+1)=W_i(t)+\eta_t\ h_c(i,\ t)[x(t)-W_i(t)] \qquad (3)\ [8]$$

Note that it is necessary that $h_c(i,t)$ and η_t tend to zero with time [9].

Each NN method is characterized by two phases: learning and test phase. During the first one, the system exploits the existent similarities among learning data in order to organize classes based on their neighbourhood and thus reduce the existent redundancy, it is important that the learning data be in large amounts and representative of all the classes. The second phase is used to determine the membership of the input data and thus classify them.

IV. Integration of Spatial Correlation with SOM

Temperature data are synchronized among all sensors that may encounter data loss. A window of five consecutive epochs is selected and the average of this window is computed. Data are reduced to one fifth of the total.

A. *Spatial Correlation*

Before starting with the application of the SOM method, we started the sensors clustering by sets of three, based on three criteria that satisfy the spatial correlation: 1) physical locations, 2) correlation coefficient, 3) data availability. The three nodes that are within a circle of 1m radius are judged as neighbors. For each neighborhood, we determine the correlation coefficients between each two sensors: if the result tends to 1, then there is a high degree of similarity between the two nodes and thus the spatial correlation is verified, else, the node with the lowest degree of similarity is excluded from the neighborhood and another one that best verifies the rules is included. Some nodes (i.e. 5, 15 ...) suffer from data loss and are excluded from the application. Table II. presents the Ids of the sensors in the 13 determined clusters.

TABLE II. Sensor IDs in each of the 13 resulted clusters

Id1	1	4	8	11	14	16	23	29	32	35	38	41	44
Id2	2	6	9	12	18	17	27	30	33	36	39	42	45
Id3	3	7	10	13	19	19	28	31	34	37	40	43	47

For a cluster of sensor ids id1, id2 and id3, the input vector of the network consists of the three absolute distances (or values differences) d12, d13 and d23 respectively between id1 and id2, id1 and id3, id2 and id3. If the spatial correlation exists, then these last should be null and situated in a quarter sphere around the origin of a 3D plane. "Fig. III" symbolizes the spatial correlation on a 2D plane of distances as axes.

B. *SOM3D*

Our network consists of a map of 32 neurons originally occupying the nodes of a hexagonal grid ("Fig. IV"). The number of neurons expresses the number of desired classes. Though we have two main classes (correct and faulty data), the increase of neurons number improves the precision and helps, as a future step, in identifying faults.

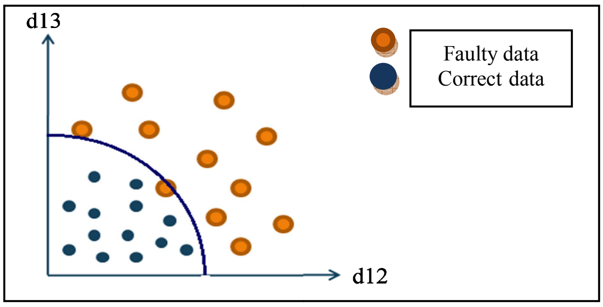

Figure III. Projection of the classification quarter sphere situated around the origine of a 2D plane of axes d12 and d13.

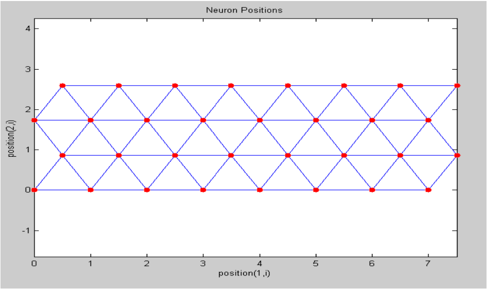

Figure IV. Topology of 32 neurons (red dots) related with neighborhood segments (blue segments) and distributed on a hexagonal grid.

The first five days are used for the learning phase. For each data input, the position of the winning neuron and its 6 neighbors are altered. One thousand iterations are enough to guarantee the network convergence where each neuron occupies a fixed position as shown in "Fig. V".

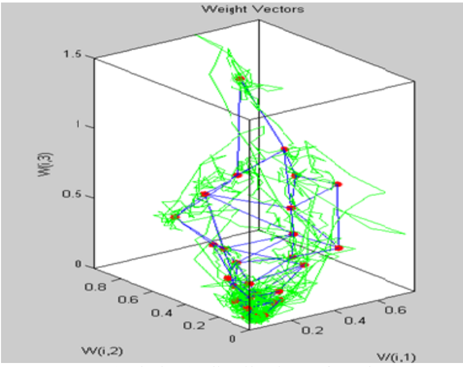

Figure V. Neurons (red dots) distribution after the convergence of the learning phase (green lines represents data and blue segments represents the connections between the initial "Fig IV" neighbored neurons).

The following step is the neurons classification based on their relative positions to the origin. The adopted threshold of classification (or the ray of the classification sphere) is equal to sqrt(3*0.5) where 0.5 is the cardinal distance of the most trustful cluster of sensors ids 8, 9 and 10. This last cluster presents the highest correlation coefficient, 0.97, that expresses the absence of data faults. All neurons belonging to this quarter sphere are tagged with +1 to reflect correct data while the others are tagged with -1 to reflect faulty data.

During the test phase, for each data vector to be classified, we compute distances between the vector and the 32 neurons. The winning neuron corresponds to the minimum distance. The data are classified corresponding to the winning neuron's tag.

C. Fault Isolation

Once the fault is detected, the second step is to try to isolate it or determine the source of failure.

Always for a set of three sensors (S1, S2 and S3) and for a certain detected fault (outside the quarter sphere):
If dij=distance(Id_i, Id_j) is null, than S_i and S_j are correlated.
If dij=distance(Id_i, Id_j)> 0.5, than S_i and S_j are not correlated.

Four cases can exist:

- If d12=0, d13>0.5 and d23 >0.5 → S1 and S2 are correlated but they are not correlated to S3, then S3 is the failure reason.
- If d13=0, d12>0.5 and d23 >0.5 → S1 and S3 are correlated but they are not correlated to S2, then S2 is the failure reason.
- If d23=0, d12>0.5 and d13 >0.5 → S2 and S3 are correlated but they are not correlated to S1, then S1 is the failure reason. "Fig VI" illustrates the three above cases.
- If d23>0.5, d12>0.5 and d13 >0.5 → the failure cannot be isolated.

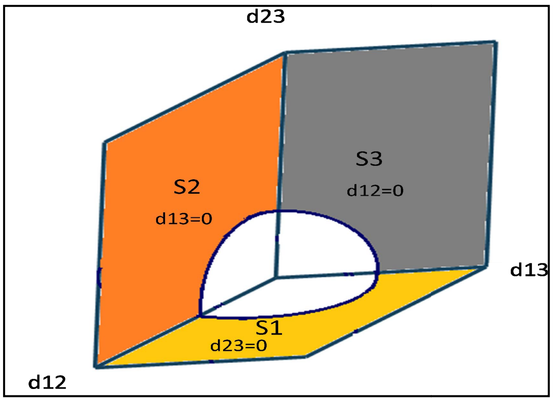

Figure VI. Illustration of fault isolation: yellow part for a faulty sensor S1, gray part for a faulty S3 and orange part for a faulty S2.

During the neurons classification phase, we also tag the faulty neurons with -2 if it belongs to the yellow part (S1 is failing), -3 if it belongs to the gray part (S3 is failing) and -4 if it belongs to the orange part (S4 is failing).

The thickness of the three parts is considered to be 0.1 and not exactly null. The same way, during the test, the additional tag of the winning neuron allows the fault isolation and the determination of the fault source.

V. Experiment Results

SOM3D applied to the temperature data of the Intel Laboratory dataset shows that 31.66% of the total data are faulty while the temporal correlation detected only 19% [5] of the faults. "Tab. III" shows the percentage of faults due to most of the sensors: we note that sensors whose ids are 1, 6, 13, 14, 42 are not trustful since 17.4, 17.2, 9.9, 26.52 and 11.8 % of their delivered data are faulty, and 23.62 % of the faults cannot be isolated by this method.

TABLE III. Faults 'percentages corresponding to some sensors delivery.

Id	Percentage(%)	Id	Percentage(%)	Id	Percentage(%)
S1	17.4	S13	9.9	S36	8.7
S2	3.55	S14	26.52	S38	1.31
S6	17.2	S16	4	S42	11.8
S9	0.78	S27	1.46	S45	2.95
S12	7.25	S30	3.58	Unknown	23.62

The limitations of spatial correlation are that we don't take into consideration the data value in itself; rather, we consider the difference between values. During the last 8 days of data acquisition, almost all sensors delivered the same abrupt value of 125 degrees: the spatial correlation is always verified and no fault is detected: here comes the importance of the temporal correlation in analyzing data value apart and lowering the misdetection rate by 17% as the experiments results show.

SOM3D's learning phase is done offline at the level of the most powered sensor since it requires a large database, memory and processing capabilities in addition to energy. The testing phase is localized at the cluster head each 31*5 s.

VI. Conclusion and Perspective

In this paper, a data fault detection approach based on both a localized neural network method, SOM3D, in conjunction with the spatial correlation and a distributed rule based method based on the temporal correlation is proposed. Temporal correlation helps in overcoming the limitations of the spatial correlation in the case of multi-sensor failures, and vice versa, spatial correlation ensures the similitude among the sensors data values. We could also profit from the SOM3D to isolate faults and determine the least trustful sensors. This will help in a future rectification step. Thus, this combination helps first in improving the detection rate about 17% more than using solely the temporal correlation method, and second, in identifying the fault reason and trying to rectify the error by relying on the most trustful sensors in a future step.

References

[1] E. Ngai, J. Liu and M. Lyu, "On the intruder detection for sinkhole attack in wireless sensor networks," ICC '06, Istanbul, Turkey, June 2006.

[2] I. Onat and A. Miri, "An intrusion detection system for wireless sensor networks," WiMob '05, 2005, pp. 253–59.

[3] T. Peng, C. Leckie and K. Ramamohanarao, "Information sharing for distributed intrusion detection systems," J. Network and Comp. Apps., vol. 30, no. 3, 2007, pp. 877–99.

[4] Y. Feng and R. Zhang, "Fault detection of WSN based on spatial correlation," Applied Mechanics and Materials Vols. 58-60, 2011, pp 1504-1510.

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

[6] R. Zhu, "Efficient fault-tolerant event query algorithm in distributed wireless sensor networks," International Journal of Distributed Sensor Networks, doi: 10.1155/2010/593849, 2010.

[7] S. Zahedi, M. Szczodrak, P. Ji, D. Mylaraswamy, M. Srivastava, R. Young, "Tiered architecture for on-line detection, isolation and repair of faults in wireless sensor networks," Proceedings of IEEE Military Communications Conference (MILCOM), 2008.

[8] 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.

[9] G.A. Barreto, J.C. Mota, L.G. Souza, R.A. Frota and L. Aguaya, "Condition monitoring of 3G cellular network through," *IEEE Trans. Neural Networks*, vol. 16, no. 5, pp. 1064-1075, Sep. 2006.

[10] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin and R. Thibaux. *http://db.csail.mit.edu/labdata/labdata.html.* Intel Lab Data, 2004.

[11] J. Yick, B. Mukherjee, D.Ghosal. "Wireless Sensor Network Survey," Science Direct, 2008.

[12] L.M. Souza, H. Vogt, M. Beigl, "A survey on fault tolerance in wireless sensor networks," 2007.

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