Distributed Faulty Sensor Node Detection in Wireless Sensor Networks based on Copula Theory

Farid Lalem
Lab-STICC UMR CNRS 6285
Université de Bretagne Occidentale, Brest, France
20, Avenue Victor Le Gorgeu, 29238 Brest, France
Farid.Lalem@univ-brest.fr

Ahcène Bounceur
Lab-STICC UMR CNRS 6285
Université de Bretagne Occidentale
20, Avenue Victor Le Gorgeu, 29238 Brest, France
Ahcene.Bounceur@univ-brest.fr

Reinhardt Euler
Lab-STICC UMR CNRS 6285
Université de Bretagne Occidentale
20, Avenue Victor Le Gorgeu, 29238 Brest, France
Reinhardt.Euler@univ-brest.fr

Mohammad Hammoudeh
Faculty of Science & Engineering
Manchester Metropolitan University
Manchester, UK
M.Hammoudeh@mmu.ac.uk

Rahim Kacimi
IRIT-UPS
University of Toulouse
118 route de Narbonne,
Université de Toulouse, France
kacimi@irit.fr

Sanaa Kawther Ghalem
Industrial computing and networking Laboratory-RIIR
University of Oran 1, Ahmed Benbellal
31000, Oran, Algeria
ghalemsanaa@gmail.com

ABSTRACT

Wireless Sensor Networks (WSNs) are arising from the proliferation of Micro-Electro-Mechanical Systems (MEMS) technology as an important new area in wireless technology. They are composed of tiny devices which monitor physical or environmental conditions such as temperature, pressure, motion or pollutants, etc. Moreover, the accuracy of individual sensor node readings is decisive in WSN applications. Hence, detecting nodes with faulty sensors can strictly influence the network performance and extend the network lifetime. In this paper, we propose a new approach for faulty sensor node detection in WSNs based on Copula theory. The obtained experimental results on real datasets collected from real sensor networks show the effectiveness of our approach.

Keywords

Wireless Sensor Networks; Copulas; Faulty sensors; Failure nodes.

1. INTRODUCTION

A Wireless Sensor Network (WSN) consists of low cost wireless nodes, working on batteries, that perform a collaborative effort in order to perceive physical or environmental conditions, such as temperature, humidity, light, sound, vibration, pressure, motion, pollutants, etc. [1].

This feature of sensor networks provides a wide range of applications that include military surveillance, habitat monitoring, seismic detection, security and health applications, etc.

Also, sensor nodes are prone to faults, often unreliable and suffer from inaccuracy and incompleteness because of their intrinsic natures or the harsh environments in which they are used [9].

Furthermore, faulty sensor nodes may generate faulty data, which could affect the analysis of the data, prevent from taking correct decisions and, beyond, lead to a waste of limited resources and reduce the network lifetime [12].

In order to ensure the network performance and extend the network lifetime, the WSN should be able to detect the faulty sensor nodes, to take actions to avoid the spreading of erroneous data over the network, and to maintain the WSN resources at a high level [8].

For this purpose, the aim of the work presented here is to develop an approach for faulty sensor node detection in WSNs, by using Copula theory which is a powerful tool to model and analyze multivariate distributions.

In the following, we propose a new fully distributed approach for faulty sensor node detection using Copula theory. Based on historical measurements of data, we assign to each sensor node a local polygon arising from the application of Copula theory. Faulty sensor nodes are then detected by calculating the probability that a sensed value falls outside this polygon during a period of time $T$. If the calculated probability is less than a predefined threshold, the sensor node is considered good, otherwise the sensor node is declared faulty and an alert will be sent to the sink.

The remainder of the paper is organized as follows. In the next section, we present an overview of related work. Section 3 describes basic concepts of Copula theory. In Section 4, the proposed approach to detect faulty sensor nodes is shown. Simulation results and a performance evaluation
are presented in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

Detecting nodes with faulty sensors can strictly influence the network performance, and extend the network lifetime. Faulty sensors have to be excluded or replaced within the network. So, the main objective of our research is to propose an efficient distributed approach to detect faulty sensor nodes in WSNs. We review in this section some models that have been proposed in the literature.

In [7], a faulty sensor node detection method for wireless sensor networks was proposed (FIND). FIND detects nodes with faulty sensor readings based only on their relative sensing results. The main idea of FIND is to consider a sensor node faulty if its readings violate the distance monotonicity significantly. FIND is based on the fact that the sensing readings monotonically change as the distance becomes larger from the nodes to the event. FIND ranks the nodes based on their sensing readings as well as their physical distances from the event. Then, if there is a significant mismatch between the sensor data rank and the distance rank, the sensor node is considered faulty. However, this approach is centralized and thus generates a high communication overhead.

In [10], the authors proposed a strategy based on modeling a sensor node by a fuzzy inference system (FIS), where a sensor measurement of a node is approximated by a function of the real measurements of the neighboring nodes. The sensor node is declared faulty if the difference between the real sensed value at a node and the estimated value given by its corresponding FIS model is greater than a given threshold. However, this approach generates a high computational complexity which is not suitable for WSNs.

In [11], an online statistical detection technique for faulty sensors is presented, which applies a non-parametric statistical method in order to identify the sensors that have the highest probability to be faulty.

The authors in [15] proposed a local technique for the identification of faulty sensors in sensor networks. To identify the sensors that give a false reading, each sensor first calculates the difference between his own reading and the median reading from the neighboring readings. Then, each sensor collects all differences from its neighborhood and standardizes them. A sensor is faulty if the absolute value of its standardized difference is sufficiently greater than a preselected threshold.

In [2], the authors proposed a localized fault detection algorithm to identify the faulty sensors. By using a local comparison with a modified majority voting, each sensor node makes a decision based on comparisons between its own sensing data and its neighbors’ data, while considering the confidence level of its neighbors. However, this approach requires a large number of messages to be exchanged between neighboring nodes to identify the faulty sensors, which is very costly in terms of energy consumption.

3. INTRODUCTION TO COPULA THEORY

Copula theory is very popular in high-dimensional statistical applications, because it can easily model and estimate the multidimensional distribution of random variables by estimating the marginal distributions and copula separately. In WSNs, detecting faulty sensors with univariate data attributes can easily be done by noting that the single data attribute is abnormal with respect to other data instance attributes. However, in multi-dimension attributes it is difficult to detect faulty instances because individual attributes may not show anomalous behavior although when taken together they can display anomalous behavior. Exploiting dependencies between different attributes of sensor readings allows us to unlock these difficulties and to propose a highly accurate detection method. This is why we decided to use the Copula theory approach since it allows us to model the dependence relationship within multivariate sensed data.

3.1 Mathematical Foundations of Copula theory

In this paper, and for the sake of simplicity, we only consider the case of bivariate Copulas, and the multivariate Copula is just an extension of the bivariate case. To this end we recall some basic definitions and theorems that will be useful later.

3.1.1 Copula Definition

The notion of Copula has been introduced by Sklar in 1959 [17], motivated by the work of Fréchet in the 1950s [13].

Formally, a bivariate Copula [23] is a joint distribution function whose marginals are uniform on [0, 1].

A Copula $C : [0, 1]^2 \rightarrow [0, 1]$ is a function that satisfies the following three conditions:

- $C(u, 0) = C(0, v) = 0$ for each $u, v \in [0, 1]$,
- $C(u, 1) = u$ and $C(1, v) = v$ for each $u, v \in [0, 1]$,
- $C$ is a 2-increasing function, i.e., for each $0 \leq u_i \leq v_i \leq 1$, $C(u_1, v_2) - C(u_1, v_2) - C(u_1, u_2) + C(u_1, u_2) \geq 0$.

3.1.2 Sklar’s theorem (bivariate case):

Let $H(x, y)$ be a joint cumulative distribution function (CDF) with marginal CDF $F$ and $G$. There exists a Copula $C$ such that for all real $(x, y)$, we have $F(x) = U$ and $G(x) = V$, where $U = (u_1, ..., u_n)$ and $V = (v_1, ..., v_n)$ are two variables uniformly distributed on $[0, 1]$. The function $H(x, y)$ can be written in terms of a single function $C(u, v)$ as follows:

$$H(x, y) = C(F(x), G(y))$$ (1)

If $F$ and $G$ are continuous, then the Copula $C$ is unique; otherwise, $C$ is uniquely determined on $(\text{range of } F) \times (\text{range of } G)$. Conversely, if $C$ is a Copula, $F$ and $G$ are CDFs, and then $H(x, y) = C(F(x), G(y))$ is a joint CDF with $F$ and $G$ as marginals.

Schematically: if we have the marginal of each variable, we simply join them with a Copula function having the desired dependence properties to obtain the joint distribution. Figure 1 shows how to obtain the joint distribution function of all variables.

3.2 The Empirical Copula

To model the observed dependence between random variables, we can use the empirical Copula structure to evaluate the suitability of a chosen Copula for the estimated parameter.
It is necessary to introduce the notion of rank before giving the formula of an empirical Copula. Given a set of samples $x_1, ..., x_n$ from a random variable $X$, the rank $R_i$ of $x_i$ is defined to be the number of observations that are less than or equal to $x_i$. Then, the smallest observation of $X$ has rank 1 while the largest has rank $n$.

Let $U=(u_1, ..., u_n)$ and $V=(v_1, ..., v_n)$ be two variables uniformly distributed on $I = [0, 1]$.

We denote by $X,Y$ two random variables, such as $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$, and we let $R_i$, $S_i$ be the rank of $x_i$, $y_i$ the rank of $y_i$.

For a specific sample $(x_i, y_i)$ chosen from $X,Y$, one can approximate the corresponding couple $(u_i, v_i)$ using the ranks $R_i$ and $S_i$ of $x_i$ and $y_i$ among $x_1, ..., x_n$ and $y_1, ..., y_n$ as follows:

\[
\begin{align*}
u_i &= R_i \quad n+1 \\v_i &= S_i \quad n+1
\end{align*}
\]

Thereby, using these ranks, we can construct an empirical distribution function for the random variables $X$ and $Y$ [6].

Formally, the calculation of the empirical Copula is given by the following equation [4]

\[
C_n(u, v) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{R_i \leq u, S_i \leq v}
\]

The function $\mathbb{1}_{a \leq b}$ is the indicator function, which is equal to 1 if $a \leq b$ and equal to 0 otherwise.

### 3.3 Copula Families

There is a wide range of Copula families, according to which dependence structure is expressed by a Copula:

- Dependence in small values.
- Dependence in the extreme values.
- Tail dependence.
- Positive or negative dependence.

Basically, we have two large families of Copulas, the Archimedean Copula family and the Elliptical Copula family. The first Copula family is categorised into:

- Gumbel Copula: Positive dependence and more accentuated on the upper tail.
- Frank Copula: positive as well as negative dependence.
- Clayton Copula: Positive dependence, especially on low-intensity events.

- Copula HRT: Dependence on extreme events of high intensity (dependence structure inverse to the Clayton Copula).

The second family applies to symmetrical distributions and it includes two types of Copulas: the Gaussian Copula and the Student Copula. In the following, we give a quick presentation of the Gaussian Copula, because it is shown in Section 5 to be the family of Copulas that best fit empirical copulas.

The Gaussian Copula is derived from the multivariate Gaussian distribution. Let $\Phi$ denote the standard univariate Normal distribution, whose formula is given by:

\[
\Phi(x) = \int_{-\infty}^{x} \phi(t) \, dt
\]

where $\phi(t)$ is the density probability of the random variable $t \sim N(\mu, \sigma)$ with mean of the distribution $\mu = 0$ and standard deviation $\sigma = 1$:

\[
\phi(t) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{t^2}{2} \right\}
\]

And we have $\Phi_{\Sigma,m}$ the m-dimensional Gaussian distribution with correlation matrix $\Sigma$ with 1 on the diagonal and correlation coefficient $\rho \in (-1,1)$ otherwise. Then, the Gaussian m-copula with correlation matrix $\Sigma$ is given by:

\[
C_{\rho}(u_1, ..., u_m) = \Phi_{\Sigma,m}(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m))
\]

whose density is:

\[
c_{\rho}(\phi(x_1), ..., \phi(x_m)) = |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \zeta^T (\Sigma^{-1} - I_m) \zeta \right\}
\]

where $X = (x_1, ..., x_m)$ and $U = (u_1, ..., u_m)$. Then, by using $u_i = \phi(x_i)$ we can equivalently write:

\[
c_{\rho}(u_1, ..., u_m) = |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} \zeta^T (\Sigma^{-1} - I_m) \zeta \right\}
\]

where $\zeta = (\phi^{-1}(u_1), ..., \phi^{-1}(u_m))^T$.

Therefore, from the formula (7), the bivariate Gaussian Copula is defined by:

\[
C_{\rho}(u, v) = \Phi_{\Sigma}(\Phi^{-1}(u), \Phi^{-1}(v))
\]

where $\Sigma$ is the correlation matrix, which is a 2 $\times$ 2 matrix with 1s on the diagonal and correlation coefficient $\rho$ otherwise.

$\Phi_{\Sigma}$ denotes the CDF for a bivariate normal distribution with zero mean and covariance matrix $\Sigma$. Then, from formula (8) and after simplification, its joint bivariate density is given by:

\[
c_{\rho}(\phi(x_1), \phi(x_2)) = \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left\{ -\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1 - \rho^2)} \right\}
\]

where $X = (x_1, x_2)$ and $U = (u_1, u_2)$ such as $u_i = \phi(x_i)$.

Figure 2 shows an example of a Gaussian Copula with value $\rho = -0.90$. 

- Figure 1: The dependence structure.
3.4 The dependograms

The Dependograms allow us to determine graphically the dependence structure between two random variables. Also, we obtained our dependograms by a scatter plot of uniform marginals \((u, v)\) extracted from a sample or resulting from simulations of a theoretical Copula. In Figure 3, we present some dependograms which are well known for some Copula families.

4. THE PROPOSED APPROACH FOR FAULTY SENSOR DETECTION

In this section, we will present the proposed method for faulty sensor detection based on Copula theory in a WSN. First, we build a model in off-line manner by the application of Copula theory on historical measurements of a good sensor node. Then, this model will be uploaded into sensor nodes to be used for the on-line detection with a predefined threshold.

In the following, we explain our model for the bivariate case.

4.1 Building the model (off-line)

The flowchart presented in Figure 4 describes the process followed, and summarizes all phases needed to build the proposed model, to ensure reliability and efficient detection with high accuracy. The different phases that occur during the different steps are given as follows:

- Step 1: Based on historical measurements of sensed data, we begin by taking \(N\) samples of two random variables \(X\) and \(Y\) which represent two attributes of sensed data such as temperature and humidity. Then, let \(D = (x, y)\) be the dataset that will be used for the construction of the proposed model based on a bivariate Copula, where \(x = (x_1, x_2, ..., x_N)\) and \(y = (y_1, y_2, ..., y_N)\).

- Step 2: In this step, we calculate the empirical Copula \(C_1\) of \(D\). Given a random sample \((x_1, y_1), ..., (x_n, y_n)\) from \(D\), we denote by \(R_i\) the rank of \(x_i\), and by \(S_i\) the rank of \(y_i\).

   Then, we calculate the empirical bivariate Copula \(C_1\) of \(D\) using the formula (4).

- Step 3: In this step, we need to find the closest Copula \(C_2\) that best fits the empirical Copula \(C_1\). In our case, we do this by graphical adequacy which is explained in 3.5.1 above. Initially, we start by a scatter plot of the pairs \((u_i, v_i)\) derived from equations (2) and (3), for all \(i \in \{1, ..., n\}\) with ranks derived from the learning
4.2 On-line detection process of faulty sensors

Figure 5 illustrate the flowchart of the process of on-line detection of faulty sensors.

In this step, each sensor node evaluates its sensed data with respect to the model built in the last stage and a predefined threshold $Th$. Initially, after receiving $n$ sensed values during a period of time $T$, which varies according to the type of application, the node calculates the probability that a randomly sensed value falls outside the polygon $H$ at the end of this period. If the calculated probability is less than a predefined threshold $Th$, the sensor node is considered good otherwise the sensor node is declared faulty and an alert will be sent to the sink.

**How to calculate the threshold $Th$?** The threshold $Th$ is the ratio between the number of sensed values that are outside the polygon $H$, and the total number of sensed values of the dataset used to build the model.

**How to calculate the probability that a point falls outside the polygon $H$?** We start by giving an example to show how this probability can be calculated. The following Figure 6 will help us to understand the method of calculating the probability that a random point $A$ falls inside the polygon $H$. In our case, the distribution is obtained from the application of Copula theory.

Then, we denote by $F(x,y)$ the cumulative distribution

\[ F(x,y) = \begin{cases} 
0 & \text{if } x < a \text{ or } y < b \\
1 & \text{if } x > c \text{ or } y > d \\
\text{otherwise} & 
\end{cases} \]

Then, we denote by $F(x,y)$ the cumulative distribution

\[ F(x,y) = \begin{cases} 
0 & \text{if } x < a \text{ or } y < b \\
1 & \text{if } x > c \text{ or } y > d \\
\text{otherwise} & 
\end{cases} \]
function (CDF) of random variables X and Y and by \(D_i\) a straight line that belongs to the polygon \(H\), for \(i \in \{1, 2, 3, 4, 5\}\), where the equation of \(D_i\) is given by: \(y = \alpha_i x + \beta_i\).

As illustrated in Figure 6, we calculate the probability that a random point \(p\) falls inside the polygon \(H\). This probability is given by:

\[
P(p \in H) = P(C_1 \land C_2 \land C_3 \land C_4 \land C_5)
\]

where \(C_i\) is the event that the point \(p\) is in the side where the polygon is situated with respect to the segment \(D_i\) of the polygon \(H\).

Mathematically, calculating the probability \(P(A \in H)\) is very efficient, but in our case of WSNs, and depending on the type of application, the number of segments that form the polygon can be very large, which leads to a high computational complexity. Therefore, this method is not suitable for WSNs.

To resolve this problem, we propose to calculate the probability \(P(A \in H)\) using the Monte Carlo simulation. We know that the probability of a random variable \(X\) is defined by: \(P(X \leq x) = \int_H f(x)dx\) where \(f(x)\) is the probability density function of the random variable \(X\).

In our case, this probability is estimated by the method of Monte Carlo and the formula is given by:

\[
\tilde{P} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\arg H}
\]

The function \(\mathbb{1}_{\arg H}\) is the indicator function, which is equal to 1 if \(\arg H\) is true and equal to 0 otherwise.

Finally, we have:

\[
\tilde{P}(x \notin H) = 1 - \tilde{P}
\]

Then, if the calculated probability is less than a given \(Th\), the sensor node is considered as good otherwise the sensor node is declared faulty.

5. PERFORMANCE EVALUATION

5.1 Results and Discussion

To present our proposed approach, we used the dataset of Intel lab [3], which contains real measurements of data collected every 31 seconds from 54 sensors, and deployed in the Intel Berkeley Research lab between February 28th and April 5th 2004. Figure 7 shows the deployment of the sensor nodes in this laboratory. We used the node number 16 to test our approach in the period cited above, since the obtained copula is very close to the one presented by Figure 3.3. We took all pairs (temperature, humidity) from the node 16, to obtain 21249 samples.

In the following, we show the results obtained using the empirical bivariate Copula, which represents the scatter plot of 20610 pairs of the empirical Copula \(C_1\).

- **Step 1**: We start with the dataset \(D = (x, y)\) containing \(N = 21249\) samples of pairs (temperature, humidity).

- **Step 2**: In this step, we calculate the empirical bivariate Copula \(C_1\) of \(D\) using the formula (4). Then, by making a scatter plot of the pairs \((u_i, v_i)\) corresponding to the sample \((x_i, y_i)\) in the dataset \(D\), we obtain Figure 9(a). In this figure, it is clear that the data, encircled in red(area \(A\)), will have to be removed from the dataset because they do not follow the behavior of the majority of sensed data. By removing the area \(A\) from Figure 9(a), we obtain Figure 9(b) which represents the scatter plot of 20610 pairs of the empirical Copula \(C_1\).

- **Step 3**: In this step, we need to find the Gaussian Copula \(C_2\) that best fits the empirical Copula \(C_1\). Using the graphical adequacy presented in section 3.5.1, we conclude, that the Copula that best fits the empirical Copula \(C_1\), is the Gaussian Copula, obtained by comparing the dependograms of the empirical Copula from Figure 9(b), with the dependograms of the Gaussian Copula displayed in Figures 2 and 3. Now, we know...
that the theoretical Copula \( C_2 \) that best fits the empirical Copula \( C_1 \) is Gaussian, and after this, we need to estimate the parameter that calibrates the theoretical Copula \( C_2 \) with the empirical Copula \( C_1 \), which is the correlation coefficient \( \rho \) in our case. To estimate the correlation coefficient \( \rho \) of the Gaussian Copula \( C_2 \), for each value of \( \rho \) ranging between \(-0.80 \) and \(-0.99 \), we calculate the difference of surface between the empirical Copula \( C_1 \) and the average of surface of 100 theoretical Copulas generated from \( C_2 \) with a sample of the same size as \( C_1 \). Finally, we choose the \( \rho \) which minimizes this difference of surface. As displayed in Figure 10, the \( \rho \) which minimizes the difference of surfaces of the empirical Copula and the theoretical Copula is \( \rho = -0.91 \), and thus, the Copula that best fits the empirical Copula \( C_1 \) is the Gaussian Copula \( C_2 \) with correlation coefficient \( \rho = -0.91 \).

• **Step 4**: We generate a sample \( Z \) of 20610 pairs of \((u_i, v_i)\), \( i = (1, \ldots, 20610) \) from the Gaussian Copula \( C_2 \), with correlation coefficient \( \rho = -0.91 \). The obtained samples are presented by Figure 2.

---

**Figure 9**: The scatter plot of the empirical Copula: (a) the original one and (b) without the area \( A \).

**Figure 10**: The scatter plot of the values of difference of surface according to different values of \( \rho \).

• **Step 5**: In this step, we calculate the convex hull \( H_2 \) of the generated sample \( Z \), to obtain \( K = 28 \) pairs \((u_i, v_i)\), which represent the vertices of the calculated convex hull. Figure 11 shows the red points that are selected to be vertices of the convex hull of the sample \( Z \), and by joining these points, we obtain the convex hull of the sample \( Z \) which is plotted in blue.

**Figure 11**: The convex hull of the generated sample \( Z \).

• **Step 6**: Now, we have all points which are vertices of the convex hull \( H_2 \) in the Copula space.

Then, for each point of \( H_2 \) from the Copula space, we calculate the corresponding point in \( D \)-space. As a result, we get the polygon \( H \) in \( D \)-space, as displayed in Figure 12.

• **Step 7**: Finally, the polygon \( H \) will be uploaded into sensor nodes in order to be used in on-line detection as a signature of the good values of the sensor.
Finally, from Figure 12, the threshold $Th$ is the ratio between the red points of sensed values that falls outside the polygon $H$, and the total number of sensed values of the dataset used to build the model.

6. CONCLUSION

In WSN applications, faulty sensors must be excluded and replaced within the network to guarantee the quality of service and the correctness of the decision to be taken, and also to avoid faulty data to be exchanged over the network, which reduces the communication overhead and keeps energy consumption low. In this paper, we have proposed an efficient approach for the detection of faulty sensor nodes based on Copula theory. We have shown all steps needed to build a model from historical measurements of a good sensor node. This model will be used as a signature with a predefined threshold to detect faulty sensors efficiently and in on-line manner. As future work, we intend to implement the proposed model within real sensor nodes.

References


