Outlier Detection in Wireless Sensor Networks Based on Copula theory

Farid. Lalem[a,*], Ahcène. Bounceur[a], Abdelkader . laouid[1], Reinhardt. Euler[a], Habib. Aissaoua[c]

[a] *Lab-STICC CNRS laboratory, University of Western Brittany, Brest, France*
[b] *University of Eloued, Research, Algeria*
[c] *University of sétif 1, Research, Algeria*

Abstract

Wireless sensor network has progressively emerged as a leading technology to develop new wireless applications that can revolutionize and improve our daily lives. Nonetheless, many WSN applications are mission-critical such as military and disaster area applications. Also, the quality of the sensed data stream collected by sensor nodes is affected by anomalies that occur due to various reasons like noise and errors, events, and malicious attacks. Thereby, anomaly detection in sensor readings efficiently is a major concern and imperative for making right decisions. In this paper, we address these challenges and we propose a new fully on-line distributed approach for anomaly detection in WSN based on Copula theory. Our experimental results on real datasets collected by real sensor network prove the efficiency and effectiveness of the proposed approach and, show that the proposed approach can achieve high detection accuracy with a low false alarm rate .

Keywords: Wireless Sensor Network, Copulas, Anomalies, Outliers.

1. Introduction

A wireless sensor network (WSN) consists of spatially distributed autonomous sensors that perform a collaborative measurement process to monitor physical or environmental conditions, such as temperature, humidity, light, sound, vibration, pressure, motion or pollutants.etc [1].

Furthermore the purpose of using a WSN is not only to collect data from the deployment area, but especially the analysis of these data at the convenient moment which allows to make of important decisions [19].

However, the raw measurements collected by sensor nodes is often unreliable and suffer from inaccuracy and incompleteness due to various reasons, noise and errors, events, and malicious attacks. Other reasons are related to the environment include the harshness and difficulties of the deployment area may also lead to incorrect data[6].

Therefore, ensuring data reliability and accuracy efficiently is a necessary task for decision-making process and conserving the limited resources of the network [8]. To achieve this goal, a robust fault detection scheme for WSN is essential.

In this paper, we address these challenges by proposing a new on-line fully distributed approach for anomaly detection based on Copula theory. Firstly, each node has a local polygon built by the application of Copula theory on historical measurements of data. Anomalies are detected in the incoming data by applying the input data stream containing measurements of monitoring physical phenomenon to this polygon, if the sensed values do not belong to this polygon, they are considered as an anomaly.

The remainder of the paper is organized as follows. In the next section, we present some definitions and basic concepts about anomaly detection in WSNs and an overview of related work. Section 3, describes basic concepts of copula theory. In Section 4, the proposed approach to detect anomaly with on-line manner is shown. Performance evaluation are presented in Section 5. Finally, Section 6 concludes the paper.

*Corresponding author
Email address: Farid.Lalem@univ-brest.fr (Farid. Lalem)

2. Background and related work

2.1. Related work

Actually, anomaly detection techniques in wireless sensor network are an important research field. Also, several schemes have been proposed in the literature by researchers to detect anomalies in sensor data with distributed or centralized manner.

The authors in [17] proposed two local techniques for the identification of faulty sensors as well as the event identification in sensor networks.To identify the sensors that gives the false reading, each sensor first calculates the difference between his own reading and the median reading from the neighbouring readings. Then, each sensor collects all differences from its neighbourhood and standardizes them. A sensor is faulty if the absolute value of its standardized difference is sufficiently greater than a preselected threshold.
The authors of [18] proposed a distance-based technique to calculate global top n outliers in a structure of aggregation tree, each tree node transmits some useful data to its parent after collecting all the data sent by his children, and then the sink floods the outliers to all nodes of the network for verification. If any node finds that it has two kinds of data which may modify the global result, it will send them to its parent in an appropriate time interval. This procedure is repeated until all nodes in the network are agree with the global result calculated by the sink. However, this technique takes into account only one-dimensional data.
The authors of [12] [13] proposed a distributed anomaly detection algorithm based on the data clustering using hyperspherical clusters (called fixed width clustering) to detect global anomalies. This technique assumes hierarchical network structure. The data at each node are clustered using hyperspherical clusters which creates a set of clusters with fixed width. Clustering is then performed based on Euclidean distance between the data vectors. Sensor nodes then sends cluster summaries that can be sufficiently represented by its centroid and the number of data vectors it contains, intermediate sensor nodes then merge the cluster summaries before communicating with other nodes. This process continues recursively up to the sink node, where an anomaly detection algorithm is applied to its merged clusters to identify the anomalous cluster. The cluster is declared as an anomalous cluster, If the average inter-cluster distance of any cluster is greater than one standard deviation of mean of inter-cluster distances of all the clusters. Summary of anomalous global clusters are communicated to all the nodes in the network. Then, each node, using the anomalous global cluster, identifies the corresponding globally anomalous data vectors at the node.
The authors of [16] proposes an integrated data compression and anomaly detection algorithm in WSN which can detect anomalies accurately using Discrete Wavelet Transform (DWT) and a single class Support Vector Machine (OCSVM). In this technique, DWT was used for compressing the data measurements at each node. Then, OCSVM is used to classify data into correct and anomalous data. However, the major drawback of this scheme is the high communication overhead caused by transmitting the whole data to the sink and the high computational cost of DWT .
The authors of [2] proposed an anomaly detection algorithm based on S transform and SVM. S transform is applied in order to reduce the size of the transmitted data and to extract the significant components of the time series data without losing the significant features of the data, and then SVM is used to classify data into original and anomalous. However, this approach generates high computational cost because it needs to classify every extracted components for final decision.

3. Introduction to Copula Theory

In WSNs, detecting anomaly with univariate data can easily be done by noting that the single data attribute is abnormal with respect to other data instances attributes. However, in multivariate WSNs it is difficult to detect anomaly because individual attributes may not show anomalous behaviour but when taken together they can display anomalous behaviour. Then, exploiting dependency between different attribute of sensor reading permit us to unlock this difficulties and allow us to propose a high accurate detection method. That's why, we decided to use the copula theory approach which allow us to captures the dependence relationship in a multivariate data.

3.1. Copula Theory and Mathematical Foundation

In this paper, for simplicity reasons we study only the theory of bivariate copula, because the multivariate theory is an extension of the bivariate case.

And we need only 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 [15], motivated by the work of Fréchet in the 50 s [9].

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

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

- $C(u,0) = C(0,v) = 0$ for each u , $v \in [0\,,1]$,
- $C(u,1) = u$ et $C(1,v) = v$ for each u , $v \in [0\,,1]$,
- C is an $2-$increasing function, i.e. for each $0 \leq u_i \leq v_i \leq 1$,

  $C(v1,v2) - C(v1,u2) - C(u1,v2) + C(u1,u2) \geq 0$.

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 set $F(x) = U$ and $G(x) = V$. Such as u , $v \in [0\,,1]$. The function $H(x,y)$ can be written in terms of a single function $C(u,v)$ such that:

$$H(x,y) = C(F(x),G(y)) \tag{1}$$

If F and G are continuous, then the copula is unique; otherwise, C is uniquely determined on (range of F) $\times$ (range of G). Conversely, if C is a copula, F and G are c.d.fs, then $H(x,y) = C(F(x),G(y))$ is a joint c.d.f with F and G as margins .

Schematically: if we have the marginal of each variable, simply, we join them with a copula function having the desired dependency properties to obtain the joint distribution, figure 1 show how to obtain the join distribution function of all variable.

3.2. The Empirical Copula

To model the observed dependence between random variable, we can use the empirical Copula structure to evaluate the suitability of a chosen Copula with the estimated parameter.

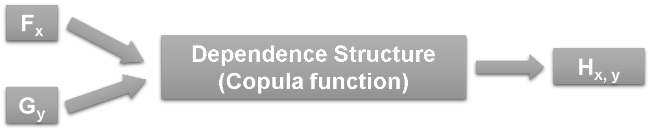

Figure 1: The dependence structure.

It is necessary to introduce the notion of rank before giving the formula of empirical copula. Given a sample $x_1, ..., x_n$ from a random variable X, the rank R_i of x_i is defined by the number of observations that are less than or equal to it. Then, the smallest observation of X has rank 1 while the largest has rank n.

Let U=$u_1, ..., u_n$ and V=$v_1, ..., v_n$ two variable uniformly distributed on I = $[0\,,1]$.

we denote by $(X$,$Y)$ two random variable, such as $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$, and let R_i the rank of x_i, and S_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 $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$:

$$u_i = \frac{R_i}{n+1} \tag{2}$$

$$v_i = \frac{S_i}{n+1} \tag{3}$$

Thereby, using this ranks, we can construct an empirical distribution function for the random variable X and Y [7].

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

$$C_n(u,v) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\frac{R_i}{n+1} \leqslant u, \frac{S_i}{n+1} \leqslant v) \tag{4}$$

The function $\mathbf{1}$ is the indicator function , which equals 1 if arg is true and 0 otherwise.

3.3. The dependogrammes

The dependogrammes allow us to understand graphically the dependency structure between two random variables, also, we obtained this deprogramme by a scatter plot of uniform margins (u,v) extracted from a sample or resulting from simulations of a theoretical copula.

3.4. How to choose the "best" copula that fit the data

When handling with bivariate data, we assume that we have a finite subset of copulas, and we are interested in knowing which one of them fits best the data, in this case we have to choose between graphical adequacy or analytic method.

3.4.1. graphical adequacy

We compare between empirical dependogram and theoretical dependogram .

3.4.2. Analytic adequacy

There are many method proposed in the literature which aims to find the best copula that fits best the data and we call this method Goodness of fit (GOF) test, among them we cite the Komogorov-Smirnov test [10], in this test, simply, we compare the empirical distribution function with the theoretical distribution function of each family and cheek if they both come from the same distribution.

4. The proposed Approach for Anomaly and Event Detection

In this section, we will present the proposed method of anomalies detection based on Copula theory in a WSNs. First, we will present the general idea of the approach. Then, we will present the steps to follow to build the model that will be used for the on-line detections.

Our proposed approach is divided into two stages, in the first stage, we build the model that will be used to detecting anomalies in off-line manner, thereafter, in the second stage, the model obtained has been uploaded into the sensor node to be used for on-line detection.

In the following, we explain the generic proposed model in the bivariate case.

4.1. building the model(offline)

The flowchart presented in fig 2 describes the process followed, and summarizes all phases needed for building the proposed model, to ensure reliable end efficient detection with high accuracy. The different phases that occurs during the different step are given as follows :

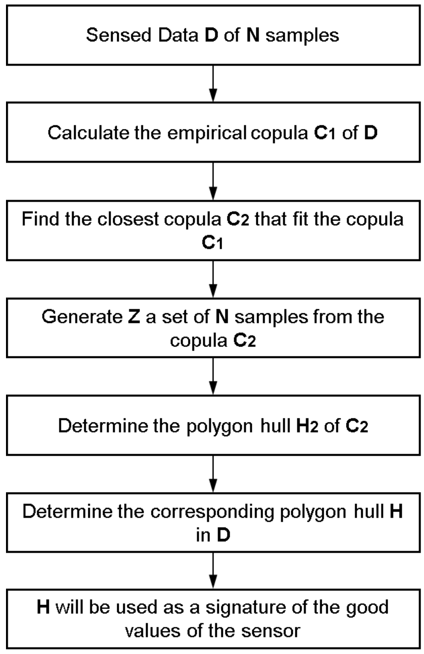

Figure 2: Offline Processing of sensors data

- *Step 1*: Based on historical measurements of sensed data, firstly, we begins by choosing N samples of two random variable X and Y witch represent two attribute of sensed data such as temperature and humidity. Then, let $D = \{X, Y\}$ the data set that will be used for the construction of the proposed model based on bivariate copula. with $X = (x_1, ..., x_n)$ and $Y = (y_1, ..., y_n)$.

- *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 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 fit the best the empirical copula C_1. In our case, we do this by graphical adequacy which is explained above in 3.4.1. Initially, we start by a scatter plot of the pairs $(u_i$,$v_i)$ derived from equations (2) and (3), for all $i \in \{1, ..., n\}$ of ranks derived from the learning data set $(x_i\ ,y_i)$, and by comparing the dependogram of the empirical copula C_1 with the dependograms of copula family cited in **??**, we find the closet family of the theoretical copula C_2 that fit the best the empirical copula C_1, also, the estimation of the copula parameters for calibrate the theoretical copula with the empirical copula, which is done in our case, by calculating the difference of the surface between the empirical copula and the theoretical copula, generated each time with a new parameter, and finally we calibrate the theoretical copula C_2 with the parameter that minimizes this difference of surface .

- *Step 4*: Once the copula family and its parameter is known, we generates a uniformly distributed sample $Z = \{U, V\}$ with $U = \{u_1, ..., u_n\}$ and $V = \{v_1, ..., v_n\}$, which is a set of N samples from the theoretical copula C_2 with the same size of the empirical copula C_1 by using the theoretical formula that correspond to C_2 .

- *Step 5*: In this step ,we calculate the polygon hull H_2 of Z using the algorithm of Eddy [5] for the calculation of the convex hull of a planer set of points. Then, we get all point that are vertices of the convex hull. in our case, by applying this algorithm on Z, we get the points H_2=(u_j,v_j)$\in$ Z for all $j \in \{1, ..., k\}$, k is the number of vertices of the polygon hull of Z.

- *Step 6*: Determine the corresponding polygon hull H in D. After calculating the subset of points H_2 that are vertices of the polygon hull of the samples Z, which is generated by the theoretical copula C_2, then, for each point (u_j,v_j)$\in$ H_2, we calculate the corresponding points (x_i,y_i)$\in$ D using the inverse transform sampling[14].

  Then, for each sampled value of (u_j,v_j) $\in H_2$, we calculate the corresponding value (x_i,y_i) in D space, which is given by $x_i = F^{-1}(u_j)$ and $y_i = F^{-1}(v_j)$.

  As a result, we get k points from D which are vertices of a polygon H.

- *Step 7*: Finally, the polygon H which is obtained from the application of copula theory and the convex hull of a planer set of points on the historical sensed data D, will be uploaded on sensors node to be used in on-line detection as a signature of the good values of the sensor .

4.2. detection processe(on-line)

To illustrate the process of on-line detection, we give an overview in the following flowchart 3.

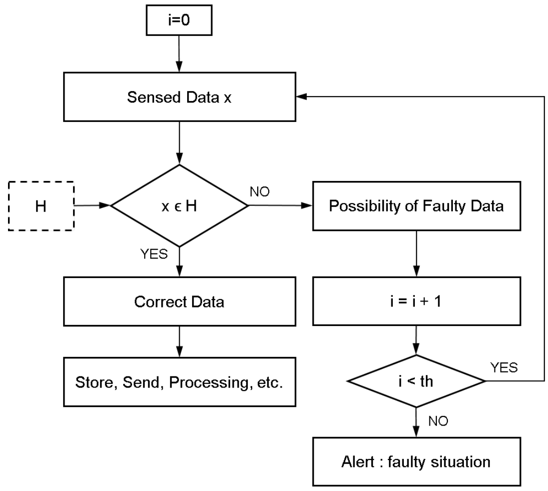

Figure 3: Online Detection of Anomaly

In this step, each node evaluate its sensed data with respect to the model built in the least stage. Initially, the node checks whether a new measurement falls inside or outside the polygon H using the simple algorithm of [11], in the case where it falls inside the polygon H, this is a correct data and will be processed,stored or transmitted. Otherwise, there is a possibility to faulty data, in that case, the counter i will be incremented, subsequently, if the value of i is less than a predefined threshold, the node pass to other instance of data, else, there is a faulty situation and an alert will be sent to the sink. In our case, we consider that the measurements that belong to the boundary of the polygon H are considered as correct data.

5. Performance Evaluation

5.1. Performance Metrics

To evaluate the proposed model in on-line detection , we will use, the detection rate (DR) and the false alarm rate (FAR)also known as false positive rate (FPR)as a metrics of evaluation. The detection rate is the ratio of correctly detected anomalous data to the total number of anomalous data. And, the false alarm rate is the ratio of the number of normal data which is incorrectly detected as anomalous data to the total number of normal data.

5.2. Result And Discussion

To show the effectiveness of our proposed approach, we used the data set of Intel lab [3], which contains a real measurements of data collected once every 31 seconds from 54 sensors, deployed in the Intel Berkeley Research lab. In this data set, we select the node number 16 for testing our approach. We take all pairs of (temperature, humidity) from the node 16. We have obtained 90 % of detected outliers with 10 % of false alarms,which is a very good score compared to other existing approaches.

6. Conclusion

In this paper, we have proposed an effective and efficient anomalies detection approach for detecting anomalous reading with on-line manner in WSN. We start by building the model that will be used in on-line detection based on Copula theory, after this, we upload the model built on sensor node to be used in on-line detection. Experimental result with real dataset, show that, the effectiveness of our proposed approach is very high by achieving 90 % of DR and ACC and lower FAR. In addition, the proposed approach is very efficient for utilizing the limited network resources such as memory and energy consumption.

[1] Ian F Akyildiz, Weilian Su, Yogesh Sankarasubramaniam, and Erdal Cayirci. Wireless sensor networks: a survey. *Computer networks*, 38(4):393–422, 2002.

[2] Anshuman Bhargava and AS Raghuvanshi. Anomaly detection in wireless sensor networks using s-transform in combination with svm. In *Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on*, pages 111–116. IEEE, 2013.

[3] Intel Lab Data, 2004.

[4] Paul Deheuvels. La fonction de dépendance empirique et ses propriétés. un test non paramétrique d'indépendance. *Acad. Roy. Belg. Bull. Cl. Sci.(5)*, 65(6):274–292, 1979.

[5] William F Eddy. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software (TOMS)*, 3(4):398–403, 1977.

[6] Eiman Elnahrawy. Research directions in sensor data streams: solutions and challenges. *Rutgers University, Tech. Rep. DCIS-TR-527*, 2:D3, 2003.

[7] Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

[8] Shah Ahsanul Haque, Mustafizur Rahman, and Syed Mahfuzul Aziz. Sensor anomaly detection in wireless sensor networks for healthcare. *Sensors*, 15(4):8764–8786, 2015.

[9] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[10] Marek Omelka, Irène Gijbels, Noël Veraverbeke, et al. Improved kernel estimation of copulas: weak convergence and goodness-of-fit testing. *The Annals of Statistics*, 37(5B):3023–3058, 2009.

[11] Joseph o'Rourke. *Computational geometry in C*. Cambridge university press, 1998.

[12] S. Rajasegarar, C. Leckie, M. Palaniswami, and J.C. Bezdek. Distributed anomaly detection in wireless sensor networks. In *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, pages 1–5, Oct 2006.

[13] Sutharshan Rajasegarar, Christopher Leckie, and Marimuthu Palaniswami. Hyperspherical cluster based distributed anomaly detection in wireless sensor networks. *Journal of Parallel and Distributed Computing*, 74(1):1833–1847, 2014.

[14] Ludger Rüschendorf. On the distributional transform, sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009.

[15] M Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[16] Saowaluk Takianngam and Wipawee Usaha. Discrete wavelet transform and one-class support vector machines for anomaly detection in wireless sensor networks. In *Intelligent Signal Processing and Communications Systems (ISPACS), 2011 International Symposium on*, pages 1–6. IEEE, 2011.

[17] Weili Wu, Xiuzhen Cheng, Min Ding, Kai Xing, Fang Liu, and Ping Deng. Localized outlying and boundary data detection in sensor networks. *Knowledge and Data Engineering, IEEE Transactions on*, 19(8):1145–1157, 2007.

[18] Kejia Zhang, Shengfei Shi, Hong Gao, and Jianzhong Li. Unsupervised outlier detection in sensor networks using aggregation tree. In *Advanced Data Mining and Applications*, pages 158–169. Springer, 2007.

[19] Yang Zhang, Nirvana Meratnia, and Paul Havinga. Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 12(2):159–170, 2010.