Anomaly detection through outlier and neighborhood data in Wireless Sensor Networks

Aymen ABID
CES-Lab, ENIS University of Sfax
Sfax, Tunisia
Email: aymen.abid.mail@gmail.com

Abdennaceur KACHOURI
LETI-Lab, ENIS University of Sfax
Sfax, Tunisia
Email: abdennaceur.kachouri@enis.rnu.tn

Adel MAHFOUDHI
CCIT, University of Taif
Taif, Saudi Arabia
Email: a.mahfoudhi@tu.edu.sa

***Abstract*—Anomaly detection through finding outlier measurements is an important issue for monitoring application using a large databases gathered by Wireless Sensor Network (WSN) like medicine and military.**

In this paper, we evaluate a detection of outliers based on the distance between the current measurement and its neighbors.

Our detailed evaluation supports a synthetic database generated with random values inserted into a real database from Intel Berkeley lab. In each round test, we increment the size of the learning window in order to have more outliers measurements.

The study of results highlights the importance of accuracy of detection that its average is 89%. Moreover, the detector provides a low false alarm rate with an average of 10% and a sufficient detection rate that can reach 100%.

Keywords*—*Big Data, Anomaly Detection, Wireless Sensor Networks, Outliers, Nearest Neighbor

I. Introduction

The large amount of data and its large size leads us to make treatments and analysis in order to organize, understand, filter, group, detect errors that exist and other possible arrangements. This implies a dynamic programming of big multidimensional databases [1] in order to identify certain information and exploit them [2]. This major analysis is called ”Knowledge Discovery in Databases” (KDD). Several areas are involved including machine learning, data mining, knowledge management, numerical analysis and outlier detection. We view that learning and data processing analysis in WSN may be successively or in parallel programming or with dynamic plan:

1) Learning → data processing
2) Learning $\|$ data processing
3) Learning $\rightleftharpoons$ data processing

In fact, in dynamic plan, the learning and outlier detection is done simultaneously. Indeed, the result of each execution of detection can update previous decision for each data. That is not the case of successive plan that make learning and detection separately, or the conjunction plan that make both simultaneously but without an update.

Data analysis in WSN can be used to identify normal and abnormal data or sensor. This is guaranteed by process of outlier detection that identify different anomalies.

Several studies investigate the detection of aberrant data and sensors within the WSN e.g. [3]. Anomaly detection by detection of outlier data techniques can be organized into clustering and classification, statistical, artificial intelligence and nearest neighbor techniques.

In this work, we focus on nearest neighbor technique using parallel learning plan for anomaly detection by the localization of outlier data.

KNN (k nearest neighbor) is frequently reviewed and implemented in pattern recognition and data mining. Also it is recently used for the detection of outliers using neighborhood information.

Authors of [4] try to model neighborhood relationship within a distance function. They define distance metric by neighborhood information in order to detect outliers in discrete and continuous data sets.

Authors of [5] aims to analyze multivariate data sets through a mutual graph of connectivity for K neighbors. This is let them to identify clusters and outliers in one step if the null hypothesis of no cluster structure or no outliers is rejected.

Authors of [6] propose an authorship verification scheme based on K-nearest neighbors classifiers. Using these feature vectors of classification, a majority vote system is applied to generate a decision about authorship for a meta-data document.

Authors of [7] present its formulation for outlier identification using a nested-loop algorithm. This algorithm computes for each point p the distance of its k^{th} nearest neighbor. And so, the n top most an outlier points are those having biggest distances.

In a related development with [7], [8] try to optimize the use of the nested-loop algorithm by an approximate nearest neighbor search. From this, it creates a canonical distance-based outlier detection algorithm to search outliers.

So, the main question of these methods is how to fix the K, the threshold n of top outliers or the threshold of neighborhood distance D. Indeed, with the foresight of neighborhood principle, outliers can have different at least three definitions. By homogeneity and density criteria, outliers are with less than k neighbors in the database [7]. By cardinality and ranking criteria, outliers are the n objects with highest distance values to their respective k^{th} nearest neighbor [9]. By distance criterion, outliers have the highest average distance D according to their k nearest neighbors [10].

This entire previous works respect one of this criterion and definition. The constraint of all is to define the k or the n or D or altogether. In our work, we discuss an outlier detector based

978-1-4673-8526-8/16/$31.00 ©2016 IEEE

on neighborhood data without a consideration of predefined K or n of top outlier or a threshold of accepted distance.

The rest of paper is organized as follows. In section II, we define our detector. Experimental results are in the section III. Finally a conclusion and tendency are presented.

II. Proposed Method

Proposed detector is an unsupervised Data Nearest for Outlier Detection (DNOD) that can analyze a learning data gathered by sensors in order to localize outlier measurements. As shown in figure 1, the data matrix of learning measurements is the only input of the detector. It not have any supervision of data labels or number of neighbors K or number of normal data to have outlier n or a distance threshold of accepted neighbors D.

The first step S1, nearest neighbors process, sort firstly measurements of all time slots (TS) for all sensor nodes (SN) in one data vector. After that, it calculates Euclidean distances between each current point in the sorted vector and its predecessor and successor. The current point is linked to its predecessor or successor having the smallest distance. It can also have an affiliation to the two neighbor points if the two computed distance are equal.

In the step S2, point-affiliations are analyzed. Each data in the vector not having a link is declared outlier. A point can be a simple measurement, a multidimensional data from a sensor node SN or time slot TS. This is in relationship with the input data of the detector.

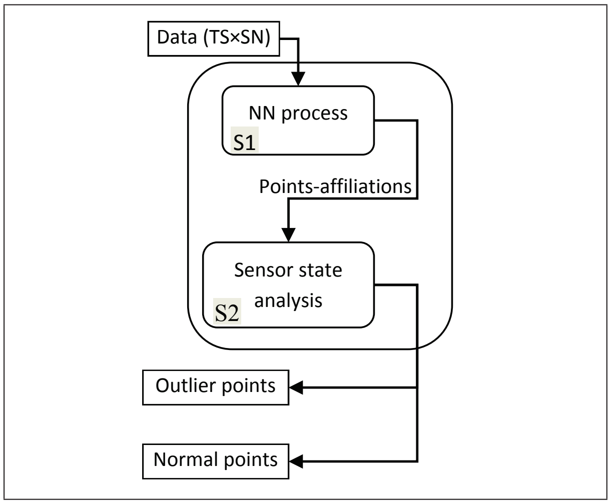

Fig. 1. Steps of the Data Nearest for Outlier Detection (DNOD)

III. Performances

A. Outlier metrics

Outlier metrics evaluate the detector based on this identification and classification of data into outlier and normal. Some question in mind can be provoked such as: Can-it find all outliers? Are-they all well declared outliers? Have-we omitted some abnormal? To respond to this, there are different ways and metrics. Some of them are defined by a confusion matrix as given in the table I [11].

TABLE I. Confusion matrix of outlier detection

	Predicted outlier	**Predicted normal**
Actual outlier	True positive TP	False negative FN
Actual normal	False positive FP	True negative TN

In fact, the detection rate is:

$$DR = \frac{TP}{(TP+FN)} \times 100, \quad (1)$$

The false alarm rate is :

$$FAR = \frac{FP}{(FP+TN)} \times 100, \quad (2)$$

Finally the accuracy is:

$$\begin{aligned} ACC &= \frac{TP+TN}{(P+N)} \times 100. \\ with & \\ P =& \ TP+FP \\ N =& \ FN+TN \end{aligned} \quad (3)$$

B. Configuration of the Simulation

The simulation of our algorithm has been executed with 1000 time slots (TS) from the Intel Berkeley base [12]. This base is used in several works specially in recent as [2] and [13]. This real base RDB is collected in an area of 3545m according to their locations and range of nodes (Tx-Range) is 40 m. The size of a data-packet, measured by the sensors and sent to their cluster heads, is 4000 bits.

In the same period (30s), the BS can receive many data from the same sensor or none. But we will save the latest data for simulation, because its the best to determine the state of sensor. At the end of each period, we have a vector of measurements called time slot TS. We obtain at last, 34375 values of temperature that we will explore by our detection process.

Outlier study is in Matlab 7.0.1 in parallel with detection process using a synthetic base SDB generated from the RDB, because the RDB have only few outlier points. The SDB contains in each TS a different synthetic outlier point generated using a random value ($RandVal$) from ran2 [14] in addition with a constant value deviated by 70% from the minimum measurement of RDB ($MinRDB$). So the 'Synthetic Random Outlier Value' is:

$$SROV = MinRDB - \frac{(MinRDB \times 70)}{100} + RandVal \quad (4)$$

We notice that this deviation can be presented statistically using the mean μ and σ the standard deviation. In the simulation, we use a window of time series TS that extend in each execution by one TS. With this manner the number of outliers increment linearly with one erroneous measurement.

C. Results of the Simulation

In this subsection, we treat relationships between predicted and actual states of data; outlier or normal.

Concerning FP in figure 2, it has a logarithmic trend alongside the enlargement of the window. This gives an inverse FAR, especially with linear TN.

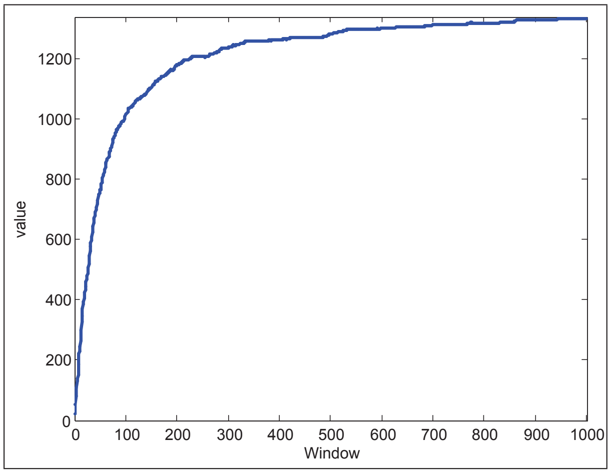

Fig. 2. False Positive curve through the size of the learning window

In fact, TN in figure 3 has the most number of measurements and well increases with the augmentation of learning window size.

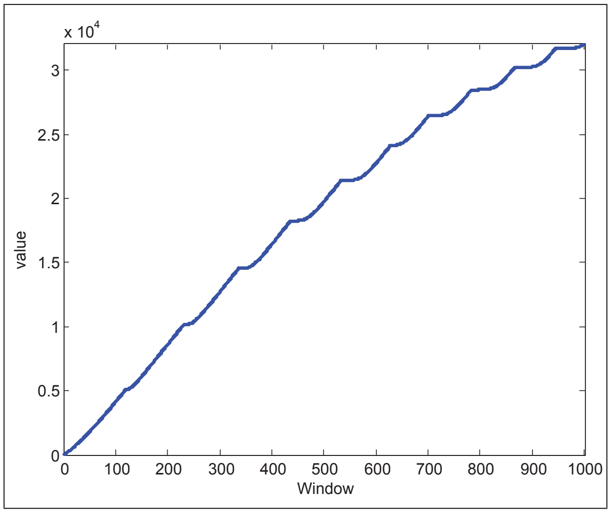

Fig. 3. True Negative curve through the size of the learning window

Thus as shown in figure 4, FAR have an inverse logarithmic curve that quickly goes down under 10% from 200TS.

Overall, the DR is close to 50%. Nevertheless, it is better for a small window below 100TS.

This medium percentage is due to the linear augmentation of FN. This linear incrementation of FN is figured in figure 6.

Also, TP as mentioned in figure 7 records progressively an important number from measurements really outlier. But it is not enough to increase DR.

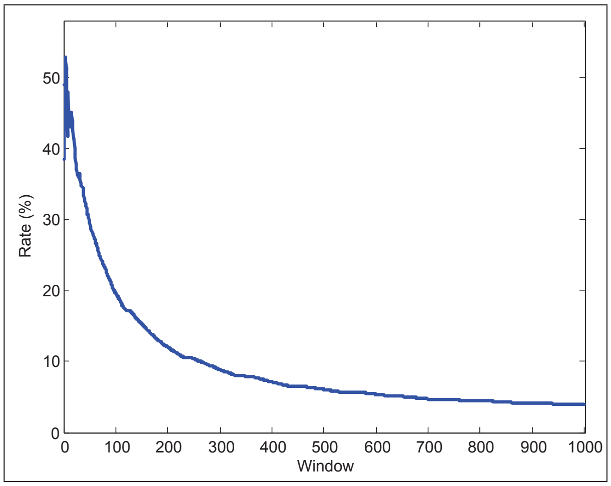

Fig. 4. False Alarm Rates curve through the size of the learning window

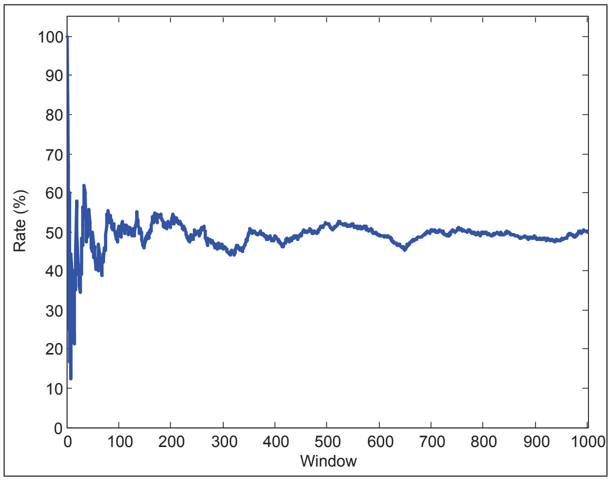

Fig. 5. Detection Rates curve through the size of the learning window

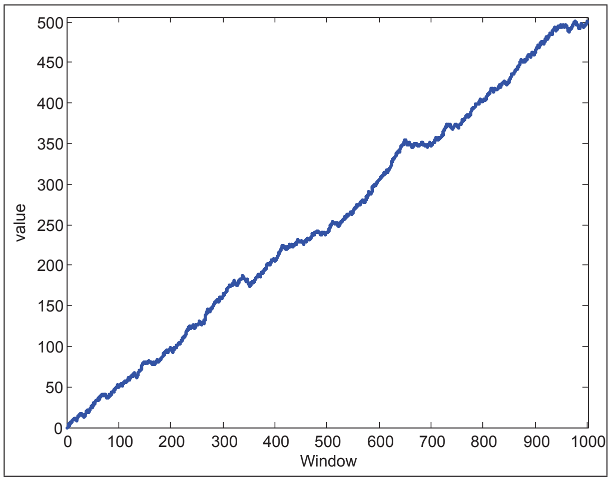

Fig. 6. False Negative curve through the size of the learning window

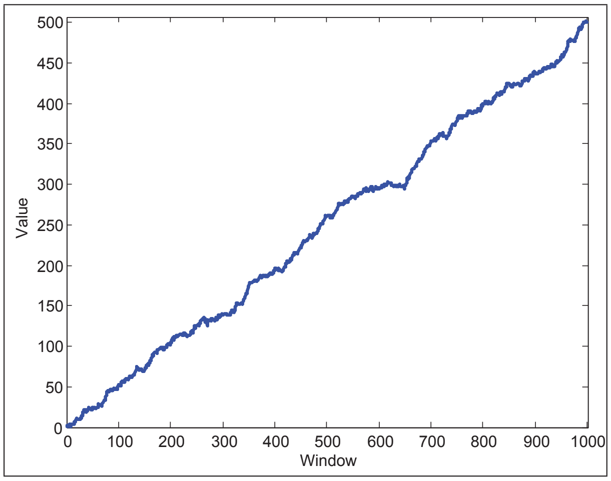

Fig. 7. True Positive curve through the size of the learning window

But for a global metric that use all indicators (TP, TN, FP and FN) in its rates, we find good results. Indeed, we show in figure 8 that the accuracy of the detector is upper than 90% and stable since 500TS.

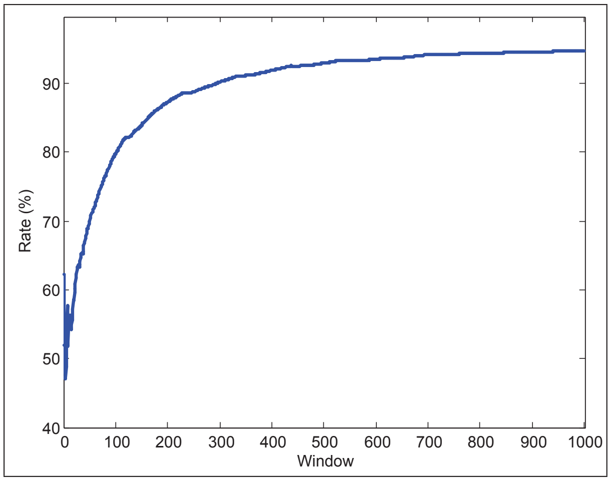

Fig. 8. Accuracy rates curve through an increasing size of the learning window

According to figure 9, positive classification of outlier with respect to the total number of outliers confirms the robustness of our detector. Indeed, the good classification of data that are not outliers into normal increase quickly with logarithmic tendency at first, as shown in figure 10. After that it have linear and incremental tendency since 20000 measurements that correspond to upper than 445TS.

In numbers as shown in table II, ”DR” managed to give a 100%. Also, ”ACC” arrives to establish 95%. Respectively, ”FAR” may down to 4%.

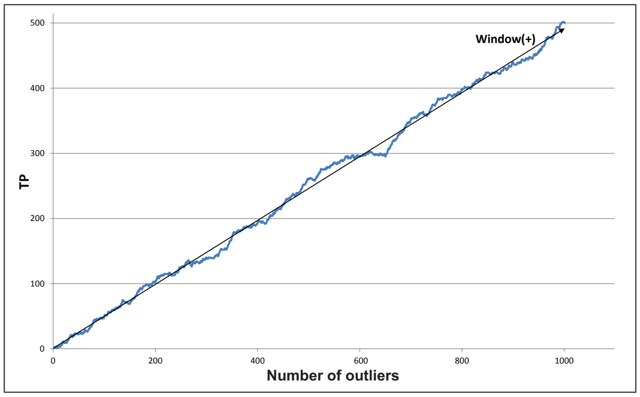

Fig. 9. Curve of True Positive in relationship with the number of existing outlier and according to an increasing window size within each simulation

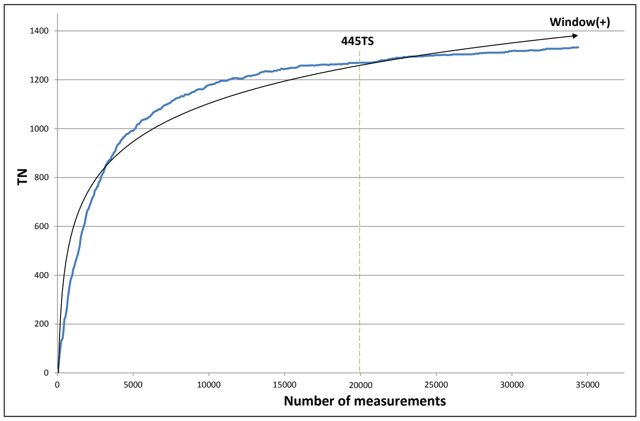

Fig. 10. Curve of True Negative in relationship with the number of measurements for each simulation with an increasing window size

IV. Conclusion

In order to detect anomalies, we propose a nearest neighbor process for anomaly detection by finding outlier measurements in data gathered by the WSN.

We present the performance of a detailed experimental study with synthetic base produced from a real-life data and outliers generated randomly. The simulation is with increasing learning window to have more outliers in each round in conjunction with testing of detector capabilities. Detection performances confirm the robustness of the Data Nearest for Outlier Detection (DNOD) with fair detection rate (DR), low false alarm rate (FAR) and good accuracy (ACC).

We plan testing with more complicated situations. Besides, we expect that this strategy may not need a specific configuration for other types of applications and even for a multidimensional detection in WSN.

TABLE II. Global results of our outlier detection (DNOD)

	DR	FAR	ACC
MIN	13%	4%	47%
AVG	49%	10%	89%
MAX	100%	53%	95%

References

[1] G. V. KUMAR, C. SREEDHAR, and K. NISHANTH, “Wireless sensor network for energy-efficient gas monitoring system using context-adaptive multimodal,” *IJARCET*, 2014.

[2] X. Luo and X. Chang, “A novel data fusion scheme using grey model and extreme learning machine in wireless sensor networks,” *International Journal of Control, Automation and Systems*, vol. 13, no. 3, pp. 539–546, 2015.

[3] A. ABID, A. KACHOURI, and A. MAHFOUDHI, “Anomaly detection in wsn: critical study with new vision,” in *International Conference on Automation, Control, Engineering and Computer Science-ACECS*. IPCO, 2014.

[4] Y. Chen, D. Miao, and H. Zhang, “Neighborhood outlier detection,” *Expert Systems with Applications*, vol. 37, no. 12, pp. 8745–8749, 2010.

[5] M. Brito, E. Chavez, A. Quiroz, and J. Yukich, “Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection,” *Statistics & Probability Letters*, vol. 35, no. 1, pp. 33–42, 1997.

[6] O. Halvani and M. Steinebach, “An efficient intrinsic authorship verification scheme based on ensemble learning,” in *Availability, Reliability and Security (ARES), 2014 Ninth International Conference on*. IEEE, 2014, pp. 571–578.

[7] S. Ramaswamy, R. Rastogi, and K. Shim, “Efficient algorithms for mining outliers from large data sets,” in *ACM SIGMOD Record*, vol. 29, no. 2. ACM, 2000, pp. 427–438.

[8] G. H. Orair, C. H. Teixeira, W. Meira Jr, Y. Wang, and S. Parthasarathy, “Distance-based outlier detection: consolidation and renewed bearing,” *Proceedings of the VLDB Endowment*, vol. 3, no. 1-2, pp. 1469–1480, 2010.

[9] E. M. Knorr and R. T. Ng, “Finding intensional knowledge of distance-based outliers,” in *VLDB*, vol. 99, 1999, pp. 211–222.

[10] F. Angiulli and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in *PKDD*, vol. 2. Springer, 2002, pp. 15–26.

[11] K. Andrea, G. Shevlyakov, N. Vassilieva, and A. Ulanov, “A new measure of outlier detection performance,” in *Machine Learning and Data Mining in Pattern Recognition*. Springer, 2014, pp. 190–197.

[12] S. Madden. (2004) Intel lab data. Intel Berkeley Research lab. [Online]. Available: http://db.lcs.mit.edu/labdata/labdata.html

[13] A. Appice, A. Ciampi, and D. Malerba, “Summarizing numeric spatial data streams by trend cluster discovery,” *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 84–136, 2015.

[14] A. ABID, A. KACHOURI, and A. MAHFOUDHI, “Assessment of anomalies detectors,” in *Tunisian Workshop on Embedded Systems Design TWESD*, 2014.