A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks

Article *in* Wireless Personal Communications · September 2017

DOI: 10.1007/s11277-017-4961-1

2 authors:

Hossein Saeedi Emadi
International University of Imam Reza
4 PUBLICATIONS **12** CITATIONS
SEE PROFILE

Sayyed Majid Mazinani
International University of Imam Reza
61 PUBLICATIONS **125** CITATIONS
SEE PROFILE

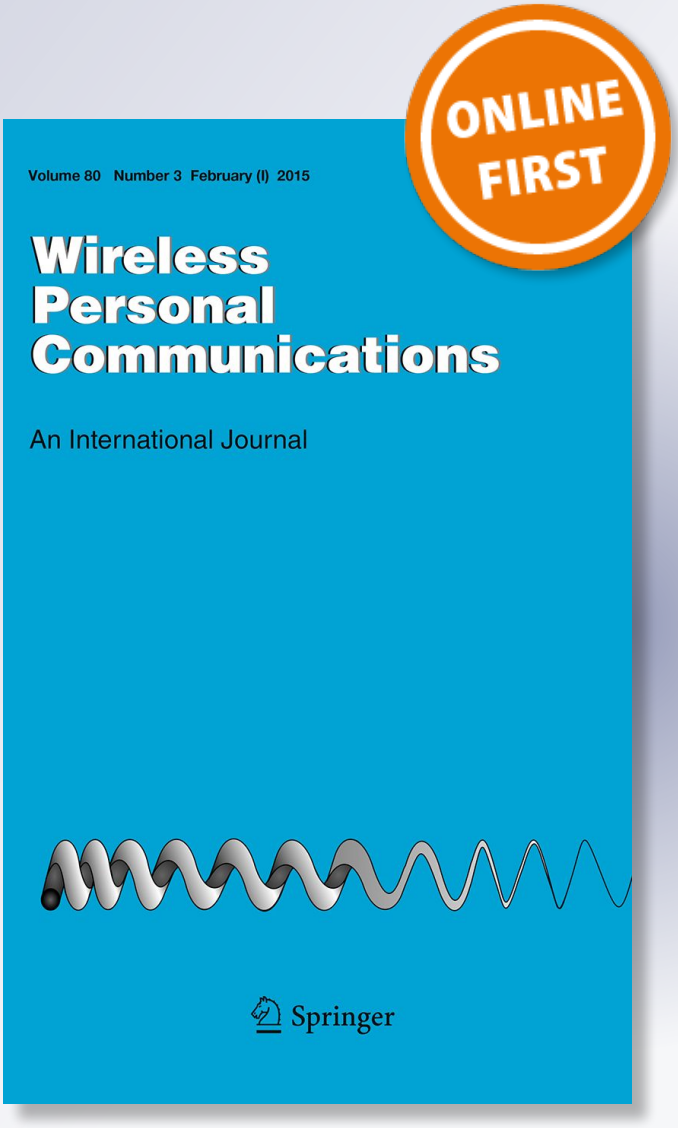
A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks

Hossein Saeedi Emadi & Sayyed Majid Mazinani

Wireless Personal Communications
An International Journal

ISSN 0929-6212

Wireless Pers Commun
DOI 10.1007/s11277-017-4961-1



Springer

Wireless Pers Commun
DOI 10.1007/s11277-017-4961-1

A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks

Hossein Saeedi Emadi[1] · **Sayyed Majid Mazinani**[1]

Abstract Anomaly is an important and influential element in Wireless Sensor Networks that affects the integrity of data. On account of the fact that these networks cannot be supervised, this paper, therefore, deals with the problem of anomaly detection. First, the three features of temperature, humidity, and voltage are extracted from the network traffic. Then, network data are clustered using the density-based spatial clustering of applications with noise (DBSCAN) algorithm. It also analyzes the accuracy of DBSCAN algorithm input data with the help of density-based detection techniques. This algorithm detects the points in regions with low density as anomaly. By using normal data, it trains support vector machine. And, finally, it removes anomalies from network data. The proposed algorithm is evaluated by the standard and general data set of Intel Berkeley Research lab (IRLB). In this paper, we could obliterate DBSCAN's problem in selecting input parameters by benefiting from coefficient correlation. The advantage of the proposed algorithm over previous ones is in using soft computing methods, simple implementation, and improving detection accuracy through simultaneous analysis of those three features.

Keywords Anomaly detection · DBSCAN · SVM · Wireless sensor network

1 Introduction

Wireless sensor networks (WSN) are comprised of small-sized devices which are called sensors. These sensors are able to process and transmit the data they receive from the environment. These sensors can be employed in places such as military environments, animal habitats, farm lands, factories, and power plants [1]. Due to tough environmental

✉ Sayyed Majid Mazinani
smajidmazinani@imamreza.ac.ir

Hossein Saeedi Emadi
h.saeedi@imamreza.ac.ir

[1] Electrical Engineering Department, Imam Reza International University, Mashhad, Iran

Published online: 11 September 2017

conditions, signal interference, low quality sensors, and other influential factors, sensors suffer from anomaly in receiving information [1]. Anomaly can cause changes and defects in data collected by wireless sensors. Given that they are not labeled, it is hard to detect anomalies in these networks. That is why anomaly detection in wireless sensors has become an important subject of intensive research. In these networks anomaly detection is done by analyzing the sensed measurements [2].

Due to the unknowability of their kind, detecting anomalies in wireless sensor networks (WSN) is difficult. Previous methods such as static clustering [3–6], support vector machine [7], and segment-based approaches [8] were confronted with the problem of the density of data and the absence of any general method for detecting the anomalies. The aforementioned methods were given to establish a threshold between the normal and abnormal data by calculating the threshold and using statistical methods to detect anomalies. To detect normal data in a feature which was exposed to environmental influences more than other features, these methods would consider, up to a certain threshold, some detections as normal and others as anomalous. These methods, however, suffered from inability to determine the exact place of the separating threshold between the normal and anomalous data such that in different situations and because of natural factors, data was considered above the threshold while the algorithm and the system detected them as anomalies. The other problem was with selecting the feature. By considering one feature, the algorithm was unable to detect the anomalous data in other features. In this paper, we have employed a density-based technique to determine the anomalies. We implement this method by selecting the three features on which anomalies have the most impact. Then, using spatial clustering algorithm in noisy regions (DBSCAN), normal and anomalous data are identified. And finally by means of support vector machine (SVM) they are with respect to the density of data in WSNs, these methods use machine learning algorithms in place of previous statistical approaches. Employing machine learning approaches in determining the exact place of the fault and in the configuration of the algorithm result in intelligent detection systems. This approach analyzes simultaneously more than one feature and, as a result, is more accurate in distinguishing between normal and anomalous data. In Sect. 2, we will analyze the proposed algorithm. In Sect. 3, results of evaluation, and in Sect. 4, conclusion will be presented. In the present paper, we will finally reach the following conclusions:

- Anomaly detection via evaluating several features using DBSCAN algorithm.
- Improving detection accuracy via classification using support vector machine (SVM).

2 The Proposed Detection Algorithm

This section deals with briefly analyzing the proposed algorithm called HSE in 6 stages and discusses its different aspects in more detail.

Algorithm 1: Network traffic classification steps

1: Selecting the features of temperature, humidity, and voltage from network traffic matrix.
2: initializing the input parameters of DBSCAN algorithm (MinPts, Epsilon)
3: Operating DBSCAN on data.
4: If CC $>$ 1, then it goes to step 2 until CC is minimized
5: Normal and abnormal data are labeled.
6: Data classification using SVM in network traffic capture

2.1 Feature Selection

Feature selection is a crucial step in the process of detection. In this section the most influential features during the process of detection must be selected. In this paper, the IRLB dataset has been used which enjoys 8 features, including history, time, epoch, moteid, temperature (T), humidity (H), and voltage (V). Among these features, temperature (T), humidity (H), and voltage (V) have the most impact on detection. Figure 1 shows the input data of sensor 3 with temperature as its feature; Fig. 2 presents the input data of sensor 3 with humidity as its feature; and Fig. 3 presents the input data of sensor 3 with voltage as its feature. These figures present sensors with different features in accordance with the order in which sensor 3 receives data.

2.2 DBSCAN Algorithm

The start of this algorithm is the arbitrary point of data which is not visited. The next step is to locate the points that are in $\in$ distance from the neighboring point p. The objective is to

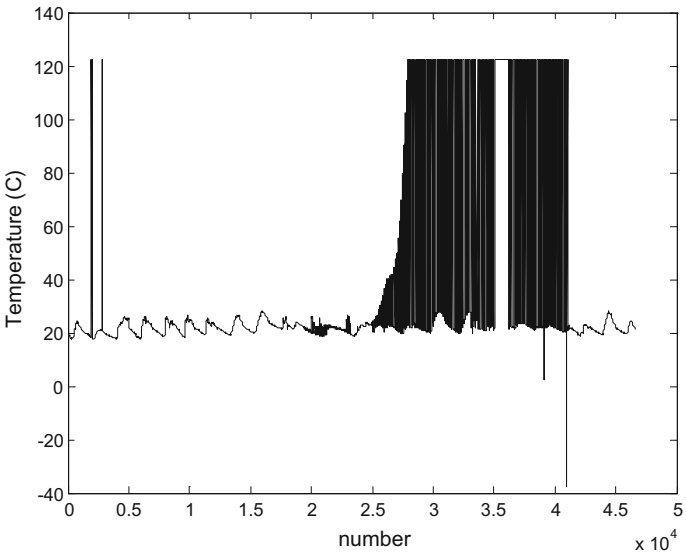


Fig. 1 Data received from sensor 3 with temperature as its feature

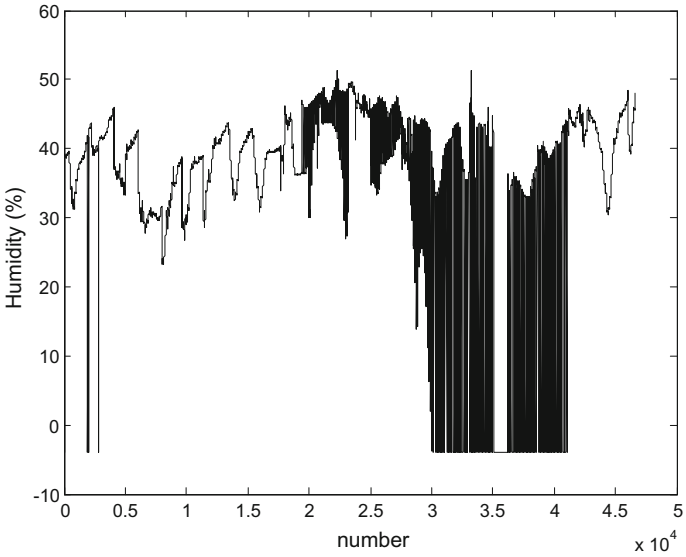


Fig. 2 Data received from sensor 3 with humidity as its feature

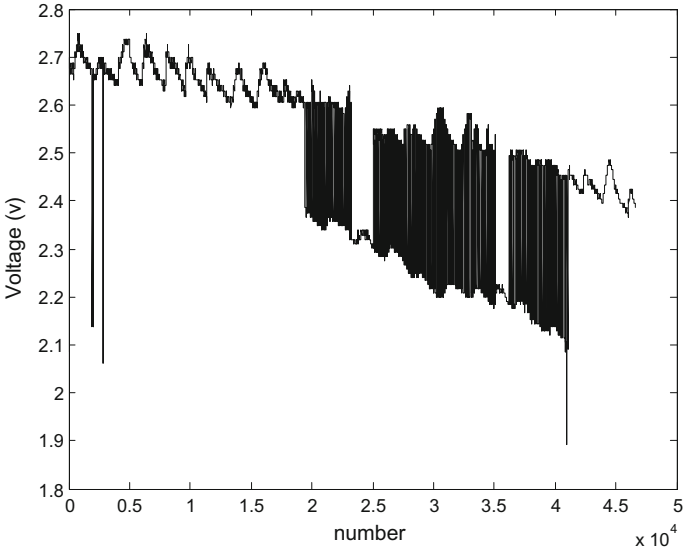


Fig. 3 Data received from sensor 3 with voltage as its feature

obtain the total number of data points within the neighborhood radius of $N\varepsilon(p)$ based on the formula (1):

$$N\varepsilon(p) = \{q \in D | dist(p,q) \leq \varepsilon\} \tag{1}$$

The comparison carried out between $N\varepsilon(p)$ and *MinPts* is meant to determine the state of data points in such a way that if $MinPts > N\varepsilon(p)$, then it will detect p as noise. And if

$MinPts \leq N\varepsilon(p)$, then it will dedicate p to another cluster. This method will continue for all the points that fall within the neighborhood of p until all cluster points are detected. After this cluster, the algorithm is continued until all the points are labeled as seen [9].

In this algorithm the distance between the two points is calculated based on Euclidean distance; with respect to the feature selection of the previous step, this is obtained from the following formula ():

$$dist(p,q) = \sqrt{(q_t - p_t)^2 + (q_h - p_h)^2 + (q_v - p_v)^2} \tag{2}$$

2.3 Coefficient Correlation (CC)

In DBSCAN algorithm, Epsilon and MinPts are the two parameters. In order to use DBSCAN to detect anomalies, the accuracy of the input parameters needs evaluating. Therefore, with the assumption that in the density-based approach we need at least two clusters to detect anomalies [2], we have the following equation for the CC, using input parameters and the number of clusters. CC illustrates the relation between the input and output according to the assumption of density-based anomaly detection. To select the best input parameter in DBSCAN algorithm, we need an equation that helps to obtain the best E and MinPts by imposing restrictions and repeating the algorithm.

$$CC = \frac{MinPts - E}{k * 10}. \tag{3}$$

Equation (3) illustrates coefficient correlation, where CC is the coefficient correlation, E is neighborhood radius, and $MinPts$ is the minimum number of points within a neighborhood radius, and K is the number of clusters. Figure 4 shows the final clustering. This relation is obtained by performing the algorithm in different parameters and comparing their obtained results.

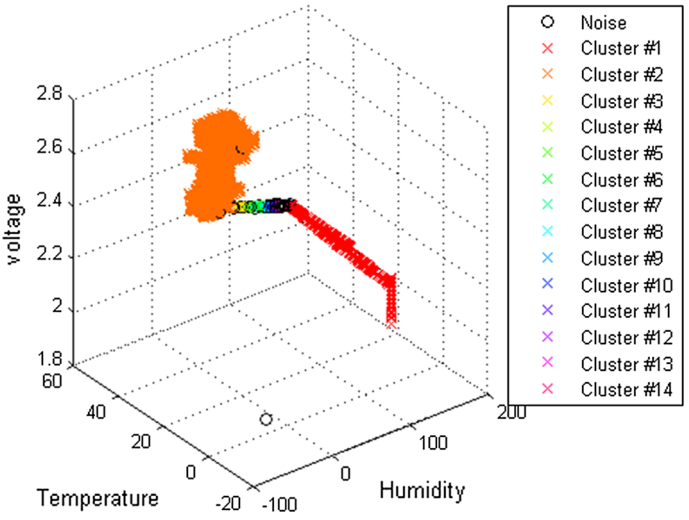


Fig. 4 Data clustering in the 3 features using DBSCAN

Table 1 Illustrates performing DBSCAN algorithm with different input parameters

Epsilon	MinPts	Cluster number	Coefficient correlation	ER	Accuracy detection (%)
5	20	2	0.75	0	25
3	30	8	0.33	0	66.2
3	20	3	0.56	0.0014	43.3
1	30	10	0.29	0.0019	70.1
1	25	16	0.15	0.0011	84.9
1	20	19	0.1	0.0006	89.9
1	15	16	0.08	0.0007	91.2
1	10	16	0.05	0.005	94.3
1	5	9	0.044	0.002	95.5

Table 1 shows the performing of the algorithm in different input parameters and the obtained results. In Table 1, Epsilon as neighborhood radius and MinPts as the minimum number of points within a neighborhood radius are DBSCAN input parameters. Cluster number is the number of clusters after performing DBSCAN, ER is the error of classification, and accuracy detection is the final detection percentage.

2.4 Data Labeling

The core idea behind density-based detection is that the majority of data are normal, and that they are accumulated in a cluster with high density [2]. After the best input parameters of DBSCAN algorithm are selected, we perform the algorithm, and the cluster with the highest density is considered as normal while clusters with low density are detected as anomaly and are labeled as such. In Fig. 4, the orange regions which enjoy the highest density level are considered as normal data region while other regions are detcted as anomaly. The square regions are noise.

2.5 Support Vector Machine (SVM)

After dertermining the normal data, we require a pattern to separate anomalous data from normal data. In this article, to classify traffic, we have employed a powerful classification algorithm called SVM which is tasked with constructing a pattern to separate data into normal and anomalous categories. Figure 5 illustrates data separation.

3 Experimental Results

In order to analyze the proposed algorithm, IBRL and Data Set are used. The total data accumulated by 54 wireless sensors during 3 days equals 3.2 billion records [10]. The position of sensors is presented in the following Fig. 6.

As it was mentioned in Feature Selection, each received data should consist of eight features. History, time, epoch, and moteid are correct. Temperature is in Celsius. Relative corrective temperature varies from 0 to 100%. Light is measured in lux (the measurement of 1 lux to moonlight, 400 lux to illuminance, and 1000 lux to direct sunlight are related).

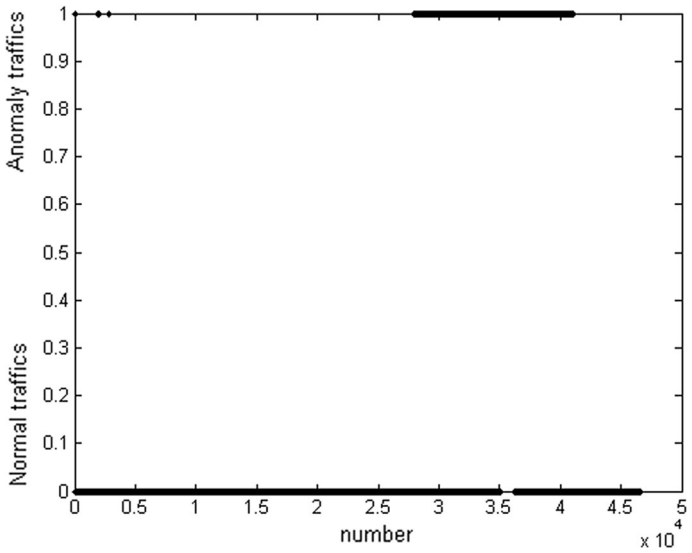


Fig. 5 Normal (0) and anomal (1) traffics detection

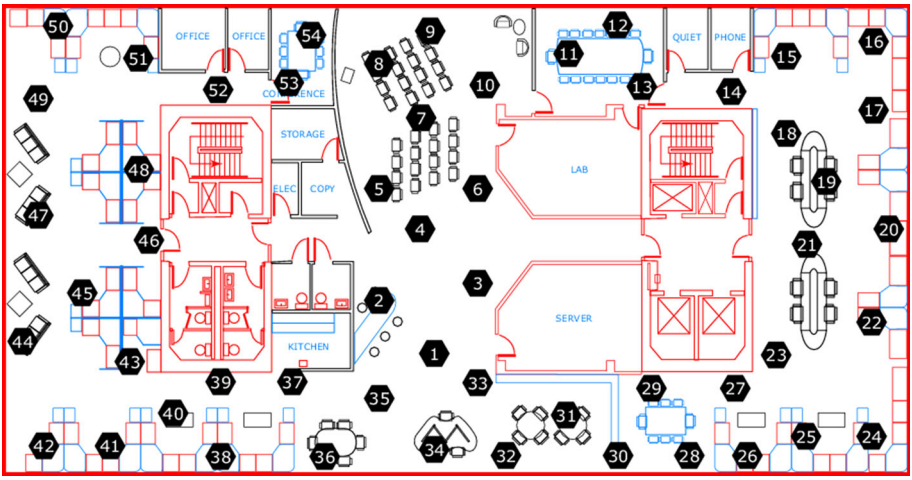


Fig. 6 Position of wireless sensor nodes [10]

Voltage is between 2 and 3 and is calculated in volts. The batteries are lithium cells which maintain a stable voltage throughout their lifetime. These variables are very sensitive to temperature. In our diagram only the three features of temperature, humidity, and voltage are presented.

The proposed algorithm was implemented on a Core i5 Laptop with 1.8 GHz CPU and 4 GB RAM, using Matlab 2014a software environment. Furthermore, SVM toolboxes are also used in the mentioned software. Table 1 shows the results obtained from the impact of input parameters on the accuracy of detection. Accordingly, Epsilon, the neighborhood radius, MinPts, and the minimum number of points in the neighborhood radius, are

Table 2 Accuracy detection of HSE algorithms

Type	Coefficient correlation	ER	Accuracy detection (%)
I	0.026	0.0028	97
II	0.028	0.0002	97.1
III	0.05	0.0012	94.8
IV	0.057	0.0028	94.1
V	0.03	0.0009	96.8

Table 3 Accuracy detection of segment-base

Type	Accuracy detection (%)
Type-I	94
Type-II	97
Type-III	87
Type-IV	89
Mixed	93.6

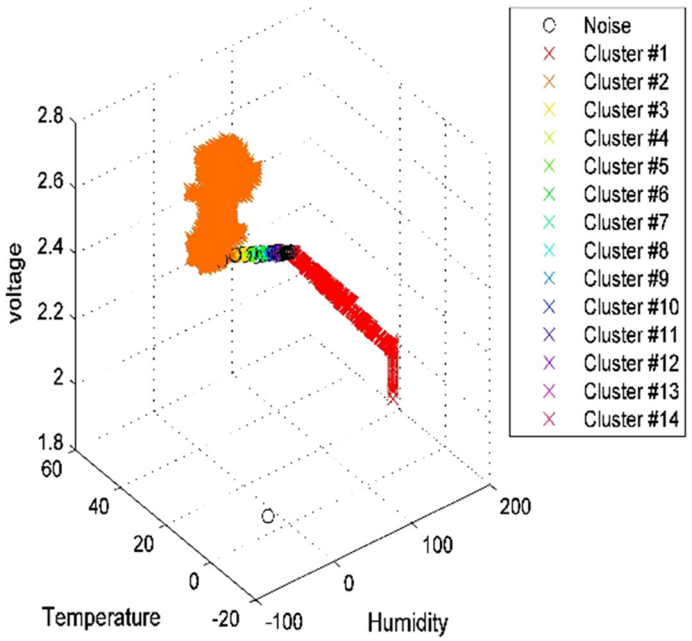


Fig. 7 Related to type 1

DBSCAN input parameters. Cluster number is the number of clusters after implementing DBSCAN, ER is the classification error, and accuracy detection is the final detecting percentage. To yield a fair evaluation, we have implemented the algorithm on some randomly selected nodes and have included the results in Table 2. To have a better understanding of the figures, we have also included each type Table 3 includes the results of evaluations of segment-based [8] approach in different states of 0.04 False Positive Rate

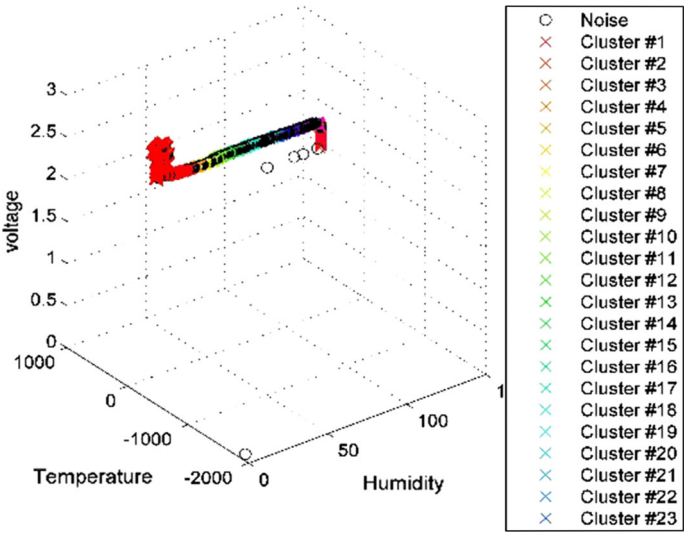


Fig. 8 Related to type 2

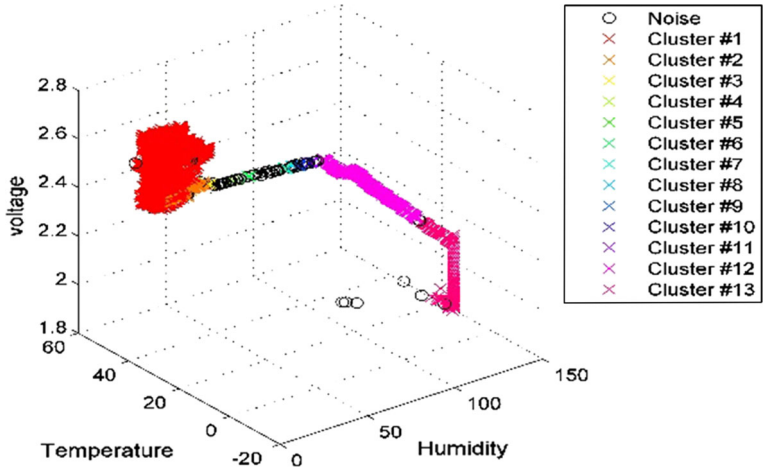


Fig. 9 Related to type 3

anomaly. When comparing Tables 2 and 3, we understand that the results obtained from the proposed algorithm, called HSE, are better than those obtained from the segment-based approach (Figs. 7, 8, 9, 10, 11).

4 Conclusion

This paper deals with the problem of anomaly detection in WSNs. Anomaly detection in these networks is confronting many challenges. One of these challenges is to separate the normal data from anomalies. In this paper, we have achieved a comprehensive method of detection using machine learning techniques. This method, by simultaneously comparing

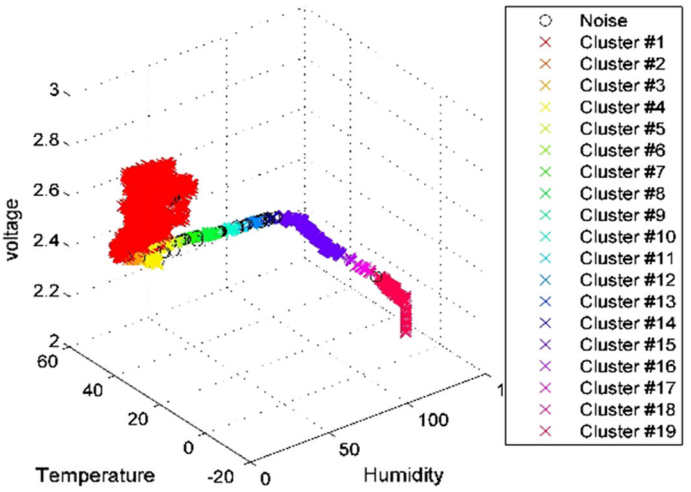


Fig. 10 Related to type 4

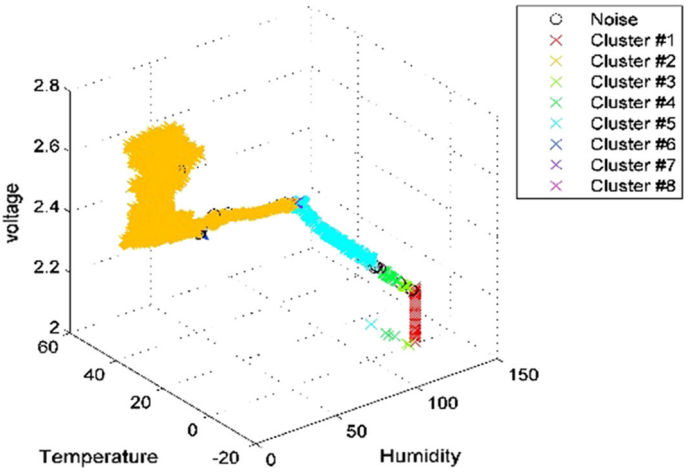


Fig. 11 Related to type 5

temperature, humidity, and voltage in various dimensions, increases detection accuracy in comparison with previous methods. The proposed HSE algorithm solves DBSCAN's problem of selecting input parameters by benefiting from coefficient correlation. A possible future research for the writers of this paper is to use the proposed algorithm to detect the anomalies brought about by the forms of denial of service attacks on the WSNs.

References

1. Yick, J., Mukherjee, B., & Ghosal, D. (2008). Wireless sensor network survey. *Computer Networks, 52*(12), 2292–2330.

2. Xie, M., Han, S., Tian, B., & Parvin, S. (2011). Anomaly detection in wireless sensor networks: A survey. *Journal of Network and Computer Applications, 34*(4), 1302–1325.
3. Liu, F., Cheng, X., & Chen, D. (2007). Insider attacker detection in wireless sensor networks. In *Proceedings INFOCOM'07* (pp. 1937–1945).
4. Xie, M., Hu, J., & Tian, B. (2012). Histogram-based online anomaly detection in hierarchical wireless sensor networks. In *Proceedings of IEEE 11th international conference trust, security and privacy in computing and communications (TrustCom)* (pp. 751–759).
5. Xie, M., Hu, J., Han, S., & Chen, H.-H. (2013). Scalable hypergrid KNN-based online anomaly detection in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems, 24*(8), 1661–1670.
6. Subramaniam, S, Palpanas, T., Papadopoulos, D., Kalogeraki, V., & Gunopulos, D. (2006) Online outlier detection in sensor data using non-parametric models. In *Proceedings 32nd international conference very large data bases* (pp. 187–198).
7. Rajasegarar, S., Leckie, C., Bezdek, J. C., & Palaniswami, M. (2010). Centered hyperspherical and hyperellipsoidal one-class support vector machines for anomaly detection in sensor networks. *IEEE Transactions on Information Forensics and Security, 5*(3), 518–533.
8. Xie, M., Hu, H., & Guo, S. (2015). Segment-based anomaly detection with approximated sample covariance matrix in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems, 26*(2), 574–583.
9. Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. *International Journal of Advanced Science and Technology, 31*(1), 59–66.
10. http://db.csail.mit.edu/labdata/labdata.html.

Hossein Saeedi Emadi was born in 1990 in marvdasht, fars Province. He received his B.Sc. degree in electronic engineer from Hadaf University in June 2012. He studied digital electronic engineer in Imam Reza International University of Mashhad for M.Sc. degree, in February 2016. His filed of interest are Data mining, VHDL, SVM, RF Design and artificial intelligence.

Sayyed Majid Mazinani was born in Mashhad, Iran on 28 January 1971. He received his Bachelor degree in Electronics from Ferdowsi University, Mashhad, Iran in 1994 and his Master degree in Remote Sensing and Image Processing from Tarbiat Modarres University, Tehran, Iran in 1997. He worked in IRIB from 1999 to 2004. He also received his PhD in Wireless Sensor Networks from Ferdowsi University, Mashhad, Iran in 2009. He is currently assistant professor at the faculty of Engineering in Imam Reza International University, Mashhad, Iran. He was the head of Department of Electrical and Computer Engineering from 2009 to 2012. His research interests include Computer Networks, Wireless Sensor Networks and Smart Grids.