EAI Endorsed Transactions
on Scalable Information Systems

Research Article **EAI.EU**

Machine Learning Based Hybrid Model for Fault Detection in Wireless Sensors Data

P. Raghu Vamsi[1] and Anjali Chahuan[2,]*

[1]Assistant Professor, Department of CSE, Jaypee Institute of Information Technology, Noida, India.
[2]Assistant Professor, Department of CSE, Inderprastha Engineering College, Ghaziabad, India.

Abstract

Wireless Sensor Networks (WSN) refers to a group of spatially deployed and dedicated sensors for sending, recording, and monitoring the physical conditions of the environment and transmitting the collected data to a central location. The major challenge is to extract high level knowledge from such data. Detecting abnormality in such data can help finding the faulty sensor and also the sensor collecting the most interesting reading from the dataset. This paper proposes a machine learning based hybrid model for knowledge discovery that works best with multivariate time-series data. The Intel Berkeley Research lab (IBRL) dataset is one of the most trending dataset collected by a WSN is considered for the study. The spatial-temporal correlation was also taken as reference to find anomalies in the dataset using three models- 1)Histogram Based Outlier Score (HBOS), 2) Minimum Covariant Determinant (MCD) and 3) Isolation Forests (IF). Further, the electrical configuration about components of WSN has been used to find faults among the outliers found in the dataset. The results show that the proposed hybrid model with Isolation Forest outperformed with a precision of 94.86%. The experiment was also able to spot the least trustful or faulty sensors among the deployed sensors in IBRL dataset.

Keywords: Anomaly Detection, Outliers, Fault Detection, Wireless Sensor Networks, Internet Of Things (IOT), Intel Berkeley Research lab (IBRL), Knowledge Discovery, Time-series data, Pattern Recognition, Histogram Based Outlier Score (HBOS), Minimum Covariant Determinant (MCD), Isolation Forests (IF).

Received on 21 August 2019, accepted on 01 November 2019, published on 05 November 2019

doi: 10.4108/eai.13-7-2018.161368

1. Introduction

Anomaly detection is a main field of research in the context of many industrial and many applications domains nowadays [1]. Anomaly is considered to be any data instance that stands different from the rest of the data. Anomaly detection, especially in the case of big data is considered to be the pre-processing step to get data that is free from errors [2]. Nonetheless, proceeding further with any kind of data without anomaly detection is not a reliable method. The data that records from the same source follow a definite pattern. Any sort of deviation from the definite pattern brings on an idea of an anomaly. However, anomaly should not always be considered error. It may also show some interesting patterns that stand different from the rest of the data [13]. Data collected from WSN are quite interesting and show different patterns [4]. Anomaly detection in WSN is the process to diagnose those data instances that deviate from the rest of the data patterns based on a single or more than a single measure. WSNs comprise of many small, economical sensor nodes, put together for sensing via short-range wireless communication. They act as a subset of the Internet of things (IoT) when they send data to a cloud over an Internet. They have strong resource constraints in terms of energy, memory, computational capacity, and communication bandwidth [3]. They are usually deployed in remote and harsh environments [4]. Moreover WSNs are sensitive to faults and malicious attacks which cause inaccurate and unreliable sensor readings, which are continuously being collected over IoT [5]. Sensors periodically collect the data and send the data to the base

*Corresponding author. Email: anjisingh.chauhan@gmail.com station (BS). BS records the data using a timestamp. Such data are known as Time series data or the temporal data [18]. Such data not only focuses on the values of its attributes but also its distribution based on time stamps. The patterns in such data are not only value centric, but also depend on the temporal continuity [22][26-32]. In this context, time forms the contextual variable with respect to which all analysis is performed. Precisely, temporal data from WSNs are very much prone to outliers, due to their electrical configurations and sensor precautions [16]. The recorded values are dependent on various physical conditions and any deviations in those conditions bring outliers in the data.

1.1 Motivation

The WSN data is generally multivariate time series data and to deal with anomalies in such data require much more computations, as the outlier detection depends on several features of the dataset. Also, in such data, even the timestamp contributes in detecting anomalies. Consequently, traditional anomaly detection techniques are not directly applicable to WSN data due to their specific perquisites, dynamic nature, and resource limitations [7]. The main objective of outlier detection in WSNs is to identify outliers in the streaming data in an online manner for IOT systems with high detection accuracy along with maintaining the utilization of the network to a minimum [8]. However, anomaly detection in the sensor data also calls for removal of erroneous sensors and the anomalies that have shown some significant change over time. Different machine learning models have shown some great results for anomaly detection, but since sensor data is multivariate and models dealing with such data are minimum. Also, for a better and accurate detection of anomaly, unsupervised methods like, One Class Support Vector Machine (OCSVM), Isolation Forests (IF), K-Nearest Neighbor (KNN), Local Outlier Factor (LOF) etc can be used with temporal data [11]. Since, time-series data comprises of continuous temporal data streams; so, unsupervised machine learning methods work better with them, as the new anomalous entries could be identified with each changing data stream. Supervised learning methods do not perform well and hence, cannot be used for continuous temporal data [16].

To this end, this paper proposed machine learning based hybrid model for fault detection in wireless sensor data. Point outliers were found in the time-series multivariate data using unsupervised machine learning models. The proposed anomaly detection is based on Spatio-temporal correlation. An intensive simulation for the models based on IBRL dataset [25] that demonstrates that the approach to find anomalies and faults outperforms both in effectiveness and efficiency.

The rest of this paper is organized as follows. After receiving literature in section 2, Section 3 presents the background and definitions. Section 4 presents the proposed work. Section 5 presents the simulation study on IBRL dataset. Finally, Section 6 concludes the paper with future work.

2. Literature Survey

Anomaly detection is a well studied subject in sensor networks. Various researches in [1, 2], have come up with a lot of approaches to deal with anomaly detection. These researches have summed up the trivial problem of anomaly detection in a way that all the scenarios and approaches regarding the same are presented. Jing Gao et al. [21] have presented the anomaly detection research specifically in time-series data. All these works have been a great motivation as well as guidance to carry forward the particular research. However, since, data around the world is increasing day by day and so are the problems within the data, lot more continuous research is required in this field.WSN and IOT sensors generates big data and required a lot of computation within the data. Also, because of the continuous streaming nature of this data, pattern recognition from such data becomes more important. Gionani et al. [18] has dealt with sensor data using filtering approach. This approach involved high computational cost gives an idea of improving the methodology. Similarly, researches in [3-5] have worked on the processing in WSN data and gave an idea about how the WSN data should be observed. When dealing with time series, it was observed that windowing the data was the most used concept to find the anomalies in temporal data used in [20-22].

However, windowing work perfectly with univariate data. This way, the value of the single feature could be plotted and using averaging techniques, the anomalies could be found. Windowing also advocated the technique of forecasting using Auto-Regressive Integrated Moving Average (ARIMA) in [18]. Other deep learning methods like, auto-encoders, Long short –term memory (LSTM), Deep belief networks (DBN) could also be applied to forecast the next time window in time series by learning the previous one[19]. The anomalies were found with the help differences between the actual and forecasted plots. Such plotting only helps with a single feature or multivariate data to work. This created a constraint for such models to use multivariate data and it was inferred that anomaly detection in such data involved finding the point outliers [11-12].

Research in [20-23] has applied different models such as DBSCAN and One- Class SVM to find outliers in multivariate data. But due to time and space complexity, these were not considered for large-scale data like the IBRL data. Romi et al. [24] worked upon different techniques to find point outliers that were low on the computational complexity. These techniques involved distance based models, parametric, non parametric models and domain based models. Using the most efficient and simplest models from this research, the paper proposes three types of models on time series data, i.e., probabilistic models, and distance based model and domain based models. These types of models had many models under each category. The paper picks up the best models from the research like Histogram-Based Outlier Score(probabilistic) [14], Minimum covariant Determinant (Distance based) [15] and Isolation Forests (Domain based) [9, 17]. The proposed approach applies these models on time-series data to fetch for the best fit model to find outliers in temporal data collected by WSN.

Research that involved work on IBRL data to find outliers was also considered. Asma Fawzy et al. [5] proposed a clustering based approach on IBRL data. However, the methodology, deals with neighbourhood information to find anomalies in a single feature (i.e. Temperature). Hence, such a big data was constrained to just a single feature. Bosman et al. [10] also performed similar kind of research using correlation techniques in each feature. Both researchers were focusing on single feature each, which does not tell the behaviour of all the sensors. Biased results for some features may create False Positive results. This way of disregarding any feature of this data might give ambiguous findings. Hence, the proposed approach deals with all the features as well as sensors together. The research advocates the idea of less False Positive rate and also taking complete data together because the data from a WSN hold great importance and the pattern in such data is correlated. Hence, disregarding any of the feature or the sensor could bring in ambiguous results. The findings of this paper will be the faulty sensors out of all the 54 sensors in IBRL data.

3. Assumptions and Dataset Description

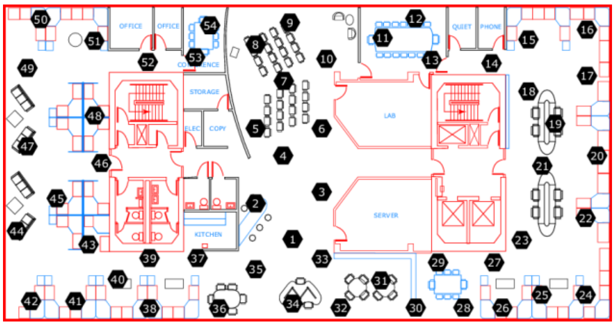

Figure 1. Sensors placement in Intel Lab [25]

The dataset contains information about the data collected from 54 sensors deployed (as shown in Figure 1) in the Intel Berkeley Research lab (IBRL) between February 28th and April 5th, 2004 [25]. This dataset was collected with epoch duration of about 30, making it a total 65,000 epochs and about 2.3 million readings. Two readings from the same epoch number were produced from different motes at the same time. Along with it, one more file from the same dataset had helped to perform spatial analysis. The file contained the x and y coordinates of sensors (in meters relative to the upper right corner of the lab) available in [25] is used. Sensors are numbered from 1 to54 as Mote Ids; data from some sensors might be missing or truncated. Temperature is in degrees Celsius. Humidity is temperature corrected relative humidity, ranging from 0 to100. Light is in Lux. Voltage is expressed in volts, ranging from 2 to 3; and have remained fairly constant over their lifetime. Variations in voltage are highly correlated with temperature. No outlier labels were present in the dataset, so, to check the model with the detection rate and accuracy, 10% artificial anomalies were infused.

4. Proposed Method

In this section, the proposed approach is introduced in detail. Anomalies showcase different forms of abnormalities within the dataset. Some data anomalies can be junk data being the errors in the dataset, while for some dataset these anomalies can impart important information in the form of novelties. As mentioned, Intel lab data consists of 54 sensors, which are collecting data continuously. The proposed methodology takes in context the electrical configuration of WSN to find the errors in the dataset. In WSN, Voltage plays a major role in collecting the reading by the sensors. A great fluctuation in voltage can give an erroneous reading which also needs to be accounted. So, while performing anomaly detection three features will be taken into consideration i.e., Temperature, Humidity and Light. The feature 'Voltage' is extracted later to find the faults. Hence, following this approach the anomalies found are further classified to find if the anomalous reading is an interesting data or an error.

As, sensor data tends to be correlated in both time and space. The proposed approach classifies the sensors using spatial clusters to assist measuring the spatial correlations. The approach uses time stamps between reading for anomaly detection in sensors to measure the temporal correlations The proposed methodology goes in the following sequence: ALGO 1 -> ALGO 2 -> ALGO 3 as showed in the Table 1. The stepwise description of the proposed hybrid model is given below.

4.1 Data pre-processing

This phase is the most important phase of the proposed methodology as it consists of most of the data manipulations that are necessary to find the anomaly in the respective dataset.

Data cleaning

Being a large dataset, IBRL data require data cleaning. Many of the readings were having null values (NaN), which would lead to ambiguous results. Also, the sensor data are mentioned with fifty-four sensors. However, reading consisted of more than fifty-four sensors and had recorded some random values. Hence, it was necessary to remove these extra sensors from the dataset, due to their non-existence. Also, all the null values in different columns were being replaced by their corresponding columns' mean value.

Correlation

When dealing with more than one feature for anomaly detection, it is more important to know the correlation among the features given in the dataset. So, in order to find the correlation between the four features i.e, Temperature, Voltage, Humidity, Light, the Pearson correlation coefficient [10] was used as a measure of the linear correlation between any two features. Here, 1 is a perfect positive correlation. – 1 is a perfect negative correlation, and 0 means no correlation. A low correlation would be $r < 0.25$; a high correlation means $r > 0.75$. These correlations gives an idea of which features will contribute the most to the anomaly score of their respective readings. The heat map in Figure 2 gives the Temperature, Humidity, and Voltage as the most correlated features in the dataset. Hence, these features will contribute at the most while predicting the outliers.

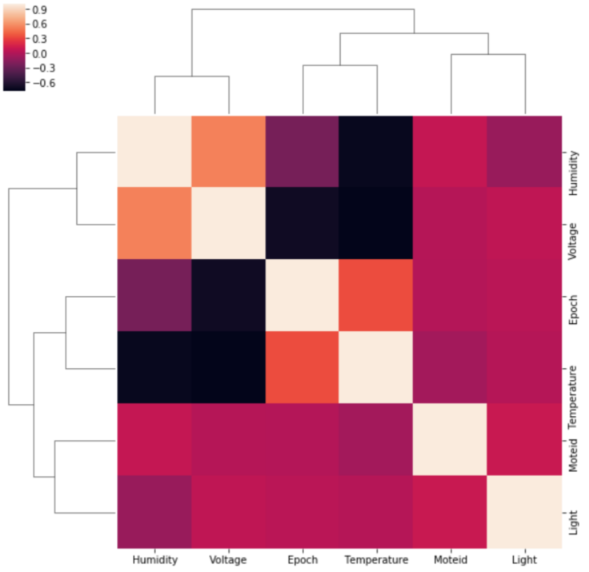

Figure 2. Correlation heat map of attributes in the dataset

Temporal modification

The IBRL data is a time series data of continuous data streams. The timestamp in the data consists of both date and time. Since, we are dealing with the temporal correlations, it is necessary that reading falling in the same timeframe are analyzed together. So, based on the hours in the day and climatic changes, the time stamps in a day were divided into three phases and named as 'Time of the Day' (TOD). Reading between 12 am to 9 am and after 5 pm was assigned the same TOD. Similarly, readings between 9 a.m. to 1 p.m. were assigned same TOD. And, the reading between 1 pm to 5 pm was assigned the same TOD. TOD acted as windows for our anomaly detectors. Justifying the temporal correlation within the data, the anomalies were found in each TOD indicating the abrupt values within the same time.

Spatial clustering

The IBRL dataset has its sensors scattered over a wide area for different coordinates. This information of coordinates of each sensor was used to perform spatial clustering [20]. Spatial clustering was done using K-Means algorithm to find the clusters that group according to the closest sensors in the same space. Using elbow method on the location data of the sensors at IBRL, three clusters were fetched. Different sensors were falling in different clusters. These clusters were used to deal with spatial correlation and hence analyzing the nature of neighbouring sensors.

4.2 Anomaly Detection

This section deals with the models that were used to find anomalies in each TOD found during temporal modification in the IBRL dataset. Three unsupervised machine learning models were taken into consideration to detect the anomalies in the multivariate data.

HBOS (Histogram- Based Outlier Score)

Histogram- Based Outlier Score (HBOS) is a statistical unsupervised anomaly detection algorithm [14].This algorithm is computationally less expensive than most of the proximity- based anomaly detection methods. Due to its low complexity, this algorithm is highly suitable for large-scale datasets [11]. HBOS work on discretionary data by offering a standard fixed bin width histogram as well as dynamic bin width [5]. Anomaly detection usually involves huge gaps in the value ranges. Due to the fact that outliers are far away from normal data, it is recommended to use the dynamic width mode, especially if distributions are unknown, besides the number of bins k needs to be set. A rule of thumb given in [5] is setting k to the square root of the number of instances N. Now, for each dimension d, an individual histogram has been computed (regardless of their bins), where the height of each single bin represents a density estimation.

The histograms are then normalized such that the maximum height is 1.0. This gives an equal weight of each feature to the outlier score. Finally, the HBOS of every instance p is calculated using the corresponding height of the bins where the instance is located by (1).

$$HBOS(p) = \sum_{i=0}^{d} \log(\frac{1}{hist_i(p)}) \qquad (1)$$

The score is a multiplication of the inverse of the estimated densities of each feature.

Minimum Covariant determinant (MCD)

This technique is used to identify the "best fit" mean and covariance in the presence of up to half of the anomalies in the dataset. This technique takes a number h between n/2 and n of non-anomalous data points as a parameter, and the algorithm finds the ellipsoid of least volume that covers h points, leaving out the n h largest outliers. It also finds the mean and covariance of the remaining h points. On receiving next data points, Mahalanobis distance is computed which score the abnormality of the new data. It assigns high score for very anomalous data, low score for relatively common data, according to the distribution of the DataStream. The DataStream of IBRL data falls under symmetric distribution and hence is suitable for the MCD model of anomaly detection. The Formula for Mahabolonis distance is calculated by (2).

$$MD(x) = \sqrt{(x-\bar{x})^t S^{-1}(x-\bar{x})} \qquad (2)$$

Here $MD\sqrt{x_i}\sqrt{}$ indicates how far away x_i is at the centre of the dataset, relative to the size of the dataset. Here $x\sqrt{}$ is the sample mean and S the sample covariance matrix.

Isolation Forests (IF)

Isolation Forest consists of multiple isolation trees, namely Isolation Tree, which is created by choosing attributes and the values of attributes randomly. At each node in the isolation trees, the instances set is divided into two parts based on the chosen attributes and its values. The attributes are selected randomly and the split value for this selected attribute is selected randomly as well between the minimum value and maximum value of this selected attribute. The anomalous instances are those objects that their attributes values are very different from the normal instances and are easier to be divided than normal instances'. In the process of isolation, they are also closer to the root and more easily divided than the normal instances. In order to alleviate the effects imported by the random characteristic in the process of building the isolation forest, the average depth of the anomalous score of the instance. The lower score of the instance has, the higher probability is there that it is an anomaly.

Fault Detection

Faults in the sensors were found using the electrical hardware irregularities. In a Wireless Sensor Network (WSN), while collecting data from sensors, voltage plays a major role. In IBRL dataset, the voltage was mentioned constant throughout the time readings were taken. Also, a mandatory condition for a WSN to work properly is that the fluctuation in voltage while collecting data should be negligible. However, if the fluctuation occurs, it might give out some erroneous or faulty reading. Since, Voltage readings are normally distributed in the given dataset, so finding the most extreme points of its normal distribution will give out the fluctuating reading of voltage in IBRL data. Following this, whenever an anomaly was detected corresponding to a fluctuating voltage, it indicated a faulty reading that occurred due to the voltage fluctuations.

Measuring sensor trustfulness

Based on the findings in each TOD for anomalies and faults, the sensors that recorded most of the anomalous values of different features of IBRL data, were checked for faults according to fault detection phase. Taking into consideration the percentage of outliers and faults, the most anomalous sensors in the respective WSN were found. The more the sensor had faulty reading, the less it was to be trusted for any analysis. The cluster having most number of faults was also fetched.

Table 1. The proposed hybrid method

```
PREREQUISITES:
AnModels: List of all models(IF, MCD, HBOS)
Sensors: List of all 54 mote Id associated with the sensors.
Features: List of features in the dataset (Temperature, Humidity, Light)

ALGO 1:
Step 1: Pre-process the data with data cleaning, correlation tests and temporal modification to produce a list of Time Zones containing subsets from the IBRL dataset.
Step 2: Perform spatial clustering using K-Means to get a list of clusters.

ALGO 2:
Input: Subsets of IBRL dataset under each Time Zone
Output: Anomalous and faulty data instances.

Step 1: Determine outlier points(data instances) using all the items in Features.
  For every Time Zone in IBRL dataset
    Begin
          Perform Anomaly detection using each AnModels to get labels
          Outliers= labels for each data instance
          Assign Outliers to dataset
          Find extreme 'Votlage' instances of Voltage feature, distributed normally to get labels for 'Voltage'
          VAnom= labels for each Voltage instance
          Assign VAnom to dataset
         For each label i in Outliers
          Begin:
            For each label k in VAnom
              Begin:
                If i==k==1
               Then classify the data instance as a fault
               Else classify the data instance as an outlier
           End
End

ALGO 3:
Input: Labeled dataset with faults and outliers
Output: Top 20 most anomalous and faulty sensors.

Step 1: For each Mote id in Sensors
         Begin:
            Calculate outlier percentage using (3)
            Calculate fault percentage using (4)
                 Get the top 20 sensors having maximum fault and outlier percentage.
          End
```

5. Simulation Study

This section presents the effectiveness of the proposed hybrid model and its effect on detecting anomalies. Also, as 54 sensors mentioned in IBRL dataset [25] are placed in a WSN, the proposed approach finds the least trustful sensor based on the percentage of faults it is depicting. After fetching the results, the sensor recording the most number of faulty data is found. Proper remedies can be applied to the faulty sensor to reduce the erroneous data in such a big and important dataset.

5.1 Simulation setup

In order to ease the process of anomaly detection, the proposed approach uses a fraction of IBRL data on the basis of dates, i.e. from '1-3-2004' to '15-3-2004.' Also since there was no ground labels available in the IBRL dataset to check for accuracy of the predicted outliers by the different models, 10% of artificial outliers were infused to the dataset, such that accuracy metrics could be calculated using the labels of these artificial outliers.

Table 2. Cluster map

Cluster id	Sensor Mote id
0	4-13, 48-54
1	1-3, 32-47
2	14-31

5.2 Results Analysis

The proposed approach deals with finding out the spatial clusters based on the locations of the sensors. Three clusters were found using the K-Means algorithm on the basis of x and y coordinates given in the IBRL dataset. Figure 3 shows the three clusters of sensors, labeled in different colours respectively. These clusters which are showing the spatial correlation between the sensors help in analyzing similar sensors and data generated from them, on the basis of distance between those sensors. The cluster map showing the particular Sensor mote id with their Cluster id can be seen in Table 2.

Furthermore, after applying the proposed model, outliers within the data were collected by each model, i.e., IF, MCD and HBOS. They were analyzed based on two metrics, such that the most efficient model can be used for anomaly detection within the IBRL dataset. The paper uses the Precision score and Area Under the Curve (AUC) score as metrics, to compare different models. These metrics are most suitable as they describe properties that are naturally expected from a good anomaly detection system. The comparative metrics can be seen in Figure 4. Analyzing these figures, it was found that Isolation forest outperformed out of the three models in detecting the anomalies. Since, the three models used to detect anomalies were based on different fundamentals; it was also found that domain based anomaly detection works best for finding anomalies in Time-Series data.

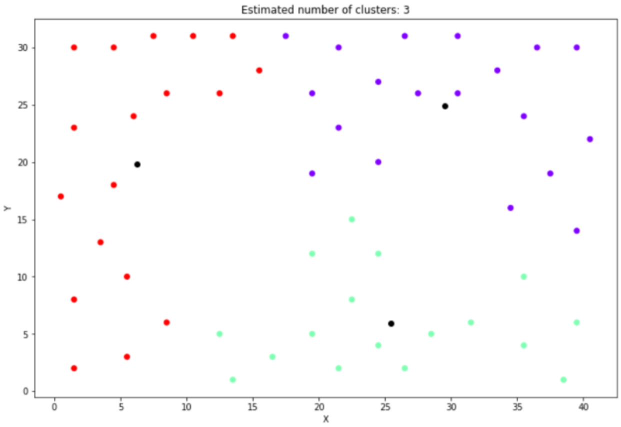

Figure 1. Clusters of nodes formed by K-means

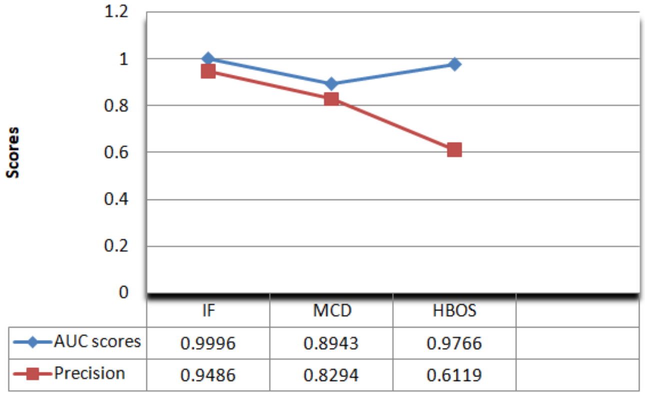

Figure 2. Comparison between models used for validation metrics

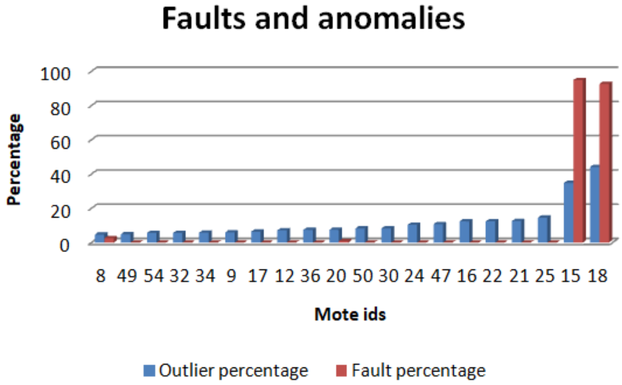

Figure 3. Top 20 Sensors having highest fault and outlier percentage

In this way, anomalies were detected in each TOD of the IBRL data. Following the outlier detection, fault detection was done to find the faulty sensors. The plot in Figure 5 shows the top sensors having highest outlier percentage in their recorded data. The sensors having Mote id 8,49,54,32,9,17,12,36,20,50,30, 24,47, 16, 22,21, 25, 15, 18 were found to be having most number of outliers. Later, out of these outliers, the sensor was tested against faults and hence, fault percentage was found.

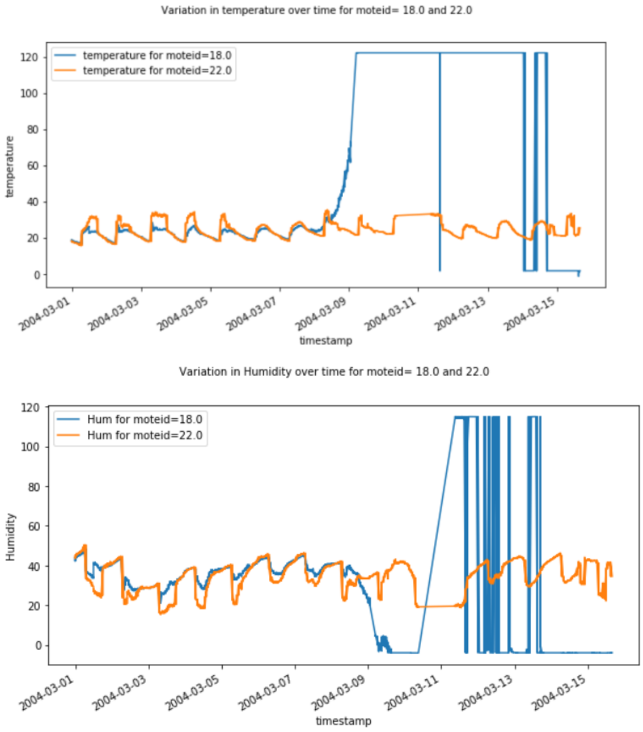

Figure 4. Depiction of faulty sensor readings of Temperature and Humidity with respect to a normal sensor

Fault percentage is the calculation of faults within the outliers found in each sensor. Each sensor was analyzed through Equation 3 and 4, to find the least trustful sensors. However, by analyzing the plot, it was observed that node 15 and 18 were the one collecting the most number of outliers. But, they also possess the highest fault percentage. This way, **15 and 18** were the least trustful sensors. However other sensors from top 20 sensors showing the highest percentage of outlier percentage gave an idea that node **21, 22 and 25** shows the highest outlier fraction but no fault percentage giving out novelties within the data.

$$Percentage_{outlier} = \frac{No\ of\ anomalies\ \ found}{total\ \ no.of\ readings} \times 100 \qquad (3)$$

$$Percentage_{faults} = \frac{No\ of faults\ \ found}{total\ \ no.of\ outliers} \times 100 \qquad (4)$$

Figures 6 show the faulty pattern of data recorded by the sensor 15 (i.e. Sensor containing most of the faults) plotted with the two most correlated features of the IBRL data with respect to other non faulty sensor. Sudden spikes in temperature and humidity readings showed the anomalous pattern in their data.

Most of the sensors from top 20 sensors that showed a significant amount of anomalies as well as faults fall in the same cluster (Refer Table 1). This showed that due to spatial correlation and transfer of data in the closer sensors in space, similar behaviour of sending out anomalous data reading was seen.

6. Conclusion and Future Work

This paper presented a hybrid method for anomaly as well as fault detection in time series WSN data. In the proposed model, it is observed that Isolation Forests performed well in detecting anomalies. The major challenge addressed in the proposed approach is to find anomalies in the multivariate time-series data. The method of choosing and categorizing windows of data with respect to Time of Day (TOD) helped to find trends in the IBRL data. As part of future we, we attempt to extend the proposed model for detecting the in-network data anomalies in the WSN.

References

[1] Hart, J., & Kamber, M. (2001). Data mining: concepts and techniques. M or an Kaufmann Publishers, 200(1), 223-259.

[2] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.

[3] Tan, P. N. (2006). Knowledge discovery from sensor data. Sensors Magazine.

[4] Akyildiz, I. F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). A survey on sensor networks. IEEE Communications magazine, 40(8), 102-114.

[5] Fawzy, A., Mokhtar, H. M., & Hegazy, O. (2013). Outliers detection and classification in wireless sensor networks. Egyptian Informatics Journal, 14(2), 157-164.

[6] Gaber, M. M. (2007). Data stream processing in sensor networks. In Learning from Data Streams (pp. 41-48). Springer, Berlin, Heidelberg.

[7] Zhang, Y., Meratnia, N., & Havinga, P. J. (2010). Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 12(2), 159-170.

[8] Rassam, M. A., Zainal, A., & Maarof, M. A. (2013). An efficient distributed anomaly detection model for wireless sensor networks. AASRI Procedia, 5, 9-14.

[9] Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. IFAC Proceedings Volumes, 46(20), 12-17.

[10] Bosman, H. H., Iacca, G., Tejada, A., Wörtche, H. J., & Liotta, A. (2017). Spatial anomaly detection in sensor networks using neighborhood information. Information Fusion, 33, 41-56.

[11] Goldstein, M., & Uchida, S. (2016, February). A Comparative Study on Outlier Removal from a Large-scale Dataset using Unsupervised Anomaly Detection. In ICPRAM (pp. 263-269).

[12] Stojanovic, N., Dinic, M., & Stojanovic, L. (2017, December). A data-driven approach for multivariate contextualized anomaly detection: Industry use case. In 2017 IEEE International Conference on Big Data (Big Data) (pp. 1560-1569). IEEE.

[13] Nguyen TL, Le TA, Pham C. The Internet-of-Things based Fall Detection Using Fusion Feature. (2018) 10th International Conference on Knowledge and Systems Engineering (KSE) 2018 Nov 1 (pp. 129-134). IEEE.

[14] Goldstein M, Dengel A. (2012) Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. KI-2012: Poster and Demo Track.59-63.

[15] Fauconnier C, Haesbroeck G. (2009) Outliers detection with the minimum covariance determinant estimator in practice. Statistical Methodology. 6(4):363-79.

[16] Hosseini M, Borojeni HR. (2018) A Hybrid Approach for Anomaly Detection in the Internet of Things. InProceedings of the international conference on smart cities and the internet of things 2018 Sep 26 (p.3). ACM.

[17] Marteau PF, Soheily-Khah S, Béchet N. (2017) Hybrid Isolation Forest-Application to Intrusion Detection. arXiv preprint arXiv:1705.03800.

[18] Giannoni F, Mancini M, Marinelli F. (2018) Anomaly Detection Models for IoT Time Series Data. arXiv preprint arXiv:1812.00890.

[19] Zhang A, Song S, Wang J, Yu PS. (2017) Time series data cleaning: From anomaly detection to anomaly repairing. Proceedings of the VLDB Endowment. 10(10):1046-57.

[20] Khaleghi A, Ryabko D, Mary J, Preux P. (2016) Consistent algorithms for clustering time series. The Journal of Machine Learning Research. 17(1):94-125.

[21] Gupta M, Gao J, Aggarwal CC, Han J. (2014) Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering. 26(9):2250-67.

[22] Chatfield, C. (2003). The analysis of time series: an introduction. Chapman and Hall/CRC.

[23] Khan MA, Khan A, Khan MN, Anwar S. (2014) A novel learning method to classify data streams in the internet of things. In National Software Engineering Conference 2014 Nov 11 (pp. 61-66). IEEE.

[24] Domingues R, Filippone M, Michiardi P, Zouaoui J. (2018) A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition. 74:406-21.

[25] Intel Lab Sensor Data: http://db.csail.mit.edu/labdata/labdata.html

[26] Tsai, Feng-Ke, Chien-Chih Chen, Tien-Fu Chen, and Tay-Jyi Lin. (2019) Sensor Abnormal Detection and Recovery Using Machine Learning for IoT Sensing Systems. In 6th International Conference on Industrial Engineering and Applications (ICIEA), pp. 501-505. IEEE.

[27] Kuo, Yu-Hsuan, Zhenhui Li, and Daniel Kifer. (2018) Detecting Outliers in Data with Correlated Measures. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 287-296. ACM.

[28] Lee, Tae Jun, Justin Gottschlich, Nesime Tatbul, Eric Metcalf, and Stan Zdonik. (2018) Greenhouse: a zero-positive machine learning system for time-series anomaly detection. arXiv preprint arXiv:1801.03168.

[29] Sun, X., Wang, H., Li, J. and Zhang, Y., (2012) Satisfying privacy requirements before data anonymization. The Computer Journal, vol.55, no.4, pp.422-437, 2012.

[30] Sun, X., Wang, H., Li, J. and Zhang, Y., (2011) Injecting purpose and trust into data anonymisation. Computers & security, 30(5), pp.332-345.

[31] Zhang, J., Li, H., Liu, X., Luo, Y., Chen, F., Wang, H. and Chang, L., (2015). On efficient and robust anonymization for privacy protection on massive streaming categorical information. IEEE Transactions on Dependable and Secure Computing, 14(5), pp.507-520.

[32] Kabir, E., Mahmood, A., Wang, H. and Mustafa, A., (2015) Microaggregation sorting framework for k-anonymity statistical disclosure control in cloud computing. IEEE Transactions on Cloud Computing, pp 1-10.