

**Masters Second Year (2020-2021)**

**Computer Science for Aerospace**

**Report**

**On**

**Faulty Sensor Detection Using Machine learning**

**Under Master Thesis**

**Submitted by:**

1. Pratvi Kalavadiya (21911764)

**Supervised by**

**Prof. Rahim KACIMI**

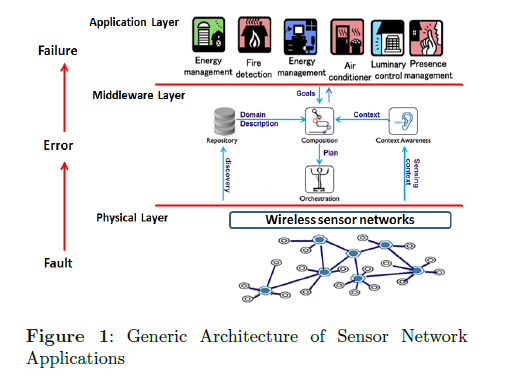
[**kacimi@irit.fr**](mailto:kacimi@irit.fr)

**INDEX**

|  |  |  |
| --- | --- | --- |
| *Sr. no.* | *Contents* | *Page no.* |
|  |  |  |
| 1. | **INTRODUCTION** | **2** |
| 2. | **STATE OF THE ART** | **3** |
| 4. | **WORK FLOW AND ROAD MAP** | **5** |
| 5. | **PLANNING AND IMPLEMENTATION** | **6** |
| 6. | **REFERENCES** | **9** |
|  | **CONCLUSION(Add after finalization)** |  |

1. **INTRODUCTION**

Revolution in technology introduced Internet of things (IoT) to the world which makes world more digitalized. The impact can be seen in our everyday lives. That digitization seems almost impossible without Wireless sensors. Mostly to collect real world data like environment conditions temperature, humidity or military surveillance we use WSN. Generally, sensors are used in WSN are two types: event base and weather analysis. When sensors are used for weather analysis it’s continuously collecting data. It is said that there will be 15 billion sensors by the year 2025 [1].according to authors of [5] its research that says only 49% data are useful in WSN. Collected data most of the time are corrupted or faulty and many reasons are present. That is why it is compulsory to ensure that collected data is accurate. This work usually done manually. Now next step will be to make it more reliable and that’s why machine learning is hot topic especially for WSN field. Many machine learning techniques can be used to automate the system. There will be two tasks find the suitable model to find faulty sensors in the network using machine learning technique and implement system for the University Paul Sabatier WSN. This WSN exists to collect weather data.



1. **STATE OF THE ART**

In this session I would like to present this topic in more detailed manner. This session is more about the researches done till now and some researches which still are going on. So that, we can get the perfect idea for the outcome we are expecting for our problem. Here, I want to mention all possible researches related to our problem which has been done for similar problem as we can get inspiration to get correct direction from that formal work.

Authors of [1] say among many sensors WSN it is hard to find that which sensor is faulty which can be identified by using virtual sensors with machine learning techniques. Techniques will be fast enough and fully automated that no human interaction required. Machine learning algorithm will find near faulty sensor. Dataset will be trained by the previous faulty sensors record. In this framework Meta data is important. Secondly, time is very much affected factor. Interval of update and comparison between correlated sensors will predict the faulty sensor. Results at the end are good and successfully done the practical for the solar power plants.

[2] Authors are more focused on PLS (Partial Least Square). PLS Regression can cope up with change in process characteristics and static model automatically. But this is only for narrow changes not for the high change.JIT (Just in Time) model can cope with both changing process characteristic and nonlinearity of the sensor. It’s generally not that useful as it’s trained locally and correlation between variables does not taken into account. Authors of this paper did case study for both PLS, JIT and Co-JIT for 6 months with changing the window to measure all the data. In conclusion it’s clear that co-JIT performed very well when there were changes rapidly. JIT and PLS couldn’t perform well in that environment.

[3] Paper is actual description of paper [1] that how we can gather the data and how important it is to have sensors data virtually. There are different options to get the data using TinyML, TinyLime and SensorML. These all are compatible with only homogenous data but not with the heterogeneous data collection. In order to achieve the goal 4 things are important. **Input data types, aggregator (to calculate desired measurement), resulting data type, aggregation frequency**. Queries can be two types. One is one time and persistent. There will be four functions which will use above mentioned parameters as inputs and will shows the output. Now as shown in the below figure iteration will be done.

In paper [4] authors focused about on-the-fly statistical approach to identify faulty sensors reading and identify and classify data and system faults. In this paper they used Hidden Markov Models to detect and identify system fault types. HMM that represents the coherence among the hidden changes of detected phenomenon and the changes of data and system faults, and through a structural reasoning between the two HMMs.

Paper [5] is more interesting and this briefs about faults types and reasons. According to research sometimes only 49% collected data can be useful interpretation. First step is to processing the data locally once it is collected. Modeling data is base for any fault detection system. It helps data to be in expected behavior. There are many reasons sensors give bad result especially sensors provide temperature and humidity.

Three other factors that leads sensor to provide faulty data.

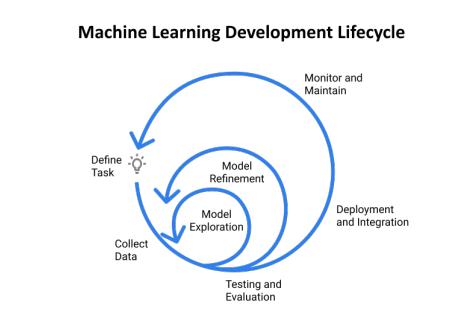
1. Age of the sensor

2. Noise present in the environment

3. Different sensor modalities.

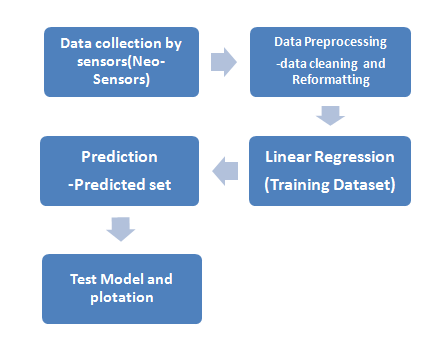
This paper gives broad view about the factors that leads sensors to produce faulty data and gives the importance of the variables and methods of linear models.

Paper [6] briefs about SVM (Support vector machine) and linear regression for wireless Body area network (WBAN) to find abnormal record. This paper provides depth of how exactly both SVM and linear regression is coupled. First SVM finds outliers especially abnormal data by finding optimum hyper plan. It then leads solution to classification. Classification leads to data mining where algorithm includes k-means Fuzzy c-means .Linear regression on the other hand helps to predict the missing data and experimental results will validate the outcome. Here linear regression uses tree approach. This model works in two phases: training and detection. Last but not the least is data mining and its importance is mentioned in paper [7] this paper is providing generic model to find the sensor fault detection. There are five different layers which are responsible for the desired output from machine learning techniques. Five different steps of any machine learning technique should have is Data acquisition, data transform, semantic and rule reasoning, learning layer and action layer. This experiment uses 8000 weather sensor nodes data of US. This study has studied the different features like Air Pressure, Temperature, humidity etc. This technique uses cluster so sensors with faulty data have different reading and it will be placed in different cluster so finding faulty sensor is easy. General Machine development life cycle is:



**3. Work Flow and road map**

After completing the state of the art now we have ground to choose the perfect solution. As machine learning is an evolving field and there are lots of techniques which can be explored for the required problem. Mainly when we talk about machine learning we have two types: supervised and unsupervised learning. We focused on supervised learning we have labeled data. Mostly for supervised learning there is regression and classification. In unsupervised learning is used when data is unlabeled. Two main techniques for unsupervised learning is clustering and dimensionality reduction. Now we know our data is labeled data and as mention our aim is to provide such a model which is able to find faulty sensors from the WSN. This model should able to give quick result and ensures the collected data quality. As seen in paper [5] only 49% of data is accurate in WSN we will here try to maximize the accuracy with finding faulty sensor node in the network. Now, as its known when we have labeled data it’s obvious to go with supervised learning. Now as from the research we have explored more on regression with [5], [6], [7] techniques were linear regression optimization, SVM and linear regression and decision tree respectively. For our problem I want to focus on model which provides support from analysis, unnecessary data removal to prediction and finally to faulty sensor node finding.



Now implementation of the model will lead us to have the virtual network. Final model will use linear regression. First data will be compared to the nearest node. From outliers data we can check node and check for values that node is measuring. If vales by particular node are wrong more than twice then its faulty sensor. For next step we can predict the values and compare those values to actual values by the sensors. Again check that node is not measuring correct values more than twice declare that sensor as faulty.

**4. Implementation Linear Regression**

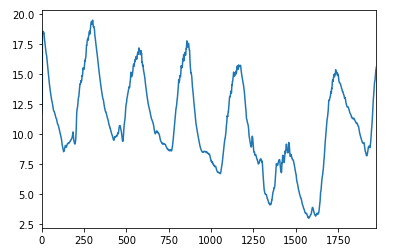
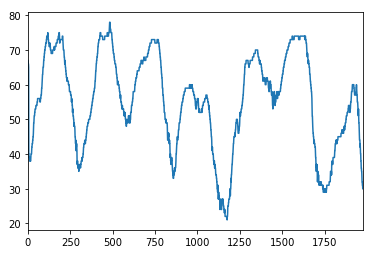
As mentioned and according to research linear regression is useful when data is labeled. Here

Linear regression is useful to predict the data and minimize the error in the existing models. We have here data-set of collected data for more than 6 months for WSN network implemented by University Paul Sabatier-III using neo sensors.

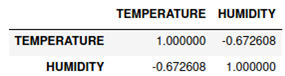
Implementation: First

Reformation of data is already done and data collected by Neo Sensors at network are

Temperature: Humidity:

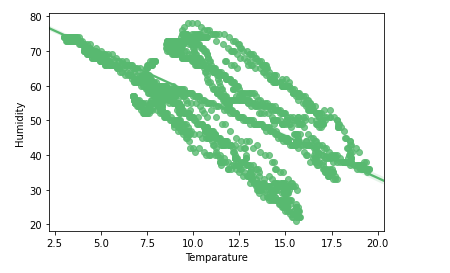


To find out the correlation on different variables:



Correlation varies between +1 to -1, +1 indicates strong positive correlation, Correlation coefficient of -1 signifies perfect negative relationship, and correlation of 0 means that no relationship exists between variables.

Here we have minus values strong negative relation between two data sets.



It is clearer when we plot the values slop is negative so strong negative relation which means both values are independent of each other.

Next step is to remove the outliers from the dataset.

To do so we have to choose the threshold to keep the dataset clean.(if needed)

**REFERENCES**

1 E. B. Ilyas, M. Fischer, T. Iggena and R. Tönjes, "Virtual Sensor Creation to Replace Faulty Sensors Using Automated Machine Learning Techniques," *2020 Global Internet of Things Summit (GIoTS)*, Dublin, Ireland, 2020, pp. 1-6, doi: 10.1109/GIOTS49054.2020.9119681.

2 Koichi Fujiwara, Manabu Kano, Shinji Haseb, Akitoshi Takinami, “Soft‐sensor development using correlation‐based just‐in‐time modeling” ISBN 9780444532275 ISSN 1570-7946

3. S. Kabadayi, A. Pridgen and C. Julien, "Virtual sensors: abstracting data from physical sensors," *2006 International Symposium on a World of Wireless, Mobile and Multimedia Networks(WoWMoM'06)*, Buffalo-Niagara Falls, NY, 2006, pp. 6 pp.-592, doi: 10.1109/WOWMOM.2006.115.

4.E. U. Warriach, M. Aiello and K. Tei, "A Machine Learning Approach for Identifying and Classifying Faults in Wireless Sensor Network," *2012 IEEE 15th International Conference on Computational Science and Engineering*, Nicosia, 2012, pp. 618-625, doi: 10.1109/ICCSE.2012.90.

5. Kevin Ni , Nithya Athreya Ramanathan profile imageNithya Ramanathan , Mohamed Nabil Hajj Chehade profile imageMohamed Nabil Hajj Chehade , Laura Kathryn Balzano profile imageLaura Balzano , Sheela Nair profile imageSheela Nair , Sadaf Zahedi profile imageSadaf Zahedi , Eddie W Kohler profile imageEddie Kohler , Gregory J Pottie profile imageGreg Pottie , Mark H Hansen profile imageMark Hansen , Mani B. Srivastava” Sensor network data fault types”

6. Salem, Osman,et al. "Anomaly Detection in Medical Wireless Sensor Networks using SVM and Linear Regression Models." *IJEHMC* 5.1 (2014): 20-45. Web. 16 Jan. 2021. doi:10.4018/ijehmc.2014010102

7 A. C. Onal, O. Berat Sezer, M. Ozbayoglu and E. Dogdu, "Weather data analysis and sensor fault detection using an extended IoT framework with semantics, big data, and machine learning," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 2037-2046, doi: 10.1109/BigData.2017.8258150.

8. C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann and F. Silva, "Directed diffusion for wireless sensor networking," in *IEEE/ACM Transactions on Networking*, vol. 11, no. 1, pp. 2-16, Feb. 2003, doi: 10.1109/TNET.2002.808417.