Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

DaRoN: A Technique for Detection and Removal of Noise in IoT Data by using Central Tendency

V. A. Jane[1], Dr. L. Arockiam[2],
[1,2] Department of Computer Science, St. Joseph's College Tiruchirappalli.

Abstract

The Internet of Things (IoT) is a significant technology that offers well-organized and trustworthy solutions for the innovation of many domains. Agriculture is one of the most concerned fields in IoT, where IoT based solutions are used to automate the maintenance and monitoring process with least human intervention. Large scale IoT based agricultural environment generates a large amount of data every moment. The agro-production environment is complex and there are numerous discrepancies in the collected raw data that cannot be directly traced by analysis and mining. To handle these inconsistencies in IoT agricultural data, this paper proposes a technique called **Detection and Removal of Noise(DaRoN)**. The proposed technique removes the null values, error values, repeated values, incomplete values, and irrelevant values using measures of central tendency. In addition, a comparative analysis was performed with existingnoise removal techniques and the performance is measured using the Support Vector Machine(SVM) classifier. In this proposed research work, noisy data is eliminated to enhance classification accuracy. The DaRoN technique will be useful for improving the quality of collected data in agricultural environment.

Key Words
Noise, Data cleaning, IoT, Preprocessing, Noise removal, Smart Agriculture.

Section I: Introduction

IoT is a predominant technology which makes many applications smarter using its features [1]. In the past, gathering data in agriculture environment was a difficult task especially in monitoring systems but IoT removes all those strenuous part with the help of sensors. Here, sensors play a vital role in data collection and generates enormous data every day. These data contain missing values, noise, outliers, and duplicate values [2]. If any one of the above is present in the collected data, then it will reduce the quality of outputs. Among which, Noise is one of the most considerable one and it is defined as meaningless information like, corrupted values, repeated values, error values, null values etc., These problems occur due to the reasons such as connection error, detection error, and collision problem in IoT [3]. If the dataset contains noisy values, then many problems will occur during the analytical process.

Noise is classified into two types such as point noise and continuous noise. The Point noise has sudden deviation from other data points.So this could be identified easily. The Continuous noise is difficult to identify because the deviation gets increased from point to points. For removing these types of noise, mean, median and mode methods are used. Noise can also be categorized based on the occurrences in the dataset. If the noise occurs in the class column then it is called *class noise.* If the noise occurs in the attribute column then it is called *attribute noise*. In contrast to *class noise*, *attribute noise* is more harmful because it directly affects the data. Thus, noise in the dataset will affect the accuracy of the analytics model [4]. So, there is a need for data pre-processing.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

Pre-processing techniques [5] are categorized into various types such as data cleaning, data integration, data transformation and data reduction. This paper focuses on noise removal and it comes under the process of data cleaning.The rest of the paper explains more details about the proposed technique and the paper is organized as follows, section II explores the related works on the relevant area, Section III describes the methodology of the proposed work, section IV summarizes the results & discussion and section V concludes the work.

Section II: Related Work

Peter et al.,[6] overviewed the role of Data mining in IoT. This work discussed about all the technologies, methods and algorithms related to the Data mining process with respect to various IoT applications. Also, it described the role of data management in smart environments.

Kun et al., [7] proposed clustering-based particle swarm optimization (CPSO) approach to handle data in the DSM (Data Stream Mining). In which, sliding window technique was used for data segmentation and SFX (Statistical Feature Extraction) was used for variable partitioning. The proposed approach was implemented using five types of IoT data set (Home, Gas, Ocean, and Electricity). The results were evaluated, and the proposed approach improved the accuracy but increased the complexity of algorithms and the over fitting problem.

HumaJamshed et al.,[8] discussed about various pre-processing techniques for mining and analysis tasks. In this work, the important methods of data pre-processing were described which includes data cleaning, data transformation, data reduction and data integration. The author proposed a technique for the same and explained with simple text data case study. The proposed technique dealt with noise removal, tokenization, and normalization. The paper concluded that the advanced techniques like machine learning improved the effectiveness of pre-processing.

Asiya et al., [9] compared the performance of noise cancellation techniques in IoT enabled Telecardiology System. The techniques which were taken for comparison were LMS (Least Mean Square), NLMS (Normalized Least Mean Square), CLLMS (Circular Leaky Least Mean Square) and VSS-CLLMS (Variable Step Size CLLMS). Baseline Wander (BW) elimination (lowest frequency in ECG (ElectroCardioGram)). VSS-CLLMS method achieved high SNRI (Signal to Noise Ratio Improvement). Theauthors focused on ECG data pre-processing with filtering mechanisms.

Liu et al., [10] proposed a technique to handle noise in IoT data by using anomaly detection technique. The proposed technique measured the rate of change and deviation by using a sliding window and statistical techniques. Also, it identified the noise in the dataset based on neighbour behaviour and erroneous data removal process was difficult if error was identified in the continuous neighbour. Here, the identification process consumed more time.

Wang et al.,[11] proposed a framework for wind data pre-processing and prediction. In this proposed work,Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEDMAN)technique used to remove noise in the wind data and MTO (multi-tracker optimizer) was used for error detection. Finally, neural network layers were utilized for model building. The proposed CEEDMDAN technique was suitable only for limited sized datasets and while large datasets were considered it increased the mean error.

Sanyall et al.,[12] proposed a scheme to handle the veracity problems (Noise, Missing values, Outliers and redundancy) in IoT sensor data. The proposed scheme consisted of two parts, first part dealt with data aggregation using cluster method and the second part dealt with data pre-processing using robust dominant subspace estimation and tracking methods. Random

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

outputs generated by dominant subspace selection increased outliers so the overall performance was decreased.

Sáez et al.,[13] presented a method Iterative Class Noise Filter (INFFC) that combined many classifiers for detecting noise in an iterative manner.The filtration method was introduced to identify the noise by eliminating the process of noise detection at each new iteration.

Garcia et al.,[14] improved noise detection using an ensemble of noise filtering methods. The proposed approach Meta Learner (MTL) reduced the redundant data in the dataset, as well as eliminated the irrelevant data. For that, Meta features were created from corrupted datasets and provided a meta-learning model that predictednoisy data.

Section III: DaRoN Proposed Technique

In the agricultural scenario, irrigation system requires constant monitoring without human intervention. To automate this process, the proposed DaRoN technique uses IoT sensors to collect data and stores the collected data in cloud. Later the collected data are pre-processed using the measures of central tendency and the performance of the pre-processed dataset is analysed using Support Vector Machine (SVM) classifier. In traditional noise handling techniques, there are 3 phases namely,Robust (detection of any analysis errors to make the data standardized), Filtering (using various measures to remove noise), and Polishing (Replacing error values)[15]. The novelty of the proposed DaRoN is that it combines the 3 phases into one to produce a Noiseless dataset.

Robust and filtering are done using pre-defined conditions and polishing is done by using measures of central tendency. The work flow of the DaRoN technique is given below in figure 1.

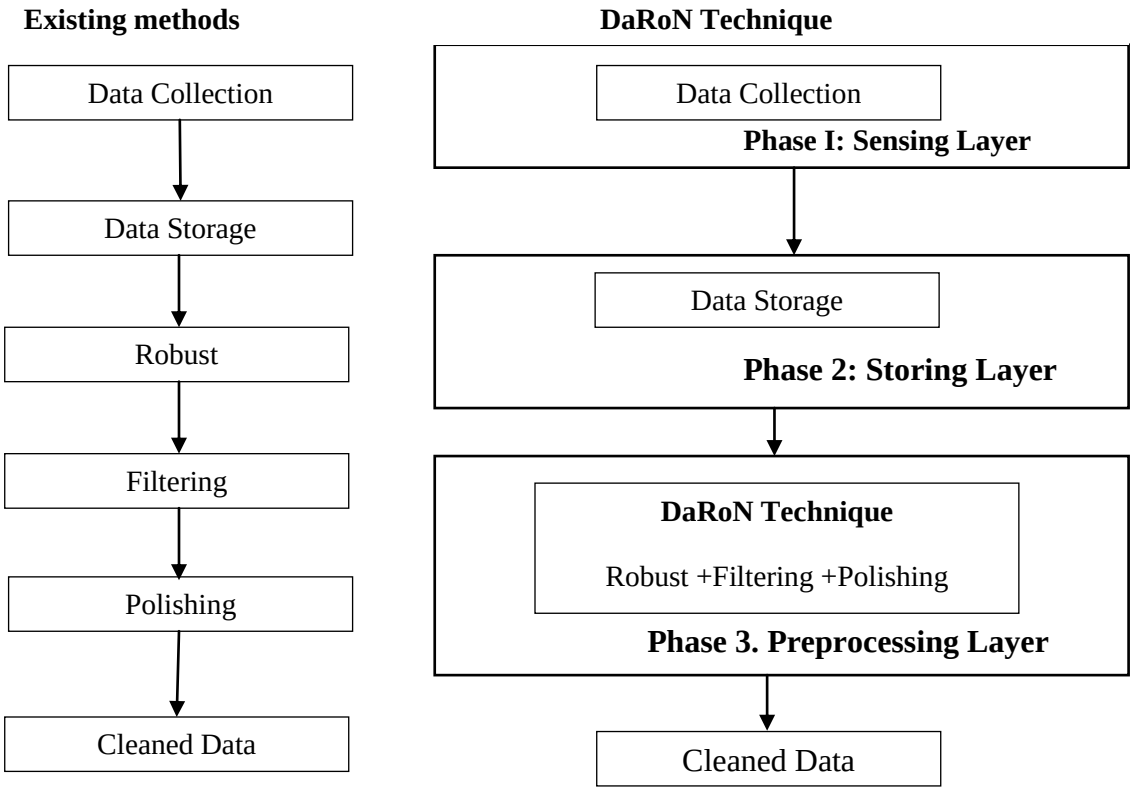

Figure 1: Work flow of the Existing vs. DaRoN technique

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

From the workflow of the proposed work, the pre-processed agricultural dataset yields noiseless cleaned dataset which increases the performance of the classifier.

Phase I: Sensing Layer

Phase one deals with the data collection that is done by using various IoT sensors in agricultural environment. There are five sensors used namely, humidity sensor, temperature sensor, soil moisture sensor, wind speed sensor and rain sensor. Sensors are placed in different places and connected to the cloud. Each sensor plays a unique role in monitoring the environment and continuously collects data. Humidity sensor collects the data about moisture level in the air. This data will be useful in determining whether irrigation is needed or not. Soil moisture sensor measures the percentage of water present in the soil. In this work, both the humidity and soil moisture sensor data are taken together to make irrigation decisions. The temperature sensor is generally used to measure temperature level from time to time. The rain sensor is used to collect rain level. The primary work of this sensor is to shut down the entire irrigation system during heavy rain fall. Data from these sensors are collected automatically and sent to the server directly for further processing.

Phase 2: Storing Layer

Second layer is storing layer, which is used for data storage purpose. Basically, data can be stored on local devices, but to handle large data, cloud storage is the best. So, the proposed technique uses cloud to store the data. Many open-source clouds are available, one of them is ThinkSpeak cloud server which provides open-source computing model, where data can be stored and retrieved remotely with the help of internet. The stored data is maintained, operated, and managed by a service provider. In ThinkSpeak, an account is created and built with various fields such as soil field, humidity field, temperature field, and rain field to store their respective information. After that, the stored data is forwarded to the preprocessing layer.

Phase 3. Preprocessing Layer

In this layer, the proposednoise removal technique is used.This novel technique uses the measures of central tendency. Traditional Noise removal techniques use three phasessuch as robust, filtering and polishing.But, the proposed DaRoN technique combines these three phases in a single phase by using the measures of Central tendency which gives better performance. The proposed technique selects thenearestmean value to replace repetitive and null values. Nearest Mode value is selected to remove Point Noise.All replacementsare done with respect to Time Details (Td).The Central tendency measures are listed below.

Mean (μ) $= \frac{\text{sum of all elements}}{\text{Total number of Elements}}$

Median (M) $= L + h\frac{((fm - f1))}{((fm - f1) - (fm - f2))}$

Mode (Z) $= \frac{(n+1)}{2}$

Let $L = \{L_1, L_2 \ldots L_n\}$, where, L_1, L_2 …$L_n$are different locations.

Each location has various sensors that are T_n, S_n, H_n, R_n and W_n where n denotes number oflocations, T_n − Temperature sensor, S_n – Soil moisture sensor, H_n – Humidity sensor, R_n – Rain sensor, W_n – Wind Sensor, and the values of each sensor from 1… n.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

If location number is one then the set of L_1 is, L_1= {t_1, s_1, h_1, r_1, w_1}. Similarly, L_2, L_3, L_4 and L_5 sets are defined. In the proposed work, 5 different locations are considered, so the total number of sensors in each category can be written as,

T= {t_1, t_2, t_3, t_4, t_5},
S ={s_1, s_2, s_3, s_4, s_5},
H = {h_1, h_2, h_3, h_4, h_5},
R = {r_1, r_2, r_3, r_4 r_5}
W = {w_1, w_2, w_3, w_4, w_5}

Therefore, L can be written as L = {T, S, H, R, W}

DaRoN Technique for noise detection and removal

for (int i = 0; i ≤ 25; i+=2) **// One observation per two hour**

collect r_1(Td[i])

for (int i=0; i<n; i++)

if(r_1 (Td[i]) < r_1(Td[i+1])) **//Checking Redundant values based on TimeTd**

remove r_1(Td[i])

compute rest of R, and all elements in T,W, H, S

if(compare r_1 with R (μ), R(M), & R(Z) **// Checking point noise and error value**

replace with R(μ), R(M), & R(Z)// **Common for rest of R and T, W, S, H**

if (r_1> 0) **// M,Z,μ are Calculated with respective to Td[i] value**

compute rest of R, and all elements in T,W, H, S

else

compute rest of R, and all elements in T,W, H, S

end if

end for

Section IV: Result and discussion

This section describes the performance of the proposed DaRoN Technique using the conventional measures such as precision, F1 score, recall and accuracy.Table 3 shows the details of the collected data. Finally, the cleaned dataset is applied to the SVM classifier for analysing the performance of the proposed DaRoN technique.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

Table 3 Collected Data Details

Type				**Amount**			
Total Data (Rows)				19,520(6 Months Data)			
Noise				9,760			
Point Noise	Repetitive	Collision	Null	4782	4326	118	652
Features(Columns)				32			

The existing pre-processing methods such as Iterative Class Noise Filter (INFFC), Meta Learner (MTL) and (CEEDMAN) are applied on the collected dataset and fed to the classifier after cleaning. Then the proposed DaRoN Technique is compared with existing techniques based on the performance metrics. The proposed DaRoN technique enhances the accuracy than others and the comparison results are shown in Figure 2.

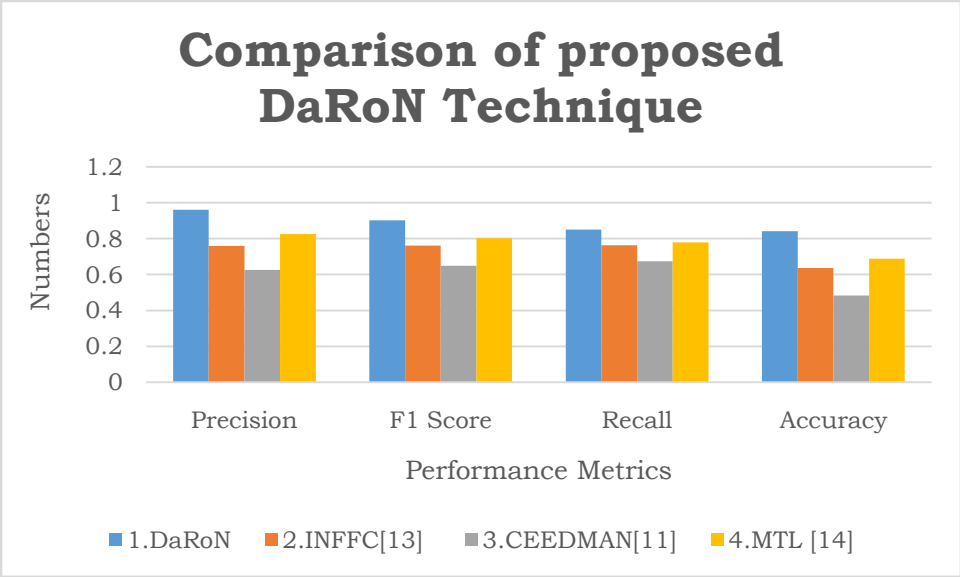

Figure 2:Comparison Result

Section V: Conclusion

In IoT agricultural data needs pre-processing for efficient decision making. The raw data collected from IoT environment has inconsistency issues which affect the efficiency and accuracy of decision making. So, refinement of data is needed. The proposed DaRoN handles the noisy data efficiently. It consists of three layers. First layer collects data from sensors placed in various locations, the collected data is stored in the second layer, and the third layer performs data cleaning process. The proposed technique detects noisy data and replaces it based on pre-defined conditions and measures of central tendency. Finally, the results were compared with existing methods and the proposed technique outperformed others by improvingthe classification accuracy. In future, issues like missing values and outliers may also be considered to further improve accuracy.

Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 2, 2021, Pages. 3197 – 3203
Received 20 January 2021; Accepted 08 February 2021.

References:

[1] Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W," A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", *Sensors*, 19(20), 4536, 2019, doi: 10.3390/s19204536.

[2] Assahli, S., Berrada, M., &Chenouni, D,"Data pre-processing from Internet of Things: Comparative study", *Wireless Technologies, Embedded and Intelligent Systems (WITS)*, 2017.

[03] Morais, C. M. de, Sadok, D., &Kelner, J, "An IoT sensor and scenario survey for data researchers*", Journal of the Brazilian Computer Society*, doi:10.1186/s13173-019-0085-7,2019.

[04]Zhong, Y., Fong, S., Hu, S., Wong, R., & Lin, W. "A Novel Sensor Data Pre-Processing Methodology for the Internet of Things Using Anomaly Detection and Transfer-By-Subspace-Similarity Transformation", *Sensors*, 19(20), 4536. doi:10.3390/s19204536, 2019.

[05] García-Gil, D., Luengo, J., García, S., & Herrera, F.,"Enabling Smart Data: Noise filtering in Big Data classification", *Information Sciences*. doi:10.1016/j.ins.2018.12.002, 2018.

[06] Peter Wlodarczak, Mustafa Ally, Jeffrey Soar, "Data Mining in IoT", *In Proceedings of 2nd Int. Workshop on Knowledge Management of Web Social Media, Leipzig, Germany, August 2017 (KMWSM '17)* , ISBN 978-1-4503-4951, https://doi.org/10.1145/3106426.3115866, 2017.

[07] Lan, K., Fong, S., Song, W., Vasilakos, A., &Millham, R, " Self-Adaptive Pre-Processing Methodology for Big Data Stream Mining in Internet of Things Environmental Sensor Monitoring", *Symmetry*, 9(10), 244, doi:10.3390/sym9100244, 2017.

[08] Jamshed, Huma& Khan, M. &Khurram, Muhammad &Inayatullah, Syed &Athar, Sameen, "Data Preprocessing: A preliminary step for web data mining". 206-221, 2015, Doi: 10.17993/3ctecno.2019.specialissue2.206-221, 2019.

[09] AsiyaSulthana ,Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOTEnabled Telecardiology System", *International Journal of Engineering & Technology*, 7 (2.17) (2018) 74-78, 2018.

[10] Liu, Y., Dillon, T., Yu, W., Rahayu, W., &Mostafa, F, "Noise removal in the presence of significant anomalies for Industrial IoT sensor data in manufacturing", *IEEE Internet of Things Journal*, 1–1. doi:10.1109/jiot.2020.2981476, 2020.

[11] Wang, Jianzhou; Wang, Ying; Li, Zhiwu; Li, Hongmin; Yang, Hufang, "A combined framework based on data preprocessing, neural networks and multi-tracker optimizer for wind speed prediction", *Sustainable Energy Technologies and Assessments*, 40, 100757–. doi:10.1016/j.seta.2020.100757, 2020.

[12] Sanyal, Sunny; Zhang, Puning," Improving Quality of Data: IoT Data Aggregation Using Device to Device Communications", *IEEE Access*, Vol.6, 67830–67840, , doi:10.1109/ACCESS.2018.2878640, 2018.

[13] Sáez, J. A., Galar, M., Luengo, J. & Herrera, F., "INFFC: an iterative class noise filter based on the fusion ofclassifiers with noise sensitivity control", *Information Fusion*, 27, 19–32, 2016.

[14] Garcia, L. P., de Carvalho, A. C. & Lorena, A. C. 2016a. "Noise detection in the meta-learning level. *Neurocomputing*, 176, 14–25, 2016.

[15] Choh Man Teng, **"**A Comparison of Noise Handling Techniques",*FLAIRS-01 Proceedings*, 2002.