Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

CID: Central Tendency Based Noise Removal Technique For Iot Data

V. A. Jane[1*] , Dr. L. Arockiam[2],

[1*, 2] Department of Computer Science, St. Joseph's College (Affiliated to Bharathidasan University), Tiruchirappalli.

Abstract
The Internet of Things (IoT) is a crucial technique that enables well-organized and reliable solutions for various areas' development. Agriculture is one of the most worried IoT areas, with IoT-based solutions being used to automate the management and evaluation system with the least amount of human participation. Every second, a large-scale IoT-based agricultural ecosystem creates a tremendous quantity of data. The agro-production ecosystem is complicated, and there are several inconsistencies in the raw data that assessment and mining cannot be directly tracked. This research presents a strategy called Detection and Removal of Noise (CID) to deal with these anomalies in IoT agriculture data. Utilizing measurements of central tendency, the suggested approach eliminates null values, incorrect values, repeating values, unfinished values, and inappropriate values. In addition, a comparison of current noise reduction strategies was carried out, and the effectiveness was assessed using the Support Vector Machine (SVM) classification. To improve accuracy of classification, noisy data is removed in this suggested investigation. The CID approach will help improve the quality of data obtained in agricultural settings.

Key Words Noise, Data cleaning, IoT, Pre processing, Noise removal, Smart Agriculture.

Section I: Introduction

IoT is a popular technology that uses its capabilities to make numerous applications smarter [1]. Collecting information in the agricultural area was previously a challenging operation, particularly in surveillance devices, but the Internet of Things (IoT) eliminates all of those demanding parts with the use of sensors. Sensors play a critical role in data collecting and create massive amounts of data on a daily basis. There are missing values, distortion, anomalies, and duplicated values in this information [2]. If any of the aforementioned are contained in the obtained data, the output quality may suffer. Noise is one of the most significant, and it is described as useless data such as corrupted values, repeating values, incorrect values, null values, and so on. These issues arise as a

result of IoT-related issues like connection errors, detection errors, and collisions [3]. Several issues may arise throughout the analysis procedure if the dataset includes noisy data.

Point noise and continuous noise are two different forms of noise. The Point noise deviates sharply from the rest of the data. As a result, this might be clearly spotted. Since the divergence increases from point to point, continuous noise is hard to detect. The mean, median, and mode approaches are employed to remove this sort of noise. The incidences in the data may also be used to characterize noises. Class noise is defined as noise that happens in the class column. When noise appears in the attribute column, it is referred to as attribute noise. Unlike class noise, attribute noise is more destructive since it influences the data directly. As a result, noise in the database may influence the analytics model's accuracy [4]. As a result, data pre-processing is required.

Data cleansing, data integration, data conversion, and data reduction are some of the sorts of pre-processing procedures [5]. This article is about noise reduction, which is part of the data cleaning procedure. Section II examines similar works in the relevant domain, Section III defines the procedure of the proposed study, Section IV summarizes the findings and discussion, and Section V ends the work.

Section II: Related Works

The function of data mining in IoT was discussed by Peter et al.,[6]. This paper covered all of the techniques, methodologies, and processes connected to data mining in relation to different IoT systems. It also explained the significance of data management in smart settings.

To manage data in the Data Stream Mining (DSM), Kun et al. [7] suggested a clustering-based particle swarm optimization (CPSO) technique. Data segmentation was done using the sliding window approach, and variable partition was done using Statistical Feature Extraction (SFX). The suggested method was tested with five different kinds of IoT data sets (Home, Gas, Ocean, and Electricity). The results were analyzed, and the suggested method enhanced efficiency while increasing computational overhead and the overfitting issue.

Various pre-processing strategies for mining and analytical activities were explored by Huma Jamshed et al., [8]. The essential techniques of data pre-processing, such as cleaning the data, data conversion, data reduction, and data aggregation, were detailed in this study. The researcher offered a method for doing so, which he demonstrated using a simple text data case study. Noise reduction, tokenization, and normalization were all addressed by the suggested method. According to the findings, modern approaches such as machine learning boosted the efficacy of pre-processing.

In an IoT-enabled Telecardiology platform, Asiya et al. [9] analyzed the accuracy of noise cancelling approaches. LMS (Least Mean Square), NLMS (Normalized Least Mean Square), CLLMS (Circular Leaky Least Mean Square), and VSS-CLLMS (Variable Step Size CLLMS) were the approaches used for comparison. Removal of baseline wander (BW) (minimum frequencies in ECG (Electro Cardio Gram)). The VSS-CLLMS approach produced a high SNRI (Signal to Noise Ratio Improvement). The researchers concentrated on filtering strategies for ECG data pre-processing.

Using an outlier detection methodology, Liu et al. [10] suggested a method for dealing with distortion in IoT data. Using a sliding window and analytical measures, the suggested methodology calculated the variation and divergence. It also recognized distortion in the dataset based on neighbourhood activity, making the task of removing incorrect data more challenging if an issue was found in the continuous neighbourhood. The identifying procedure took longer in this case.

Wang et al., [11] established a framework for pre-processing and forecasting wind data. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) approach was used to eliminate noise from wind data, and MTO (multi-tracker optimizer) was employed to find errors in this suggested study. Eventually, neural network layers were used to construct the model. The suggested CEEMDAN approach was only acceptable for small datasets, and when bigger datasets were investigated, the mean error worsened.

Sanyall et al., [12] suggested a technique to deal with IoT sensor data authenticity issues (noise, null values, anomalies, and repetition). The suggested technique was divided into two sections: the first dealt with data aggregation using the clustering technique, and the other with data pre-processing employing resilient dominant subspace calculation and monitoring techniques. Outliers rose as a result of the randomized outputs created by dominating subspace selections, lowering overall effectiveness.

Sáez et al., [13] proposed the Iterative Class Noise Filter (INFFC), which iteratively merged several classifiers for distortion detection. The filtering approach was developed to define the distortion by removing the noise detection step from every iteration.

Garcia et al., [14] used an aggregation of noise filtering approaches to enhance noise identification. Meta Learner (MTL) was developed as a method for reducing duplicate data and eliminating unnecessary data in a database. Meta features were produced from damaged dataset for this purpose, and a meta-learning framework that forecasted noisy information was developed as a result.

Section III: CID Proposed Methodology

Irrigation systems in agriculture needs regular monitoring without human interference. The suggested CID approach collects information from IoT devices and saves it in the cloud to automate this procedure. The obtained data is then pre-processed utilizing central tendency metrics, and the efficiency of the pre-processed data is evaluated using a Support Vector Machine (SVM) classification. Robust (identification of every analytical flaws to make the data standardized), Filtering (utilizing different approaches to reduce noise), and Polishing (changing incorrect values) are the three steps of classic noise treatment techniques [15]. The suggested CID is unique in that it blends the three stages into one to provide a noise - free data.

Pre-defined criteria are used for robust and filtration, and measurements of central tendency are used for refining.

The pre-processed agricultural data generates a noise - free cleansed data as a result of the suggested work's methodology that improves the classifier's effectiveness.

Phase I: Sensing Layer

The first stage focuses on data collecting in an agricultural setting utilizing different IoT devices. The humidity sensor, temperature sensor, soil moisture sensor, wind speed sensor, and rain sensor are among the 5 sensors employed. Sensors are installed in various locations and are linked to the cloud. Every sensor monitors the surroundings in its own way and captures information in real time. A humidity sensor captures information about the amount of water in the air. This information would be valuable in assessing whether or not watering is required. The proportion of water in the soil is measured using a soil moisture sensor. Both humidity and soil moisture sensor data are used in this study to determine irrigation recommendations. Temperature sensors are often used to detect temperature levels on a regular basis. The rain sensor is used to measure the amount of rain falling. This sensor's principal function is to turn off the whole irrigation system during intense rainfall. The data collected from the sensors is periodically gathered and transferred to the cloud for more analysis.

Phase 2: Storing Layer

The storing layer is the second layer, and it is used to store information. Information may be kept locally, but cloud storage is the most efficient way to manage enormous amounts of information. As a result, the information is stored in the cloud using the suggested method. Several open-source clouds are accessible; one of them is the Think Speak cloud server that offers an open-source computation paradigm for storing and retrieving data over the network. A service provider is responsible for maintaining, operating, and managing information. In Think Speak, an account is formed and developed with numerous fields for storing information like soils, moisture, heat, and rainfall. The information is then sent to the pre-processing phase.

Phase 3. Pre processing Layer

The proposed noise reduction approach is applied in this layer. The metrics of central tendency are used in this unique method. Conventional noise reduction methods consist of three stages: robust, filtering, and polishing. The suggested CID approach, on the other hand, integrates all three stages into a single stage by employing measurements of central tendency, resulting in improved effectiveness. To substitute repetitious and null entries, the suggested approach uses the nearest mean values. To eliminate Point Noise, the Nearest Mode value is used. All changes are completed in accordance with the time details (Td). The measurements of central tendency are mentioned 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 …L_n}, where, L_1, L_2 …L_n are different locations.

Every location contains numerous sensors that are T_n, S_n, H_n, R_n and W_n in which n indicates count of locations, 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 every sensor from 1… n.

If location number is one then the set of L_1 is, L_1= {t_1, s_1, h_1, r_1, w_1}. Likewise, L_2, L_3, L_4 and L_5 sets are defined. In the suggested method, 5 diverse locations are assumed, hence the overall count of sensors in every category shall be represented 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}

Hence, L is represented as L = {T, S, H, R, W}

CID method for noise identification and elimination

for (int i = 0; i ≤ 25; i+=2) **// One observation every 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 examines the suggested CID Method's effectiveness utilizing standard metrics like precision, F1 score, recall, and accuracy. Table 3 summarizes the information gathered. Lastly, the cleansed data is fed into a SVM classifier to evaluate the suggested CID method's efficiency.

On the acquired datasets, known pre-processing techniques like Iterative Class Noise Filter (INFFC), Meta Learner (MTL), and CEEDMAN are used and provided to the classifiers following cleansing. On the basis of performance measures, the suggested CID Method is then contrasted to current approaches. The suggested CID approach outperforms the competition in terms of accuracy, as demonstrated in Figure 1.

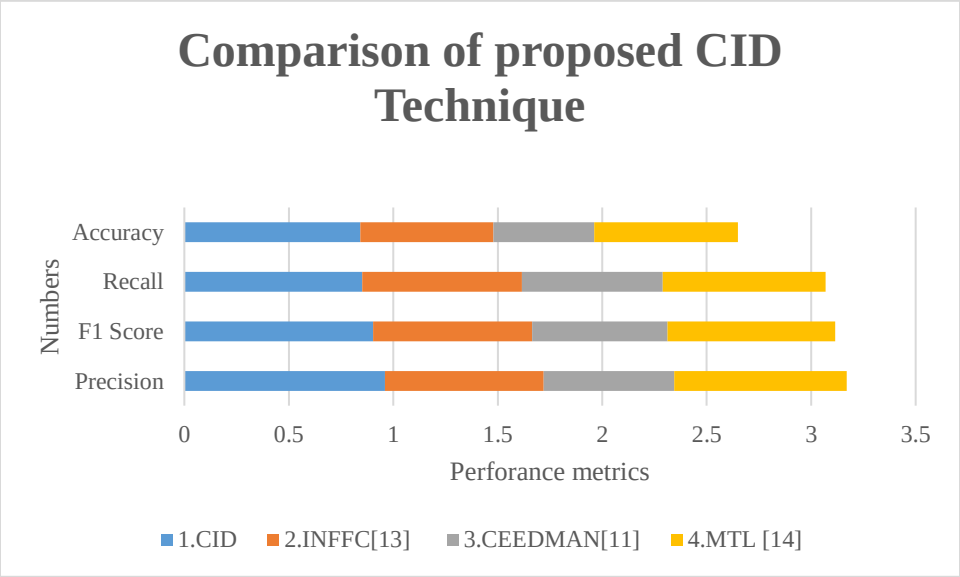

Figure 1: Comparison Result

Section V: Conclusion

For effective decision-making in the IoT, agricultural data must be pre-processed. The inconsistencies in the raw dataset obtained from the IoT ecosystem have an impact on decision-making efficiency and reliability. As a result, data refining is required. The suggested CID effectively manages noisy data. It is made up of three layers. The first layer gathers information from sensors located in the different areas, the second layer stores the gathered information, and the third layer cleans the information. The suggested method identifies noisy data and substitutes it with new data based on pre-defined parameters and central tendency metrics. Lastly, the findings were compared to current approaches, and the new strategy surpassed the competition by increasing accuracy of classification. Missing data and outliers might well be taken into account in the future to enhance accuracy.

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] Asiya Sulthana ,Md Zia Ur Rahman, "Efficient adaptive noise cancellation techniques in an IOT Enabled 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 of classifiers 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.