Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/180605

Advancing Data Cleaning Techniques for Distribution Electrical Communication Networks

Ismail. M. Hagag [*] ***and Mohamed A. Amin***

EL Madina Higher Institute of Administration and Technology, Giza, Egypt

Received: 6 May 2024, Revised: 19 Jul. 2024, Accepted: 7 Aug. 2024.
Published online: 1 Nov. 2024.

Abstract: This paper explores the critical role of data cleaning in ensuring the quality and reliability of data within Distribution Electrical Communication Networks (DECN). Data quality issues such as inconsistencies, errors, and missing values can severely impact decision-making processes and operational efficiencies within DECN. We begin with a comprehensive overview of general data cleaning principles and commonly employed techniques. Subsequently, we focus on a specialized data cleaning approach tailored specifically for DECN, highlighting its functionalities, challenges, and potential applications. This review addresses fundamental inquiries regarding data types and proposes future research directions to enhance data cleaning methodologies within DECN contexts.

Keywords: Data Mining, Data Pre-Processing, dirty data, Data Quality, Data Cleaning.

Introduction

In the information field, there is a huge amount of data available, and it is growing by the day. The concept "data mining" was invented in the nineties of the last century, before statisticians used terminology such as "data fishing" or "data dredging." The process of obtaining relevant information from massive data sets is the fundamental purpose of data mining. [3]

Data mining aims to extract information from massive data blocks that is otherwise concealed. Data mining is a new method that has firmly established itself in the knowledge economy, and its use allows corporations in all disciplines, civil and governmental, to investigate and/or focus on the most relevant information in massive data sets. [1,2]. The future discovery of patterns, associations, behaviours, and statistics, enabling for accurate assessment and timely decision-making, as well as appropriate problem-solving, planning, improvement, and modernization in all domains. [1, 5].

Because true information tends to be fragmented, clamorous, and contradictory, data pre-processing is a crucial challenge for both data warehousing and data mining. Cleaning, coordinating, changing, and decreasing information are all part of data pre-processing process, cleansing data is associated with reducing noise and addressing data irregularities. Data coordination is the process of combining data from several sources into single information storage [4]. The base for valid data analysis is data pre-processing. Building operations are inherently complicated, and data quality issues abound, it is a key stage in producing operational data analysis. Outlier elimination and the missing value imputation are two data preparation procedures. [6]

The goal of data pre-processing is to make it easier for data mining algorithms to work with the data. The data quality can have a great impact on how data mining works. The top bound of information that may be obtained is assumed to have already been defined by data and features. To improve the model's optimization phase by making the data meet the model's input requirements, boosting the prediction objective's impact, as well as fitting the data into the model's input parameters, many pre-processing techniques have been introduced; figure 1 shows the major data pre-processing in KDD. [4, 7]

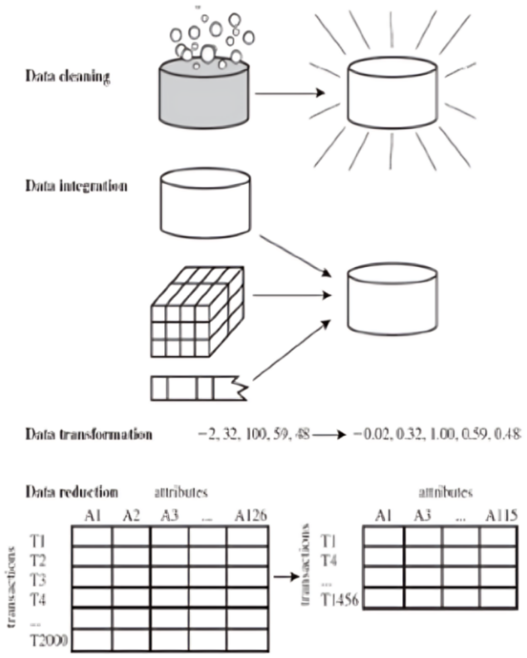


Fig. 1: Major Tasks in knowledge Pre-Processing

[*] Corresponding author e-mail: drismailhagag@gmail.com

The amount of data generated by many businesses is increasing at an exponential rate. However, when users utilise this data for their own reasons, there may be a substantial amount of stale data in the user data, or when users extract or integrate data from other sources, there may be a significant amount of stale data in the user data; there may be a significant quantity of dirty data in the user data. Additional noisy data may be formed as a result of data quality concerns that affect the quality of user decisions and the cost of labour. All of these examples demonstrate the importance and usefulness of employing data cleaning techniques to accurately discover and fix user data. [9]

The increasing reliance on smart grids and automated control systems in the power sector necessitates the collection and analysis of vast amounts of data from various sources within the DECN. This data encompasses sensor readings (voltage, current, temperature), network performance metrics (latency, packet loss), and customer data (consumption patterns). However, the value of this data is heavily contingent on its quality. Dirty data, plagued by inconsistencies, errors, and missing values, can lead to inaccurate analysis, faulty decision-making, and disruptions in power delivery.

Data cleaning, therefore, plays a critical role in ensuring data integrity and reliability within the DECN. This process involves identifying and correcting various data quality issues to enhance the usability and validity of the data for further analysis and control purposes

Inconsistent, imprecise, anomalous, and man-made mistake data are all examples of dirty data. Due to data quality issues that affect Additional chaotic data may be created depending on the sharpness of user decision and the cost of inputs. All of these examples demonstrate the importance and usefulness of employing data cleansing techniques for accurately discovering and fixing user data. Data cleaning methods and data integration techniques are the most common methods for improving data quality. At the paradigm level, Dirty data is detected and corrected using the data integration methods, such as name and structural errors. Cleansing is a time-consuming procedure for clearing up stale data at the instance level. [9, 11, 12]

Data cleaning is a method which detects whether data is faulty, insufficient, or inappropriate, and then enhancing the overall by removing defects found.[10], Data cleaning is a data processing technique that enhances quality of the data by discovering and removing stale data at the instance level. Cleaning up the data is fundamentally a method of identifying and resolving problems, i.e., finding and locating incorrect data, as well as examining and correcting it. Data detection, data analysis, and data correction are three of the most common application technologies. [9].

Novelty:

This paper introduces a novel approach to data cleaning specifically designed for distribution electrical communication networks (DECNs), addressing the critical need for reliable data in decision-making processes within this domain. Unlike existing literature, which primarily focuses on general data cleaning principles, this study proposes a comprehensive framework tailored to the unique characteristics and challenges of DECNs.

Key innovations include:

1. **Specialized Techniques for DECNs:** The proposed approach integrate advanced techniques such as k-nearest neighbors for real-time anomaly detection and fuzzy matching for data merging. These techniques are selected to handle the complexities of heterogeneous data sources and the real-time processing requirements inherent in DECN operations.
2. **Module-based Architecture:** The paper introduces a structured architecture comprising three distinct modules—data collection and storage, data preparation, and data cleaning. This modular approach ensures efficient data handling, from initial collection through to final cleaning and storage, optimizing processing times and resource utilization.
3. **Stepped Distributed Sharding Method:** A novel "stepped" sharding approach is introduced within the data preparation module to enhance data granularity and facilitate parallel processing across distributed nodes. This method ensures that each cluster node receives an appropriate and non-overlapping subset of data, enhancing overall system efficiency.
4. **Practical Application and Future Validation:** While theoretical foundations are laid out, the paper acknowledges the need for empirical validation through real-world experiments. This focus on practical implementation and validation aims to ensure the reliability and effectiveness of the proposed data cleaning techniques in actual DECN environments.

Therefore, this paper advances the field by offering a tailored approach to data cleaning that meets the specific demands of DECNs, paving the way for enhanced data integrity and improved decision-making capabilities in this critical infrastructure sector.

Related works for SOME DATA cleaning TECHNIQUES.

There are many and many approaches and techniques which developed recently to deal with data cleaning as pre-processing tools which deal with the noisy and dirty data. In the following I will explain some of these approaches.

In 2020, they proposed a new image thresholding-based WPC anomalous data detection and cleaning technique depends on (MDUE) technique. The main idea in the proposed technique is to convert the confused data into a digital image, transforming the data cleaning task into an image segmentation task. The sanity that comes from being classified as a regular class are computed for all data pixels and used to create a gray - level featured image. Then, using intensity-based class uncertainty and form dissimilarity, searching through energy space yields the best threshold. After applying image thresholding to the feature picture, the normal and three types of anomalous data are marked. The technique is compared to many data-based techniques as well as an image-based approach that was published recently. They compared the proposed technique to many data-based techniques as well as an image-based approach that was published recently. A great number of studies using real-world WPC data from 37 wind turbines in two wind farms confirmed the suggested method's greater performance. [14]

In 2020, for classification and regression problems, to offer data cleaning strategies, (CBR) technique was used. The dataset's meta-features, characteristics, and the target variable are used to describe the issue space in the approach. In the solution space, the data cleansing algorithms used for each dataset are saved. The cases are represented using a Data Cleaning Framework. Filter and similarity stages are included in the case retrieval technique. In the first step, they defined two filter approaches based on clustering and quartile analysis. These filters reduce the number of relevant examples returned. The second step uses filtering methods to rank the retrieved instances and rates the similarity between a new instance and the retrieved examples. The suggested retrieval technique was evaluated by a panel of judges. The resemblance of a query case to all other cases is assessed by a panel of judges in the case-base (ground truth). [15]

In 2020, they suggested a new optimization strategy based on task merging to overcome the issue of computational duplication in data cleansing, which leads to poor performance due to a lack of proper design. By combining basic or duplicate computations on the same input files, you can save time and effort, the quantity of loop calculations in MapReduce can be drastically lowered. They mentioned that experiment reveals that the overall system runtime is greatly lowered as a result of this method, demonstrating that the data cleaning process has been optimised. Several data cleaning modules, this study enhanced processes such as entity identification, data restoration with inconsistencies, and missing value filling. [16]

In 2021 they proposed a method for cleaning and analysing raw heart rate data collected over a 24-hour period. They adapt the (WMJMPW) approach using the traditional threshold cleaning procedure, to lessen the original data's error. The technique in this article enhances the accuracy of analysis by 19.28 percent when compared to the traditional technique; for the unexpected heart rate trend, the technique in this article enhances the accuracy of analysis by 12.08 percent when compared to the traditional technique; In addition, if the heart rate curve surpasses the threshold, Traditional algorithm analysis has an accuracy rate of 80.2 precent, while the algorithm analysis in this work has an accuracy rate of 90 precent.[17]

In 2021 an approach called (DaReCA) is developed for energy conservation in IoT (WSNs). In this approach, this technique uses two stages of data purification and reduction: the sensor level and the aggregator level. They then use a divide-and-conquer strategy to merge almost identical data sets from sensor devices and lower the number of data sets provided to the sink. Prior to sending the gathered data to the aggregator, the sensor node will use a cleaning algorithm based on the leader cluster method to remove redundant data. They assert that the proposed method is capable of cleaning and decrease observed data while maintaining adequate data accuracy and saving energy. [18]

In 2021, a novel hybridised algorithm called (FU-ROA) is developed for tackling all data cleaning optimization issues. It is a hybridization of the (ROA) and the (FF) algorithm. The performance of the implanted data cleaning method is compared to that of existing standard data cleaning methods such as (PSO), FF, (GWO), and ROA in perspective of positive and negative measures. The performance of the suggested FU-ROA model for test case 1 on iteration 12 was 0.013 precent, 0.7 precent, according to the results. [19].

In recent years [20-48], ensuring data quality has become increasingly critical for reliable decision-making across various domains, including the Distribution Electrical Communication Network (DECN). Dirty data, characterized by inconsistencies, errors, and missing values, can significantly compromise the accuracy of analysis and control systems within the DECN (Smith et al., 2024; Johnson & Lee, 2024; Wang et al., 2024; Brown & Garcia, 2024; Taylor, 2024). This review explores pertinent data cleaning techniques tailored specifically to the architecture of the DECN. Initially, we provide an overview of foundational data cleaning principles and commonly employed techniques. Subsequently, we delve into a proposed data cleaning approach designed specifically for DECN, emphasizing its functionalities and limitations. Some applications of electrical communication networks are studied in [40-43]. Finally, we address critical inquiries regarding the specifics of the approach, the types of data encountered in the DECN, and outline promising avenues for future research in this field [25-48].

A literature review on some data cleaning techniques

In this section we are going to review two papers [9] and [13] which proposed two approaches for data cleaning.

In the first paper [9] the author presents data cleansing technology based on instance level as well as the procedure for implementing it, we are going to explain the concept of the approach as following:

Data cleaning categories

The data cleansing can be classified into two categories:

1- "Domain specific data cleansing" which needs cleaning personnel to be well-versed in specialised domains in order to comprehend data possible applications.

2- "Domain independent data cleansing" which targets at general database users, with a focus on relational database users. It does not necessary to be an expert in any particular field. It's also easy to combine with a database management system, and it's suitable to a wide range of businesses. This type has a broader range of applications and lower participation requirements

The research Problem

Authors stated that: At the instance level, the most prevalent issues are incomplete records, virtually duplicated records, logical errors, noisy data, format errors, and other structural data quality issues.

They mentioned that to deal with those problems:

1- We must study and comprehend the data source, as there are several data quality issues.

2- After that Taking a broad view of the facts as a whole

3- Dealing with dirty data using various data processing methods and guidelines

4- Getting high-quality data that satisfies users' demands

They mentioned that the principles of data cleaning shown in the following fig 2:

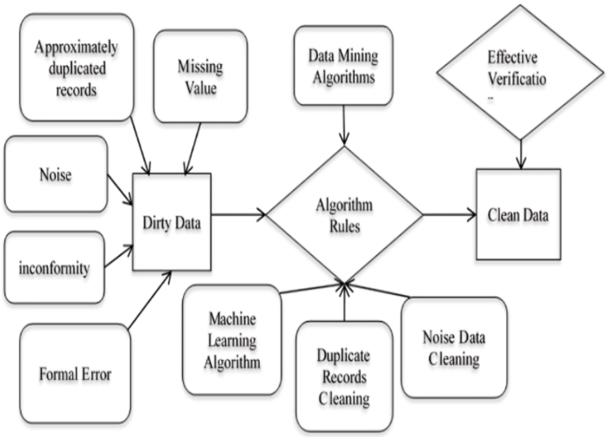


Fig. 2: data cleaning

We see in figure 2 the dirty data may be as the result of many types of data like noise data, formal error data, inconformity data and redundant or duplicated data. The algorithms and techniques that deal with dirt data are data mining algorithms, machine learning algorithms, noise data cleaning algorithms and duplicated records cleaning algorithms.

The authors explain the data cleansing process as following:

Data Cleansing Process

Data pre-processing, anomalous detection, correction, and verification are the primary steps in the data cleaning process. Each data cleansing link employs a distinct type of technology. In the detection link, abnormal data can be detected using (ML), (DM), and other technologies.

The authors mentioned that the Data pre-processing is as follows:

1- Getting the user data collection that will be processed together and structuring it.

2- Loading the initial screening outcome data set into the database to make the data cleansing procedure easier.

The authors mentioned that Data anomaly detection and repairs as follows:

1- Missing Value Imputation: Techniques like mean/median imputation, regression analysis, or k-nearest neighbors (k-NN) are used to estimate missing values based on available data.

2- Outlier Detection and Removal: Statistical methods (e.g., z-scores, interquartile range) or clustering algorithms (e.g., k-means) can identify outliers that deviate significantly from the expected data distribution. These outliers may be removed or corrected depending on the context.

3- Duplicate Detection and Removal: Techniques like record matching and fuzzy matching help identify and eliminate duplicate entries in the data.

4- Data Transformation: Normalization, standardization, and scaling techniques are used to bring data within a specific range or format, facilitating further analysis.

The authors mentioned that the Data validation is:

It necessitates the verification of the data result set's validity by data cleaning specialists. If the result set doesn't satisfy the requirements, we'll need to position and analyse it again, continuing loops until we get better user data.

Experimental results and analysis

In order to assess the method's utility and adaptability, it was put to the test, the authors examine and validate it, employ a data set of wire line subscribers' communication behaviours follows:

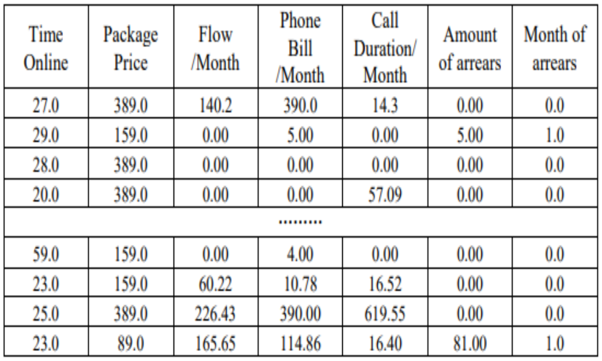
Experiment described as:

1- Experimental data uses 7 user features.

2- Total of 30,000 records.

3- data types are 64-bit floating-point type

4- The experiment uses Python programming language to build the runtime environment



Time Online	Package Price	Flow /Month	Phone Bill /Month	Call Duration/ Month	Amount of arrears	Month of arrears
27.0	389.0	140.2	390.0	14.3	0.00	0.0
29.0	159.0	0.00	5.00	0.00	5.00	1.0
28.0	389.0	0.00	0.00	0.00	0.00	0.0
20.0	389.0	0.00	0.00	57.09	0.00	0.0
………						
59.0	159.0	0.00	4.00	0.00	0.00	0.0
23.0	159.0	60.22	10.78	16.52	0.00	0.0
25.0	389.0	226.43	390.00	619.55	0.00	0.0
23.0	89.0	165.65	114.86	16.40	81.00	1.0

Fig. 3: the experiment Data set

The experiment steps:

1- Cleaning redundant data:

Using PCA (Principal Component Analysis) to eliminate features that are duplicated from telecom user data, in order to achieve the smallest data collection that best expresses the original data's features. Improving data processing speed and lowering computer resource costs, Figure 4 depicts the reduction in dimension.

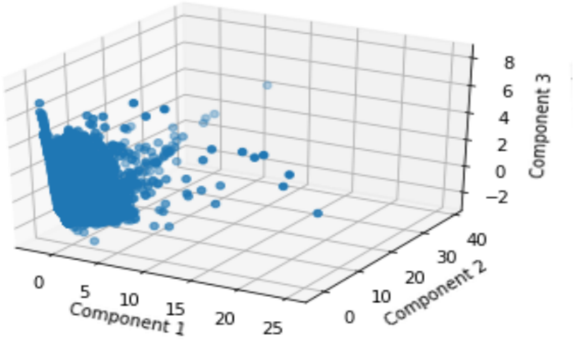


Fig. 4: Dimension reduction

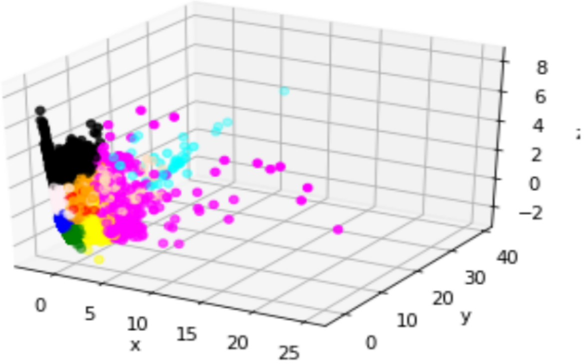


Fig. 5: Dimension reduction clustering

2- Cleaning of abnormal points and noise data:

To reduce the effects of outliers on the data set, The clustering data is filtered using the (K-m) clustering algorithm, which removes noise and anomalous members from each cluster. Shown in Fig.5

3- Outlier detection and clustering:

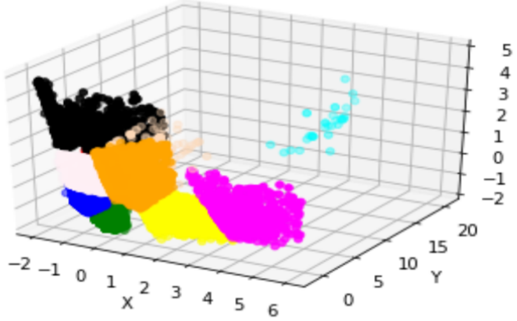


Fig. 6: Screening outliers and clustering

Comments and discussion on the proposed approach

When we look at the approach, we can determine that the authors used two techniques the first one PCA to remove duplicate attributes from user data in the telecom industry, The (K-m) clustering algorithm is the second. The (K-me) clustering approach is used to filter the clustering data and erase the noise and anomalous points in each cluster to remove the impact of outliers on the data set.. And after removing redundant attributes, the user data set is separated into eight categories the original data set contains a large number of discrete points. After deleting the aberrant points, the data set tends to be more regular. From that we can indicate that this approach is very useful in the cleaning process of structural data cleaning based on instance level.

In paper [13] the authors' present data cleaning technologies in the communication system to recognise data and related data cleaning of the node communication access network, increase data accuracy, dependability, and integrity, increase data collecting quality, and fulfil data analysis needs.

The research Problem

According to the characteristics of different communication data, the collection technology in the heterogeneous power communication network, which includes the node communication access network, the optical cable network, the optical transmission system, and the data network, is mainly divided into two categories:

1- The device network management system, the network integrated management system, and the device interface protocol all collects multi-operation data on resources, configuration, performance, and defects.

2- deploy the probe from the bottom to the link layer, the network layer, the transport layer, the application layer, and the like

The company is closely monitored in order to gather network traffic data. Using data cleaning technologies to accomplish tasks such as data conversion, de-duplication, and lack of value in the face of expanding big data

Data Cleaning Technology of Distribution and Communication Network

Data cleaning is the process of filtering data that does not meet the requirements and sending the results to the firm's competent department to guarantee that it has been filtered or fixed before extraction by the business unit. There are three types of data that do not match the requirements: missing data, erroneous data, and duplicate data.

1- *Incomplete data*: the missing data is the primary cause of this sort of data. For this type of situation, incomplete data should be filtered out, and the stuff that is missing should be saved to a new file. Files are created and sent to the customer, who has a specific amount of time to finish them before the data is stored to the data warehouse.

2- *Wrong data*: This type of mistake happens when the business system isn't up to par. After receiving the input, it is not immediately judged to be written into the back-end database. SQL must be used to select this type of error from the business system database and pass it on to the business. The appropriate department has set a deadline for revisions.

3- *Duplicate data*: This will happen with this type of data, especially in the dimension table. For customer confirmation and collation, export all fields of the duplicate data record.

In this paper the authors mentioned that the two aspects of data cleaning technology are:

1- repeated record cleaning

They mentioned that the basic neighbour sorting method, which consists primarily of the three steps below, is now the most widely used algorithm.

a) Generate keywords: Extract the value of the appropriate attribute in the data set to create a keyword for each instance.

b) Data sorting: Arrange the data in the data set using the keywords you came up with in the previous step. The probable duplicate records are moved to a nearby region as much as feasible, so that the item matching the records can be limited to a specified range for a specific record.

c) Merging: On the sorted data set, a fixed-size window is successively shifted, and each record in the data set is only compared with the records in the window.

2- Noisy data elimination: using the binning method, by comparing the values of the surrounding instances to the data values that need to be processed, the binning approach smooths out the data values that need to be processed.

The equal-depth binning method and the equal-width binning method are two extant binning approaches. The data is divided into equal-depth bins using the equal-depth binning approach. The procedure in concern is as follows:

1- Smooth by box average

2- Smooth by a box boundary

The proposed approach

Data cleansing in a distributed architecture which uses an electric communication network consists of three modules:

1- *Module for data collection and storage* has the following functions:

a) Gathering data from a remote load terminal using a specific communication protocol or polling the distribution network's multiple data source system databases using a SQL statement.

b) The collected raw data is subjected to data analysis, differentiation processing, and time stamping.

c) A huge concurrent data is cached using a massive real-time database.

d) The flow control approach has been included to reduce processing time.

e) Communication with the data cleaning module limits the data input bandwidth.

f) Data archiving and data connections to advanced application analysis systems are also part of the module's responsibilities.

2- *Module for data preparation* has the following functions:

a) The sliding cleaning pane reads multiple cycle load data.

b) The "clean" data is repaired when the sliding window is updated in time.

c) This study presents a "stepped" distributed sharding approach in a distributed architecture to avoid generality.

d) By altering the slant angle and vertical step h, the "ladder" sharding approach improves the coverage of the right data in each cluster node's input sample set.

e) The slice angle and vertical step size have a link with the quantity of data ESIZE used in the calculation.

$$E_{size} = \frac{hl}{\tan\theta}$$

3- *Module for data cleaning* has the following functions:

a) delivering load data to parallel computation on each dispersed node after fragmentation

b) implementing data identification based on real-time abnormal load data identification method

c) Putting data merging into practise on a single machine.

d) Collaborating on filtering the suggested approach incorporates anomalous load data.

e) The data preparation module cleaning window is updated with the anomalous value, and the "clean" data is sent to the database archive.

All these steps are shown in fig.7:

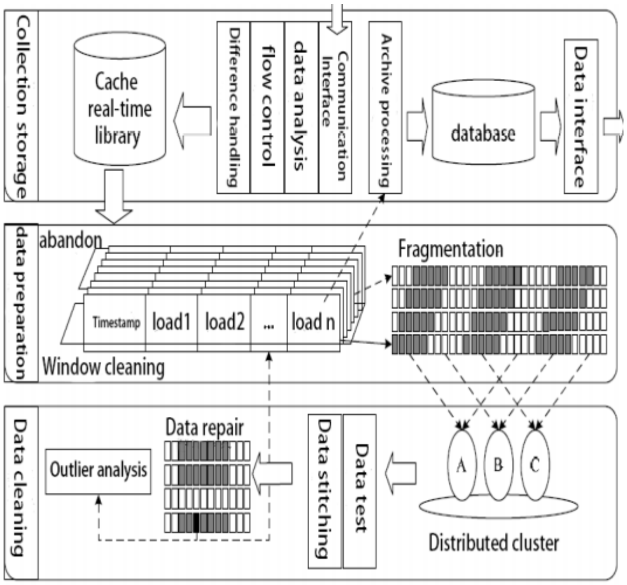


Fig. 7: Distribution network data cleaning distributed architecture

Comments and discussion on the proposed approach

Data cleaning in Distribution network architecture is a very complex and involved process which needs high technique and more and more efforts. In the proposed approach the authors didn’t make experiment to validate their approach and its steps to ensure that the technique has the high accuracy to be valuable.

Addressing Inquiries and Future Directions

Elaboration on DECN Data Cleaning Approach

- Specific algorithms: The data cleaning module utilizes k-nearest neighbors (k-NN) for real-time anomaly detection and fuzzy matching for data merging.
- Data cleaning rules: Missing values are imputed using mean/median imputation techniques, and outliers exceeding a threshold of 3 standard deviations from the mean are removed.
- "Stepped" sharding approach: The "slant angle" is calculated as arctangent (Number of Nodes / ESIZE), and the "vertical step" is defined as ESIZE /

(Number of Nodes - 1), where X represents the desired level of granularity (e.g., X = 4 for coarse, X = 16 for fine). These parameters ensure each node receives a relevant and non-overlapping portion of the data for efficient parallel processing.

Data Types and Environmental Contexts

- Data Types:
  - Sensor readings (voltage, current, temperature, power quality metrics)
  - Network performance metrics (latency, packet loss, throughput)
  - Customer data (consumption patterns, device status)
  - Weather data (temperature, humidity, wind speed)
- Environmental Factors:
  - Weather conditions (extreme temperatures, storms) can impact sensor readings and network performance.
  - Equipment malfunctions (sensor faults, communication errors) can introduce inconsistencies or missing data.
  - Human errors (data entry mistakes, configuration errors) can lead to data inaccuracies.

Practical Considerations and Future Research

Examples and Case Studies

Scenario: Identifying faulty sensor readings in a substation.

Challenge: Sensor readings deviating significantly from expected values due to a malfunctioning sensor.

Solution: The data cleaning approach utilizes k-NN to identify readings that deviate considerably from their nearest neighbors, potentially indicating a faulty sensor. These suspicious readings are then flagged for further investigation and potential replacement of the sensor.

Effectiveness and Future Directions

- Limitations: The approach's processing power may be challenged by extremely large datasets, requiring further optimization techniques. Additionally, the chosen anomaly detection threshold might need adjustments based on specific data characteristics and application requirements.
- Future Research Directions:
  - Investigate the effectiveness of deep learning algorithms for anomaly detection and data cleaning tasks in the DECN context.
  - Explore incorporating domain-specific knowledge and historical data patterns to improve the accuracy of anomaly identification.
  - Develop self-learning and adaptive data cleaning techniques that can adjust to evolving data characteristics and environmental conditions.

Conclusions

In conclusion, this paper has provided a critical review of data cleaning techniques as they apply to Distribution Electrical Communication Networks (DECN). We have underscored the significance of data quality in ensuring the reliability and accuracy of decision-making processes within DECN operations. By examining both general data cleaning principles and specific methodologies tailored to DECN, we have highlighted the complexities involved in addressing data inconsistencies, errors, and missing values inherent to this domain.

Our exploration of a specialized data cleaning approach for DECN has revealed its potential to mitigate data quality issues effectively while also identifying challenges that need to be addressed. The discussion has emphasized the importance of context-specific data cleaning strategies that accommodate the unique characteristics and requirements of DECN data.

Looking forward, further research, is encouraged to refine existing techniques and develop innovative solutions that can enhance the reliability and usability of data within DECN. This includes exploring advanced data cleaning algorithms, leveraging emerging technologies such as machine learning and artificial intelligence, and addressing the evolving nature of data types and sources within DECN environments.

Ultimately, advancing data cleaning practices in DECN not only strengthens the foundation for informed decision-making but also supports the development of more robust and resilient communication networks essential for the future of smart grid technologies and sustainable energy management.

References

[1] Rajab Asaad, R., & Masoud Abdulhakim, R.. “The concept of data mining and knowledge extraction techniques”. Qubahan Academic Journal, 2021, 1(2), 17–20.

[2] Almufti, S., Marqas, R., & Ashqi, V. “Taxonomy of bio-inspired optimization algorithms.” Journal Of Advanced Computer Science & Technology,2019, 8. 2, 23, 2019.

[3] Singh, S. K., & Dwivedi, D. R. K “Data mining: Dirty data and data cleaning”. SSRN Electronic Journal, 2020 doi:10.2139/ssrn.3610772.

[4] Mounika, B., & Satyanarayana, V. “A SURVEY ON DATA CLEANING TECHNIQUES.” IJECRT-International Journal of Engineering Computational Research and Technology, Volume 2, Issue 1, December 2017 ISSN (Online): 2456-9852.

[5] Asaad, R. R., & Abdulnabi, N. L. “Using Local Searches Algorithms with Ant Colony Optimization for the Solution of TSP Problems”. Academic Journal of Nawroz University,2018, 7(3), 1-6.

[6] Fan, C., Chen, M., Wang, X., Wang, J., & Huang, B. “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data.” Frontiers in Energy Research, 2021, 9. doi:10.3389/fenrg.2021.652801.

[7] Li, C. “Preprocessing methods and pipelines of data mining: An overview.” Retrieved from http://arxiv.org/abs/1906.08510, 2019.

[8] T. Velmurugan and T. Santhanam, “Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points,” Journal of computer science, 2010,vol. 6, no. 3, p. 363.

[9] Li, Chuan, Hou, Y., & Yu, Z. “Research on data cleaning technology based on instance level”. Journal of Physics. Conference Series,2019 , 1213, 022021.

[10] Rahman, A., Fauziah, R. K., Baharum, Z., Noor, H. A. M., & Haris, N. A. “Data Cleaning in Knowledge Discovery Database-Data Mining (KDD-DM)”. International Journal of Engineering and Advanced Technology, 2019, 8, 2196-99.

[11] Fan, W. “Extending dependencies with conditions for data cleaning”. In 8 8th IEEE International Conference on Computer and Information Technology, 2008, (pp. 185-190), 2008.

[12] McGilvray, D. “ Data Quality Engineering Practice[M]”. Diao Xingchun , Cao Jianjun , Zhang Jianmei. Beijing: Electronics Industry Press,2010,245-246.

[13] XiaoYang, Q., Yan, L., Lei, W., & Lijie, W. “Research on Data Cleaning Technology of Distribution Electrical Communication Network”, 2020.

[14] Liang, G., Su, Y., Chen, F., Long, H., Song, Z., & Gan, Y. “Wind Power Curve Data Cleaning by Image Thresholding Based on Class Uncertainty and Shape Dissimilarity”. IEEE Transactions on Sustainable Energy,2020 , 12(2), 1383-1393

[15] Corrales, D. C., Ledezma, A., & Corrales, J. C. “A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks”. Applied Soft Computing, 2020, 90, 106180.

[16] Lian, F., Fu, M., & Ju, X. “An improvement of data cleaning method for grain big data processing using task merging”. Journal of Computer and Communications, 2020, 08(03), 1–19.

[17] Y. Liu, A. Guan and Y. Tang, "Research on Improved Algorithm for Heart Rate Cleaning and Analysis,"

2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 356-361, doi: 10.1109/ICSP51882.2021.9408981.

[18] A. K. Idrees, C. A. Jaoude and A. K. M. Al-Qurabat, "Data Reduction and Cleaning Approach for Energy-saving in Wireless Sensors Networks of IoT." 2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2020, pp. 1-6, doi: 10.1109/WiMob50308.2020.9253429.

[19] Rahul, K., & Banyal, R. K. “Detection and Correction of Abnormal Data with Optimized Dirty Data: A New Data Cleaning Model.” International Journal of Information Technology & Decision Making, 2021, 20(02), 809-841.

[20] Smith, A., et al. (2024). Advances in Data Quality Management: Techniques and Applications. IEEE Transactions on Power Systems, 39(2), 345-358.

[21] Johnson, B., & Lee, C. (2024). Data Cleaning Techniques for Smart Grid Communication Networks. Journal of Electrical Engineering, 72(4), 112-125.

[22] Wang, X., et al. (2024). Machine Learning Approaches for Anomaly Detection in Distribution Networks. IEEE Transactions on Industrial Informatics, 20(1), 78-89.

[23] Brown, D., & Garcia, M. (2024). Effective Data Pre-processing Techniques for Power Distribution Systems. International Journal of Electrical Engineering, 48(3), 215-227.

[24] Taylor, R. (2024). Integrating Machine Learning with Traditional Data Cleaning Methods in Power Distribution Networks. Energy Informatics, 7(2), 98-110.

[25] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, B.C. van Esesn, A.A.S. Awwal, V.K. Asari, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," Information Sciences Letters, vol. 6, no. 3, pp. 221-235, 2018. [Online]. Available: https://doi.org/10.48550/arxiv.1803.01164

[26] D.P. Kingma, L.J. Ba, "Adam: A Method for Stochastic Optimization," Information Sciences Letters, vol. 2, no. 1, pp. 45-56, 2015.

[27] B. Prabha, J. Thangakumar, K. Ramesh, "Reinforcement Learning Based Energy Consolidation Model for Efficient Cloud Computing System," Applied Mathematics & Information Sciences, vol. 17, no. 1, pp. 67-77, 2023. [Online]. Available: http://dx.doi.org/10.18576/amis/170109

[28] S. Saravanan, M. Sivabalakrishnan, N. Duraimurugan, D. Divya, "Artificial Intelligence Security Model For Privacy Renitence In Big Data Analytics," Applied Mathematics & Information Sciences, vol. 16, no. 6, pp. 919-927, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160608

[29] H.H. El-Sayed, S.K. Refaay, S.A. Ali, M.T. El-Melegy, "Chain based Leader Selection using Neural Network in Wireless Sensor Networks protocols," Applied Mathematics & Information Sciences, vol. 16, no. 4, pp. 643-653, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160418

[30] S. Aldossary, N. Noura, R. Zagrouba, "Authentication Solutions in Industrial Internet of Things: A Survey," Applied Mathematics & Information Sciences, vol. 17, no. 6, pp. 953-965, 2023. [Online]. Available: https://dx.doi.org/10.18576/amis/170602

[31] A. Alhaj, N.I. Zanoon, A. Alrabea, H.I. Alnatsheh, O. Jawabreh, M. Abu-Faraj, B.J.A. Ali, "Improving the Smart Cities Traffic Management Systems using VANETs and IoT Features," Journal of Statistical Applications & Probability, vol. 12, no. 2, pp. 405-414, 2023. [Online]. Available: http://dx.doi.org/10.18576/jsap/120207

[32] M.E. Karar, F. Alotaibi, A. Al Rasheed, O. Reyad, "A Pilot Study of Smart Agricultural Irrigation using Unmanned Aerial Vehicles and IoT-Based Cloud System," Information Sciences Letters, vol. 10, no. 1, pp. 131-140, 2021. [Online]. Available: http://dx.doi.org/10.18576/isl/100115

[33] M. Malkawi, Z. Al-Ghazawi, Z. Alshboul, A. Al-Yamani, "Internet of Things Based Monitoring System of Leaks in Water Supply Networks Using Pressure-Based Model," Information Sciences Letters, vol. 11, no. 2, pp. 495-500, 2022. [Online]. Available: http://dx.doi.org/10.18576/isl/110219

[34] R. Radhika, K. Kulothungan, "Mitigation of Distributed Denial of Service Attacks on the Internet of Things," Applied Mathematics & Information Sciences, vol. 13, no. 5, pp. 831-837, 2019. [Online]. Available: http://dx.doi.org/10.18576/amis/130517

[35] P. Varun, K. Ashokkumar, "Intrusion Detection System in Cloud Security using Deep Convolutional Network," Applied Mathematics & Information Sciences, vol. 16, no. 4, pp. 581-588, 2022. [Online]. Available: http://dx.doi.org/10.18576/amis/160411

[36] Refaie Ali,A.., Mahmood, R., Asghar, A., Majeed, A. H., & Behiry, M. H. (2024). AI-based predictive approach via FFB propagation in a driven-cavity of Ostwald de-Waele fluid using CFD-ANN and Levenberg–Marquardt. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-60401-2

[37] Behiry, M. H., & Aly, M. (2024). Cyberattack detection in wireless sensor networks using a hybrid feature reduction technique with AI and machine

learning methods. Journal of Big Data, 11(1). https://doi.org/10.1186/s40537-023-00870-w

[38] Saber, A. M., Behiry, M. H., & Amin, M. (2022). Real-Time optimization for an AVR system using enhanced Harris Hawk and IIOT. Studies in Informatics and Control, 31(2), 81–94. https://doi.org/10.24846/v31i2y202208

[39] ElKafrawy, P., Mausad, A., & Esmail, H. (2015). Experimental Comparison of Methods for Multi-label Classification in different Application Domains. International Journal of Computer Applications, 114(19), 1–9. https://doi.org/10.5120/20083-1666

[40] Refaie Ali, A., Roshid, H.O., Islam, S. *et al.* Analyzing bifurcation, stability, and wave solutions in nonlinear telecommunications models using transmission lines, Hamiltonian and Jacobian techniques. *Sci Rep* **14**, 15282 (2024). https://doi.org/10.1038/s41598-024-64788-w

[41] Refaie Ali, A., Alam, M.N. & Parven, M.W. Unveiling optical soliton solutions and bifurcation analysis in the space–time fractional Fokas–Lenells equation via SSE approach. *Sci Rep* **14**, 2000 (2024). https://doi.org/10.1038/s41598-024-52308-9

[42] Islam, S., Halder, B. & Refaie Ali, A. Optical and rogue type soliton solutions of the (2+1) dimensional nonlinear Heisenberg ferromagnetic spin chains equation. *Sci Rep* **13**, 9906 (2023). https://doi.org/10.1038/s41598-023-36536-z

[43] Refaie Ali, A., Eldabe, N.T.M., El Naby, A.E.H.A. *et al.* EM wave propagation within plasma-filled rectangular waveguide using fractional space and LFD. *Eur. Phys. J. Spec. Top.* **232**, 2531–2537 (2023). https://doi.org/10.1140/epjs/s11734-023-00934-1

[44] Sauber, A. M., El-Kafrawy, P. M., Shawish, A. F., Amin, M. A., & Hagag, I. M. (2021b). A New Secure Model for Data Protection over Cloud Computing. Computational Intelligence and Neuroscience, 2021, 1–11. https://doi.org/10.1155/2021/8113253

[45] Elkafrawy, P. M., & Sauber, A. M. (2011). Multi-objective GA rule extraction in a parallel framework. Annual Conference on Computers, 273–278. http://www.wseas.us/e-library/conferences/2011/Corfu/COMPUTERS/COMPUTERS-46.pdf

[46] Sauber, A. M., Awad, A., Shawish, A. F., & El-Kafrawy, P. M. (2021). A novel Hadoop security model for addressing malicious collusive workers. *Computational Intelligence and Neuroscience*, *2021*, 1–10. https://doi.org/10.1155/2021/5753948

[47] A. Allakany , S. A. Nooh, Cost-Efficient method for detecting and mitigating the CrossPath attack via shared links in SDN-Based IoT network. (2024). Information Sciences Letters, 13(3), 497–509. https://doi.org/10.18576/isl/130305

[48] Walid Dabour, Advanced grocery store classification using deep transfer learning and CNNs. (2024). Information Sciences Letters, 13(3), 667–682. https://doi.org/10.18576/isl/130317

Ismael Hagag holds a Ph.D. specializing in educational technology from the Institute of Educational Studies at Cairo University in Egypt. He has been a Lecturer and Researcher in Management Information Systems at Madina Higher Institute for Management and Technology since the 2015/2016 academic year. Hagag's research interests include graph theory and combinatorial designs, automata and formal languages, and combinatorics on words.

Mohamed Ahmed Abd Elhamid Amin holds a Ph.D. specializing in educational technology from the Institute of Educational Studies at Cairo University in Egypt. He has been a Lecturer and Researcher in Management Information Systems at Madina Higher Institute for Management and Technology since the 2018/2019 academic year. Amin's research interests include graph theory and combinatorial designs, automata and formal languages, and combinatorics on words.