Ensuring Data Quality in Artificial Intelligence: Managing Transformations in Nuclear Power Plant Operational Data

Jae Min Kim*, Gwi Sook Jang, Seo Ryong Koo
**Korea Atomic Energy Research Institute, 111, Daedeok-daero 989 beon-gil, Daejeon 34057, Republic of Korea,*
*Corresponding author: *jaemink@kaeri.re.kr*

***Keywords :** abnormal operation, feature selection, artificial intelligence, data quality management*

1. Introduction

In recent years, there has been a significant increase in research focused on applying artificial intelligence (AI) to the challenges faced in nuclear power plant (NPP) operations. The development of AI systems that assist operators in decision-making processes requires not only high-quality data but also data that is appropriately transformed to facilitate effective learning by AI models.

The performance and validation of systems utilizing AI, such as the one being developed, are heavily influenced by the quality of data. To address this, a common approach is to improve the quality of data, which in turn enhances the reliability of AI models. Data that is produced and managed according to special management processes will not only lay the foundation for selecting high-quality data but also become a necessary process when regulatory agencies demand higher requirements.

Transforming data into a form that is suitable for AI learning is a critical step in the development process. However, it is essential to ensure that these transformations are justified and do not introduce unnecessary modifications that could compromise data integrity or reduce the reliability of the AI models. The emphasis on data quality management throughout this process is crucial in verifying that the transformations applied to the data are beneficial and do not inadvertently degrade its utility.

If standards for managing data quality are applied to the production and management of NPP operation data, it could enable the use of a common language across various groups engaged in AI-related research. This would allow for the comparison of model performance and the assessment of the training data used on a common basis.

Consequently, this could serve as foundational work that regulatory agencies might utilize in future applications at actual power plants. To achieve this, it is necessary to develop methods to verify the validity of the operational data produced.

2. Data Quality Management

Efforts to develop international standards related to data quality began in 2006 under the Technical Committee of the International Organization for Standardization (ISO), known as 'ISO 8000'. Initially, the focus was on master data commonly used in applications like Enterprise Resource Planning (ERP). This approach aimed at standardizing data values and formats, thereby improving the quality of the data itself and can be characterized as a 'data-centric' quality management method.

However, this method alone faced challenges as it often led to the repetition of the same errors. To overcome these limitations, a 'process-centric' approach to data quality management became necessary. In this approach, tracking the flow of data, such as the transformation or transfer of erroneous data, allows for the correction and identification of root causes of data errors, thereby preventing the recurrence of the same data errors.

Transforming data to make it more suitable for AI learning, such as through feature selection or other techniques, is an important aspect of preparing the data. However, these transformations must be carefully managed to ensure they enhance, rather than detract from, the data's quality. This involves scrutinizing the transformed data to verify that it still accurately represents the original operational conditions, and that no essential information has been lost or distorted in the process.

We plan to manage the data for training the AI model to be used in the NPP operator support system by applying this process. This management is based on ISO-8000 as reported by the Korean Standards Association.

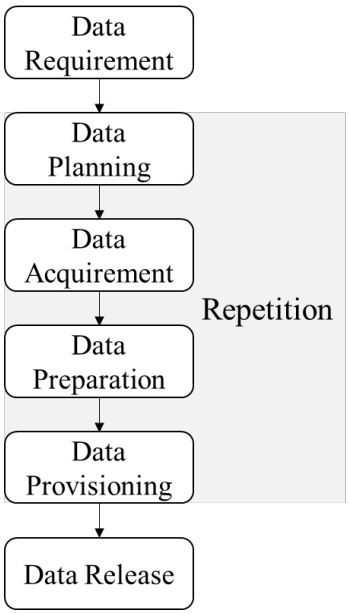

Fig. 1. Data treatment along with data lifecycle.

Figure 1 illustrates the data lifecycle from analyzing user requirements to release. Most stages are subjective to user opinions and can be readily applied across various industrial sectors. However, labeling and data quality assessment require methods that adequately reflect the characteristics of the specific industry where this process is applied.

The data preparation and data provisioning phases are particularly relevant in this context. Considering the characteristics of NPP operation data, it involves collecting data that reflects the fluid and equipment characteristics of each system in the time series. Considering these features of the data, the aim is to develop metrics for assessing the quality of data according to operating scenarios.

To enable this, two approaches can be explored: the traditional data mining method utilizing statistical analysis and the utilization of AI models to identify key features within the data. However, when employing AI models, it is important to note that distinguishing features to effectively differentiate given labels and finding features that represent specific events are two different tasks.

As a foundational step in developing the intelligent operator decision support system utilizing AI, indicators will be selected to determine whether the operation scenarios, including simulated abnormal events, are adequately reflected in the data collected.

3. Example: AI training for abnormal operation

The application of data transformation techniques, such as feature selection, must be validated to ensure that they serve their intended purpose in improving AI model performance. This involves not only applying these techniques but also rigorously testing the outcomes to confirm that the model's performance is enhanced as a result. In the context of NPP operations, this means assessing whether the transformed data contributes to more accurate predictions and better decision-making capabilities for operators.

Before the data is used for AI model training, it undergoes a preparation phase where its quality and relevance are thoroughly reviewed. This phase includes evaluating whether the data accurately reflects the operational conditions it is meant to represent, especially after any transformations have been applied. The goal is to ensure that the AI model is learning from data that is both representative of real-world scenarios and free from unnecessary modifications that could introduce biases or errors.

In situations where the evaluation reveals that the transformations were unnecessary or even detrimental to model performance, it may be necessary to revise the data preparation strategy. This might involve reselecting features, adjusting the intensity of transformations, or even reverting to a simpler data structure that better aligns with the AI model's learning requirements. The key is to strike a balance between optimizing data for AI training and maintaining the integrity of the original data.

This process also involves continuous iteration and refinement. As the AI model is trained and its performance is evaluated, insights gained from these evaluations are fed back into the data preparation process. This feedback loop helps to progressively improve both the data quality and the model's accuracy, ensuring that the system evolves to meet the operational demands of the NPP environment.

If this research follows data quality management procedures, the data requirement is data recording the changing state of an NPP after an abnormal event occurs. During the data planning stage, activities such as selecting abnormal events, operation times, points of abnormal event injection, and labeling for abnormal situations are performed.

In the data acquisition phase, data is produced according to the planned abnormal operation scenarios. NPP operation data is collected through simulators due to the limited quantity available from actual operation data and the restricted implementation of events.

Before using the produced data for training an AI model, there is a data preparation phase to review whether the planned data has been well collected. This phase allows for the examination of whether the data showed a different response due to the occurrence of an abnormal event, or if there were any omissions or outliers due to errors.

Once reviewed, the data is used for training the AI model, and data provisioning follows, based on the model results. This involves revising the data plan, including producing additional data or adjusting the intensity of abnormal events, and repeating this process as needed to achieve desired outcomes.

The data used in this manner should be managed in a distributable form for potential use in future research.

To avoid security issues, sensitive content should be removed, a distribution management ledger created, and a document outlining the data management procedures should be written to make it easy to understand.

4. Conclusions

This paper highlights the importance of both data quality management and the careful application of data transformations in the development of AI-based NPP operator support systems. While transforming data to improve its suitability for AI learning is necessary, it is equally important to ensure that these transformations do not introduce unnecessary modifications that could compromise data quality.

The success of AI systems in this context depends on a careful balance between enhancing data for model training and maintaining the integrity of the original information. The strategies discussed in this paper emphasize the need for rigorous validation of data transformations and continuous monitoring of data quality throughout the AI model's lifecycle. Future research will continue to explore the most effective data transformation techniques and how they can be integrated into robust data quality management practices to optimize AI performance in NPP operations.

ACKNOWLEDGEMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry, & Energy (MOTIE) of the Republic of Korea (No. 20224B10100130).

REFERENCES

[1] ISO TC184/SC 4, “Data quality - Part 2: Vocabulary,” ISO8000-2:2022, 2022.

[2] ISO/IEC JTC1 SC42 WG2 N1635, “Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 1: Overview, terminology, and examples,” ISO/IEC WD 5259-1, 2022.

[3] ISO/IEC JTC1 SC42 WG2 N1635, “Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework,” ISO/IEC WD 5259-1, 2022.

[4] Gyu-Hee Jeong, “Analysis of Artificial Intelligence Data Quality Certification Standards and Implementation of Data Certification Process,” Korean Standards Association, 2022.