Development of a Clustering Methodology for Deriving Representative Accident Scenarios Using Advanced Dynamic Time Warping Algorithm

Byun, Hyeonho [a], Kim, Hyeonmin [a*]
[a]*Korea Atomic Energy Research Institute, 34057 111 Daedeok-daero 989beongil, Yuseong-gu,*
[*]*Corresponding author: hyeonmin@kaeri.re.kr*

***Keywords :** PSA, code simulation, UMAP, DTW

1. Introduction

To develop new reactor designs, such as small modular reactors (SMRs), it is necessary to conduct probabilistic safety assessment (PSA) to verify the safety of the reactor. In particular, PSA involves classifying representative accident scenarios for each initial event (IE) and evaluating the overall risk to validate the reactor's safety. Recently, dynamic PSA (DPSA) has been proposed as a more advanced approach to risk evaluation compared to the conservative methods of traditional binary event tree-based risk assessments [3]. DPSA diversifies event branches based on the timing and consequences of events and comprehensively evaluates the risks associated with each branch.

Based on DPSA, various accident scenarios can be derived, representative accident scenarios can be identified, and optimal response systems can be developed. However, DPSA is characterized by the theoretical possibility of generating an almost infinite number of thermohydraulic progression scenarios following an IE. Performing thermohydraulic analyses on all these nearly infinite scenarios to develop an optimal response system would be highly inefficient. Therefore, it is essential to establish a methodology for identifying representative scenarios with similar thermohydraulic progressions, and there by analysis can focus on representative scenario.

This study aims to develop a clustering methodology to identify the structural relationships among various thermohydraulic data, and visualize them in a lower dimensional space (e.g., three-dimension), and assign similar accident scenarios to the same cluster. Through this approach, representative accident scenarios can be derived effectively.

2. Methodology

This study focuses on deriving accident simulation results from MAAP through Monte Carlo simulation-based random sampling. The generated accident scenarios are then analyzed using UMAP based on soft-DTW for dimensionality reduction and DBSCAN for clustering the accident scenarios. All research materials were implemented using Python, with UMAP employed via the umap-learn library, DBSCAN via the sklearn library, and soft-DTW implemented manually as a custom metric, as it is not natively supported.

2.1 Random scenario sampling

Random Scenario Sampling (RSS) refers to the process of stochastically determining the states of major systems and components in a nuclear reactor over time to define accident scenarios. Using the Monte Carlo method, the operational success, failure, or failure timing of each component is determined stochastically. This study used MAAP code to simulate the reactor’s accident scenario, however, given that MAAP involves hundreds of reactor components, RSS must be very limited to key Structures, Systems, and Components (SSCs) rather than all components.

When performing random sampling for SSCs, their operational success or failure is determined based on existing PSA failure data. For example, if an SSC is in standby mode and needs to operate, its operation or failure is probabilistically determined using the failure probability on demand. Secondly, for components already in operation, the time of failure is determined using operational failure rate data analyzed by former PSA study.

If an operator's intervention is required, it also becomes subject to RSS. The operator's required action time is calculated probabilistically, along with the success or failure of the action. Finally, nuclear reactors involve interconnected systems, where the operation of certain elements is conditionally dependent on the proper functioning of others. This interdependency is designed using combinations of AND/OR gates. Thereby, we can formulate various accident scenarios according to the failure probability and failure time of SSCs. This study simulated 6,400 scenarios by using suggested random sampling method.

2.2 Soft Dynamic Time Warping algorithm

Soft DTW is an extension of DTW, made differentiable using a mathematical trick. In this section, the general characteristics of the DTW algorithm are explained first,

followed by a discussion of soft-DTW. Although UMAP supports various custom metrics, the custom metric must be capable of calculating both the distance and gradient between two time series. Therefore, this study applied soft-DTW instead of DTW [1].

DTW is a metric designed to better compare two time series data with different speeds or lengths. It achieves this by matching the elements of the two datasets in a way that minimizes the cumulative distance, finding the optimal warping path that minimizes the distance between the two. The definition of DTW distance between two time series data is as follows:

$$D_{i,j} = E_{i,j} + min(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1}) \quad (1)$$

However, $D_{i,j}$ is not differentiable function because of the second term in (1). Therefore, a differentiable function is suggested in order to flexibly use the concept of DTW. The soft-DTW replaces the second term in (1) by using the concept of *softmin*. The *softmin* enables the *min* to be soft, and thereby (1) can be differentiable

$$softmin_r = -rlog(\Sigma exp(-a_i/r)) \quad (2)$$

2.3 Scenario clustering algorithm

This section introduces the UMAP algorithm. UMAP reduces data to a latent space for visualization purposes only, so labels were assigned to each data point using Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

The goal of UMAP is to find an optimally embedded set E consisting of three-dimensional coordinates, where E minimizes the edgewise cross-entropy for each PCT dataset. To achieve this, the algorithm computes the distances to neighboring data points for every data point using a defined metric, a process referred to as k-nearest neighbor searching. In this study, the soft-DTW metric method is applied during this process. After the k-nearest neighbor searching, a weighted directed graph is constructed by computing the weights of edges between data points. Finally, the embedded result with the lowest cross-entropy is derived [4, 5].

DBSCAN is an algorithm based on the idea that spatially adjacent data points are likely to belong to the same cluster. If the density of data points around a certain point is high, it assigns a cluster, while low-density points are treated as noise. DBSCAN requires two parameters: the radius (epsilon) and the minimum number of data points within the radius [2].

3. Results

This section presents the results of classifying 6,400 accident scenarios using PCT outputs generated through Monte Carlo-based simulations, UMAP with a soft-DTW metric, and DBSCAN. The classification results using the soft-DTW metric were analyzed.

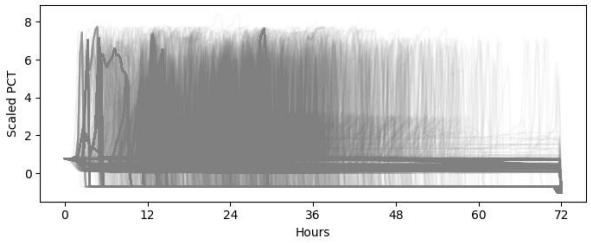

Fig. 1. Scaled PCTs corresponding to the LOOP scenarios

Typically, there are 10 to 20 types of IEs that can lead to accidents in PWRs. However, each IE is further subdivided into sub-scenarios depending on the component failure of related safety systems and the timing of such failures. Therefore, instead of deriving and analyzing scenarios for every type of IE, it is possible to select on representative scenario, apply the clustering method developed in this study, and validate it. If successful, this methodology can be applied to other scenarios as well. Hence, this study performed a case study using the LOOP, a scenario has similarities both large and small reactors.

PCTs were calculated over a 72-hour period following the IE at 50-second intervals. Fig. 1 shows the scaled PCT results, where the average value of the entire PCT during a single scenario was normalized to 1. This represents the scaled PCTs for all 6,400 accident scenarios. The target reactor type was OPR-1000, the required design parameters are taken from the PSA technical report carried out by KAERI [6]

The final result of UMAP and DBSCAN is shown in Fig.2. Data treated as noise by DBSCAN were assigned a label of -1, which is why the minimum value on the color bar on the right side of Fig.2 is -1.

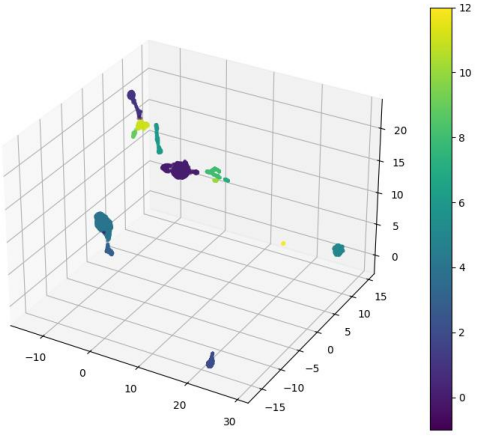

Fig. 2. Clustered results of LOOP scenario

The most effective method for deriving representative accident scenarios for each cluster is to identify the data points located near the cluster's centroid. However, if the cluster has a skewed shape (e.g., a curved line), the

centroid may lie outside the cluster structure, potentially leading to inappropriate representative accident scenarios. However, the results using the Soft-DTW metric confirmed that the centroids of all clusters were located within their respective cluster structures.

Fig. 3 shows all representative scenarios classified using the soft-DTW metric. Excluding cases classified as noise, a total of 13 clusters were identified. For each cluster, the five PCT results closest to the centroid were plotted.

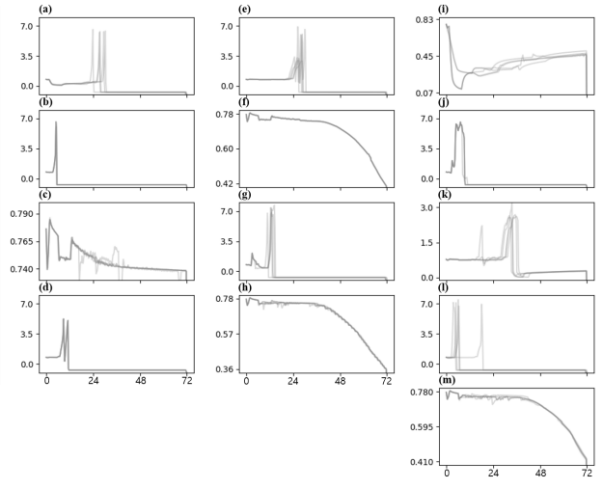

Fig. 3. Representative PCT curves located near to centroid of each cluster

Out of the total 13 clusters, 6 are scenarios without CD (c, f, h, i, k, m), while the remaining 7 (a, b, d, e, g, j, l) correspond to scenarios where CD occurred (Fig. 3).

After analyzing the similarities in the accident progression of the five representative scenarios for each cluster, it was confirmed that the representative accident scenarios within the same cluster share the same event heading sequence, as summarized in Table II.

Table II: Summary of representative scenario corresponding to each cluster

Cluster index	Scenario summary
(a)	AC power recovery, HPSI on, TDAFW runs to fail, MDAFW backup TDAFW, PSV stuck, CSR fails to start, recirculation on, seal-LOCA triggered, CD
(b)	MACST fail, CD
(c)	AC power immediate recovers after IE, TDAFW on, OK
(d)	AC power fails to recover, HPSI runs to fail, TDAFW runs to fail, MDAFW runs to fail, CD
(e)	AC power fails to recover, HPSI on, TDAFW runs to fail, MAFW runs to fail, CD
(f)	AC power immediate recovers after IE, TDAFW on, OK
(g)	AC power repeatedly alternates between recovery and failure, HPSI fails to start, TDAFW on, PSV stuck, seal-LOCA triggered, CD
(h)	AC power recovery, TDAFW on, CSR runs to fail, recirculation on, seal-LOCA triggered, OK
(i)	AC power recovery, HPSI on, TDAFW runs to fail, MDAFW runs to fail*, PSAV stuck, CSR runs to fail*, recirculation on, OK *The failure occurs beyond 40 hours. Therefore, the PCT increases gradually due to the lack of cooling system, but the simulation completes in an OK state
(j)	AC power fails to recover, HPSI on, TDAFW runs to fail, fails to MDAFW backup, PSV stuck, CSR on, recirculation on, CD
(k)	AC power recovery, HPSI on, TDAFW runs to fail, MDAFW runs to fail, CSR runs to fail, recirculation on, seal-LOCA triggered, OK
(l)	AC power fails to recover, HPSI runs to fail*, TDAFW runs to fail*, MDAFW fails to start, PSV stuck, CSR fails to run*, recirculation on, CD at 6 hours *The failure occurs earlier than CD
(m)	AC power recovery, HPSI runs to fail, TDAFW on, CSR runs to fail, recirculation on, seal-LOCA triggered, OK

The derived representative accident scenarios are expressed as distinguishable event combinations, except for (c) and (f), which have exactly the same event sequence. While (c) and (f) can be differentiated based on the shape of the PCT, the types and timings of the safety systems that were activated are identical. However, the scenario summaries presented in Table II are based on an analysis of MAAP's input files, so if the analysis had been extended to include all safety systems of the reactor for activation/failure analysis, different results might have been obtained.

4. Discussion

The main assumption in this study is that PCT, a thermohydraulic variable, represents the overall status of the reactor. This choice was made because PCT is the most important variable for determining the presence or absence of CD. The second reason is that currently, UMAP does not have a method for calculating the informatic distance between time series data of more than two dimensions, so only PCT, a one-dimensional time series, was used to classify LOOP SBO accident scenarios. While UMAP does have a method to evaluate the similarity between image data, it flattens 2D image data into one dimension for classification, so

algorithmically, it is not much different from classifying one-dimensional data.

5. Conclusions

In conclusion, this study introduces a novel clustering methodology that effectively identifies representative accident scenarios by analyzing the structural relationships among thermohydraulic data. By leveraging techniques such as UMAP for dimensionality reduction and soft-DTW for similarity measurement, the proposed approach offers a way to visualize complex accident scenarios in a lower-dimensional space, making it possible to group similar scenarios together. This methodology enhances the efficiency of probabilistic safety assessments (PSA) for advanced reactor designs, such as small modular reactors (SMRs), by reducing the computational burden of analyzing an almost infinite number of potential thermohydraulic progressions. By classifying and clustering accident scenarios based on their thermohydraulic behavior, this approach paves the way for developing optimal response systems and improving the overall safety assessment process in reactor design.

ACKNOWLEDGEMENT

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (No. GTL24031-000).

REFERENCES

[1] Keogh, Eamonn, and Chotirat Ann Ratanamahatana. "Exact indexing of dynamic time warping." Knowledge and information systems 7 (2005): 358-386.
[2] Schubert, Erich, et al. "DBSCAN revisited, revisited: why and how you should (still) use DBSCAN." ACM Transactions on Database Systems (TODS) 42.3 (2017): 1-21.
[3] Devooght, Jacques, and Carol Smidts. "Probabilistic dynamics as a tool for dynamic PSA." Reliability Engineering & System Safety 52.3 (1996): 185-196.
[4] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction."
[5] Hozumi, Yuta, et al. "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets." Computers in biology and medicine 131 (2021): 104264.
[6] KAERI, Probabilistic Safety Assessment for Ulchin Units 3&4 [Level 1 PSA for Internal Events : Main Report]