An Approach to Machine Learning-based Categorization of Source Term Behaviors of Level 2 PSA Scenarios

Kyungho Jin, Jaehyun Cho[*]
Korea Atomic Energy Research Institute,
(34057) 111, Daedeok-daero 989, Daejeon, Republic of Korea

[*]Corresponding author: chojh@kaeri.re.kr

1. Introduction

In general, a number of severe accident (SA) scenarios developed in level 2 probabilistic safety assessment (PSA) are characterized into several categories because grouping similar scenarios has the advantage of reducing computing costs before analyzing their source terms. To date, this categorization has been completely reliant on *qualitative* methods such as logical trees or expert judgements due to the lack of source term data: after grouping SA scenarios qualitatively, only one representative scenario of each category is analyzed to evaluate its source term behavior.

Recently, the Korea Atomic Energy Research Institute (KAERI) developed a severe accident (SA) simulation scheme to efficiently analyze source terms for a large number of SA scenarios rather than analyzing the limited number of scenarios [1, 2]. This approach, called *exhaustive simulation*, aims to construct massive source term database for all possible scenarios in level 2 PSA. Thanks to the exhaustive simulation, *quantitative* source term categorizations can now be accomplished.

Therefore, this paper, inspired by the exhaustive simulation, proposes a data-driven categorization method using the constructed source term database. Specifically, two unsupervised learning methods are employed in this paper: one is clustering to categorize unlabeled SA scenarios, and the other is an autoencoder structure to reduce dimensionality of multivariate time series source term data before clustering.

This paper is organized as follows. Section 2 briefly introduces the source term database constructed by the exhaustive simulation. Section 3 describes overall clustering framework with dimensionality reduction techniques. Finally, the categorization results with source term characteristics are compared to the conventional method in Section 4.

2. Source term database constructed by the exhaustive simulation

As mentioned in Sec. 1, KAERI constructed a source term database for the OPR1000 using exhaustive simulation scheme [1, 2]. Specifically, the source terms for a total of 690 SA scenarios that occupy 99% of total core damage frequency were analyzed. Each source term data consists of the time-dependent accumulated release fraction of 18 radioactive materials with different lengths.

Note that this paper used source term data for a total 658 SA scenarios after preprocessing because 32 scenarios were difficult to utilize for clustering (e.g., most of them have a value of zero). Furthermore, not all variable but 3 important variables (CsI, CsOH, and Cs_2MoO_4) that are known as major contributors to the accident consequence [3] were used to simplify the data-driven structure. Table I shows a description of the source term data used in this paper.

Table I. Descriptions of the source term database used in this paper

No. SA scenarios	N = 658
Notation of *i*-th SA scenario	S_i, i = 1, 2, 3, ..., N
Type of data	Time-dependent accumulated release fraction, denoted by x_{t_i}
No. used variables	d = 3 (CsI, CsOH, and Cs_2MoO_4)

3. Clustering framework with dimensionality reduction

Although the exhaustive simulation can provide massive and quantitative information for source term categorization, it is difficult to directly utilize this type of data for a data-driven method because of its inherent features. The issues to be resolved in the source term database can be summarized as follows:

- SA scenarios are unlabeled.
- Source term data is multivariate and time-dependent data having high dimensionality.

SA scenarios have no specific distinctions for categorization. In other words, we do not know which category the 2nd scenario belongs to. Therefore, unsupervised learning method should be considered to group unlabeled SA scenarios. In this regard, clustering algorithm [4, 5] is employed in this paper.

Although clustering can handle unlabeled SA scenarios for source term categorizations, high dimensionality of data should be carefully dealt with because it degrades clustering performance due to *the curse of dimensionality*. Therefore, an autoencoder structure based on artificial neural network (ANN) is used to reduce the dimension of source term data in this paper. Overall framework of the proposed method for quantitative source term categorization can be illustrated in Fig. 1.

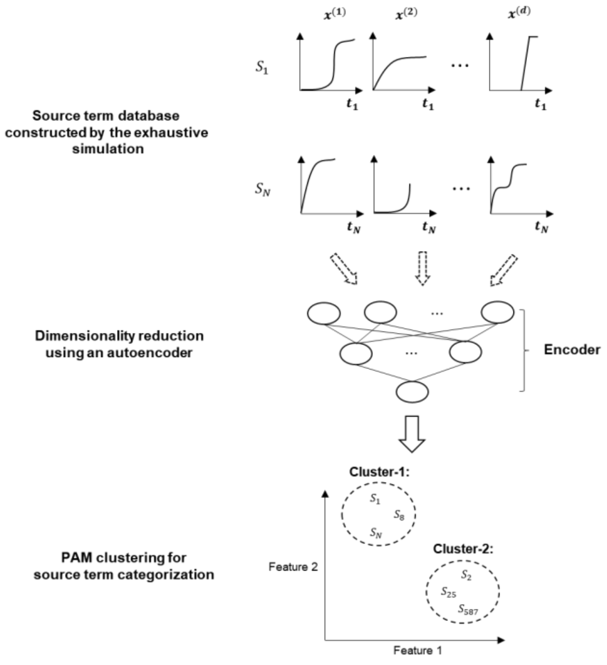
$x^{(1)}$
$x^{(2)}$
$x^{(d)}$
S_1
t_1
Source term database constructed by the exhaustive simulation
S_N
t_N
Dimensionality reduction using an autoencoder
Encoder
Cluster-1:
S_1
S_8
S_N
Cluster-2:
S_2
S_{25}
S_{587}
Feature 2
Feature 1
PAM clustering for source term categorization

Fig. 1. Clustering framework with dimensionality reduction for source term categorization

3.1 Autoencoder for dimensionality reduction

One way to avoid the curse of dimensionality is to reduce the data dimension and extract the major features from the time series data. Although this method is likely to involve some information loss while reducing dimension, it is fact that it is more efficient to enhance clustering performance rather than using high dimensionality data as it is [6].

The dimensionality reduction of multivariate time series data can be simply achieved by several well-known method such as principal component analysis (PCA) or an autoencoder. Especially, in this paper, an autoencoder structure was employed. Fig. 2 shows the example structure of an autoencoder which is based on ANN.

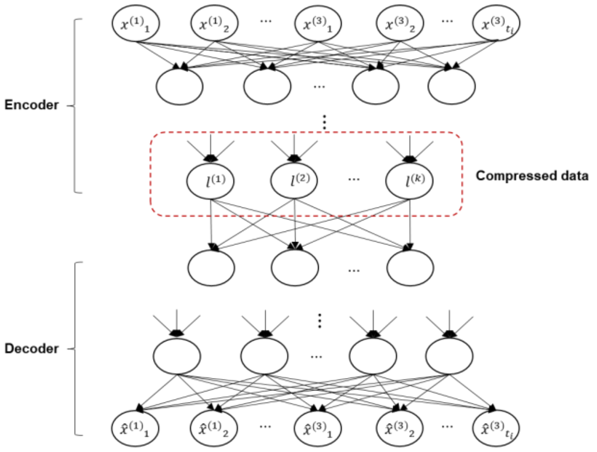

Fig. 2. Example of the autoencoder structure.

In Fig. 2, the input layer receives time series data (S_i) as input. This sequential data is compressed while passing through the hidden layers. After encoding, the original high dimensional data can be compressed: the dimension of the original data ($N \times t \times d$) can be reduced to ($N \times k$) at the end layer of encoder. Therefore, the compressed data L can be represented as follows:

$$L = \begin{bmatrix} l_1^{(1)} & \cdots & l_1^{(k)} \\ \vdots & \ddots & \vdots \\ l_N^{(1)} & \cdots & l_N^{(k)} \end{bmatrix} \tag{1}$$

It should be noted that the compressed data L contains the key features of the high dimensional time series data. As mentioned in Sec. 1, since an autoencoder is the unsupervised learning method, the decoder receives the dimensionally reduced data L to reconstruct the inputs $\widehat{x_1^1}, \widehat{x_2^1}, \dots, \widehat{x_{t_i}^d}$. This model is trained by minimizing the error between original and reconstructed data.

3.2 Clustering: PAM algorithm

Clustering is a typical unsupervised learning method that can classify unlabeled data. In other words, clustering is a way of grouping similar data based on their similarity or dissimilarity without a specific distinction. In this paper, K-medoids, one of the famous clustering algorithms is employed because of its low-sensitiveness to outliers. It is also known as partitioning around medoids (PAM) algorithm.

PAM is a partitional clustering method based on medoids. When the number of clusters n_c is determined, n_c data points are randomly selected as the medoids (e.g., $l_1, l_{15}, \dots, l_{127}$ data points in the compressed data L). After calculating the similarity between the data points and the medoids, each data point is assigned to its nearest medoids. Next, new medoids, which are not identical to the previous medoids, are selected. The similarities between the data points and the new medoids are calculated again and the cost of each medoid is evaluated. These steps are repeated until the difference in the cost does not decrease. The clustering procedure with the compressed data L and PAM is summarized in Table II.

Table II. PAM clustering with the compressed data L obtained from the autoencoder

Step	PAM algorithm
I	Determine n_c
II	Randomly select n_c data points as medoids $U = \{u_1, u_2, \dots, u_{n_c}\}$
III	Calculate the similarities between L and U Assign L to its nearest medoid based on the similarities
IV	Select new medoids U_{new} Calculate the similarities between L and U_{new} Assign L to its nearest medoid based on the similarities
V	Calculate the cost[1)], C If $C_U - C_{U_{new}} < 0$, then U is replaced with U_{new}
VI	Repeat (IV) to (V) until the medoids do not change

4. Comparison of source term categorization

To confirm the effectiveness of the quantitative source term categorization, the characteristics grouped by the conventional and proposed method were compared in this section.

4.1 Source term characteristics grouped by the conventional source term category

As mentioned in Sec. 1, the conventional method employs logical trees or expert judgements to categorize a number of SA scenarios. For the OPR1000, total 17 source term categories were identified by the qualitative logical tree as shown in Fig. 3[1].

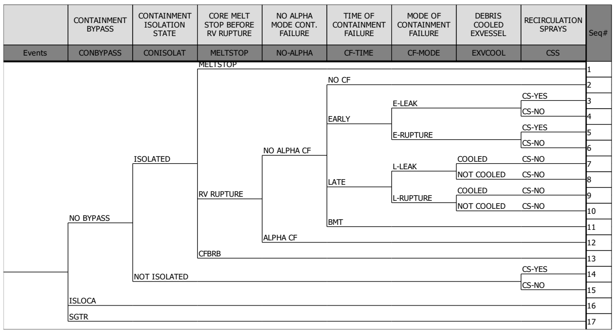

Fig. 3. Source term categories using the traditional approach [1, 2].

According to [1, 2], it was confirmed that the conventional categorization method caused significant differences in the release amount of important variable such as CsI within the same category. For example, Fig. 4 shows that the accumulated release fraction of CsI in the conventional source term category 17.

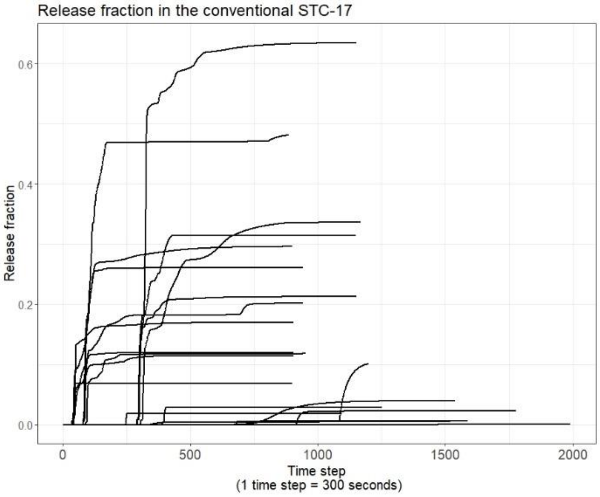

Fig. 4. Source term characteristics of CsI in the conventional source term category 17.

As shown in Fig. 4, source term characteristics are completely different even though they belong to the same category. Therefore, the conventional categorization method such as logical trees may contain large uncertainties.

4.2 Source term characteristics grouped by clustering with dimensionality reduction

This section describes the categorization results with the source term characteristics grouped by the proposed method supported by the exhaustive simulation. Fig. 5

[1] NOCF: no containment failure, ECF: early containment failure, LCF: late containment failure, BMT: basement melt-through, CFBRB: containment failure before reactor vessel breach, NOISO: containment isolation failure, BYPASS: containment bypass, ISLOCA: interfacing system loss of coolant accident (taken from [1, 2])

shows the accumulated release fraction of CsI in two categories.

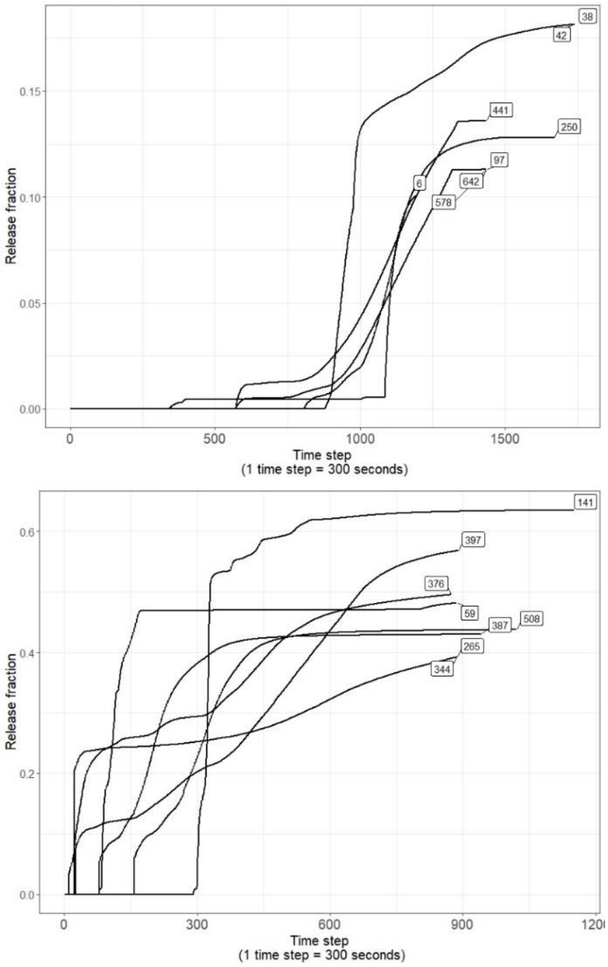

Fig. 5. Source term characteristics grouped by the proposed method: (upper) Category I, (lower) Category II

As shown in Fig. 5, similar source term behaviors can successfully be categorized by the proposed method. The initial release point in Category I is relatively late whereas the source terms in Category II released initially. Furthermore, the release amount of source terms can also be easily distinguished from other categories.

5. Conclusion

This paper proposed a quantitative source term categorization method based on two unsupervised learning with the source term database constructed by exhaustive simulation. The proposed method employed an autoencoder structure to reduce the data dimension and extract the key features from the time series data. In addition, three important variables, key contributors to accident consequences, were sorted out to reduce the size of the autoencoder. Finally, the severe accident scenarios were categorized by the PAM clustering method for the feature data compressed through the encoder, and it was confirmed that grouping by scenario with similar source term behavior was well performed.

In this paper, dimensionality reduction was performed only through the autoencoder. It is necessary to find a more optimized method by comparing with additional techniques such as PCA or discrete wavelet transform (DWT). Furthermore, the proposed method should be verified using the source term data for various NPPs and more results of accident consequence analysis.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, ICT, and Future Planning of the Republic of Korea and the National Research Foundation of Korea (NRF-2020M2C9A1061638).

REFERENCES

[1] J. Cho, S.H. Lee, Y.S. Bang, S. Lee, S.Y. Park, Exhaustive Simulation Approach for Severe Accident Risk in Nuclear Power Plants, Reliability Engineering & System Safety. (n.d.).

[2] J. Cho, S.H. Lee, Y.S. Bang, S. Lee, S.Y. Park, Exhaustive Simulation Approach for Severe Accident Scenarios, in: Transactions of the Korean Nuclear Society Virtual Autumn Meeting, 2021.

[3] J. Cho, S.H. Han, Identification of Risk-Significant Components in Nuclear Power Plants to Reduce Cs-137 Radioactive Risk, Reliability Engineering and System Safety. 211 (2021) 107613. https://doi.org/10.1016/j.ress.2021.107613.

[4] J.A. Hartigan, Clustering Algorithms, Wiley, 1975.

[5] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 2nd Edition, O'Rreilly Media, 2019.

[6] H. Zhang, T.B. Ho, Y. Zhang, M.-S. Lin, Unsupervised Feature Extraction for Time Series Clustering Using Orthogonal Wavelet Transform, Informatica. 30 (2006) 305–319.

[7] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2011.