Machine learning approach for approximation of thermal-hydraulic code using k-NN

Seunghyoung Ryu[a], Hyeonmin Kim[b], Seung Geun Kim[a], Jaehyun Cho[b]*
[a] Artificial Intelligence Application & Strategy Team
[b] Risk Assessment Research Team
Korea Atomic Energy Research Institute, 111, Daedeok-daero 989beon-gil, Daejeon, 34507, Korea

Highlights

- **Simulation of thermal-hydraulic (TH) dynamics via TH code run requires long computation time (e.g., minutes to hours)**
- **Rapid simulation is required for developments of a digital twin model, and one approach is data-driven model.**
- **In this study, we evaluate the use of *k*-nearest neighbor (*kNN*) model for the approximation of TH code results.**
- **Output is derived by averaging the results of *k* nearest input scenarios in training data.**
- **Magnitude of error varies according to the target variables; therefore, selection of proper measurement is important.**
- **The result implies that there exists outliers which cannot be well approximated via similar data.**

Problem Formulation

- **Modeling for data-driven approach**
  - $\mathrm{x} = [x_1, x_2, \ldots, x_d] \in \mathbb{R}^d$: Vector that defines the accident scenario, used as an input of TH code run.
  - $Y \in \mathbb{R}^{T \times D}$: Time series data for target output variables. T is the length of time series and D is the number of target variables.
  - $Y = f(\mathrm{x})$: TH code can be denoted as a function that takes x as an input and outputs time-series data Y from given scenario.
  - D_{train} : training data set contains pair of x and Y, $\{(\mathrm{x}_i, Y_i)\}$
  - D_{test} : test data set used for performance evaluation, $\{(\mathrm{x}_j, Y_j)\}$

- **Dataset configuration**
  - We obtain TH code run dataset by running multiple simulation with modular accident analysis program (MAAP) code.
  - Four accident scenarios are considered as follows.
  - TLOCC : Total loss of component cooling water (TLOCCW)
  - SLOCA-2 : Small loss of coolant accident (SLOCA) with diverse high-pressure pump injection and auxiliary pump injection
  - SLOCA-29 : SLOCA with changing severe accident guidance (SAG) 2 and 3
  - MLOCA : Medium loss of coolant accident (MLOC)
  - For each scenario, 2,000 sub-scenarios that having different safety system parameter is sampled from truncated normal distribution.
  - Final dataset contains 7,631 results.
  - Selected target output variables (MAAP code) are as follows.
  - PPS, PPSTRB(3), FREL(1), FREL(2)

K-nearest neighbor

- **Algorithm**
  - kNN is a well-known machine learning methodology that can be used for both regression and classification.
  - First, kNN requires a distance metric between two samples $d(x, y)$, and the number of nearest neighbors k.
  - Our goal is deriving fast approximation of Y_j in D_{test} from x_j and samples in D_{train}.
  - For all x_i from D_{train}, calculates $d(x_j, x_i)$
  - Find k nearest samples ($\{x_i\}$)
  - Calculate Y_j by averaging $\{Y_i\}$

- **Set-up**
  - Normalized input (standard scaler)
  - K-fold CV (5)
  - $k \in \{1,3,5,7,9\}$

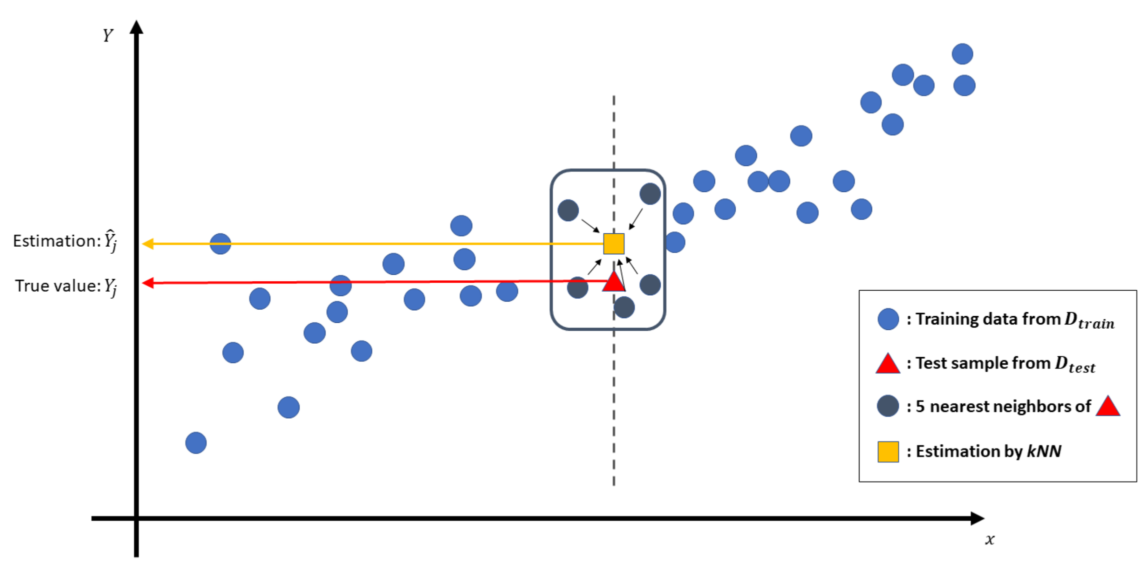

Figure 1. Illustration of using kNN (k=5) for regression problem

Experimental Results

- **Error metric**
  - Y_t : True value at time *t for sample*, $\hat{Y}_t$: Prediction of Y_t, $\bar{Y}$*: Average of* Y_t *for given sequence*
  - MSE (mean squared error) : $\frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2$
  - MAPE (mean absolute percentage error) : $\frac{1}{T}\sum_{t=1}^{T}((Y_t - \hat{Y}_t)^2 / Y_t) \times 100$
  - NAE (normalized absolute error) : $\frac{1}{T}\sum_{t=1}^{T}((Y_t - \hat{Y}_t)^2 / \bar{Y}) \times 100$
  - SMAPE (symmetric mean absolute percentage error) : $\frac{1}{T}\sum_{t=1}^{T}(|Y_t - \hat{Y}_t|)/(|Y_t| + |\hat{Y}_t|) \times 100$

- **Error results**
  - Results of FREL(1) and PPS with varying number of neighbors $k \in \{1,3,5,7,9\}$

k	MSE	NAE	MAPE	SMAPE
1	0.04	735k	17k	16.2
	1.7E-08	5.7	8.5	4.7
3	0.04	737k	17k	16.2
	1.7E-08	5.7	8.7	4.7
5	0.03	513k	12k	17.2
	2.5E-08	6.8	12.0	5.5
7	0.03	415k	12k	18.7
	3.5E-08	7.9	15.3	6.3
9	0.03	359k	12k	18.6
	4.1E-08	8.4	17.5	6.3

Table 1. Results of FREL(1); mean(gray), median(white)

k	MSE	NAE	MAPE	SMAPE
1	3.5E+12	47.4	80.6	15.2
	8.5E+11	28.8	27.3	10.3
3	3.4E+12	45.0	78.3	15.2
	7.4E+11	25.7	25.7	10.3
5	2.5E+12	44.9	76.5	15.1
	8.1E+11	26.1	30.2	10.1
7	2.2E+12	44.7	75.8	15.3
	8.0E+11	26.6	33.8	10.5
9	2.1E+12	44.8	75.8	15.4
	7.7E+11	27.0	34.2	10.6

Table 2. Results of PPS; mean(gray), median(white)

Examples of kNN results

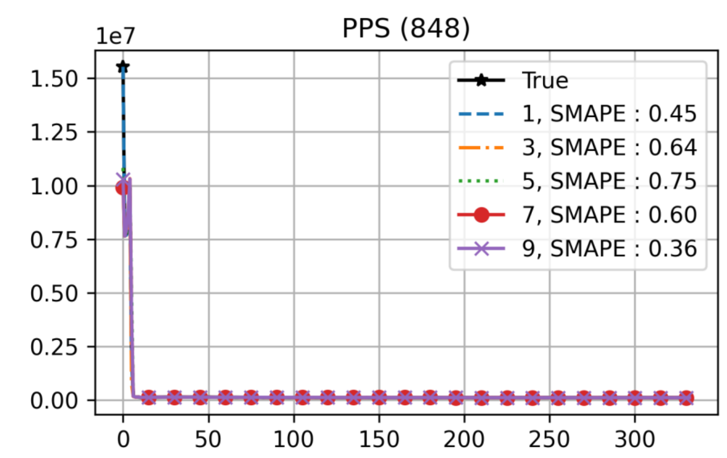

Figure 2. Results of PPS (good case)

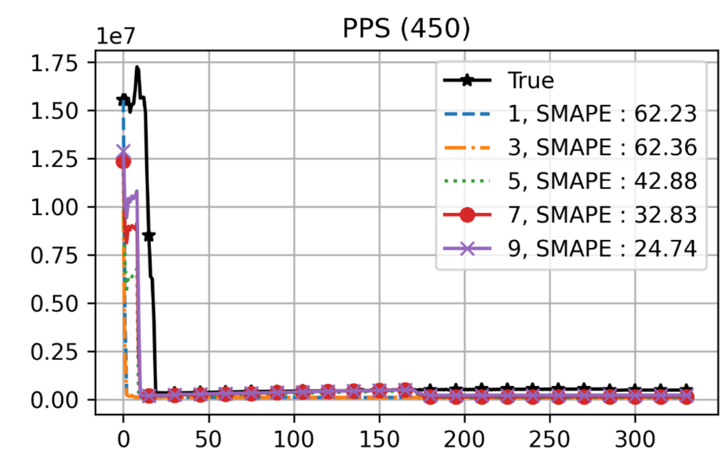

Figure 3. Results of PPS (bad case)

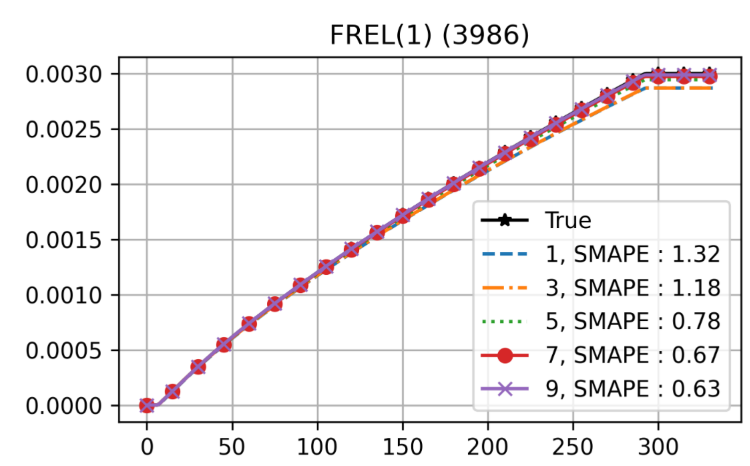

Figure 4. Results of FREL(1) (good case)

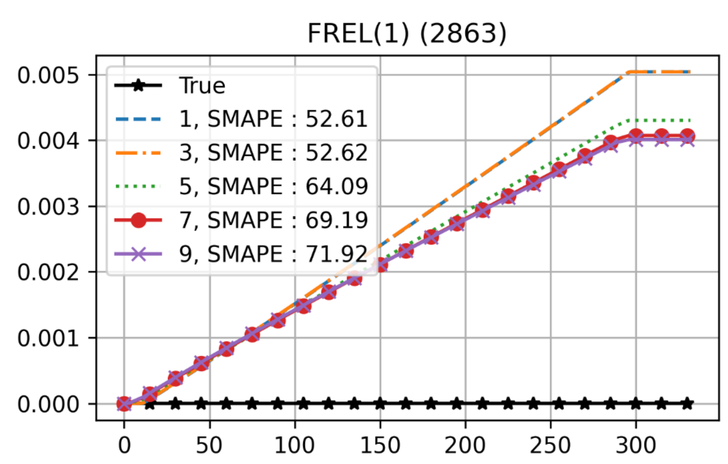

Figure 5. Results of FREL(1) (bad case)

Conclusion

- **We analyze the result of utilizing *k-NN* algorithm for the approximation of TH code results.**
- **Even if input vectors are similar (small distance), their outputs can be very different and may induce excessive prediction error with *kNN*.**
- **More advanced method will be required to modeling intrinsic dynamics from data.**
- **It becomes difficult to apply *kNN* when the number of data increases.**

Acknowledgement

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).

Reference

[1] Kim, Hyeonmin, Jaehyun Cho, and Jinkyun Park. "Application of a deep learning technique to the development of a fast accident scenario identifier." IEEE Access 8 (2020): 177363-177373.
[2] Ryu, Seunghyoung, et al. "Probabilistic deep learning model as a tool for supporting the fast simulation of a thermal–hydraulic code." Expert Systems with Applications 200 (2022): 116966.
[3] Fix, Evelyn, and Joseph Lawson Hodges. "Discriminatory analysis. Nonparametric discrimination: Consistency properties." International Statistical Review/Revue Internationale de Statistique 57.3 (1989): 238-247.