機械学習を活用した熱源・空調システムのモデリング手法に関する研究
（第 3 報）モデルの適用領域に関する考察

尾　形　　甫（新菱冷熱工業）　　福　井　雅　英（新菱冷熱工業）
矢　島　和　樹（新菱冷熱工業）　　金　子　友　昭（新菱冷熱工業）

は じ め に

機械学習や深層学習は、大量のデータを高速に学習し、データの中に潜むパターンを認識することが可能で、データ駆動によるタスクの自動化や意思決定への活用が期待される。熱源・空調分野においては、最適制御や不具合検知への活用が進められている [1) 2)]。既報では、BEMS（Building Energy Management System）に蓄積されるデータの質や量を考慮したターボ冷凍機のモデルについて検討を行った。

一般に、深層学習や機械学習の汎化性能を高めるには、できる限り多くの事象をもつデータセットでモデルを訓練する必要がある [3)]。学習データセットが網羅していない領域、いわゆる外挿領域では、予測値の不確実性が高まる。そのため、モデルを適切に利用するには、予測値が信頼できる領域を把握することが重要である。しかし、深層学習や機械学習で取り扱う現実の問題は、説明変数が多次元で複雑な問題が多いことから、サンプルが問題を適切に代表しているかを判断することは容易でない。機械学習によるモデルが信頼できる予測を行うことができる領域は Applicability Domain（以下、AD とする）と呼ばれ、データ密度に基づき判断する手法が提案されている [4)]。例えば、特徴量空間における学習データセットのサンプル密度を求め、一定数以上サンプルが密集している領域では、モデルが十分に訓練されており、推論性能が発揮できると判断する。

化学分野においては、多くの先行研究が見受けられるが、熱源機を対象とした事例はない。本報では、機械学習による熱源機モデルの適切な運用方法を確立するため、AD の決定方法について検討した。

1. 検討対象

1.1 対象機

A プラントで稼働している冷凍容量 11.4 GJ/h（900 USRT）のインバータターボ冷凍機を対象とした。機器仕様を**表－1** に示す。冷水の定格値は、流量が 453.6 m^3/h、出口温度が 6.0 ℃である。冷却水の定格値は、流量が 635.5 m^3/h、入口温度が 32.0 ℃である。

表－1　機器仕様

項目	定格値
冷凍能力	11.4 GJ/h（900 USRT）
冷水出入口温度	12.0 ℃ → 6.0 ℃
冷水流量	453.6 m^3/h
冷却水出入口温度	32.0 ℃→37.0 ℃
冷却水流量	635.5 m^3/h
主電動機消費電力	499.0 kW
定格 COP	6.3

1.2 モデル

AD を検討するモデルは、前報 [5)] と同様にニューラルネットワークで作成した。説明変数は、冷水出口温度、冷水流量、冷却水入口温度、冷却水流量、冷水出入口温度差の 5 変数とし、目的変数は消費電力とした。

1.3 サンプリングデータ

サンプルは、A プラントにおける 1 時間ごとの計測値で、期間は 32 ヵ月分である。データのスクリーニングとして冷凍機が停止中のサンプルを全て、冷凍機が起動した時刻のサンプルを 1 時間分、冷凍機が停止した時刻のサンプルを 1 時間分除去した。サンプルの合計は、22,930 点である。

2. 検討方法

2.1 概要

サンプルのうち、24 ヵ月分を学習データセット、8 ヵ月分をテストデータセットに使用した。さらに学習データセットからは､冷却水入口温度が 16.0 ℃未満のサンプルを全て削除した。データセットの内訳を表－2 に示す。テストデータセットのうち AD 外となるサンプルは、冷却水入口温度が 16.0 ℃未満に該当するサンプルである。複数のアルゴリズムを用い、学習データセットを基準にテストデータセットの各サンプルを AD 内外のいずれかに分類し、検出能力の評価を行う。

2.2 データ間の距離関数

特徴量空間におけるサンプル群の密度を推定するには、サンプル間の距離を求める距離関数が重要である。本報では、ユークリッド距離を用いた。2 点 AB 間のユークリッド距離 d_{AB} は、A 点の座標を（a_1，a_2，　…，a_n）、B 点の座標を（b_1，b_2，　…，b_n）とすると、式（1）のように表される。

$$d_{AB} = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad \cdots(1)$$

ここに、

d_{AB}	: A 点と B 点のユークリッド距離	[-]
a_n	: A 点のn次座標	[-]
b_n	: B 点のn次座標	[-]
n	: 次元数	[-]

2.3 アルゴリズム

検出には、k 近傍法（k-Nearest Neighbors：以下、k-NN とする）、One Class Support Vector Machine（以下、OCSVM とする）、Density Based Spatial Clustering of Applications with Noise（以下、DBSCAN）、Isolation Forest を用いた。k-NN は、最も近い k 個のデータを抽出するアルゴリズムである。本検討では、k＝10 とした。k-NN で各テストデータから最も近い学習データを 10 点求め、異なる指標を用いた 3 種類の検出器を作成した。第一の検出器は、10 点の積算距離を指標とする分類である（**図－1** 中(1)）。積算距離が小さいほど、テストサンプルの近辺に学習サンプルが密集しており、十分に学習ができていると判別する。パラメータとして、学習したサンプルからの外れ値と判別する積算距離 L_s を設定する。積算距離は、式(2)で表される。

$$d_{sum} = \sum_{k=1}^{10} d_k \quad \cdots(2)$$

ここに、

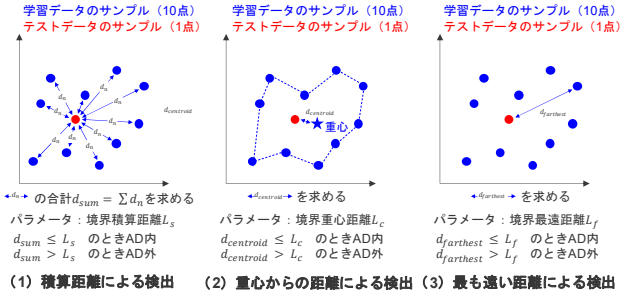
d_{sum}	: 10 点の積算距離	[-]
d_k	: k点とのユークリッド距離	[-]

第二の検出器は、10 点の重心との距離を指標とする分類である（**図－1** 中(2)）。重心からの距離が小さいほど、テストサンプルの付近に均一に学習サンプルが存在しており、十分に学習できていると判別する。パラメータとして、学習サンプルからの外れ値と判別する重心からの距離 L_c を設定する。10 点の重心は、各座標を平均して求めた。

第三の検出器は、10 点のうち最も遠い距離を指標とする分類である（**図－1** 中（3））。第一の検出器と比較し、バラつきを許容せず、より厳密にテストサンプルの付近に学習サンプルが存在することを判別する。パラメータとして、学習サンプルからの外れ値と判別する最大距離 L_f を設定する。

表－2　データセットの内訳

	学習	テスト
サンプル数	10,124	5,675
負荷率範囲	全て	全て
冷却水流量範囲	全て	全て
冷却水入口温度範囲	16～32 ℃	全て



図－1　k-NN による検出の概略図

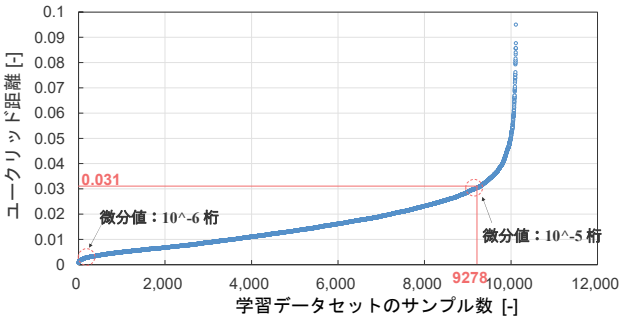
OCSVM は、教師なし学習による分類手法で、データセット中の各サンプルが 1 つのラージクラスに属するか否かを判別する。属さないサンプルは、外れ値と捉えることができる。主なハイパーパラメータは *nu* であり、サンプル数に対する外れ値の割合を指定する。

DBSCAN は、サンプルの密度に基づくクラスタリング手法で、データセット中の各サンプルが構成するクラスタとクラスタに至らないノイズを判別する。主なハイパーパラメータは *eps* と *min samples* で、密集とみなす距離とクラスタと見做すサンプル数を指定する。*min samples* は *k* と同様に 10 とした。

Isolation Forest は、教師なし学習による分類手法で、各データが孤立するまで決定木で分岐を行い、孤立するまでの分岐数が少ないと異常な値であると判別する。主なハイパーパラメータは *contamination* であり、サンプル数に対する外れ値の割合を指定する。

2.4.　パラメータ

検出器のパラメータは距離と割合の二種類に大別され、次の手順により決定した。まず学習データセットに含まれる全サンプルに対し、最も近いサンプルとのユークリッド距離を求める。結果を昇順に並び替え、**図－2** に示す。9,000 点から 10,000 点付近にかけ、急激に増加していることが確認できる。急激な増加前は、学習データセットに含まれるサンプル同士の距離が近く、密度が高いと考えられる。一方で、急激な増加以降は密度が低いと考えられる。増加点は微分値により判断し、桁数が 10^{-6} から 10^{-5} に増加した 9278 点目、距離 0.031 を境界とした。距離が 0.031 を超えるサンプル数は、846 点であり、学習データセットに含まれる全サンプルの 8.4 %にあたる。よって、距離に関するパラメータには 0.031、割合に関するパラメータには 0.084 を採用する。各アルゴリズムに設定したパラメータを**表－3** に示す。



図－2　サンプリングデータ間の距離

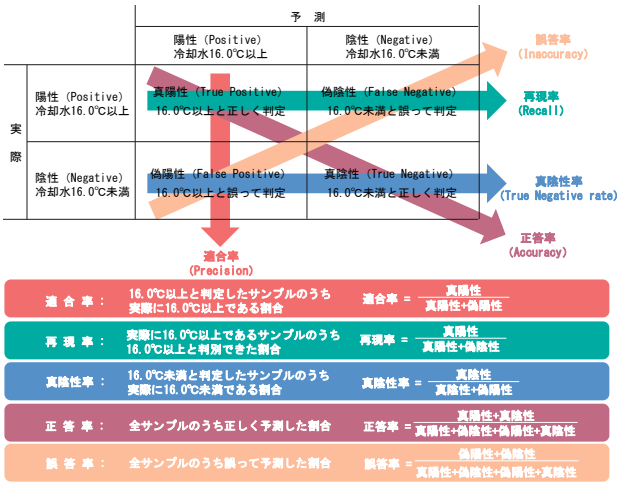
表－3　各アルゴリズムの設定パラメータ

アルゴリズム		パラメータ
OCSVM		$nu = 0.084$
DBSCAN		$eps = 0.031$
Isolation Forest		$contamination = 0.084$
k-NN	積算距離	$L_s = 0.031 * 5$
	重心距離	$L_c = 0.031$
	最遠距離	$L_f = 0.031$

3.　結果および考察

各検出器の性能は、混同行列を用いて評価した。混同行列の概略を**図－3** に示す。本報では、テストデータセットのうち、冷却水入口温度が 16.0 ℃以上のサンプルを陽性、16.0 ℃未満のサンプルを陰性として表現する。検出器により、正しく陽性と判断したサンプルを真陽性、誤って陽性と判断したサンプルを偽陽性とカウントする。同様に、正しく陰性と判断したサンプルを真陰性、誤って陰性と判断したサンプルを偽陰性としてカウントする。**表－4** に各アルゴリズムによる検出結果を示す。DBSCAN、k-NN（積算）、k-NN（最遠）は、真陰性が多く、真陽性が少ない。Isolation Forest は、真陽性が多く、真陰性が少ない。OCSVM と k-NN（重心）は、真陽性と真陰性がいずれも多く総合的に検出率が優れている。1 サンプルあたりの AD の判定にかかる計算時間は、OCSVM が最も短く、Isolation Forest、DBSCAN、k-NN の順番で続く。

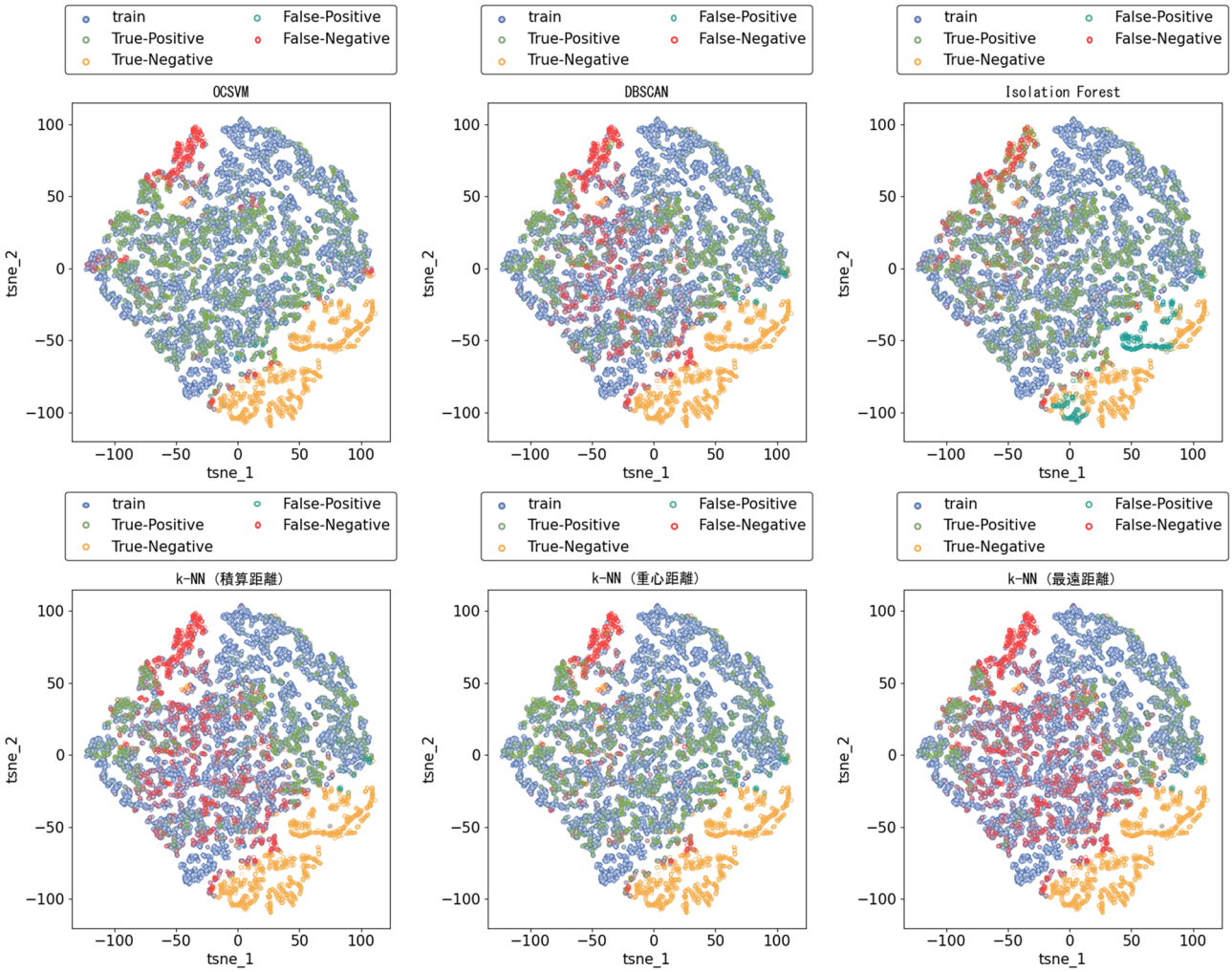
次に、検出の特徴を把握するため、t 分布型確率的近傍埋め込み法（t-distributed Stochastic Neighbor Embedding）を用いて説明変数の次元を削減し、可視化を行った。**図－4** に可視化の結果を示す。図中の右下部分は、テストデータのサンプルが密集しているが、学習データのサンプルは存在していない。冷却水温度が 16.0 ℃未満のクラスであると推察される。検出数の多少はあるものの、いずれの検出器においても陰性と検出している。図中の左上部分は、学習データのサンプルが存在するため、冷却水温度 16.0 ℃以上のクラスである。しかし検出数の多少はあるものの、いずれの検出器においても陰性と検出している。$-75 \leq tsne\_1 \leq -25$ かつ $70 \leq tsne\_2 \leq 110$ における学習データのサンプリング数をカウントすると 36 点であった（テストは 359 点）。これはデータセット作成時には、想定していなかった学習データセットの疎状態であるが、検出器により適切に検出できていることが判明した。DBSCAN、k-NN（積算）、k-NN（最遠）は、いずれも中央付近に偽陰性のサンプルがあり、真陽性の検出率が低いことが確認できる。Isolation Forest は、右下部分に偽陽性のサンプルがあり、真陰性の検出率が低いことが確認できる。OCSVM と k-NN（重心）は、陽性と陰性が適切に検出されていることから、AD の決定能力に優れていると推察される。なお、k-NN（重心）は、検出率が優れているが AD の判定にかかる計算時間が長いため、最適化など計算を繰り返す用途には、不向きである。



図－3　混同行列の概略

表－4　検出結果

	OCSVM	DBSCAN	Isolation Forest	k-NN（積算距離）	k-NN（重心距離）	k-NN（最遠距離）
真陽性数	2,725	2,045	2,833	1,571	2,706	1,166
偽陽性数	58	67	679	47	70	37
真陰性数	2,161	2,152	1,540	2,172	2,149	2,182
偽陰性数	731	1,411	623	1,885	750	2,290
適合率	0.98	0.97	0.81	0.97	0.97	0.97
再現率	0.79	0.59	0.82	0.45	0.78	0.34
真陰性率	0.75	0.60	0.71	0.54	0.74	0.49
正答率	0.86	0.74	0.77	0.66	0.86	0.59
誤答率	0.14	0.26	0.23	0.34	0.14	0.41
計算時間	< 1 ms	280 ms	21.8 ms	7,380 ms	13,800 ms	7,420 ms



図－4　可視化の結果

お わ り に

本報では、インバータターボ冷凍機の消費電力予測モデルを対象に AD の決定方法を検討し、以下の知見を得た。

1） OCSVM と k-NN（重心）で AD を決定すると検出率が高いと。

2） 用途により、AD の判別に要する計算時間を考慮する必要がある。

今後は、他のデータセットや機械について検討を進める。

謝　　辞

本研究を進めるにあたり、丸の内熱供給株式会社に運転データをご提供いただいた。ここに謝意を表する。

参 考 文 献

1) 矢崎, 他 (2020-2022). 都市型地域冷暖房の省エネルギー手法に関する研究（その 3～その 5）. 空気調和・衛生工学会大会学術講演論文集.

2) 宮田, 他 (2018-2022). 機械学習を用いた空調熱源システムの 不具合検知・診断（第 1 報～第 3 報）. 空気調和・衛生工学会論文集.

3) Aurelien Geron. (2020). scikit-learn,Keras,TensorFlow による実践機械学習. オライリー・ジャパン.pp.26－27

4) Dragos Horvath et al. (2009). Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. Journal of Chemical Information and Modeling.

5) 尾形, 他（2023）. 機械学習を活用した熱源・空調システムのモデリング手法に関する研究（第 1 報）. 空気・調和衛生工学会大会学術講演論文集