CF-004

AI 運用環境の変化に対する汎用的検知手法および要因分析手法の提案

梁宇昕† 恵木正史†

†(株)日立製作所 研究開発グループ

1. はじめに

ビジネス適用が進む機械学習モデル（本報告では AI と同義とする）の保守・運用を効率化するため，AI 固有の環境変化による精度劣化を検知する技術が求められている。環境変化とは，運用時の AI に入力されるデータの傾向が，学習時から漸進的に変化する現象を指す。学習時と運用時とでデータの傾向が異なってくると，当然ながら AI の予測精度は劣化する [1]。対策の遅れは事業損失に直結する。

しかし，従来の環境変化検知手法はいずれも，下記に示すように適用できる AI に制限があり，その拡大が課題となっていた。

① 識別問題向けの機械学習モデルに限定される
② 機械学習モデルの種別が限定される
③ 入力特徴量が連続値に限定される

そこで，本報告では XAI(eXplainable AI)技術[2]を活用した拡大手法を提案する。本手法により上記の②③の制約をなくすことができる。

2. 従来研究

既知の識別クラスとは違う傾向を持つ観測データを未知クラスとして識別することで環境変化検知する従来技術に，Extreme Value Machine(EVM) [3]と OpenMax[4]がある。

2.1 従来技術 1: Extreme Value Machine

以下では，データ$\boldsymbol{x} \in \mathbb{R}^m$を$L$個のクラスに分類する他クラス分類問題を扱う。$N$個の学習データを$\boldsymbol{x}_i, i = 1, \cdots, N$とし，このうちクラス$j$に属する学習データのインデックスの集合を$C_j$とする。

EVM では，データ$\boldsymbol{x}$がクラスjに属する尤度$P(j|\boldsymbol{x})$は，次式のような Weibull 型の関数で記述される。

$$P(j|\boldsymbol{x}) = \mathrm{argmax}_{i \in C_j} \exp\left\{-\left(\frac{||\boldsymbol{x}-\boldsymbol{x}_i||}{\lambda_i}\right)^{k_i}\right\} \quad (1)$$

また，データ$\boldsymbol{x}$が属するクラスy^*は次式のように，尤度が閾値δを超えるクラスがあればその最大値を与えるクラスと推定し，各クラスの尤度がいずれもδ未満であった場合には，未知のクラスと推定する。すなわち，EVM では未知のクラスの出現をもって環境変化と見なす。

$$y^* = \begin{cases} \mathrm{argmax}_{j \in \{1,\cdots,L\}} P(j|\boldsymbol{x}) & \text{for } \exists j\, P(j|\boldsymbol{x}) \geq \delta \\ \text{unknown class} & \text{otherwise} \end{cases} \quad (2)$$

2.2 従来技術 2: OpenMax

本節では OpenMax について概略を説明する。OpenMax は画像識別用の畳み込みニューラルネットワーク(CNN) [5]に環境変化検知機能を組み込んだ AI であり，入力された観測データに対し 2 種の情報を出力する。1 つは観測データの識別クラスの推定結果であり，もう 1 つは観測データが未知のクラスに属する尤度である。OpenMax の尤度計算手法は EVM と類似しており，観測データが各既存クラスのデータ群に所属する尤度を Weibull 関数に基づいて推定し，各クラスの尤度がいずれも閾値未満であった場合には，未知のクラスと推定する。EVM では特徴量に基づいて尤度を計算するが，OpenMax では CNN の中間層のデータに基づいて尤度を算出する。

画像データに対して EVM を適用しようとすると，ピクセル空間上での多クラス分類問題として扱う必要があり、CNN のように高次の特徴量の恩恵を受けることができないため，認識精度は高くはならない。そこで，OpenMax では CNN 内で画像データが入力層から出力層に向かって徐々に変換・加工され，抽象度が増していく過程に着目し，そのような中間層のデータに対して EVM と同等の処理を使うことで，上記問題を回避することに成功している。

3. 課題

3.1 課題 1: 環境変化検知技術の適用可能範囲拡張

環境変化の検知技術はできるだけ幅広い AI に適用できることが望ましい。しかし，従来手法はそれぞれ適用範囲が限定される問題点がある。

EVM は，連続的な特徴量での教師データの分布を想定し、分布の裾野、すなわち尤度が著しく低い事象の発生に基づいて環境変化を検知する。したがって、特徴量に男／女などカテゴリカル変数が含まれる場面では EVM は利用できない。一方 OpenMax は，CNN 特有の中間層のデータを用いるため，SVM[6]や勾配ブースティング[7]などの中間層のデータが存在しない AI には適用できない。

上記のような制限を緩和し，幅広い特徴量，幅広い AI に適用するために，EVM を拡張することが課題となる。

3.2 課題 2: 検知後の要因分析の実現

環境変化を検知すると，運用者は次のステップとしてどのような環境変化が生じたのか分析する。分析の観点は多種多様であるが，本報告では分析の初手として，環境変化が，入力データのどのような特徴に現われるかを把握すること，と捉える。そこで報告では，検知した環境変化を特徴づける，入力データの特徴を抽出することを課題とする。

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4. 提案

本報告では EVM をカテゴリカル変数に適用可能にするため、特徴量空間ではなく、後述する根拠空間で各既存クラスのデータ群が属する分布を記述することを提案する。提案手法では XAI 技術を活用し，特徴量ベクトルを AI の予測根拠となる根拠ベクトルに変換する。

このようにすることで、もともとカテゴリカルな特徴量であっても、根拠空間では貢献度という連続的な値に変換されるため、EVM を適用することが可能になる。これにより，課題 1 を解決する環境変化検知手法が実現する。

さらに，環境変化の検知を根拠空間で行っているため、変化要因を一般人が理解可能な形で数値化できる利点がある。これにより課題 2 を解決する要因分析手法が実現する。

根拠ベクトルを用いた環境変化検知手法を提案 1 で，追加分析による変化要因の数値化手法を提案 2 で述べる。

4.1 提案 1: 環境変化検知技術

本報告では Shapley 値[8]を根拠ベクトルに用いた尤度計算法を提案する。Shapley 値は協力ゲーム理論にて，複数のプレイヤが協力して得る利得を，各プレイヤの貢献度として計算し，公正分配する方法である。Shapley 値を XAI 技術に提供した応用例として，利得に AI の出力値，プレイヤに各特徴量をあてはめ，各特徴量が AI の出力値に与えた貢献度を計算する手法として SHAP [9]がある。

Shapley 値は連続値を取る数字ベクトルである。もともとカテゴリカルな特徴量であっても，それを写像した Shapley 値は裾野を持つ連続的な分布を取り、環境変化検知を行える。

また，Shapley 値の貢献度の単位は出力値の単位に共通化される。例えば，AI の出力値が尤度 (無単位)であるなら，全ての特徴量に対する貢献度もまた尤度(無単位)になる。そのため，次元が違う特徴量を同列に扱える。

さらに，AI の出力に影響を与えない特徴量の値が変動しても，その特徴量に対応する Shapley 値は変動しない特性を持つ。そのため，保守運用する AI に影響を与える環境変化を検知するためには適する。

SHAP を用いた Shapley 値の計算式を式(3)に記す。

$$\varphi_h(v) = \sum_{S \subseteq M\{h\}} \frac{|S|!\,(|M|-|S|-1)!}{|M|!}\big(v(S \cup \{h\}) - v(S)\big) \quad (3)$$

Shapley 値は，特徴量に対しそれぞれ計算される。

式(3)において，vは Shapley 値の計算対象とする AI が予測に用いる出力関数である。$\varphi_h(v_j)$は h 番目の特徴量に対し計算された貢献度, Sはプレイヤの提携，すなわち特徴量の部分集合を示す。$v(S \cup \{h\})$は提携に h 番目の特徴量を含めた場合の AI 出力値の期待値, $v(S)$は提携に h 番目の特徴量を含めない場合の AI 出力値の期待値を示す。

Shapley 値を用いた環境変化検知手法に関する数式を以下に記す。$\varphi(v)$は観測データの Shapley 値，$\varphi_i(v)$は各学習データの Shapley 値を示す。

$$P(j|\varphi(v)) = argmax_{\,i \in C_j} \exp^{-\left(\frac{||\varphi(v)-\varphi_i(v)||}{\lambda_i}\right)^{k_i}} \quad (4)$$

$$y^* = \begin{cases} argmax_{\,j \in \{1:L\}} P(j\,|x) & if\ P(j\,|x) \geq\ \delta \\ \text{unknown class} & \text{Otherwise} \end{cases} \quad (5)$$

Algorithm 1 環境変化検知アルゴリズム

Input: 学習データセット N, 観測データ x, 新規クラス閾値 δ, 出力関数 v
Output: 観測データ x の推定クラス y^*

```
maxlikelihood = δ
for each i in N do
    Shapley 値 φ(v) を式 (3) で計算
end for
L = N に含まれる既知クラスの数
for j = 1 to L do
    (P(j| φ(v)) を式 (4) で計算
    if (P(j| φ(v)) > maxlikelihood) then
        y* = j
        maxlikelihood = (P(j| φ(v))
    end if
end for
if maxlikelihood == δ then
    y* = unknown class
end if
return y*
```

図 1. 環境変化検知アルゴリズム

図 1 は提案する環境変化検知技術のアルゴリズムを示す。アルゴリズムではまず、業務環境から得た観測データ，AI, および学習データを入力として, SHAP を用いて Shapley 値を計算する。その後，Shapley 値を入力として式(4)，式(5)で未知クラスの出現を環境変化として検知する。

4.2 提案 2: 環境変化要因分析技術

環境変化を検知すると，運用者は次にどのような環境変化が生じたのか分析する。この時，未知クラスとして検出された観測データについて，検知した環境変化を特徴づける，入力データの特徴を示す貢献度の情報を運用者に提供することで，対策立案を効率化する手法を提案する。

図 2 は貢献度の情報を保守運用担当者に提供する効果を示したイメージ図である。ここでは 1 例として，化学プラントの地点 A，地点 B などに設置されたセンサの情報から，プラントの動作モードを予測する問題を考える。地点 A, 地点 B にはそれぞれ温湿度計が設置されており，それらの情報から観測データの特徴量を生成する。動作モード予測 AI はこの特徴量の情報から，プラントの動作モードを予測する。提案 1 では，この観測データに未知クラスのデータが出現しているかを評価した。提案 2 では，未知クラスに分類された観測データの貢献度を分析・表示している。簡単のため，貢献度の値をバーの色と長さで表現する。長いバーは貢献度の絶対値が大きいことを指し，色は赤色が正の値，青色が負の値を指す。

この例では環境変化で地点 B の温湿度計センサが壊れる，気温 42(℃)，湿度 0(%)という未知の観測データが出現する問題を考える。

図 2 の上部に示される環境変化前の状態では，観測データは既知クラスの動作モードに識別されるため，未知クラスに識別される尤度の値自体が低くなる。よって，その尤度の値を分配した貢献度の値も小さく，環境変化が起きていない度合いを視覚的に表現できる。

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

一方，図 2 の下部に示される環境変化後の状態では，観測データが未知クラスに識別される尤度の値が大きくなり，環境変化が起きている度合いを，運用者が視覚的に理解できるようになる。さらに，貢献度が正の方向に大きい特徴量は，未知クラスに識別する尤度を上げる要因となった特徴量であり，点検する優先度が高いとわかるため，「気温_地点 B」が 42(℃)であることと「湿度_地点 B」が 0(%)であることが，未知クラスの検知に効いていることも視覚的に理解できる。これらの情報により，異常の情報を知った保守運用担当者は，「地点 B の温湿度計が壊れていないか確認する」という対策に自律的に移ることができる。

上記の貢献度を計算するためには，AI が未知クラスとなる観測データを特徴量に基づき識別できることを前提としている。しかし，提案 1 で未知クラスの出現を検知した段階では，検知の対象とした AI が未知クラスを識別できない。そのため，既存クラスと未知クラスの識別を行う予測器である AI*を作成する必要がある。AI*は環境変化検の知を受け自動生成される。

図 3 は提案する要因分析技術のアルゴリズムを示す。AI*の再学習には，観測データに正解値を付与する必要があり，半自動化を目指す本研究の趣旨から外れてしまう。そこで，本研究では提案 1 の尤度計算において，式(5)で得たクラス分類の結果である y^* を，観測データの正解値とすることで AI*を作成する。そして，未知クラスのデータを AI^* が未知クラスと識別した時の出力値を Shapley 値で分配し，各特徴量の貢献度として計算する。

5. 評価

5.1 実験手法

化学プラントの状態を示すセンサデータから故障モードや動作モードを識別する AI を作り，新種故障モードの出現検知，および検知要因となった特徴量の特定を行った。

実験の評価には，マサチューセッツ工科大学の提供する Tennessee Eastman Process(TEP)のデータ[10]を利用した。マサチューセッツ工科大学の TEP データは，化学プラントのプロセスシミュレータを様々な条件下で実行した時に得たシミュレーションデータである。シミュレーションは正常モード 1 種と故障モード 21 種についてそれぞれ行われ，プラント含まれる装置の状態を示す 52 次元の特徴量のデータを計 960 点分シミュレートする。

TEP データは，先行研究により一部の故障モードと特徴量の依存関係が解明されている[11]。表 1 は依存関係が解明されている故障モードの一部である。例えば，「故障 6」は，原料 A の供給装置が壊れることを指す。そのため，原料 A 供給量を示す特徴量「F1」の値が極端に低くなることが「故障 6」であることと相関すると知られている。
そのため，先行研究では特徴量「F1」の値が 0.077 以下になる場合に，「故障 6」と正常モードが識別できると知られている。一方，「故障 2」と相関する特徴量は「MV4」「F10」「XB」「YB」「YF」と複数存在し，比較的検知が難しい。

本報告では，今回は 52 次元の特徴量を持つデータから，そのデータの故障モードを多クラス識別する AI を作成し，未知の故障モードの出現を環境変化として検知・分析した時の精度を評価した。

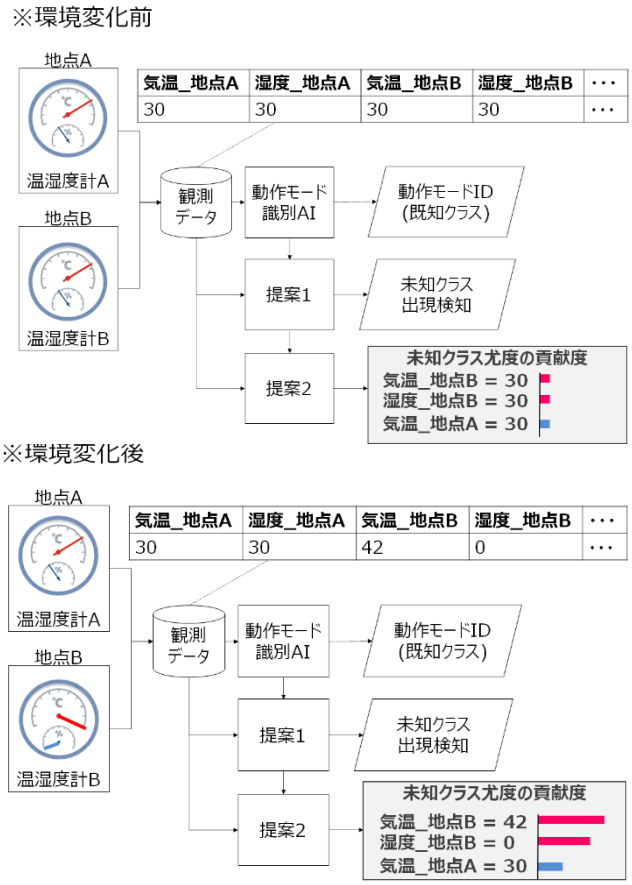

図 2. 要因分析技術適用時のイメージ図

Algorithm 2 環境変化要因分析アルゴリズム

Input: 学習データセット N, 観測データセット O, 予測関数 v, アルゴリズム 1 による観測データセット O の推定クラス y^*

Output: 貢献度ベクトル Q

1: **for** $i = 0$ to $sizeof(y^*)$ **do**
2: $O[i]$ に対し $y^*[i]$ を正解として付与する
3: **end for**
4: v を O を用いて再学習し更新する
5: Q = []
6: **for each** data in O **do**
7: Shapley 値 $\varphi(v)$ を式 (3) で計算
8: Q.append($\varphi(v)$)
9: **end for**
10: **return** Q

図 3. 環境変化要因分析アルゴリズム

表 1: TEP データの故障モードと特徴量の依存関係[11]

故障モード	依存関係
故障 2	(MV4 > 60.48) & (F10 > 0.21) & (XB > 15.07) & (YB > 21.89) & (YF < 5.41)
故障 6	(F1 < 0.077)

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

具体的には，意図的に 1 種の故障モードの学習データを AI の学習データセットから除外し，その除外した故障モードのデータと正常モードのデータを観測データとして AI に入力した時に，以下をそれぞれ評価した。

1) 環境変化検知精度: 提案 1 で除外した故障モードのデータを未知クラスとして識別する精度
2) 要因分析精度: 提案 2 で計算した貢献度において，貢献度が最大の特徴量が先行研究[11]の示す依存関係のある特徴量と一致する率

5.2 実験結果と考察

表 2 は環境変化検知精度の評価結果である。提案手法は Precision，Recall 共に 0.9 程度あり，90%の精度で環境変化を検出したと言える。この結果は，従来手法である EVM と遜色のない検知精度である。

表 3 は，要因分析精度の評価結果である。提案 2 で計算した貢献度は，故障モードと依存する特徴量を 99%以上の精度で正しく推定できており、実用的な情報を抽出できていると考えられる。

以上より，環境変化を 90%の精度で検知しており、且つ要因分析のための情報を抽出できていることから，提案手法は実用性があると考察する。

6. おわりに

ビジネス適用が進む AI の保守・運用を効率化するため，AI 固有の環境変化による精度劣化を検知する技術が求められている。環境変化とは，運用時の AI に入力されるデータの傾向が，学習時から漸進的に変化する現象を指し，対策の遅れは事業損失に直結する。

従来の環境変化検知手法はいずれも，下記に示すように適用できる AI に制限がありその拡大が課題となっていた。

①識別問題向けの機械学習モデルに限定される
②機械学習モデルの種別が限定される
③入力特徴量が連続値に限定される

そこで，本報告では XAI 技術を活用した拡大手法を提案し，本手法により上記の②③の制約をなくした。

提案手法を化学プラントの故障識別 AI に適用したところ，90%の精度で環境変化を検出し、環境変化の要因となった特徴量を特定した。これにより，提案手法の妥当性が示された。

表 2. 環境変化検知技術の精度評価結果

故障モード		提案 1	従来手法
故障 2	Precision	0.966	0.886
	Recall	0.986	0.991
故障 6	Precision	0.938	0.938
	Recall	0.890	0.957

表 3. 要因分析技術の精度評価結果

故障モード	提案 2
故障 2	99.1%
故障 6	99.9%

参考文献

[1] Žliobaitė, Indrė. "Learning under concept drift: an overview." arXiv preprint arXiv:1010.4784 (2010).
[2] Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017).
[3] Rudd, Ethan M., et al. "The extreme value machine." IEEE transactions on pattern analysis and machine intelligence 40.3 (2017)
[4] Bendale, Abhijit, and Terrance E. Boult. "Towards open set deep networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Lawrence, Steve, et al. "Face recognition: A convolutional neural-network approach." IEEE transactions on neural networks 8.1 (1997)
[6] Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.
[7] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
[8] Winter, Eyal. "The shapley value." Handbook of game theory with economic applications 3.2 (2002)
[9] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems. 2017.
[10] "Tennessee Eastman Problem Simulation Data" http://web.mit.edu/braatzgroup/links.html 2020/3.27 現在
[11] Ragab, Ahmed, et al. "Fault detection and diagnosis in the Tennessee Eastman Process using interpretable knowledge discovery." 2017 Annual Reliability and Maintainability Symposium (RAMS). IEEE, 2017.

Copyright © 2020 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.