F-005

仮想標本を用いた AI 予測根拠の校正手法に関する検討
Study on explanation correction method using augmented samples

濱本 真生 †
Masaki Hamamoto
難波 博之 †
Hiroyuki Namba
恵木 正史 †
Masashi Egi

1. はじめに

近年、予測値に対する各特徴量の貢献度を分析する手法[1]をはじめとした説明可能 AI 技術により、AI の予測根拠を可視化できるようになった。一方で、AI がデータに含まれるノイズや疑似相関を学習したときなど、AI の予測根拠と有識者の知識との間に乖離が見られたとき、テストデータについて高い予測精度を達成する AI であっても有識者からの信頼が得られず、その修正に大きな時間的コストを要するという新しい問題が生まれている。この問題に対し、好ましい予測根拠を外部から与えることで AI の予測根拠を校正する手法が提案されている[2][3]。しかし、それらはニューラルネットワークなど微分可能な機械学習モデルにしか適用できない。そのため、これまでに報告者らはどのような機械学習モデルにも適用可能な校正手法として Ensemble Based Explanation Correction (EBEC)を提案した[4]。しかし、EBEC は観測データ(実標本)が少ない領域の予測根拠を修正できない課題がある。そのため、プロセス・インフォマティクス分野などで工程の都合により化学反応中の物性値データをほとんど観測できないケースでは EBEC の適用が困難になる。本報告では、仮想的なデータ(仮想標本)を導入することで、実標本が少ない領域でも予測根拠を校正可能にする EBEC の改善手法を提案し、オープンデータを用いてその有効性を検証する。

2. AI 予測根拠の校正手法

本章では AI 予測根拠の校正手法である EBEC の概要とその課題について述べる。

2.1 EBEC の概要

図 1 に EBEC の概要を示す。予測モデルのパラメータを決定する損失関数の空間には沢山の局所解があり、それぞれの局所解では同程度に良い精度を有しながら予測根拠が異なる予測モデルが存在する。これは統計モデリングの羅生門効果と呼ばれている[5]。EBEC はこの羅生門効果を活用し、似て非なる沢山のモデル群を線形結合することで好ましい予測根拠を有する予測モデルを生成する手法である。すなわち、異なるハイパーパラメータとブートストラップ法を適用して複数の予測モデルを作成し、予測根拠の大域的特性（すなわち大域説明）が所望の特性になるようにアンサンブルの結合係数を決定する手法である。EBEC の最適結合係数$\hat{\boldsymbol{a}}$は、アンサンブルするモデル数を n とし、その結合係数のベクトルを$\boldsymbol{a} = \{a_1, a_2, \ldots, a_n\}$としたとき、次の式(1)のように求めることができる。

$$\hat{\boldsymbol{a}} = argmin\left(\sum_i (G_i(\boldsymbol{a}) - Y_i)^2 + \sum_i \sum_f \lambda_{\mathrm{expl}_f} \left(R_{i,f}(\boldsymbol{a}) - Z_{i,f}\right)^2\right) \quad (1)$$

† (株) 日立製作所　Hitachi, Ltd.

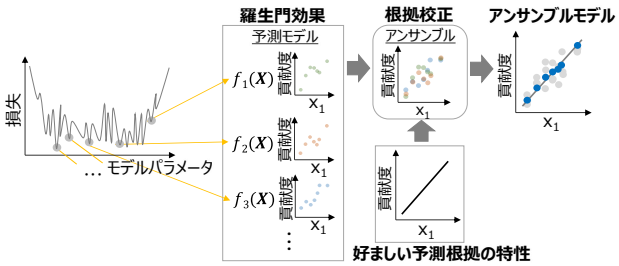

図 1　EBEC の概要

式(1)の第 1 項と第 2 項はそれぞれ予測誤差と説明誤差の大きさを示しており、これらの和を最小化する$\boldsymbol{a}$が最適結合係数$\hat{\boldsymbol{a}}$である。$G_i(\boldsymbol{a})$とY_iはそれぞれ i 番目の標本に対するアンサンブルモデルの予測値とその正解値である。$R_{i,f}(\boldsymbol{a})$、$Z_{i,f}$、$\lambda_{\mathrm{expl}_f}$はそれぞれアンサンブルモデルの予測値に対する i 番目の標本の特徴量 f の貢献度(予測モデルの予測根拠)、その好ましい貢献度(予測根拠の正解値)、および説明誤差の重み係数である。

2.2 EBEC の課題

EBEC は標本ごとに予測根拠の正解値を定義し、その正解値に近づくように結合係数を決定する手法である。そのため、図 2 (a)に示すような標本数が極端に少ない領域(希少データ領域)については予測モデルの大域説明を好ましい特性に近づけることができない課題がある。

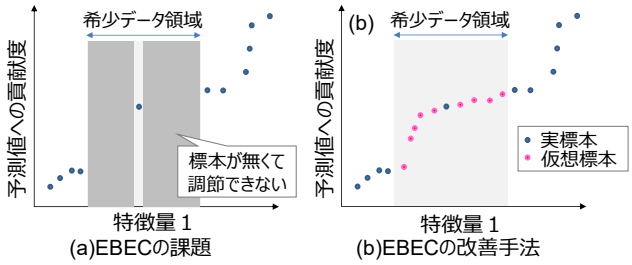

図 2　(a)EBEC の課題と(b)改善手法

3. 仮想標本を用いた EBEC 改善手法の提案

前述したように、EBEC は標本ごとに予測誤差と説明誤差を算出し、それらの和を最小化するようにアンサンブルの結合係数を定めることで予測根拠を校正するため、標本が無い領域の予測根拠をうまく調整することができない。しかし、予測モデルは任意の仮想的な入力値に対し、予測値とその予測根拠を提示することができる。本報告ではこの仮想的な入力値を仮想標本と呼ぶ。この仮想標本を用いることで、図 2 (b)のピンク点に示すように希少データ領域について、予測モデルの大域説明を可視化することができる。仮想標本には予測値に対する正解値が無いため、予測誤差を評価することはできないが、予測根拠については大

域説明の傾向からその妥当性を評価することができる。そこで、仮想標本を導入し、仮想標本については説明誤差だけを評価対象とするように式(1)を拡張することで希少データ領域の予測根拠を調整可能にする EBEC の改善手法を提案する。本報告ではこの改善手法を EBEC using augmented samples (EBEC-AS)と呼ぶ。EBEC-AS の目的関数を式(2)に示す。式(1)との違いは予測誤差に対する重み係数$\lambda_{\mathrm{pred}_i}$が追加されたことのみである。$\lambda_{\mathrm{pred}_i}$は実標本に対しては 1、仮想標本に対しては 0 を与える重み係数である。

$$\hat{\boldsymbol{a}} = argmin\left(\sum_i \lambda_{\mathrm{pred}_i}(G_i(\boldsymbol{a}) - Y_i)^2 + \sum_i \sum_f \lambda_{\mathrm{expl}_f}\left(R_{i,f}(\boldsymbol{a}) - Z_{i,f}\right)^2\right) \quad (2)$$

4. 有効性の検証

本章ではコンクリート強度を予測するデータセット[6]を用いて提案手法の有効性を実験により検証する。セメント配合密度(データセットの特徴量 1)はコンクリート強度に線形に寄与する性質があることが知られているため、本実験ではこの特性を好ましい予測根拠の特性として設定する。実験に用いた訓練データとテストデータの構成を図 3 に示す。実験の目的は希少データ領域における提案手法の予測根拠校正能力を評価することである。そこで、特徴量 1 の値が 200 以上、400 未満である領域を希少データ領域として設定し、希少データ領域外の標本に加えて、希少データ領域からランダムに抽出した 2 つの標本を訓練データ(青点)とした。その他の標本をテストデータ(ピンク点)とした。仮想標本として希少データ領域上でランダムに 100 標本を生成した。XGBoost を用いた 200 個の予測モデルを作成し、その均一アンサンブルを校正前の予測モデルとして EBEC を適用した。

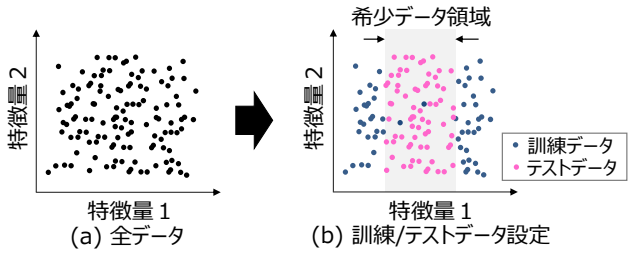

図 3　訓練データとテストデータの構成

実験結果を図 4 に示す。図 4 は校正前(Base)、従来手法(EBEC)、提案手法(EBEC-AS)の予測モデルのセメント密度(特徴量 1)に関する Shapley 値の主効果の大域説明をプロットしている。(a)と(b)は校正前モデルと従来手法による校正後モデルの訓練データに対する大域説明を示しており、EBEC の効果によってセメント密度の予測値への寄与に対する線形性が向上していることが分かる。一方、(d)と(e)は校正前モデルと従来手法による校正後モデルの全データに対する大域説明を示している。(e)の希少データ領域(灰色領域) では線形性が悪く EBEC による予測根拠の校正がうまく機能していないことが分かる。これに対し、提案手法による校正後モデルの訓練データと全データに対する大域説明を示す(c)と(f)では希少データ領域の大域説明も線形性が改善しており、EBEC による予測根拠の校正がうまく機能していることが分かる。ここで、(c)の希少データ領域内にある点群はランダムに与えた仮想標本であり、仮想標本

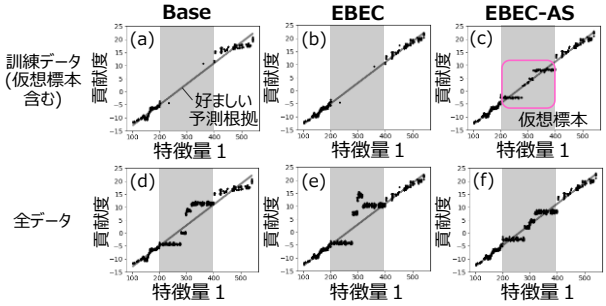

図 4　根拠校正効果の比較

に対する予測根拠の校正結果がテストデータに対してもうまく反映されていることがわかる。

希少データ領域上の訓練データをランダムに変えて本実験を 10 回試行し、テストデータに対して、予測誤差と特徴量 1 に関する説明誤差の平均 RMSE を比較した結果を表 1 に示す。ここで、説明誤差は図 4 に灰色で示す好ましい予測根拠を正解値として算出した。提案法は従来法に比べて説明誤差の RMSE を 47%低減する結果を得た。またこれにより、予測誤差の RMSE が 8%減少する効果を確認し、提案手法の有効性を確認した。

表 1　予測誤差と説明誤差の比較

手法	予測誤差 (a.u.)	説明誤差 (a.u.)
Base (校正前)	1	1
EBEC (従来法)	0.952	0.862
EBEC-AS (提案法)	0.877	0.451

5. おわりに

本報告では、観測データが少ない領域の予測根拠を校正できない EBEC の課題に対し、仮想標本を用いてその校正能力を向上する手法を提案した。オープンデータを用いた実験では、従来法に比べて観測データが少ない領域の説明誤差を 47%低減し、予測誤差が 8%減少する効果を確認した。これにより、EBEC をより幅広いケースに適用できる見込みを得た。

参考文献

[1] S. M. Lundberg, and S. I. Lee, “A Unified Approach to Interpreting Model Predictions,” Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)}, pp. 4768-4777, 2017.
[2] H. Fukui., T. Hirakawa, T. Yamashita, and H. Fujiyoshi, “Attention branch network: Learning of attention mechanism for visual explanation,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10705–10714, 2019.
[3] L. Rieger, C. Singh, W. Murdoch, and B. Yu, “Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge,” Proceedings of the 37th International Conference on Machine Learning, PMLR 119, pp. 8116-8126, 2020.
[4] M. Hamamoto and M. Egi, “Model-agnostic Ensemble-based Explanation Correction Leveraging Rashomon Effect” Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence, pp. 01-08, 2021.
[5] L. Breiman, “Statistical Modeling: The Two Cultures,” Statistical Science Vol. 16, No. 3, pp. 199–231, 2001.
[6] I. Lyse, “Tests on consistency and strength of concrete having constant water content,” Proceedings of American Society for Testing and Materials, Vol. 32, Part 2, pp. 629-636, 1932.