深層学習における予測結果の確からしさ評価技術

Reliability Evaluation Method of Predictive Results in Deep Learning

山科 勇輔*1 Yusuke Yamashina
髙尾 健司*2 Kenji Takao
岩下 信治*3 Nobuharu Iwashita
木津 哲也*4 Tetsuya Kizu
濵﨑 大夢*5 Nozomu Hamasaki

時系列の運転データを用いた異常検知や将来予測の手法として，深層学習の一つであるLSTM（Long short-term memory）の適用例が増えてきている。LSTMでは，学習時と近い運転状態であれば高い予測精度を期待できるが，学習時と異なる運転状態では著しく予測精度が低下する可能性がある。しかし，従来のLSTMでは予測結果が信頼に足るものか否かを知る術がなかったため，プラントの運転制御等の高い信頼性が求められる箇所への適用は進んでいなかった。本報では，九州大学との共同研究にて開発した，LSTMを用いた深層学習モデルにおける予測結果とその確からしさを評価する技術について紹介する。

1. はじめに

機械学習技術が発展してきており，特に深層学習では，“人が明確に規則として書き出せない”ことも訓練データが大量にあれば判断・予測ができるようになってきている。しかし，深層学習の適用には，大量かつ“適切な”訓練データを用意しなければ誤った結果を出力してしまう。例えば，自動運転での誤認識による事故など，実際にいくつかの例が報告されている(1)。

深層学習のモデルの1つであるLSTMは，リカレント機構を有するRNN（Recurrent Neural Network）の1つであり，図1に示すように，ニューロンを含むLSTMセルとリンクから構成されており，入力に近いセルから順にその出力とリンクが持つ重みを積和し，最終的な値を出力している。学習時には各リンクの重みとLSTMセル内のリンクの重みを最適化している。

しかし，従来のLSTMの学習では，正解値と予測値との誤差が最小となるように重みの最適化を行っており，すべてのLSTMセルを活用しているわけではない。つまり，学習時に最適化されない重みが残る可能性がある。このため，ニューロンの発火状況によっては，学習時に発火しなかったニューロンが運用時に発火することによって予測結果が大幅にずれる等の悪影響を及ぼすことが考えられる（図2参照，ここでは，LSTMセルをニューロンとして説明）。

一方で，画像処理分野における深層学習において，ニューロンカバレッジ（テスト入力によって発火するニューロンの数）を指標としたモデルの確からしさの評価や，カバレッジが最大となるようなテストデータを生成し再学習させることで精度が向上した結果が報告されている

本報では，時系列データを対象とした深層学習（LSTM）において，①LSTMに適応した評価指標（カバレッジ）の開発，②カバレッジ指標を用いた評価関数の開発によりロバスト性の高いモデルの構築，③実運用時の予測結果の確からしさの評価手法を開発し，ボイラを事例に有効性の検証を行った結果について紹介する。

*1 ICTソリューション本部 CIS部
*2 ICTソリューション本部 CIS部 主席技師 工博
*3 ICTソリューション本部 CIS部 主席技師
*4 三菱パワー株式会社エンジニアリング本部ボイラ技術総括部ボイラ技術部 主席技師
*5 三菱パワー株式会社エンジニアリング本部ボイラ技術総括部ボイラ技術部

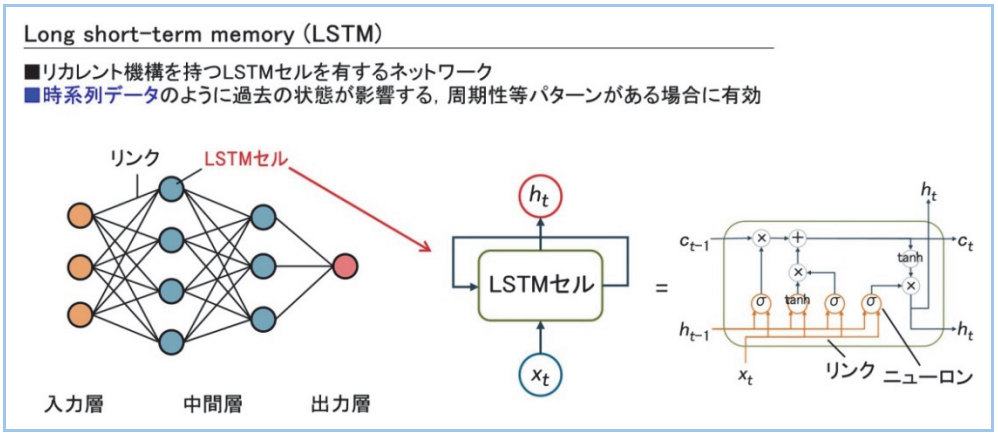

図1 LSTM モデルの概要

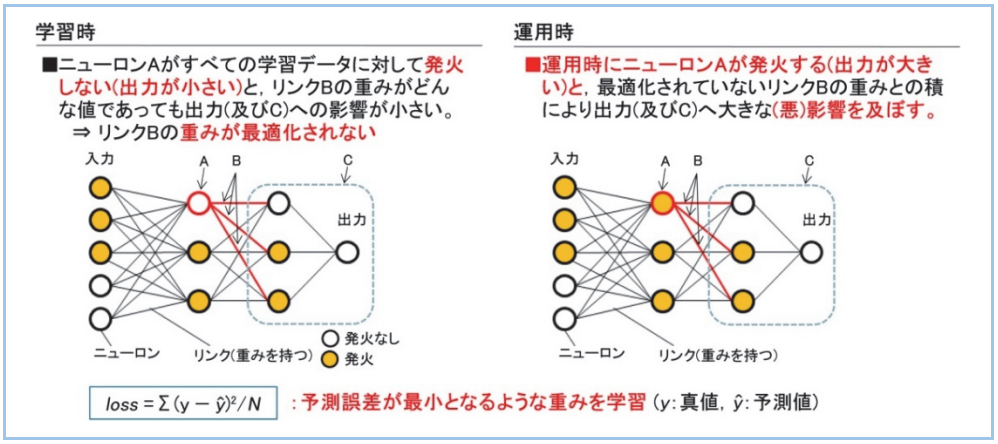

図2 深層学習手法の学習における課題

2. 深層学習における確からしさ評価技術

2.1 時系列データに対する深層学習モデルの評価指標

LSTM は通常のニューラルネットワークと異なり再帰構造であるリカレント層を有している。この再帰構造の中には2つの内部状態（図3中の隠れ状態 h_tと記憶セルc_t。h_tは次の層に伝播する値であり，c_tは同一層で時間方向に伝播する値）が存在しており，これら2種類の値がニューロンの発火に影響しているため，これらをカバレッジとして定量評価することでモデルの確からしさ（各重みが最適化されているかどうか）の指標とした。

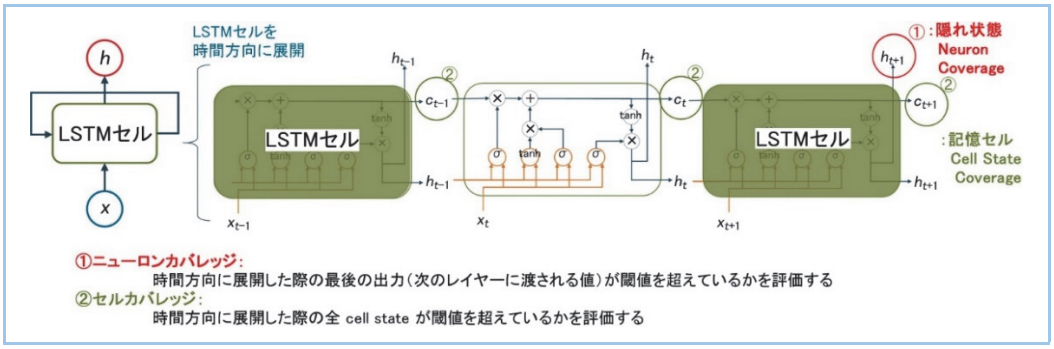

図3 時系列データに対する深層学習モデルの評価指標

2.2 カバレッジ指標を考慮した評価関数の開発

通常，ニューラルネットワークモデルの作成では，実測値と予測値との誤差から成る評価関数（時系列データに対しては，MSE（Mean Squared Error）がよく用いられる）を最小にするように重みを学習する。これに対し，前項で開発した評価指標をカバレッジ項としてloss関数に付加することで，誤差だけでなくカバレッジも考慮できる新しい評価指標を開発した。これにより，学習の段階

で最適化されないニューロンを残さないことで，運用時に最適化されていないニューロンが発火することによる悪影響を防ぐことができ，精度を安定させることができる。

学習のイメージを図4に示す。

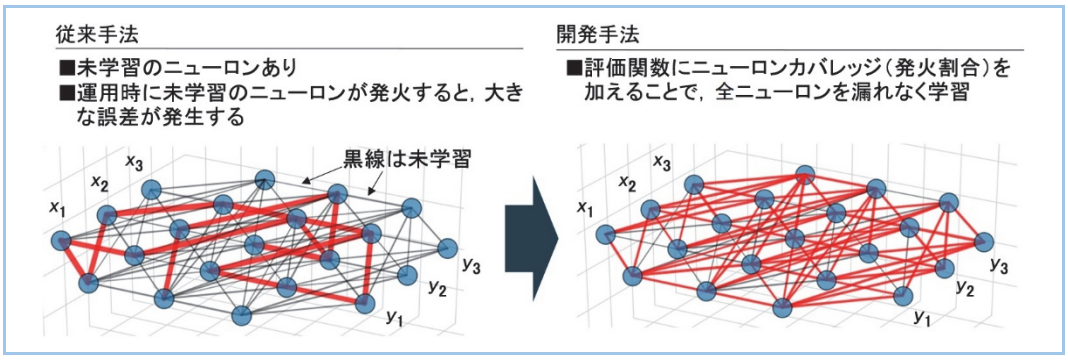

図4 カバレッジを考慮した学習手法

2.3 カバレッジ指標を用いた予測結果の確からしさ評価

一般に深層学習などの機械学習モデルでは，学習範囲外の入力に対しては予測精度が著しく悪化する可能性が高く，予測結果を制御に適用することはリスクが高い。そこで，予測値の確からしさを評価する手法も開発した。具体的には，カバレッジ指標によって各ニューロンが発火したかどうかを評価できるようになったため，図5に示すように，学習時と実運用時との発火パターンの違いから，実運用時のデータの分布が学習時と同じ分布かどうかを定量的に評価する。

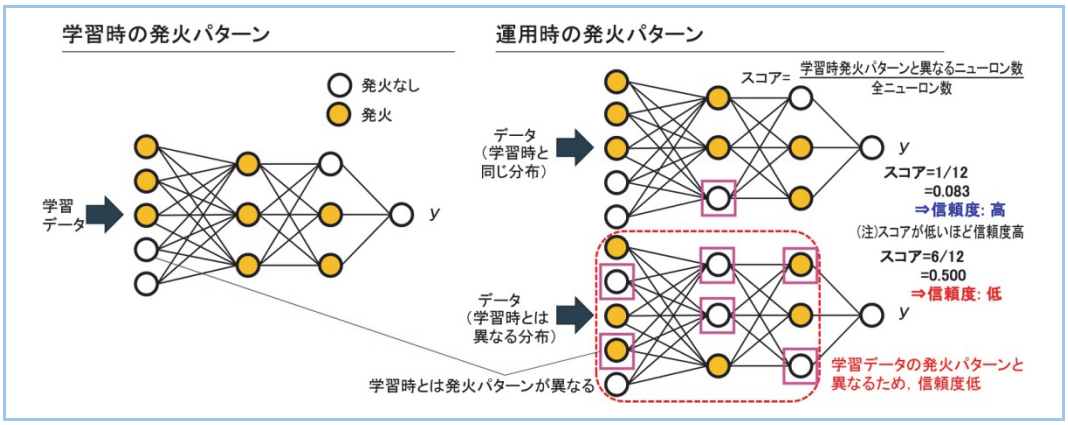

図5 カバレッジ指標を用いた予測結果の確からしさ評価

3. ボイラを対象とした開発技術の検証

ボイラ内部には複数の高圧除煤装置が設置されており，この運用間隔を最適化したいというニーズがある。最適な運用間隔の制約条件となるのは，熱交換器出口温度であり，これを閾値内に収める必要があるため，深層学習手法を用いて将来の出口温度を予測し，最適な運用間隔の策定を図っている。運転計画の観点から，2 日先の出口ガス温度を予測し，その予測値を基に除煤装置の起動タイミングを決めるため，“真値“が得られる前に予測値の確からしさの評価が必要であり，また，燃料性状による影響が大きいため，高精度かつ高い確からしさで予測する必要がある。

図6に示す 3 つの異なる期間の運転データを対象に本手法の検証を行った。検証条件として，入力変数は，ボイラ入口温度，出口温度，高圧除煤装置（SB）9台の累積停止時間（起動時に0にリセット）の計 11 変数とし，目的変数は2日後の出口温度とした。データのサンプリング時間は10 分であり，深層学習モデルとして，中間層にLSTM 層を1層とした（ユニット数 50）モデル構成とした。図6に示す3ケースのうち1つを学習データとし，他の2つを検証データとした。

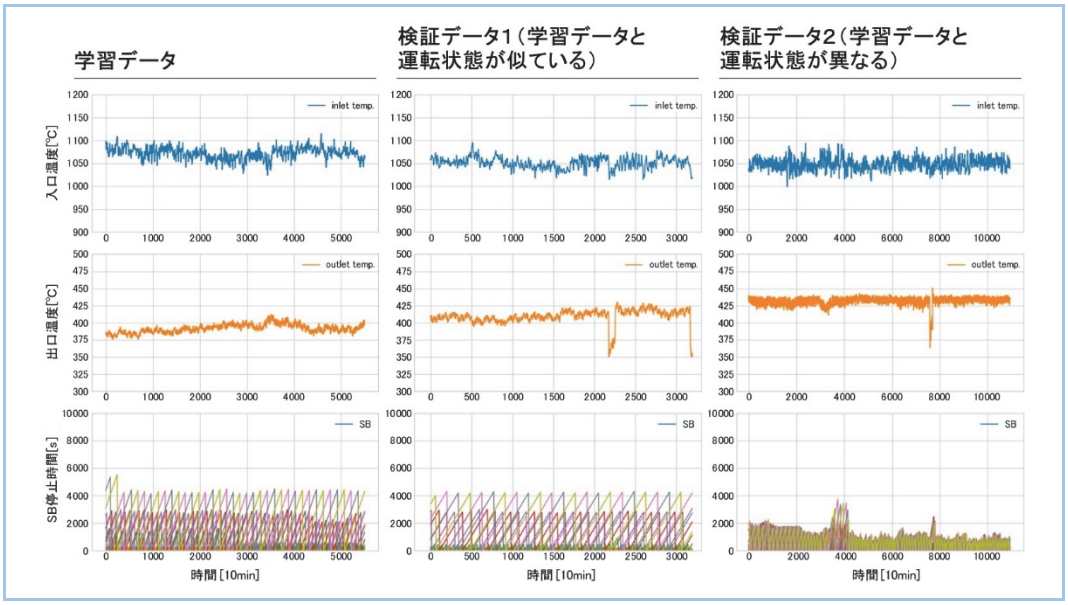

図6　ボイラガス温度実測データ

まず，従来手法にて学習を行った結果を図7に示す。学習データに対しては精度よく当てはまるモデルが構築できているが，ニューロンカバレッジは 0.54 であり，多くのニューロンが発火していないことが分かった。また，学習時と近い運転データ（検証データ1）に対してもカバレッジが増加しており，学習時に発火しなかったニューロンが発火していることを示している。そのため，一部で予測誤差（RMSE：Root Mean Square Error）が増加していることが確認できる。学習範囲外の運転データ（検証データ2，出側温度が学習時に比べ高い）に対してはカバレッジがさらに増加し，予測精度が大きく低下している。これらのことから学習時に最適化されない重みの影響で運用時に精度が悪化することを確認した。

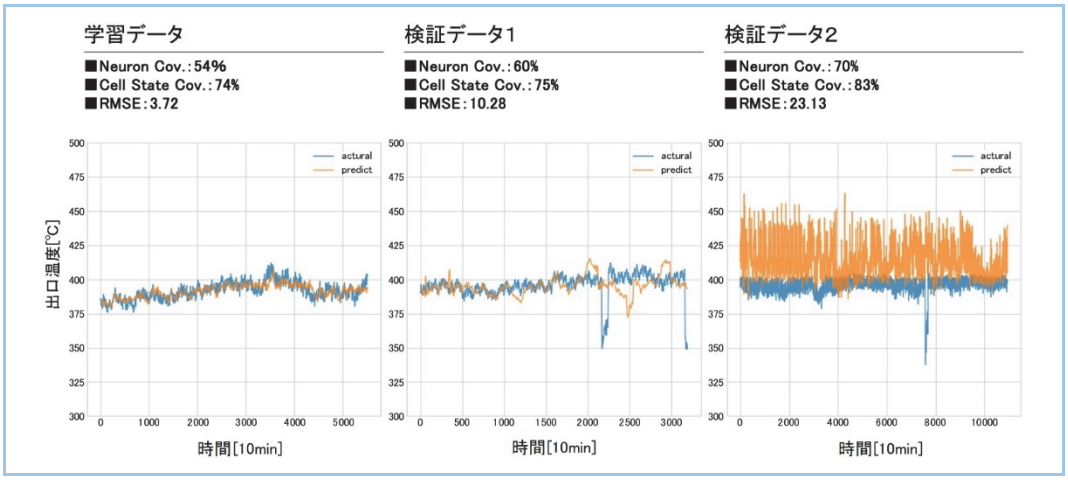

図7　カバレッジ評価結果

次にカバレッジを考慮した評価関数にて，学習・予測を行った結果を図8に示す。これより，学習時のニューロンカバレッジが1であり，すべてのリンクの重みが最適化されていることがわかる。また，検証データ1と2の両ケースで予測精度が向上しており，特に検証データ2に対しては， RMSE を 23.13℃から 6.55℃と 28％低減できる結果が得られ，精度を安定させることができている。

最後に，予測結果の確からしさについて示す。検証データ2に対する予測結果確からしさを図9に示す。予測精度が悪くなるタイミングで確からしさも低下していることが確認できる。これより，学習したモデルではカバーできない範囲のデータが入力される時の予測誤差が増大する状況を，確からしさ評価により事前に把握できることが確認できた。

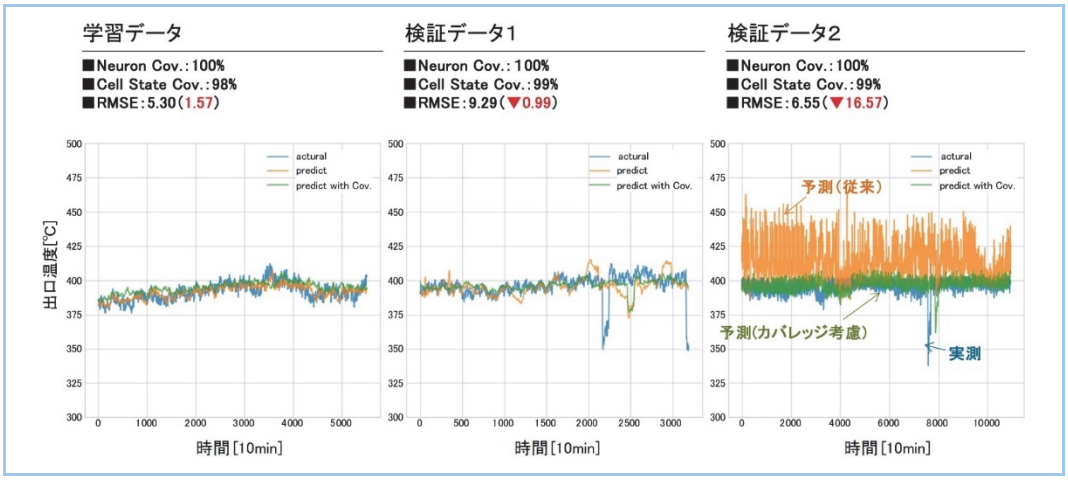

図8 カバレッジを考慮した学習結果

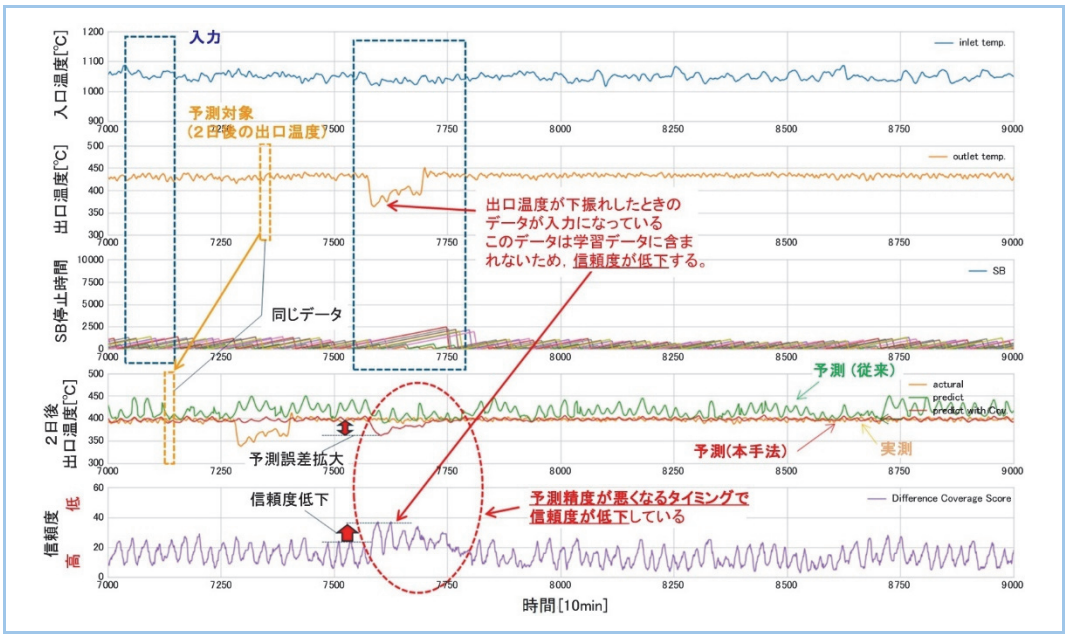

図9 カバレッジ指標を用いた予測結果の確からしさ評価

4. まとめ

本報では，時系列データを対象とした深層学習モデルに対して，カバレッジ指標を用いた学習済みモデルの評価手法，予測結果に対する確からしさ評価手法について紹介した。本技術により学習済みモデルの確からしさを実運用前に評価でき，また，実運用時も入力データに対する予測結果の確からしさも評価できることで，制御分野などへの深層学習手法の適用が可能となり，高性能な製品や深層学習を用いたサービスビジネスを提供することができる。

今後，客先火力プラントでの検証や，その他，当社製品全般の補用品の余寿命予測などへの展開を計画している。

参考文献

(1) Tian, et al., DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, ICSE, 2018
(2) Hochreiter, et al., LONG SHORT-TERM MEMORY, Neural Computation, Vol.9, No.8, p.1735-1780, 1997
(3) L.Ma, et al,. DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems, ASE, 2018
(4) L.Ma, et al., DeepMutation: Mutation Testing of Deep Learning Systems, ISSRE, 2018